

# ASAP: Eliminating Algorithm-based Disclosure in Privacy-Preserving Data Publishing

Xin Jin<sup>a</sup>, Nan Zhang<sup>a,\*</sup>, Gautam Das<sup>b</sup>

<sup>a</sup>*Department of Computer Science, George Washington University, USA, 20052*

<sup>b</sup>*Department of Computer Science and Engineering, University of Texas at Arlington, USA, 76019*

---

## Abstract

Numerous privacy-preserving data publishing algorithms were proposed to achieve privacy guarantees such as  $\ell$ -diversity. Many of them, however, were recently found to be vulnerable to algorithm-based disclosure - i.e., privacy leakage incurred by an adversary who is aware of the privacy-preserving algorithm being used. This paper describes generic techniques for correcting the design of existing privacy-preserving data publishing algorithms to eliminate algorithm-based disclosure. We first show that algorithm-based disclosure is more prevalent and serious than previously studied. Then, we strictly define *Algorithm-SAFE Publishing* (ASAP) to capture and eliminate threats from algorithm-based disclosure. To correct the problems of existing data publishing algorithms, we propose two generic tools to be integrated in their design: *global look-ahead* and *local look-ahead*. To enhance data utility, we propose another generic tool called *stratified pick-up*. We demonstrate the effectiveness of our tools by applying them to several popular  $\ell$ -diversity algorithms: Mondrian, Hilb, and MASK. We conduct extensive experiments to demonstrate the effectiveness of our tools in terms of data utility and efficiency.

### *Keywords:*

Privacy preservation, Data publishing, Algorithm-based disclosure, Algorithm-safe publishing

---

\*Corresponding author. Tel: +12029945919. Fax: +12029944875

*Email addresses:* xjin@gwu.edu (Xin Jin), nzhang10@gwu.edu (Nan Zhang), gdas@uta.edu (Gautam Das)

## 1. Introduction

### 1.1. Privacy-Preserving Data Publishing

Many organizations, such as hospitals, require publishing microdata with personal information, such as medical records, for facilitating research and serving public interests. Nonetheless, such publication may incur privacy concerns for the individual owners of tuples being published (e.g., patients). To address this challenge, privacy-preserving data publishing (i.e., PPDP) was proposed to generate the published table in a way that enables analytical tasks (e.g., aggregate query answering, data mining) over the published data, while protecting the privacy of individual data owners.

In general, a *microdata* table (denoted by  $T$ ) can contain three types of attributes: 1) *personal identifiable* attributes (e.g., *SSN*), each of which is an explicitly unique identifier of an individual, 2) *quasi-identifier* (QI) attributes (e.g., *Age*, *Sex*, *Country*), which are not explicit identifiers but, when combined together, can be empirically unique for each individual, and 3) *sensitive attributes* (SA) (e.g., *Disease*), each of which contains a sensitive value (set) that must be protected. In privacy-preserving data publishing, personal identifiable attributes are usually removed prior to publishing. QI and/or SA attributes are perturbed to achieve a pre-defined privacy model while maximizing the utility of published data.

Samarati and Sweeney [1] first defined a privacy model,  $k$ -anonymity, for PPDP. It requires each tuple in the published table (denoted by  $T^*$ ) to have at least  $k - 1$  other QI-indistinguishable tuples - i.e., tuples with the same QI attribute values. To protect individual SA information, Machanavajjhala et al [2] introduced another privacy model,  $\ell$ -diversity, which further requires each group of QI-indistinguishable tuples to have diverse SA values. Variations of the  $\ell$ -diversity include  $(\alpha, k)$ -anonymity [3],  $t$ -closeness [4],  $(k, e)$ -anonymity [5],  $m$ -invariance [6], etc. To satisfy these privacy models, numerous PPDP algorithms have been proposed [7, 8, 9, 10, 5, 11].

### 1.2. Algorithm-Based Disclosure

It was traditionally believed that, to determine whether a privacy model is properly satisfied, one only needs to look at the published table, i.e., the output of a data publishing algorithm, without investigating the algorithm itself. The recently discovered algorithm-based disclosure [12] contradicts this traditional belief as it demonstrates that privacy disclosure can be incurred

by the *design* of a data publishing algorithm. In particular, if a privacy-preserving algorithm is vulnerable to algorithm-based disclosure, then once an adversary learns the design of the algorithm, s/he may utilize this knowledge to reverse-engineer the published table to compromise additional private information. We shall discuss the details in Section 2.

Algorithm-based disclosure poses a significant threat to the privacy of published data, because the data publishing algorithm is usually considered public and may be learned by an adversary. One might argue that, given the large number of public algorithms that are available for PPDP, it is difficult for an adversary to precisely identify which algorithm has been used and thereby to launch the algorithm-based attack. This is a typical “security through obscurity” argument which counts on the secrecy of an algorithm to ensure the security of its output. However, such arguments have been repeatedly argued against and aborted in the literature of security and cryptography. As the Kerckhoff’s principle [13] in cryptography states, “The cipher method must not be required to be secret, and it must be able to fall into the hands of the enemy without inconvenience.” Similarly, we argue that, to design an effective algorithm for privacy-preserving data publishing, one must eliminate algorithm-based disclosure.

### 1.3. Existing Work and Limitations

Wong et al. [12] demonstrated the first known case of algorithm-based disclosure by showing that the minimality principle used by many existing algorithms, i.e., to perturb QI with the minimum degree possible for satisfying the privacy model, may lead to the disclosure of private SA information when the adversaries have the original QI as external knowledge. An example of this disclosure will be described in Section 2. To counteract this attack, Wong et al. proposed a new privacy model called  $m$ -confidentiality [12], which guarantees that even an adversary with knowledge of QI cannot have confidence of more than  $1/m$  on the SA value of an individual tuple. This attack was also studied in [14], with a new privacy model  $p$ -safety proposed as a countermeasure.

The new privacy models studied in the existing work, i.e.,  $m$ -confidentiality [12] and  $p$ -safety [14], are by definition safe against (at least certain types of) algorithm disclosure. In addition, some recently proposed privacy models such as differential privacy [15] are also by definition immune from algorithm-based disclosure. While defining these new privacy models and developing

their corresponding new algorithms provides a clean-slate solution for eliminating algorithm-based disclosure, limiting the investigation of algorithm-based disclosure to this realm has a number of problems.

First, the state-of-the-art PPDP calls for a proper understanding of the scope of algorithm-based disclosure for the existing data publishing algorithms. Currently, unless a data publishing algorithm is designed for an inherently algorithm-disclosure-safe privacy model such as differential privacy, it is unclear how to determine whether the algorithm is vulnerable to algorithm-based disclosure. Meanwhile, there are considerable ongoing efforts [16, 17] on developing data publishing algorithms for popular privacy models such as  $\ell$ -diversity which do not provide such definition-inherent guarantee against algorithm-based disclosure. To enable the safe deployment of these algorithms in practice, it is important to understand whether and how algorithm-based disclosure may occur for a given data publishing algorithm.

Furthermore, the wide prevalence of data publishing algorithms calls for a generic method to revise the design of an existing algorithm for eliminating algorithm-based disclosure. In the literature, for popular privacy models such as  $\ell$ -diversity, there have been not only a myriad of algorithms for publishing tabular data, but also numerous others that publish application-specific data such as location [18], social network [19], and transaction information [20]. Instead of re-inventing algorithms for all these applications, we argue that a more cost-effective way is to develop a generic method that eliminates algorithm-based disclosure from the existing algorithms.

#### 1.4. Outline of Technical Results

In this paper, we attack the problem of algorithm-based disclosure from a novel algorithmic angle. In particular, we first illustrate the challenge of identifying algorithm-based disclosure by demonstrating that the space of such disclosure is substantially larger than previously recognized. Then, we provide a testing tool to determine whether a given data publishing algorithm is subject to algorithm-based disclosure. Finally, we develop two tools, *global look-ahead* and *local look-ahead*, to revise the design of existing data publishing algorithms for eliminating algorithm-based disclosure. To recover the utility loss incurred by applying these tools, we develop *stratified pick-up*, another tool to retain a high level of utility for the published table.

Our detailed results can be stated as follows:

**First, we find that the space of algorithm-based disclosure is much broader than previously discovered.** While the previous work

identifies algorithm-based disclosure when an adversary holds external knowledge about the QI attributes, we find that other forms of external knowledge, such as the distribution of SA values and/or certain negative association rules [21] can also give rise to algorithm-based disclosure. Our further investigation even eliminates the dependency of algorithm-based disclosure on external knowledge. That is, we find algorithm-based disclosure can happen even when the adversary holds no external knowledge about the published data. To this end, we find that MASK [12], originally proposed to eliminate the previously discovered algorithm-based disclosure, actually suffers from another type of algorithm-based disclosure we discover in the paper.

**Second, we propose a testing tool for checking whether a given data publishing algorithm is vulnerable to algorithm-based disclosure.** In order to do so, we first introduce *Algorithm-SAFE data Publishing* (ASAP), a model that formally defines algorithm-based disclosure as the difference between two random worlds: a *naive* one where every possible mapping between an original table and the published table is equally likely unless such a mapping violates an adversary’s external knowledge, and a *smart* one where the mapping must also follow the data publishing algorithm. An algorithm satisfies ASAP iff it always maintains equivalence between these two worlds.

To identify the vulnerability of existing data publishing algorithms, we derive two necessary conditions of ASAP. To induce/judge immunity against algorithm-based disclosure, we derive two sufficient conditions for ASAP. The main idea is to avoid any unpublished QI-SA correlation to be used in generating the published table. The combination of these necessary and sufficient conditions forms our tool for checking whether a given data publishing algorithm is vulnerable to algorithm-based disclosure.

**Third, we develop two tools, global look-ahead and local look-ahead, for revising the design of existing algorithms to follow ASAP.** They are designed to amend the most common violation of ASAP found in existing algorithms in terms of their QI and SA perturbation strategies, respectively. To demonstrate the effectiveness of our tools, we first apply global look-ahead to revise Mondrian [8] and Hilb [11], two well-known data publishing algorithms designed to achieve  $\ell$ -diversity. Then, we apply local look-ahead to MASK [12]. We prove that all revised algorithms satisfy ASAP.

**Fourth, we devise another tool, stratified pick-up to improve the utility of published data without violating ASAP.** The idea of

stratified pick-up is to use an Anatomy [10] like technique to minimize the number of tuples in each published QI-group (i.e., set of QI-indistinguishable tuples). To demonstrate its effectiveness, we apply stratified pick-up on top of the output from algorithms altered by our first two tools, and show that they provide almost equal or even better utility than the corresponding original algorithms.

**Our contribution also includes a comprehensive set of experiments on real-world datasets.** First, we measure the magnitude of algorithm-based disclosure for MASK on Adult, a popular benchmark dataset for privacy-preserving data publishing. Also, we test the extent of algorithm-based disclosure for the state-of-the-art  $\ell$ -diversity algorithm Hilb. Then, we evaluate the effectiveness of our tools by comparing the utility of our ASAP-compliant algorithms (i.e., Mondrian++, Hilb++, MASK++) against their original counterparts on Census, another large benchmark dataset. Experimental results show that while eliminating algorithm-based disclosure, our ASAP algorithms remain efficient and achieve almost equal or even (sometimes significantly) better utility than the existing algorithms.

The rest of the paper is organized as follows. Section 2 describes two motivating examples for algorithm-based disclosure. Section 3 introduces preliminaries and notations used in the paper. Section 4 formally defines ASAP. Section 5 derives two necessary as well as two sufficient conditions for ASAP, and verifies the vulnerability of existing algorithms. We develop two generic tools in Section 6 to correct the design of existing algorithms for eliminating algorithm-based disclosure, and develop another tool in Section 7 to enhance utility. We conduct experiments in Section 8, review the related works in Section 9, and conclude in Section 10.

## 2. Motivating Examples

This section describes two motivating examples of algorithm-based disclosure. We consider two adversaries: “naive” Nash and “smart” Sam throughout the paper. Both of them hold the same external knowledge and observe the same published table. The only difference is that “naive” Nash does not know the data publishing algorithm, whereas “smart” Sam does. Both Nash and Sam want to compromise whether their friend Tom, a 37-year-old male from Japan, has AIDS or not.

For the ease of discussion, we follow the same SA settings as previous work [12, 22] - i.e., infectious disease {AIDS} is sensitive, while noninfectious

Table 1: An example of algorithm-based disclosure in  $\ell$ -diversity algorithm

	(a) microdata	(b) 2-diversity table	(c) external knowledge																																																									
row#	<table border="1" style="border-collapse: collapse; width: 100%;"><thead><tr><th>Sex</th><th>Disease</th></tr></thead><tbody><tr><td>F</td><td>gastritis</td></tr><tr><td>F</td><td>heart disease</td></tr><tr><td>F</td><td>cancer</td></tr><tr><td>F</td><td>diabetes</td></tr><tr><td>M</td><td>AIDS</td></tr><tr><td>M</td><td>AIDS</td></tr></tbody></table>	Sex	Disease	F	gastritis	F	heart disease	F	cancer	F	diabetes	M	AIDS	M	AIDS	<table border="1" style="border-collapse: collapse; width: 100%;"><thead><tr><th>Sex</th><th>Disease</th></tr></thead><tbody><tr><td>F</td><td>gastritis</td></tr><tr><td>F</td><td>heart disease</td></tr><tr><td>*</td><td>cancer</td></tr><tr><td>*</td><td>diabetes</td></tr><tr><td>*</td><td>AIDS</td></tr><tr><td>*</td><td>AIDS</td></tr></tbody></table>	Sex	Disease	F	gastritis	F	heart disease	*	cancer	*	diabetes	*	AIDS	*	AIDS	<table border="1" style="border-collapse: collapse; width: 100%;"><thead><tr><th>Name</th><th>Sex</th></tr></thead><tbody><tr><td>Amy</td><td>F</td></tr><tr><td>Eva</td><td>F</td></tr><tr><td>Grace</td><td>F</td></tr><tr><td>Helen</td><td>F</td></tr><tr><td>Jack</td><td>M</td></tr><tr><td><b>Tom</b></td><td>M</td></tr></tbody></table>	Name	Sex	Amy	F	Eva	F	Grace	F	Helen	F	Jack	M	<b>Tom</b>	M															
Sex	Disease																																																											
F	gastritis																																																											
F	heart disease																																																											
F	cancer																																																											
F	diabetes																																																											
M	AIDS																																																											
M	AIDS																																																											
Sex	Disease																																																											
F	gastritis																																																											
F	heart disease																																																											
*	cancer																																																											
*	diabetes																																																											
*	AIDS																																																											
*	AIDS																																																											
Name	Sex																																																											
Amy	F																																																											
Eva	F																																																											
Grace	F																																																											
Helen	F																																																											
Jack	M																																																											
<b>Tom</b>	M																																																											
	(d) (1M,1AIDS)	(e) (3M,2AIDS)	(f) (2M,0AIDS)																																																									
	(g) (2M,1AIDS)																																																											
	<table border="1" style="border-collapse: collapse; width: 100%;"><thead><tr><th>Sex</th><th>Disease</th></tr></thead><tbody><tr><td>F</td><td>gastritis</td></tr><tr><td>F</td><td>heart disease</td></tr><tr><td>F</td><td>cancer</td></tr><tr><td>F</td><td>AIDS</td></tr><tr><td>*</td><td>diabetes</td></tr><tr><td>*</td><td>AIDS</td></tr></tbody></table>	Sex	Disease	F	gastritis	F	heart disease	F	cancer	F	AIDS	*	diabetes	*	AIDS	<table border="1" style="border-collapse: collapse; width: 100%;"><thead><tr><th>Sex</th><th>Disease</th></tr></thead><tbody><tr><td>F</td><td>gastritis</td></tr><tr><td>F</td><td>heart disease</td></tr><tr><td>*</td><td>cancer</td></tr><tr><td>*</td><td>AIDS</td></tr><tr><td>M</td><td>diabetes</td></tr><tr><td>M</td><td>AIDS</td></tr></tbody></table>	Sex	Disease	F	gastritis	F	heart disease	*	cancer	*	AIDS	M	diabetes	M	AIDS	<table border="1" style="border-collapse: collapse; width: 100%;"><thead><tr><th>Sex</th><th>Disease</th></tr></thead><tbody><tr><td>F</td><td>gastritis</td></tr><tr><td>F</td><td>heart disease</td></tr><tr><td>F</td><td>AIDS</td></tr><tr><td>F</td><td>AIDS</td></tr><tr><td>M</td><td>cancer</td></tr><tr><td>M</td><td>diabetes</td></tr></tbody></table>	Sex	Disease	F	gastritis	F	heart disease	F	AIDS	F	AIDS	M	cancer	M	diabetes	<table border="1" style="border-collapse: collapse; width: 100%;"><thead><tr><th>Sex</th><th>Disease</th></tr></thead><tbody><tr><td>F</td><td>gastritis</td></tr><tr><td>F</td><td>heart disease</td></tr><tr><td>F</td><td>cancer</td></tr><tr><td>F</td><td>AIDS</td></tr><tr><td>M</td><td>diabetes</td></tr><tr><td>M</td><td>AIDS</td></tr></tbody></table>	Sex	Disease	F	gastritis	F	heart disease	F	cancer	F	AIDS	M	diabetes	M	AIDS
Sex	Disease																																																											
F	gastritis																																																											
F	heart disease																																																											
F	cancer																																																											
F	AIDS																																																											
*	diabetes																																																											
*	AIDS																																																											
Sex	Disease																																																											
F	gastritis																																																											
F	heart disease																																																											
*	cancer																																																											
*	AIDS																																																											
M	diabetes																																																											
M	AIDS																																																											
Sex	Disease																																																											
F	gastritis																																																											
F	heart disease																																																											
F	AIDS																																																											
F	AIDS																																																											
M	cancer																																																											
M	diabetes																																																											
Sex	Disease																																																											
F	gastritis																																																											
F	heart disease																																																											
F	cancer																																																											
F	AIDS																																																											
M	diabetes																																																											
M	AIDS																																																											

diseases {cancer, diabetes, gastritis, heart disease} are non-sensitive.

### 2.1. Example 1: Disclosure of $\ell$ -diversity Algorithms Based on QI Generalization

Consider a generalization based algorithm (e.g., [2]) which achieves  $\ell$ -diversity. Table 1a depicts a microdata table with one QI attribute *Sex* and one SA *Disease*. Table 1b is 2-diversity version of Table 1a, such that the proportion of any sensitive SA value in one QI-group is at most  $\frac{1}{\ell} = \frac{1}{2}$ .

First, let us review the case of algorithm-based disclosure discussed in [12], where both “naive” Nash and “smart” Sam know the original QI (Table 1c) through external knowledge. What Nash can do is to join Table 1c with the published Table 1b to infer that Tom belongs to the “\*”-group. Thus, from Nash’s view, the probability of Tom having AIDS is  $\frac{1}{2}$ , which does not violate 2-diversity. We now consider “smart” Sam who knows that the generalization algorithm will not generalize any group unless it violates 2-diversity. Based on this, Sam can infer that no generalization would have been conducted if the 2 males had 0 or 1 AIDS. Therefore, both males, including Tom, must have AIDS. Hence, by leveraging the algorithm-based knowledge, “smart” Sam acquires a different view from “naive” Nash, and Sam’s view violates

Table 2: An example of algorithm-based disclosure in MASK algorithm

(a) microdata

row #	Age	Sex	Country	Disease
1	46	F	Mexico	cancer
2	49	F	Mexico	heart disease
3	32	F	Mexico	heart disease
4	35	F	Mexico	AIDS
5	24	F	Japan	AIDS
6	38	F	Japan	AIDS
7	25	M	Japan	AIDS
8 (Tom)	37	M	Japan	AIDS

(b)  $k$ -anonymity table ( $k = 4$ )

Age	Sex	Country	Disease
[32 – 49]	F	Mexico	cancer
[32 – 49]	F	Mexico	heart disease
[32 – 49]	F	Mexico	heart disease
[32 – 49]	F	Mexico	AIDS
[24 – 38]	*	Japan	AIDS
[24 – 38]	*	Japan	AIDS
[24 – 38]	*	Japan	AIDS
[24 – 38]	*	Japan	AIDS

(c)  $m$ -confidentiality ( $m = 2$ )

Age	Sex	Country	Disease
[32 – 49]	F	Mexico	cancer
[32 – 49]	F	Mexico	heart disease
[32 – 49]	F	Mexico	heart disease
[32 – 49]	F	Mexico	AIDS
[24 – 38]	*	Japan	cancer
[24 – 38]	*	Japan	cancer
[24 – 38]	*	Japan	heart disease
[24 – 38]	*	Japan	AIDS

2-diversity. This is an example of algorithm-based disclosure.

Now, we show the limitation of [12] by demonstrating that algorithm-based disclosure may occur without involving any external knowledge. Note that when “naive” Nash holds no external knowledge, his view of Tom’s SA is the same as what the published table discloses, which by definition satisfies 2-diversity.

Consider the view of “smart” Sam. He can reason as follows: 1) the number of males in the table should be less than 4 but greater than 0, because otherwise no generalization would be needed; 2) if there were only 1 male, Table 1b would not be published because the algorithm would prefer an alternative FFFF\*\* (i.e., Table 1d) to attain better data utility; 3) if there were 3 males, Table 1b would again not be published because of another alternative FF\*\*MM (i.e., Table 1e) with better utility. Apparently, there is only one option left, that is, 2 males in the table. If none or only one of them had AIDS, no generalization would be needed (i.e., Table 1f and 1g). Thus, both males, including Tom, must have AIDS. One can see that the

above deduction is solely enabled by Sam’s knowledge of the algorithm and violates the requirement of 2-diversity. Thus, algorithm-based disclosure may occur without any external knowledge beyond the anonymization algorithm.

*2.2. Example 2: Disclosure of MASK Algorithm Based on SA Perturbation*

MASK [12] was the first attempt to eliminate algorithm-based disclosure. It aims to achieve  $m$ -confidentiality, which (when  $\ell = m$ ) maintains  $\ell$ -diversity even if an adversary has the original QI as external knowledge.

Consider a microdata table in Table 2a. Tables 2b and 2c depict an example of using MASK to achieve 2-confidentiality. MASK first applies  $k$ -anonymization ( $k \geq m$ ) to the microdata table (e.g., 4-anonymity in Table 2b). Then, for each group violating  $\ell$ -diversity (e.g., the “Japan” group), MASK randomly perturbs the sensitive SA values (e.g., AIDS) to non-sensitive values (e.g., cancer, heart disease), until the proportion of sensitive SA values is decreased to  $p$ , where  $p$  is the proportion of sensitive SA values from a randomly selected  $\ell$ -diversity group (e.g.,  $p = \frac{1}{4}$  in the “Mexico” group).

We now show the existence of algorithm-based disclosure in Table 2c when an adversary knows a negative association rule from common-sense, say, “Japanese have an extremely low incidence of heart disease [2, 12]”. Consider the view of “naive” Nash. He can conclude from Table 2c that Tom is in the “Japan” group, and heart disease must be a perturbed value because the heart disease rate in that group (i.e., 25%) conflicts with the negative association rule. But without knowing the MASK algorithm, “naive” Nash can only randomly guess the original value of heart disease to be AIDS or cancer<sup>1</sup>. Thus, the probability of Tom having AIDS in his view is:  $50\% \times \frac{1}{2} + 50\% \times \frac{1}{4} = \frac{3}{8}$ . This does not violate 2-confidentiality.

Now consider the view of “smart” Sam, who knows that MASK would not perturb any SA values in the “Japan” group unless the group violates 2-confidentiality after  $k$ -anonymization (i.e., Table 2b). Thus, Sam concludes that the “Japan” group should have at least 3 AIDS (out of 4 tuples). As such, in “smart” Sam’s view, the probability of Tom having AIDS is at least  $\frac{3}{4}$ , which violates 2-confidentiality. Again, knowing the algorithm empowers

---

<sup>1</sup>For the ease of illustration, we assume here that “smart” Sam has a uniform prior. Note, however, that such an assumption by no means restricts the generality of our discussion, as other distributions would work as well.

“smart” Sam to gain a different view from “naive” Nash, where Sam’s view violates  $m$ -confidentiality.

Consider another algorithm-based disclosure situation in MASK if an adversary has access to some original SA distribution. This is common in reality because data publishers may report statistics for public use. For example, in order to ease the fear of increasing cancer incidence in the community, a local hospital may announce that “only 1 out of 8 hospitalized patients (in the published table) has cancer”.

Now consider what “naive” Nash can compromise from the published Table 2c. He can confirm that MASK should have perturbed 2 SA values to cancer. However, he cannot further tell which 2 out of the 3 cancer are the perturbed values, and whether AIDS or heart disease is the original value. As such, the probability of Tom having AIDS in the view of “naive” Nash is  $33.3\% \times \frac{3}{8} + 33.3\% \times \frac{3}{8} + 33.3\% \times \frac{1}{2} = \frac{5}{12}$ . Likewise, this does not violate 2-confidentiality.

Whereas, “smart” Sam, who knows the MASK algorithm, can infer that the extra 2 cancer must be from the “Japan” group, and AIDS is their original value. The reason is: otherwise, MASK would not conduct any perturbation because both groups in the table after  $k$ -anonymization are already 2-confidentiality. Thereby, there exists algorithm-based disclosure because the probability of Tom having AIDS in Sam’s view is at least  $\frac{3}{4}$ , which violates 2-confidentiality.

As we can see, MASK is still subject to algorithm-based disclosure. And algorithm-based disclosure can exist along with various types of external knowledge, or even without external knowledge. We re-emphasize that this paper is aiming to limit the algorithm-based disclosure (i.e., the view of “smart” Sam), that is, the private information beyond what can be gained by external knowledge.

### 3. Preliminaries

#### 3.1. Privacy-Preserving Data Publishing

Consider  $T = \{t_1, \dots, t_n\}$ , a microdata table of  $n$  tuples. Each  $t_i$  consists of  $d$  QI attributes  $\langle Q_1, Q_2, \dots, Q_d \rangle$ , denoted by  $Q$  and one SA attribute, denoted by  $S$ . For example, Table 2a is a table of  $n = 8$  tuples. Each tuple has  $d = 3$  QI attributes, i.e.,  $Q = \langle Age, Sex, Country \rangle$  and 1 SA, i.e.,  $S = Disease$ . For any tuple  $t \in T$ , let  $t[Q] = \langle t[Q_1], t[Q_2], \dots, t[Q_d] \rangle$  be a vector of QI values in tuple  $t$ ; let  $t[S]$  be the SA value of  $t$ . Let

$\mathcal{D}_Q = \mathcal{D}_{Q_1} \times \mathcal{D}_{Q_2} \times \dots \times \mathcal{D}_{Q_d}$  and  $\mathcal{D}_S$  be the finite domain of  $Q$  and  $S$ , respectively. We say a tuple  $t = \langle q, s \rangle$  is in the table  $T$  (denoted by  $t \in T$ ) where  $q \in \mathcal{D}_Q, s \in \mathcal{D}_S$  iff  $\exists i \in [1, n]$  such that  $t[Q] = q$  and  $t[S] = s$ .

Before releasing the data, a data publisher takes a data publishing algorithm  $A$  to perturb the microdata  $T$ . Let  $T^*$  be the published table of  $T$ ; let  $Q^*$  be the perturbed QI attributes in  $T^*$ . The published table  $T^*$  consists of several QI-groups - i.e., partitions of tuples such that each individual tuple is indistinguishable from any others in the same QI-group. Thus, the *correlation* between QI and SA attributes in  $T^*$  is regarded as the private information to be protected. We represent such *QI-SA correlation* by  $\mathcal{S}^*(\cdot)$ , which maps  $q \in \mathcal{D}_Q$ , the QI attributes of an individual tuple, to the posterior distribution of SA for that tuple. Formally, we have the following definition:

**Definition 1.** (*QI-SA correlation*) Let  $T$  and  $T^*$  be the original and published microdata tables, respectively. Given any tuple  $t = \langle q, s \rangle \in T$ , the QI-SA correlation  $\mathcal{S}^*(q)$  in  $T^*$  with respect to  $t$  is a  $|\mathcal{D}_S|$ -component vector  $\langle \mathcal{S}^*(q)[s_1], \mathcal{S}^*(q)[s_2], \dots, \mathcal{S}^*(q)[s_{|\mathcal{D}_S|}] \rangle$ , where  $|\mathcal{D}_S|$  equals the SA domain size and  $\mathcal{S}^*(q)[s_i] = \Pr\{t[S] = s_i | t[Q] = q, T^*\}$ .

An example of  $Q^*$  and  $\mathcal{S}^*(\cdot)$  in Table 1b (and Table 2c) is shown in Table 3 (and Table 4, respectively).

We are now ready to state the  $\ell$ -diversity privacy model [2] in terms of  $\mathcal{S}^*(\cdot)$ . In particular, we adopt a simple variation of  $\ell$ -diversity [10, 12, 11] which requires that no individual SA value can be compromised with probability over  $\frac{1}{\ell}$ :

**Definition 2.** ( *$\ell$ -diversity [2]*) A published table  $T^*$  fulfills  $\ell$ -diversity iff  $\forall t = \langle q, s \rangle \in T$  and  $\forall s_i \in \mathcal{D}_S$ ,

$$\max_{j \in [1, |\mathcal{D}_S|]} \mathcal{S}^*(q)[s_j] \leq \frac{1}{\ell}.$$

### 3.2. Expression of External Knowledge

As discussed in Section 2, certain type of external knowledge may facilitate algorithm-based disclosure, though algorithm-based disclosure does not mandate the adversary's possession of external knowledge. For the ease of understanding of our ASAP model in the next section, we formalize a simple expression of external knowledge by conjunctive COUNT query. Given the microdata  $T$ , consider a COUNT query  $CQ(T)$  in the form:

Table 3:  $Q^*$  and  $S^*$  of TABLE 1b

(a) $Q^*$	(b) $S^*$				
Sex	AIDS	cancer	diabetes	gastritis	heart disease
F	0	0	0	1/2	1/2
*	1/2	1/4	1/4	0	0

Table 4:  $Q^*$  and  $S^*$  of TABLE 2c

(a) $Q^*$			(b) $S^*$		
Age	Sex	Country	AIDS	cancer	heart disease
[32 – 49]	F	Mexico	1/4	1/4	1/2
[24 – 38]	*	Japan	1/4	1/2	1/4

SELECT COUNT(\*) FROM  $T$

WHERE  $(Q_1 = q_1) \wedge \dots \wedge (Q_d = q_d) \wedge (S = s)$

Note that the selection condition (i.e., WHERE clause) does not require to include every  $Q_j (j \in [1, d])$  or  $S$  in  $T$ . We describe external knowledge  $K_e$  as arithmetic equations (or inequalities) between a pair of COUNT query answers, or between one COUNT query answer and one constant.

Consider Table 2a as the microdata. An example of external knowledge about Tom, who is a 37-year-old male from Japan, is “Tom does not have cancer”. We can express such  $K_e$  as  $CQ(T) = 0$  where  $CQ(T) = \text{SELECT COUNT(*) FROM } T \text{ WHERE } Age = 37 \wedge Sex = M \wedge Country = \text{Japan} \wedge Disease = \text{cancer}$ .

Another example of  $K_e$  is “Japanese have an extremely low incidence of heart disease”, which can be described by  $CQ_1(T)/CQ_2(T) < 0.05^2$  where  $CQ_1(T) = \text{SELECT COUNT(*) FROM } T \text{ WHERE } Country = \text{Japan} \wedge Disease = \text{heart disease}$  and  $CQ_2(T) = \text{SELECT COUNT(*) FROM } T$ .

#### 4. Algorithm-Safe Publishing

This section formalizes algorithm-based disclosure by introducing a new model called *Algorithm-Safe Publishing* (ASAP). A data publishing algorithm

---

<sup>2</sup>The value of 0.05 can be adjusted according to actual needs for reflecting the effect of “extremely low incidence”.

is vulnerable to algorithm-based disclosure when it violates ASAP. We will first define two key concepts relating to ASAP: a *naive random world* and a *smart random world*, which models the view *without* and *with* knowledge of the algorithm, respectively. Then, we will define ASAP based on the equivalence between these two worlds.

#### 4.1. Naive vs. Smart Random World

As in Section 3.1, let  $\mathcal{D}_Q$  and  $\mathcal{D}_S$  be the domains of QI and SA, respectively. Let  $\Omega$  be a finite set of all possible values in the microdata that can be calculated from  $\mathcal{D}_Q \times \mathcal{D}_S$ . When an adversary with external knowledge  $K_e$  observes a published table  $T^*$ , his/her view on the microdata table  $T$  can be modeled as a (posterior) probability distribution over  $\Omega$ , that is, a *mapping* from any  $T' \subseteq \Omega$  to a real value  $\Pr(T = T'|T^*, K_e) \in [0, 1]$ , such that  $\sum_{T' \subseteq \Omega} \Pr(T = T'|T^*, K_e) = 1$ .

First, let us consider “naive” Nash as illustrated in Section 2. In his view of any  $T' \subseteq \Omega$ ,  $T'$  is likely to be the microdata (i.e.,  $\Pr(T = T'|T^*, K_e) > 0$ ) iff: 1) any tuple  $t \in T'$  is bijectively mapped to a tuple  $t^* \in T^*$  such that the value of  $t^*[Q^*]$  is no less specific than that of  $t[Q]$ ; and 2)  $T'$  satisfies the integrity conditions imposed by  $K_e$ . To denote such a “non-zero likelihood” relationship, we use  $T' \Rightarrow T^*$ , indicating that  $T'$  could possibly be published as  $T^*$ . Nonetheless, without learning the data publishing algorithm  $A$ , “naive” Nash cannot distinguish between any  $T'$  in set  $\{T'|T' \subseteq \Omega, T' \Rightarrow T^*\}$ . According to the standard random world assumption [2, 23], Nash has to assign an equal probability to each of them. Thus, we define the view of Nash as a *naive random world*  $\mathcal{NW}(\cdot)$ :

**Definition 3.** (*naive random world*) A naive random world  $\mathcal{NW}(\cdot)$  is a probability distribution such that  $\forall T' \subseteq \Omega$ ,

$$\mathcal{NW}(T') = \begin{cases} 1/c, & \text{if } T' \Rightarrow T^*. \\ 0, & \text{otherwise} \end{cases}$$

where  $c = |\{T'|T' \subseteq \Omega, T' \Rightarrow T^*\}|$ .

To illustrate the definition, consider the previous example of Table 1 in a simple way: AIDS is the sensitive SA value (shadow background) while other SA values (no background) are indistinguishable. Suppose Nash has external knowledge  $K_e$  in the form of two rules: “Amy and Grace are unlikely to have AIDS” and “at least 1 male has AIDS”. Table 5a shows an example

Table 5: An example of naive & smart random world

(a) naive random world						(b) smart random world			
Name	QI	SA	SA	SA	SA	SA	Name	QI	SA
Amy	F						Amy	F	
Eva	F	A	A				Eva	F	
Grace	F						Grace	F	
Helen	F			A	A		Helen	F	
Jack	M	A		A		A	Jack	M	A
<b>Tom</b>	M		A		A	A	<b>Tom</b>	M	A

of this naive random world  $\mathcal{NW}(\cdot)$ . In the view of “naive” Nash, he can find a total of six values of  $T'$  after linking AIDS to the original QI attributes without violating  $K_e$ . Since Nash cannot distinguish any of these six values (i.e., tables) from another,  $\mathcal{NW}(T') = \frac{1}{6}$  has to be assigned equally to each  $T'$ . Thereby, the probability distribution on these 6  $T'$  constitutes the naive random world  $\mathcal{NW}$ .

Second, consider the view of “smart” Sam who learns the mechanism of the data publishing algorithm  $A$ . Sam is able to further distinguish each  $T'$  in the set  $\{T' | T' \subseteq \Omega, T' \Rightarrow T^*\}$  by taking each  $T'$  as the input to  $A$  and then checking whether  $A$  truly outputs  $T^*$  as the published table (if  $A$  is a deterministic algorithm), or whether  $T^*$  could be a possible output (if  $A$  is a randomized algorithm). We define the view of Sam as a smart random world  $\mathcal{SW}(\cdot)$ :

**Definition 4.** (*smart random world*) A smart random world  $\mathcal{SW}(\cdot)$  is a probability distribution,  $\forall T' \subseteq \Omega$ ,

$$\mathcal{SW}(T') = \begin{cases} \Pr\{T = T' | A\}, & \text{if } T' \Rightarrow T^*. \\ 0, & \text{otherwise} \end{cases}$$

Return to the example in Table 5a. “Smart” Sam iteratively performs the 2-diversity algorithm, by accepting each  $T'$  in 5a as input. Table 5b shows the only  $T'$  that may produce  $T^*$  according to the algorithm - All the other 5  $T'$  already satisfy 2-diversity. Thus, any further generalization of them is unnecessary. One can see that “smart” Sam can then construct the smart random world  $\mathcal{SW}$  by assigning  $\mathcal{SW}(T') = 1$  to the  $T'$  in Table 5b and  $\mathcal{SW}(T') = 0$  to all others.

#### 4.2. Definition of ASAP

One can see from the above discussion that whether or not an algorithm is vulnerable to algorithm-based disclosure is determined by whether the naive random world  $\mathcal{NW}$  equals the smart random world  $\mathcal{SW}$  given the same external knowledge  $K_e$  and published table  $T^*$ . There is *no* algorithm-based disclosure iff the two worlds are always equivalent conditioning on the same external knowledge and published table. Formally, we define Algorithm-Safe Publishing (ASAP) as follows.

**Definition 5.** (*Algorithm-Safe Publishing*) A published table  $T^*$  fulfills *Algorithm-Safe Publishing (ASAP)* iff  $\forall t = \langle q, s \rangle \in T, \forall s_i \in \mathcal{D}_S$ , there is:

$$\Pr\{t[S] = s_i | t[Q] = q, \mathcal{NW}\} = \Pr\{t[S] = s_i | t[Q] = q, \mathcal{SW}\}.$$

### 5. Checking Algorithm-based Disclosure

This section presents two necessary and two sufficient conditions for ASAP, respectively. These conditions jointly serve as an exploratory tool to screen a data publishing algorithm for algorithm-based disclosure.

#### 5.1. Necessary Condition 1: $Q^*$ -Independence

$Q^*$ -Independence, our first necessary condition, is motivated by cases where an adversary may learn the original QI through external knowledge. In particular,  $Q^*$ -Independence requires that during QI perturbation (e.g., generalization), the data publishing algorithm must not use any QI-SA correlation that cannot be inferred from the published  $\mathcal{S}^*$ . Otherwise, algorithm-based disclosure may occur. This necessary condition explains why Example 1 in Section 2 is subject to algorithm-based disclosure.

**Theorem 5.1.** ( *$Q^*$ -Independence*) Let  $T$  be a microdata table and  $T^*$  be its ASAP published table. The published QI attributes  $Q^*$  in  $T^*$  must be conditionally independent of the original SA attribute  $S$  in  $T$ , given a combination of the original QI attributes  $Q$  and the published QI-SA correlation  $\mathcal{S}^*$ , denoted by  $Q^* \perp S | (Q, \mathcal{S}^*)$ .

*Proof.* Let  $\mathcal{D}_Q$  and  $\mathcal{D}_S$  be domains of the original QI and SA, respectively. Let  $K_e$  be the external knowledge. In our case,  $K_e$  is specialized to be the original QI.

First, consider in a view of naive random world (Definition 3). Given a published table  $T^*$  and the original QI attribute  $t[Q] = q$  such that  $t = \langle q, s \rangle \in T$ , a naive adversary is unable to distinguish any  $s_i \in \mathcal{D}_S$  such that  $\mathcal{S}^*(q)[s_i] > 0$ .  $t[S] = s_i$  and  $\mathcal{S}^*(q')$ ,  $\forall q' \in Q \setminus q$  (i.e., the published SA of tuples other than  $q$ ) in the naive random world are conditionally independent, given  $\mathcal{S}^*(q)$ . Thus, we have:

$$\Pr\{t[S] = s_i | t[Q] = q, \mathcal{NW}\} = \Pr\{t[S] = s_i | t[Q] = q, \mathcal{S}^*(q)\}. \quad (1)$$

Second, consider in a view of the smart random world (Definition 4). A smart adversary can further distinguish  $s_i$  by checking whether or not the algorithm would publish a table with  $Q^*$  (perturbed from  $Q$ ) and with the QI-SA correlation  $\mathcal{S}^*$  as is. Hence, we have:

$$\Pr\{t[S] = s_i | t[Q] = q, \mathcal{SW}\} = \Pr\{t[S] = s_i | t[Q] = q, \mathcal{S}^*, Q^*, Q\}. \quad (2)$$

From the definition of ASAP (i.e., Definition 5), we have:

$$\Pr\{t[S] = s_i | t[Q] = q, \mathcal{S}^*(q)\} = \Pr\{t[S] = s_i | t[Q] = q, \mathcal{S}^*, Q^*, Q\}. \quad (3)$$

Consider  $t[S]$  as a random variable. To measure the uncertainty of  $t[S]$ , Equation (3) can be transformed from the perspective of information theory [24] as:

$$H(t[S] | t[Q] = q, \mathcal{S}^*(q), Q) = H(t[S] | t[Q] = q, \mathcal{S}^*, Q^*, Q). \quad (4)$$

where  $H(x|y)$  is conditional entropy [25] which measures the uncertainty of a random variable  $x$  given  $y$  is known.

For a microdata  $T$  with  $n$  tuples  $t_1, t_2, \dots, t_n$  where  $t_i[Q] = q_i$ ,  $i \in [1, n]$ , we have:

$$\sum_{1 \leq i \leq n} H(t_i[S] | t_i[Q] = q_i, \mathcal{S}^*(q_i), Q) = \sum_{1 \leq i \leq n} H(t_i[S] | t_i[Q] = q_i, \mathcal{S}^*, Q^*, Q). \quad (5)$$

$$\Rightarrow H(S | \mathcal{S}^*, Q) = H(S | \mathcal{S}^*, Q^*, Q). \quad (6)$$

$$\Rightarrow I(S; Q^* | \mathcal{S}^*, Q) = 0. \quad (7)$$

where  $I(x; y|z)$  is conditional mutual information [25], indicating the amount of information about either  $x$  or  $y$  provided by knowing the other, given  $z$  is

known. The equivalence between Equation (5) and (6) holds due to the following two reasons: 1)  $t_i[S] = s_i$  and  $\mathcal{S}^*(q')$ ,  $\forall q' \in Q \setminus q_i$  in the naive random world are conditionally independent, given  $\mathcal{S}^*(q_i)$ ; and 2) each tuple  $t \in T$  can be considered to be independently generated with certain distribution on  $\mathcal{D}_Q \times \mathcal{D}_S$ .

From Equation (7), we know that  $Q^*$  and  $S$  should be conditionally independent given  $\mathcal{S}^*$  and  $Q$ . Otherwise, their conditional mutual information would not be zero. Thereby, our necessary condition Theorem 5.1 is proved.  $\square$

The basic idea for the theorem proof can be stated as follows: In the perturbation of  $Q$  to  $Q^*$ , suppose a data publishing algorithm consults certain QI-SA correlation information, but does not ultimately publish it in  $T^*$ . According to the basics of mutual information [25], an adversary can recover the consulted QI-SA correlation (even though unpublished) according to the perturbation of  $Q$  to  $Q^*$ , which can be readily observed by an adversary with knowledge of QI.

Examples of existing algorithms which violate  $Q^*$ -independence include algorithms designed for  $\ell$ -diversity [2, 3, 11, 8, 9],  $t$ -closeness [4],  $(k, e)$ -anonymity [5],  $(c, k)$ -safety [23], etc. The reason can be intuitively stated as follows. All these data publishing algorithms follow a similar pattern that they gradually optimize the published table according to a utility metric (e.g., discernibility [26], classification metric [27], KL-divergence [28]), until reaching a table which violates the privacy guarantee to be achieved. At this time, they fall back to the previous table which provides the best utility without violating the privacy guarantee. Unfortunately,  $Q^*$ -independence is violated at the moment when the algorithm determines that a table violates the privacy guarantee because, to make such a decision, the algorithm must have observed certain QI-SA correlation which can never appear in the published table (as such correlation violates the privacy guarantee). As a result, according to Theorem 5.1, the unpublished, guarantee-violating, QI-SA correlation may be derived by smart Sam from observing how QI has been perturbed in the published table. Thus, the algorithms are vulnerable to algorithm-based disclosure.

## 5.2. Necessary Condition 2: $\mathcal{S}^*$ -Independence

In analogy to  $Q^*$ -independence, our second necessary condition  $\mathcal{S}^*$ -independence aims to check algorithm-based disclosure by analyzing the QI-SA correla-

tion used for generating  $\mathcal{S}^*$  in the published table. In addition to the external knowledge of QI as considered in  $\mathcal{Q}^*$ -independence,  $\mathcal{S}^*$ -independence also takes into account possible external knowledge in the form of negative association rules, e.g., “Tom is unlikely to have cancer”. Essentially,  $\mathcal{S}^*$ -independence states that no QI-SA correlation beyond what is ultimately published should be used in the perturbation of SA.

For the ease of understanding, let us first introduce a few notations. Given an individual tuple  $t = \langle q, s \rangle \in T$ , the external knowledge  $K_e$  may include a set of “unlikely” SA  $S_0 \subseteq \mathcal{D}_S$  such that for each value in  $S_0$ , it cannot be linked to that individual tuple  $t$ , i.e.,  $\forall s \in S_0, \mathcal{S}^*(q)[s] = 0$ . Otherwise, we say the published table  $T^*$  contradicts with  $K_e$ .  $\mathcal{S}^*(q)[\mathcal{D}_S \setminus S_0]$  represents the distribution of all “likely” SA in  $T^*$  that can be linked to  $q$  in  $T$ . Return to Example 2 in Section 2. External knowledge of “Tom is unlikely to have cancer” indicates that  $S_0 = \{\text{cancer}\}$  for Tom. Referring to Table 4, we have  $\mathcal{S}^*(q)[\mathcal{D}_S \setminus S_0] = \langle \mathcal{S}^*(q)[\text{AIDS}], \mathcal{S}^*(q)[\text{heart disease}] \rangle = \langle \frac{1}{4}, \frac{1}{4} \rangle$ .

**Theorem 5.2.** ( *$\mathcal{S}^*$ -Independence*) *Let  $T$  be a microdata table and  $T^*$  be its ASAP published table. For any individual tuple  $t = \langle q, s \rangle \in T$ , given a set of “unlikely” SA  $S_0$  and the published  $\mathcal{S}^*(q)[\mathcal{D}_S \setminus S_0]$  in  $T^*$ , the published QI-SA correlation  $\mathcal{S}^*(q)$  in  $T^*$  must be conditionally independent of its original SA attribute  $s$  in  $T$ , denoted by  $\mathcal{S}^*(q) \perp s | (S_0, \mathcal{S}^*(q)[\mathcal{D}_S \setminus S_0])$ .*

*Proof.* Let  $K_e$  be the external knowledge which is essentially  $S_0$  regarding an individual tuple  $t = \langle q, s \rangle \in T$ . We first consider in a view of naive random world. A naive adversary has no ability to distinguish any  $s_i \in \mathcal{D}_S \setminus S_0$ , given  $T^*$ . Hence, we have:

$$\Pr\{t[S] = s_i | t[Q] = q, \mathcal{N}\mathcal{W}\} = \Pr\{t[S] = s_i | t[Q] = q, \mathcal{S}^*(q)[\mathcal{D}_S \setminus S_0], S_0\}. \quad (8)$$

Next, in a view of smart random world, a smart adversary would further verify whether the data publishing algorithm would publish  $\mathcal{Q}^*$  and  $\mathcal{S}^*$  as is when s/he learns  $S_0$  regarding the individual  $t$ . Hence, we have the following:

$$\Pr\{t[S] = s_i | t[Q] = q, \mathcal{S}\mathcal{W}\} = \Pr\{t[S] = s_i | t[Q] = q, \mathcal{S}^*, \mathcal{Q}^*, S_0\}. \quad (9)$$

Similar to the previous proof of Theorem 5.1, we follow Definition 5 and account for the equivalence between Equation (8) and (9) in an information theoretical way:

$$H(t[S] | t[Q] = q, \mathcal{S}^*(q)[\mathcal{D}_S \setminus S_0], S_0) = H(t[S] | t[Q] = q, \mathcal{S}^*, \mathcal{Q}^*, S_0) \quad (10)$$

Since the uncertainty measured by the conditional entropy  $H(x|y)$  can be reduced to  $H(x|z)$  when  $y$  can be derived from  $z$  [25], we have the following inequality:

$$H(t[S]|t[Q] = q, \mathcal{S}^*(q), S_0) \geq H(t[S]|t[Q] = q, \mathcal{S}^*, Q^*, S_0) \quad (11)$$

Inequality (11) holds because  $\mathcal{S}^*(q)$  can be derived from  $\{t[Q] = q, \mathcal{S}^*\}$ . Therefore, combining Equation (10) and Inequality (11), we have:

$$H(t[S]|t[Q] = q, \mathcal{S}^*(q), S_0) \geq H(t[S]|t[Q] = q, \mathcal{S}^*(q)[\mathcal{D}_S \setminus S_0], S_0) \quad (12)$$

Consider another Inequality (13) as follows:

$$H(t[S]|t[Q] = q, \mathcal{S}^*(q)[\mathcal{D}_S \setminus S_0], S_0) \geq H(t[S]|t[Q] = q, \mathcal{S}^*(q), S_0) \quad (13)$$

Inequality (13) holds because  $\mathcal{S}^*(q)[\mathcal{D}_S \setminus S_0]$  can be derived from  $\{\mathcal{S}^*(q), S_0\}$ . By Inequalities (12) and (13), we then derive:

$$H(t[S]|t[Q] = q, \mathcal{S}^*(q)[\mathcal{D}_S \setminus S_0], S_0) = H(t[S]|t[Q] = q, \mathcal{S}^*(q), S_0) \quad (14)$$

$$\Rightarrow I(t[S]; \mathcal{S}^*(q)[S_0]|t[Q] = q, \mathcal{S}^*(q)[\mathcal{D}_S \setminus S_0], S_0) = 0. \quad (15)$$

Recall that  $\mathcal{S}^*(q)[S_0]$  represents the “unlikely” (specified by  $K_e$ ) SA in  $T^*$  to be linked to tuple  $t$ . By the definition of mutual information [25], Equation (15) implies at least one of the following two implications is true. First, it implicates that  $\mathcal{S}^*(q)[S_0]$  is determined by  $S_0$ . In other words,  $T^*$  *never* contradicts with  $K_e$ . However, it is impossible in reality because no one can foresee all kinds of  $S_0$  controlled by  $K_e$ . Thus, we focus on the second implication which requires that  $\mathcal{S}^*(q)[S_0]$  should be conditionally independent of the original  $t[S]$ . Note that  $\sum_{s \in S_0} \mathcal{S}^*(q)[s] = 1 - \sum_{s \in \mathcal{D}_S \setminus S_0} \mathcal{S}^*(q)[s]$ . Thus, the entire  $\mathcal{S}^*(q)[\cdot]$  should be conditionally independent of the original  $t[S]$ , given  $\{t[Q] = q, \mathcal{S}^*(q)[\mathcal{D}_S \setminus S_0], S_0\}$ . As such, Theorem 5.2 is proved.  $\square$

The basic idea of the proof is similar to that of Theorem 5.1 - According to the basics of mutual information [25], if the perturbation of SA relies on certain QI-SA correlation information that is not eventually published, then such unpublished information may be inferred by an adversary through by observing whether the published SA contradicts its external knowledge  $K_e$ , i.e., whether  $\forall s \in S_0, \mathcal{S}^*(q)[s] = 0$ .

For example, MASK [12] is vulnerable to algorithm-based disclosure due to violation of Theorem 5.2. Recall that MASK first checks whether a group

violates  $\ell$ -diversity, and will only perturb SA in those offending groups. Again, such SA perturbation demands the usage of unpublished QI-SA correlation because it is not (always) possible to determine whether a group violates  $\ell$ -diversity based solely upon the table published by MASK. According to Theorem 5.2,  $\mathcal{S}^*$ -Independence is violated.

### 5.3. ASAP Sufficient Conditions

This subsection provides two sufficient conditions to publish an ASAP table. Recall from the two necessary conditions that a fundamental cause for the vulnerability of many existing data publishing algorithms is the usage of unpublished QI-SA correlation in the perturbation of QI and/or SA. Our first sufficient condition focuses on correcting the problem of QI perturbation while prohibiting SA perturbation.

**Theorem 5.3.** (*ASAP sufficient condition 1*) *The published table from a data publishing algorithm fulfills ASAP if the algorithm satisfies both of the following two conditions: 1) QI is perturbed whereas SA is not; and 2) the QI perturbation depends on no information beyond the original QI and the published QI-SA correlation.*

*Proof.* We prove it by contradiction. Assume that a published table  $T^*$  generated from an algorithm satisfying both conditions in Theorem 5.3 is not ASAP. In the perspective of information theory, there must exist a tuple  $t = \langle q, s \rangle \in T$  and external knowledge  $K_e$  such that

$$H(t[S]|t[Q] = q, Q^*, \mathcal{S}^*, K_e) < H(t[S]|t[Q] = q, \mathcal{S}^*(q), K_e). \quad (16)$$

Note that  $H(t[S]|t[Q] = q', Q^*, \mathcal{S}^*, K_e) \leq H(t[S]|t[Q] = q', \mathcal{S}^*(q'), K_e)$  holds for all other tuples  $\langle q', s \rangle \in T$ . Since SA is not perturbed, we know that  $t[S]$  and  $\mathcal{S}^*(q')$ ,  $\forall q' \in Q \setminus q$  (i.e., the published SA of tuples other than  $q$ ) are conditionally independent, given  $\mathcal{S}^*(q)$ . Moreover, since each tuple  $t \in T$  can be considered to be independently generated with certain distribution on  $\mathcal{D}_Q \times \mathcal{D}_S$ , we have:

$$H(S|Q, Q^*, \mathcal{S}^*, K_e) < H(S|Q, \mathcal{S}^*, K_e) \quad (17)$$

$$\Rightarrow I(S; Q^*|Q, \mathcal{S}^*, K_e) > 0. \quad (18)$$

From the second condition in Theorem 5.3, we know that  $H(Q^*|Q, \mathcal{S}^*, K_e) = 0$  for all possible  $K_e$ . As such, given  $Q$  and  $\mathcal{S}^*$ ,  $S$  can provide no additional

information about  $Q^*$ . That is,

$$I(S; Q^*|Q, \mathcal{S}^*, K_e) \leq H(Q^*|Q, \mathcal{S}^*, K_e) = 0. \quad (19)$$

This is contradictory to (18). Thus, algorithms satisfying both conditions in Theorem 5.3 always publish ASAP tables.  $\square$

Intuitively, this sufficient condition states that if anyone who has access to the data publishing algorithm, the published table and the original  $Q$  can simulate<sup>3</sup> the perturbation of QI without consulting any additional (unpublished) QI-SA correlation, then the QI-perturbation process is immune from algorithm-based disclosure. For example, Anatomy[10] satisfies Theorem 5.3 because 1) Anatomy uses only SA values in the partitioning (i.e., QI perturbation), and 2) it does not perturb SA at all. Thus, Anatomy is immune from algorithm-based disclosure.

In analogy, we propose the second sufficient condition by also addressing SA perturbation.

**Theorem 5.4.** (*ASAP sufficient condition 2*) *The published table of a data publishing algorithm fulfills ASAP if the algorithm satisfies both of the following two conditions: 1) the QI perturbation depends on no information beyond the original QI; and 2) the SA perturbation depends on no information beyond the published QI-SA correlation.*

*Proof.* The proof is in analogy to that of Theorem 5.3.  $\square$

One can see that the intuitive explanation of this sufficient condition is also similar to that of sufficient condition 1 - It again assures ASAP if anyone who has access to the data publishing algorithm, the published table and the original  $Q$  can simulate the perturbation process - this time on both QI and SA. Therefore, in the rest of this paper, we say a data publishing algorithm is *simulatable* iff the algorithm satisfies either Theorem 5.3 or Theorem 5.4.

Nevertheless, it is important to point out that neither Theorem 5.3 nor Theorem 5.4 is necessary (although sufficient) for publishing ASAP tables.

---

<sup>3</sup>Note that such a simulation is *not* a duplication of the data publishing process. When the data publishing algorithm is non-deterministic, one can simulate the randomized part by using the same random number generator, but does not have to generate the same random number.

To explain it, consider a simple algorithm to guarantee  $\ell$ -diversity - by generalizing QI using existing algorithms such as Mondrian [8] and suppressing all SA from the published table. This algorithm certainly satisfies ASAP because no QI-SA correlation is ever disclosed. Nonetheless, it violates the first condition of Theorem 5.3 because SA is suppressed (i.e., perturbed), and violates the first condition of Theorem 5.4 because the perturbation of QI consults unpublished QI-SA correlation (see Section 5.1).

## 6. Eliminating Algorithm-based Disclosure

This section introduces two amendment tools: global look-ahead and local look-ahead, each of which on its own suffices to eliminate algorithm-based disclosure from vulnerable algorithms. To demonstrate the power of global look-ahead, we apply it to alter the design of two popular  $\ell$ -diversity algorithms: Mondrian [8] and Hilb [11], and prove the revised algorithm to be satisfy Theorem 5.3. As to demonstrate the power of local look-ahead, we transform the  $m$ -confidentiality algorithm MASK [12] to a simulatable algorithm by proving that the revised algorithm satisfies Theorem 5.4.

### 6.1. A Running Example

For the ease of understanding, we use the following simple microdata table as a running example throughout this section: There are 8 tuples with two QI attributes  $x$  and  $y$  and one SA. Figure 1 depicts a 2-dimensional visualization of the tuples on an  $x$ - $y$  plane. In the figure, each tuple is represented by a circle, with its  $x$  and  $y$  coordinates indicating its values of QI attributes  $x$  and  $y$ , respectively, and its painted pattern indicating the SA value. Each pattern is corresponding to a distinct SA value. Thus, there are 6 different SA values in our example.

### 6.2. Global Look-Ahead

The global look-ahead is mainly to enable a data publishing algorithm to generate a QI-perturbation (e.g., partition of tuples into QI-indistinguishable groups) only if a worst-case (i.e., most skewed) scenario of QI-SA correlation is able to achieve the predefined privacy model such as  $\ell$ -diversity. The global look-ahead guarantees that no information beyond unpublished QI-SA correlation information will be used. To illustrate it, we take two existing  $\ell$ -diversity algorithms, i.e., Mondrian and Hilb, for example.

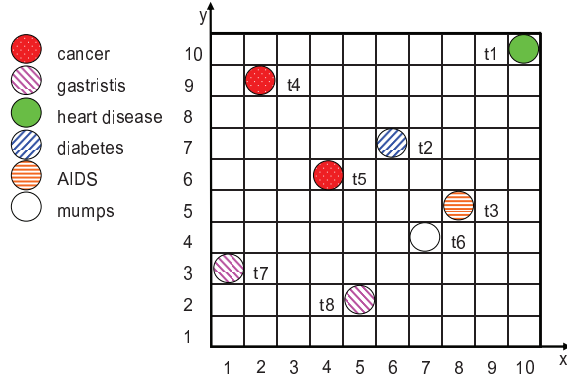


Figure 1:  $x$ - $y$  plane of 8 tuples

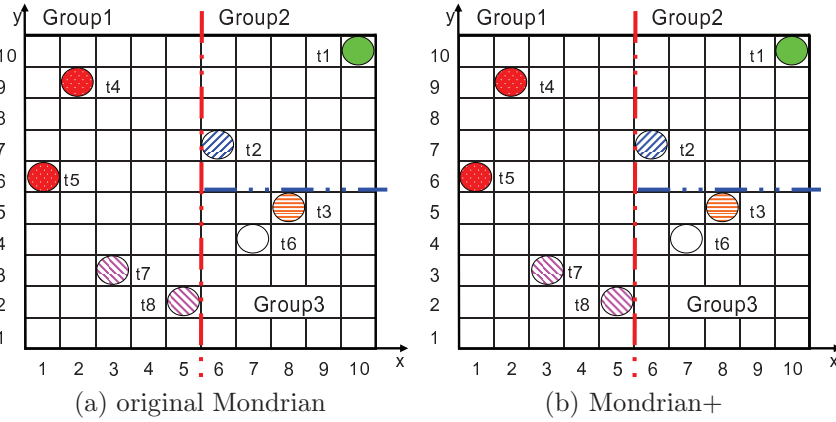


Figure 2: Fix Mondrian  $\ell$ -diversity ( $\ell = 2$ )

### 6.2.1. From Mondrian To Simulatable Mondrian+

We first review the original Mondrian [8], and then discuss how to transform it via global look-ahead to Mondrian+, a simulatable data publishing algorithm.

The original Mondrian works in a recursive fashion. A simple implementation of it begins with selecting a split attribute that has the largest range of value. Alternatively, strategies discussed in [9] can be used in the selection as well. After that, Mondrian repetitively partitions  $G$  (initially  $T$ ) into two groups  $G_1$  and  $G_2$ , where  $G_1$  and  $G_2$  include the tuples of  $G$  divided by the median coordinate on the split attribute. Let  $|\cdot|$  be the number of tuples

in the set, and  $S_{\max}(\cdot)$  be the number of tuples with the most frequent SA value in the set. If either  $|G_1| < \ell \times S_{\max}(G_1)$  or  $|G_2| < \ell \times S_{\max}(G_2)$  holds, such partitioning trial has to be *revoked*.

Figure 2a is an example of the original Mondrian algorithm. Suppose the algorithm first chooses  $x$  as the split attribute with the largest range, and partitions the 8 tuples by the median coordinate ( $x = 5$ ) into two QI-groups:  $G_1 = \{t_4, t_5, t_7, t_8\}$  and  $G_2 = \{t_1, t_2, t_3, t_6\}$ . Both  $G_1$  and  $G_2$  satisfy 2-diversity because  $|G_1| = 4 \geq \ell \times S_{\max}(G_1) = 2 \times 2 = 4$  and  $|G_2| = 4 \geq \ell \times S_{\max}(G_2) = 2 \times 1 = 2$ . For the same reason,  $G_2$  is further partitioned and published as two QI-groups *Group2* and *Group3* (see Figure 2a).

Unlike  $G_2$ ,  $G_1$  has to be published as is. The reason is that regardless of which median coordinate to be chosen (i.e.,  $y = 3$  or  $x = 2$ ), any further trial on partitioning has to be revoked due to violating 2-diversity. Take the case of  $y = 3$  as the median coordinate in  $G_1$  for example. Both QI-groups  $g_1 = \{t_4, t_5\}$  and  $g_2 = \{t_7, t_8\}$  violate 2-diversity because  $|g_1| = 2 < \ell \times S_{\max}(g_1) = 4$  and  $|g_2| = 2 < \ell \times S_{\max}(g_2) = 4$ . Note that the information of  $S_{\max}(g_1) = 2$  and  $S_{\max}(g_2) = 2$  consulted by Mondrian at this point, cannot be recovered in the ultimately published but un-partitioned  $G_1 = \{t_4, t_5, t_7, t_8\}$  (i.e., *Group1* in Figure 2a). Hence, the problem of the original Mondrian is to use unpublished QI-SA correlation information in the QI perturbation, more specifically when the partitioning has to be revoked.

To fix the problem, we follow the idea of global look-ahead to alter Mondrian to Mondrian+ with minor change. In particular, Mondrian+ revokes a partitioning from  $G$  into  $\{G_1, G_2\}$  only if either  $|G_1| < \ell \times S_{\max}(G)$  or  $|G_2| < \ell \times S_{\max}(G)$  holds. In other words, Mondrian+ allows a partitioning when each QI-group generated from the partitioning is able to satisfy  $\ell$ -diversity in the worst-case scenario looked ahead by  $S_{\max}(G)$ , i.e., when all  $S_{\max}(G)$  tuples with the most frequent SA value are all partitioned to the same QI-group. The reason is, for each  $G_1$ , we have:

$$\frac{S_{\max}(G_1)}{|G_1|} \leq \frac{S_{\max}(G)}{|G_1|} \leq \frac{S_{\max}(G)}{\ell \cdot S_{\max}(G)} = \frac{1}{\ell}. \quad (20)$$

For the same reason,  $\frac{S_{\max}(G_2)}{|G_2|} \leq \frac{1}{\ell}$  holds as well.

Figure 2b illustrates our Mondrian+ algorithm. First, Mondrian+ manages to generate two QI-groups  $G_1 = \{t_4, t_5, t_7, t_8\}$  and  $G_2 = \{t_1, t_2, t_3, t_6\}$  after acting the very same first partitioning as the original Mondrian does in Figure 2a. The reason is both  $G_1$  and  $G_2$  can achieve 2-diversity even

if the worst-case happens, i.e.,  $|G_1| = 4 \geq \ell \times S_{\max}(G) = 2 \times 2 = 4$  and  $|G_2| = 4 \geq \ell \times S_{\max}(G) = 2 \times 1 = 2$ . However,  $G_1$  is published as is without any further partitioning, because the size of any QI-group  $g$  partitioned from  $G_1$  has to be less than  $|G_1| = 4$  and thus  $|g| \geq \ell \times S_{\max}(G_1) = 2 \times 2 = 4$  cannot be satisfied. In other words, it is impossible to achieve 2-diversity in the worst-case scenario. Unlike the original Mondrian at this time point, the QI-SA correlation, i.e.,  $S_{\max}(G_1) = 2$  consulted by Mondrian+ can always be recovered from the published data. For example, we can deduce it from counting the SA values in the published *Group1*. Following the same fashion, it is easily verified that  $G_2$  would be partitioned by Mondrian+ and published as *Group2* and *Group3*.

It can be easily proved that Mondrian+ follows Theorem 5.3. First, the first condition of Theorem 5.3 is automatically satisfied because Mondrian+ only perturbs QI without the SA perturbation. Likewise, the second condition holds because the QI perturbation in Mondrian+ only uses the original QI attributes and the published SA (i.e.,  $S_{\max}(\cdot)$ ). Therefore, we have the following theorem:

**Theorem 6.1.** *Mondrian+  $\ell$ -diversity algorithm is simulatable.*

Algorithm 1 details the steps of Mondrian+ with minor change against the original Mondrian. Line 4 implements the idea of global look-ahead.  $\text{MONDRIAN}(G, k)$  in Line 5 denotes a function that invokes  $k$ -anonymity Mondrian algorithm [8] on the dataset  $G$ .  $\text{MONDRIAN}(G, k)$  can be replaced with other  $k$ -anonymity algorithms (e.g., K-OPTIMIZE [26], Datafly [29]). The time complexity is  $O(n(\log n)^2)$ , where  $n$  is the number of tuples in the microdata table  $T$ . In particular, the number of iterations from Line 2 to Line 11 is at most  $O(\log n)$ . Following the analysis in [8], Line 5 takes  $O(n \log n)$  time. Hence, the overall time complexity is  $O(n(\log n)^2)$ .

### 6.2.2. From Hilb to Simulatable Hilb+

We now discuss how to transform the state-of-the-art  $\ell$ -diversity algorithm Hilb [11] to be simulatable via the global look-ahead. The original Hilb works as follows. First, it transforms by Hilbert-curve the multi-dimensional QI space of a microdata table  $T$  to a single-dimensional space  $Q_T$ . Based on  $Q_T$  values, Hilb sorts all tuples in  $T$  in an ascending order, and bucketizes all the ordered tuples in terms of their SA values. Figure 3a shows a simple example of the original Hilb. Suppose  $t_1 \rightarrow t_8$  is the ascending order. All

---

**Algorithm 1** Simulatable Mondrian+ algorithm
 

---

```

1:  $QIGroup \leftarrow \emptyset$ .  $InputSet \leftarrow \{T\}$ .
2: repeat
3:    $G \leftarrow$  the largest group in  $InputSet$ .
4:   if  $|G_1| \geq \ell \times S_{\max}(G)$  &&  $|G_2| \geq \ell \times S_{\max}(G)$  then
5:      $\{G_1, G_2\} \leftarrow$  MONDRIAN( $G, \ell \times S_{\max}(G)$ ).
6:      $QIGroup \leftarrow \{QIGroup \setminus G\} \cup \{G_1, G_2\}$ .
7:   else
8:      $InputSet \leftarrow InputSet \setminus G$ .
9:      $QIGroup \leftarrow QIGroup \cup G$ .
10:  end if
11: until  $InputSet = \emptyset$ .
12: return  $QIGroup$ .

```

---

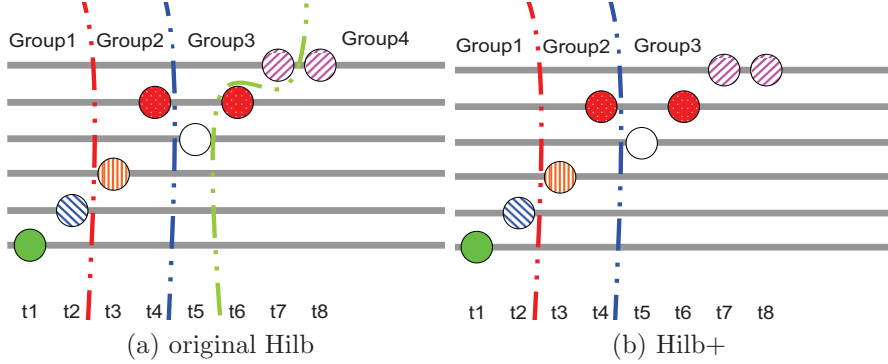


Figure 3: Fix Hilb  $\ell$ -diversity ( $\ell = 2$ )

the 8 tuples are bucketized into 6 buckets (because there are 6 different SA values) and ordered ascendingly from  $t_1$  to  $t_8$  based on their  $Q_T$  values.

Second, Hilb greedily splits out one QI-group  $G_1$  each time by picking up  $|G_1|$  (initially  $\ell$ ) tuples from distinct  $|G_1|$  buckets with lowest  $Q_T$  in  $G$  (initially  $T$ ), and progressively increments  $|G_1|$  by one when  $|G| - |G_1| < \ell \times S_{\max}(G \setminus G_1)$ . The previous greedy step has to stop when  $|G_1| > m$  where  $m$  is the number of buckets. At this point, Hilb enters the roll-back step by restoring  $|G_1|$  to  $\ell$ , runs in a similar fashion to Anatomy [30], i.e., picking up  $|G_1|$  tuples from the  $|G_1|$  largest buckets (in terms of the number of remaining tuples in the bucket), and produces the QI-group  $G_1$  if  $|G| - |G_1| \geq$

$\ell \times S_{\max}(G \setminus G_1)$ . Such a QI-group  $G_1$  at the roll-back step can be proved to always exist by incrementing  $|G_1|$  progressively by one each time. Once such  $G_1$  is found, Hilb generates it and returns to the first greedy step if there are any tuples left in the buckets.

Figure 3a illustrate an example of Hilb. Hilb starts with splitting out  $G_1 = \{t_1, t_2\}$  with the lowest  $Q_T$  in  $G = T$  from 2 distinct buckets.  $G_1$  is published as *Group1* because the remaining tuples can achieve 2-diversity, i.e.,  $|G| - |G_1| = 6 \geq \ell \times S_{\max}(G \setminus G_1) = 2 \times 2 = 4$  holds. However, after *Group2* =  $\{t_3, t_4\}$  is generated for the same reason, Hilb cannot split out any QI-group from the remaining  $G = \{t_5, t_6, t_7, t_8\}$  in the greedy step, even by incrementing  $|G_1|$  progressively until  $|G_1|$  reaches  $m = 6$ . The reason is that when  $|G_1| < 6$ , it is impossible to find a QI-group  $G_1$  with the lowest  $Q_T$  such that  $|G| - |G_1| \geq \ell \times S_{\max}(G \setminus G_1)$ . When  $|G_1|$  is incremented to 6, Hilb enters the roll-back by restoring  $|G_1| = 2$ . It locates  $|G_1|$  distinct buckets that have the largest number of tuples, and generates  $\{t_5, t_7\}$  as to be *Group3* picked from these buckets. The remaining  $\{t_6, t_8\}$  is then published as *Group4*.

As we can see, the problem of Hilb is to consult the unpublished QI-SA correlation (i.e.,  $S_{\max}(G \setminus G_1)$ ) at the time of incrementing  $|G_1|$ , as well as at the time of transiting from the greedy to the roll-back step when  $|G_1| = m$ . Let us focus on  $G = \{t_5, t_6, t_7, t_8\}$  and simply consider the greedy step when  $|G_1| = 3$ , for example.  $S_{\max}(G \setminus G_1) = S_{\max}(\{t_8\}) = 1$  is the only information that is consulted by Hilb to fail the trial of splitting out  $G_1 = \{t_5, t_6, t_7\}$  and to drive incrementing  $|G_1|$  to 4. However,  $S_{\max}(\{t_8\}) = 1$  can never be recovered from the published QI-groups in Figure 3a. Therefore, like Mondrian, Hilb violates the second condition of Theorem 5.3 while the first condition is automatically satisfied.

Our basic idea of altering Hilb to Hilb+ by using the global look-ahead is to split out a QI-group  $G_1$  from  $G$  only when the remaining tuples  $G \setminus G_1$  can achieve  $\ell$ -diversity in the worst-case scenario, i.e.,  $|G| - |G_1| \geq \ell \times S_{\max}(G)$ . Instead of using  $S_{\max}(G \setminus G_1)$ ,  $S_{\max}(G)$  can be recovered in the published data, no matter whether  $G_1$  is ultimately split out or not. Figure 3b illustrates the procedure Hilb+. *Group1* and *Group2* are generated sequentially in the same fashion because  $|G| - |G_1| = 8 - 2 = 6 \geq \ell \times S_{\max}(G) = 2 \times 2 = 4$  and  $|G| - |G_1| = 6 - 2 = 4 \geq \ell \times S_{\max}(G) = 4$  is satisfied, respectively. In other words, the remaining tuples after *Group1* (or *Group2*) is split out can satisfy 2-diversity in the worst-case. Unlike Hilb in Figure 3a, Hilb+ chooses to publish the entire  $G = \{t_5, t_6, t_7, t_8\}$  as *Group3*, rather than split out any  $G_1$  from it. The reason is that the remaining tuples cannot satisfy

2-diversity in the worst-case, i.e.,  $(|G| - |G_1|) = 2 < \ell \times S_{\max}(G) = 2 \times 2 = 4$ . Note that at this point, Hilb+ does not need to greedily repeat the process by incrementing  $|G_1|$ , because there is no  $|G'_1|$  such that  $|G'_1| > |G_1|$  and meanwhile  $(|G| - |G'_1|) \geq \ell \times S_{\max}(G)$  can be satisfied. Clinging to  $S_{\max}(G)$  to look-ahead a worst-case scenario, Hilb+ is able to satisfy the second condition as well as the first condition of Theorem 5.3. Therefore, we can have the following theorem:

**Theorem 6.2.** *Hilb+  $\ell$ -diversity algorithm is simulatable.*

---

**Algorithm 2** Simulatable Hilb+ algorithm

---

```

1:  $QIGroup \leftarrow \emptyset$ .  $G \leftarrow \{T\}$ .
2: Apply Hilbert curve to transform multi-dimensional QI space of  $G$  into
   1-D dimensional space  $Q_T$ . Sort all the tuples in  $G$  in ascending order of
    $Q_T$ .
3: Split sorted tuples in  $m$  buckets based on SA values.
4: frontier  $\mathcal{F} \leftarrow$  set of first record in each bucket.
5: repeat
6:    $|G_1| \leftarrow \ell$ .
7:   if  $(|G| - |G_1|) < \ell \times S_{\max}(G)$  then
8:      $|G_1| \leftarrow |G|$ .
9:   end if
10:   $G_1 \leftarrow$  set of  $|G_1|$  tuples of  $\mathcal{F}$  with lowest  $Q_T$ .
11:   $G \leftarrow G \setminus G_1$ .
12:  Update  $\mathcal{F}$ .
13:   $QIGroup \leftarrow QIGroup \cup G_1$ .
14: until  $G = \emptyset$ .
15: return  $QIGroup$ .
```

---

Algorithm 2 describes the details of Hilb+  $\ell$ -diversity algorithm. Line 1-Line 4 pre-processes the microdata  $T$  by Hilbert curve transformation, sorts and bucketizes tuples as with the original Hilb. Line 6-Line 14 describes the greedy procedure of splitting out a QI-group  $G_1$  from  $G$ . Line 7 implements the global look-ahead. To avoid using the unpublished QI-SA correlation, i.e.,  $S_{\max}(G \setminus G_1)$ , there are no trials to increment  $|G_1|$  or to invoke the roll-back step as explained before. Refer to the analysis of Hilb in [11]. The overall time complexity of Algorithm 2 is at most  $O(n \log n)$ , where  $n$  is the number of tuples in  $T$ .

### 6.3. Local Look-Ahead

Global look-ahead addresses a large class of existing algorithms that perturb QI of the microdata data while leaving SA intact. Local look-ahead, on the other hand, deals with existing algorithms that perturb the SA values before publishing a microdata table.

Before perturbing SA, the existing algorithms (e.g., MASK [12]) need to first check the original SA distribution of a QI-group, in order to determine whether the privacy guarantee (e.g.,  $\ell$ -diversity) is violated. Unfortunately, as we discussed in Section 5.2, when the privacy guarantee is indeed violated by the original distribution, then the checking process itself uses certain QI-SA correlation that cannot be finally published, therefore incurring algorithm-based disclosure. Local look-ahead enables query of the original SA distribution without using any QI-SA information that will not eventually be published. To achieve this, local look-ahead retrieves the frequency of one SA value at a time, *after* ensuring that the retrieval will never lead to a violation of the privacy guarantee even in the worst-case scenario. To make it more concrete, we demonstrate as follows how to use local look-ahead to eliminate algorithm-based disclosure from the  $m$ -confidentiality algorithm MASK [12].

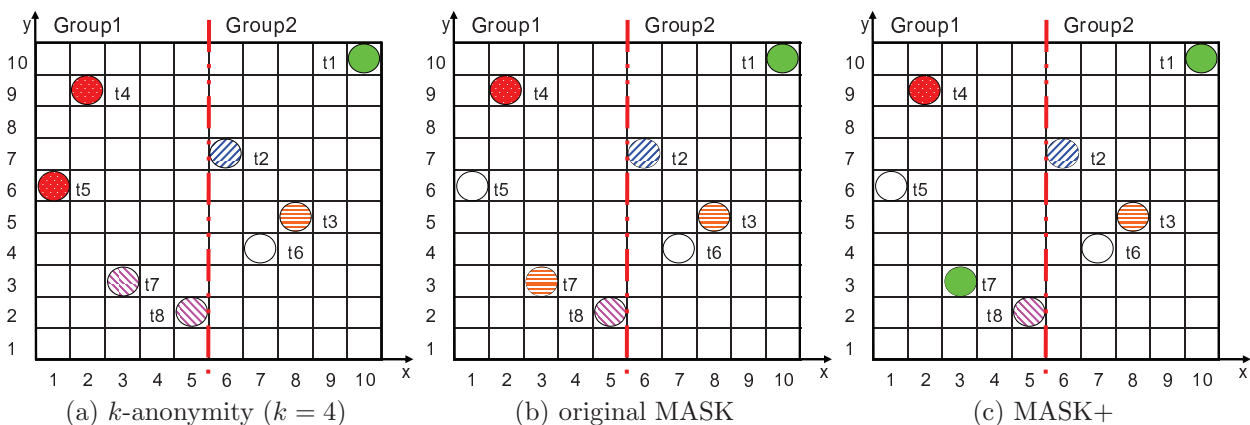


Figure 4: Fix MASK ( $m = 3$ )

We start with a brief review and example of the original MASK algorithm. As discussed in Section 2.2, to achieve  $m$ -confidentiality, MASK first applies an existing  $k$ -anonymization algorithm (e.g., [9, 26, 29]) to partition the

microdata table into a number of  $k$ -anonymous QI-groups, and then perturbs the SA distribution of QI-groups which violate  $\ell$ -diversity with  $\ell = m$ . The parameter  $k$  must be at least  $m$  and can be specified as an input parameter<sup>4</sup>. Figures 4a and 4b illustrate an example of the two steps when  $m = 3$ . After the first step, Group 1 :  $\{t_4, t_5, t_7, t_8\}$  and Group 2 :  $\{t_1, t_2, t_3, t_6\}$  are two QI-groups which satisfy  $k$ -anonymization with  $k = 4$ . In the second step, since only one QI-group, Group 1, violates 3-confidentiality (as  $|\text{Group 1}| = 4 < m \times S_{\max}(\text{Group 1}) = 3 \times 2 = 6$ , where  $S_{\max}(\cdot)$  is the maximum frequency of an SA value in a given QI-group), MASK perturbs the SA distribution of Group 1 to satisfy 3-confidentiality while publishing Group 2 as is, as shown in Figure 4b. One can see that the information of  $S_{\max}(G_1) = 2$ , which was used to decide the perturbation of SA for Group 1, is not eventually published in Figure 4b. Thus, MASK consults unpublished QI-SA correlation and therefore violates the second necessary condition of ASAP (i.e., Theorem 5.4).

To fix the problem, we transform MASK to MASK+ by local look-ahead. A key observation here is that, while deciding whether to perturb the SA distribution of a QI-group  $G$ , one cannot directly query  $S_{\max}(G)$  because, as long as the returned result violates  $m$ -confidentiality, ASAP is already violated. To address this problem, local look-ahead *progressively* queries the frequency of each SA value in an *ascending order* of its frequency (i.e., from the least to the most frequent SA). To avoid reading any SA frequency that violates  $m$ -confidentiality in  $G$ , after each query we “look ahead” a worst-case scenario of the SA frequencies yet to be read, and stop if the worst-case scenario violates  $m$ -confidentiality.

Formally, let  $\mathcal{D}_S$  be the domain of SA and  $f_1, \dots, f_{|\mathcal{D}_S|}$  be the frequency of SA values in an ascending order - i.e.,  $f_i$  is the number of tuples in a given QI-group  $G$  which feature the  $i$ -th least frequent SA value  $s_i \in \mathcal{D}_S$ ,  $i \in [1, |\mathcal{D}_S|]$  (break ties arbitrarily). Let  $f_i = 0$  for any  $s_i$  not appearing in  $G$ . We suppose in the following that  $|\mathcal{D}_S| > m$  holds (as a side note, it is trivial when  $|\mathcal{D}_S| = m$ , because each SA value from  $\mathcal{D}_S$  must occur in each QI-group with relative frequency at exactly  $1/m$ ). Note that MASK+ can always read  $f_1$ , if  $|\mathcal{D}_S| > m$ , because it never violates  $m$ -confidentiality. Otherwise, it implies that all the other SA frequencies violate  $m$ -confidentiality, leading to  $\sum_{i=1}^{|\mathcal{D}_S|} f_i > |G|$ , which is impossible by the fact of  $\sum_{i=1}^{|\mathcal{D}_S|} f_i = |G|$ . Suppose

---

<sup>4</sup>We assume  $\lfloor \frac{|G|}{m} \rfloor \times |\mathcal{D}_S| \geq |G|$  for any  $k$ -anonymous QI-group  $G$  because otherwise MASK cannot achieve  $m$ -confidentiality.

$f_i$  is the next SA frequency to be read. MASK+ reads  $f_i$  if and only if  $(|G| - \sum_{j=1}^{i-1} f_j - f_{i-1}) / (|\mathcal{D}_S| - i) \leq \lfloor |G|/m \rfloor$ . Otherwise, MASK+ switches to perturbing the SA distribution of  $G$  which shall be discussed next. Since  $f_i \geq f_{i-1}$ ,  $(|G| - \sum_{j=1}^{i-1} f_j - f_{i-1}) / (|\mathcal{D}_S| - i)$  essentially defines the worst-case scenario because:

$$f_i \leq \frac{|G| - \sum_{j=1}^{i-1} f_j - f_i}{|\mathcal{D}_S| - i} \leq \frac{|G| - \sum_{j=1}^{i-1} f_j - f_{i-1}}{|\mathcal{D}_S| - i} \leq \left\lfloor \frac{|G|}{m} \right\rfloor \quad (21)$$

As such, no information used by MASK+ implies a violation of  $m$ -confidentiality.

We now illustrate this process by an example. Consider the 4-anonymous Group 1 in Figure 4a. It has the following ascending order of SA frequencies,  $\langle f_{AIDS}, f_{diabetes}, f_{heart\ disease}, f_{mumps}, f_{cancer}, f_{gastritis} \rangle = \langle 0, 0, 0, 0, 2, 2 \rangle$ . MASK+ can always safely read the 1<sup>st</sup> least frequent SA, i.e.,  $f_{AIDS} = 0$ , due to  $|\mathcal{D}_S| > m$ . Also, MASK+ can safely read the 2<sup>nd</sup> least frequent SA,  $f_{diabetes} = 0$ , because the worst-case scenario cannot exceed  $\lfloor |G|/m \rfloor = \lfloor 4/3 \rfloor = 1$  (i.e.,  $(|G| - f_{AIDS} - f_{diabetes}) / (|\mathcal{D}_S| - 2) = 4/4$ ). In other words, there is no chance for AIDS and diabetes in Group 1 to violate 3-confidentiality. Nonetheless, MASK+ will not read the 3<sup>rd</sup> least frequent SA, i.e.,  $f_{heart\ disease}$ , because  $(|G| - f_{AIDS} - f_{diabetes} - f_{diabetes}) / (|\mathcal{D}_S| - 3) = 4/3 > \lfloor 4/3 \rfloor$ , which means that heart disease in Group 1 may violate 3-confidentiality in the worst-case scenario. At this point, MASK+ decides to perturb the SA distribution of Group 1.

We now consider how MASK+ perturbs the SA perturbation. An important requirement here is that all information “read” in local look-ahead must be derivable from the perturbed (and therefore published) SA distribution. Again, let  $\langle f_1, \dots, f_{|\mathcal{D}_S|} \rangle$  be the ascending order of SA frequencies in a QI-group  $G$ . Suppose  $f_i$  is the last SA frequency successfully read by MASK+ before it decides to perturb SA in  $G$ . The QI-group  $G'$  perturbed from  $G$  must satisfy that its top  $i$  least frequent SA values and their frequencies, which is the only information used in SA perturbation, must remain exactly the same as the original  $G$ .

To achieve this, the SA perturbation process can be described as follows. Let  $f'_j$  ( $j \in [1, |\mathcal{D}_S|]$ ) be the published frequency of the SA value corresponding to  $f_j$ . To start, we make no change to  $f_1, \dots, f_i$  and assign the value of  $f_i$  to  $f'_{i+1}, \dots, f'_{|\mathcal{D}_S|}$  - i.e.,  $\langle f'_1, \dots, f'_i, f'_{i+1}, \dots, f'_{|\mathcal{D}_S|} \rangle = \langle f_1, \dots, f_i, f_i, \dots, f_i \rangle$ . If  $|G| - \sum_{k=1}^{|\mathcal{D}_S|} f'_k > 0$  holds, following a descending order from  $|\mathcal{D}_S|$  to  $i + 1$ ,  $f'_j$  is iteratively updated to be  $\min(\lfloor |G|/m \rfloor, |G| - \sum_{k=1}^{|\mathcal{D}_S|} f'_k + f_i)$  un-

til  $|G| - \sum_{k=1}^{|\mathcal{D}_S|} f'_k = 0$ . At this point, the published SA distribution satisfies  $m$ -confidentiality, and retains all the information used by local look-ahead.

Return to the running example of Group 1 in Figure 4a. Recall that MASK+ decides to perturb SA after reading  $f_{AIDS} = 0$  and  $f_{diabetes} = 0$ . In the perturbation process, MASK+ starts with  $\langle f'_{AIDS}, f'_{diabetes}, f'_{heart\ disease}, f'_{mumps}, f'_{cancer}, f'_{gastritis} \rangle = \langle 0, 0, 0, 0, 0, 0 \rangle$ . Since  $|G| - \sum_{k=1}^{|\mathcal{D}_S|} f'_k = 4 > 0$ , it then updates  $f'_{gastritis}$  to be  $\min(\lfloor G/m \rfloor, |G| - \sum_{k=1}^{|\mathcal{D}_S|} f'_k + f_i) = \min(\lfloor \frac{4}{3} \rfloor, 4 - 0 + 0) = 1$ . This procedure is repeated until  $f'_{mumps}$  is updated to 1, after which  $|G| - \sum_{k=1}^{|\mathcal{D}_S|} f'_k = 4 - 4 = 0$ . Figure 4c shows the result of applying MASK+ to Figure 4a. Unlike MASK, the only information used by MASK+ in perturbing SA, i.e.,  $f_{AIDS} = 0$  and  $f_{diabetes} = 0$ , can still be readily learned from Figure 4c.

---

**Algorithm 3** Simulatable MASK+  $m$ -confidentiality algorithm

---

- 1:  $QIGroup \leftarrow \emptyset$ .
  - 2:  $\mathcal{D}_S \leftarrow$  the SA domain of  $T$ .
  - 3:  $G \leftarrow k$ -anonymous groups by applying any  $k$ -anonymization algorithm on  $T$  where  $k$  is defined by the user.
  - 4: **repeat**
  - 5:      $G_1 \leftarrow$  the next  $k$ -anonymous group from  $G$ .
  - 6:      $\langle f_1, \dots, f_{|\mathcal{D}_S|} \rangle \leftarrow$  the ascending SA frequency order in  $G_1$ .
  - 7:      $i \leftarrow 1$
  - 8:     **repeat**
  - 9:          $i \leftarrow i + 1$ .
  - 10:     **until**  $i = |\mathcal{D}_S| - 1$  or  $\frac{|G| - \sum_{j=1}^{i-1} f_j - f_{i-1}}{|\mathcal{D}_S| - i} > \lfloor \frac{|G|}{m} \rfloor$ .
  - 11:     **if**  $i < |\mathcal{D}_S| - 1$  **then**
  - 12:         Perturb SA in  $G_1$  to  $G'_1$  such that the top  $i - 1$  least frequent SA values and their frequencies in  $G_1$  retain exactly the same as in  $G'_1$ .
  - 13:          $G_1 \leftarrow G'_1$ .
  - 14:     **end if**
  - 15:      $QIGroup \leftarrow QIGroup \cup G_1$ .
  - 16: **until**  $G = \emptyset$ .
  - 17: **return**  $QIGroup$ .
- 

Algorithm 3 describes our simulatable MASK+ algorithm altered from MASK to achieve  $m$ -confidentiality.

We now discuss why MASK+ satisfies both conditions of Theorem 5.4.

The satisfaction of the first condition is straightforward because MASK+ only uses the original QI information in its QI perturbation. Nonetheless, we discuss the satisfaction of the second condition as follows.

Note that the SA information used by MASK+, i.e., the real frequencies of a number of SA values as well as the frequency order, is actually published by the algorithm for the following reason: Consider  $m$ -confidentiality as the privacy guarantee. What MASK+ uses in terms of SA information includes two parts: 1) safe-value frequencies: i.e., the real frequency of a number of SA values which MASK+ reads with a “safety” guarantee of not violating  $m$ -confidentiality (we refer to these SA values as Type-1 values), and 2) unsafe-value identities: i.e., the fact that the other SA values (which we refer to as Type-2 values) have frequencies greater than or equal to the frequencies of SA values read by MASK+. According to the SA-perturbation process used by MASK+, both types of information can be readily learned from the published table, as the frequencies of Type-1 SA values remain the same in the published table, while the frequencies of Type-2 SA values are still greater than or equal to the frequencies of all Type-1 SA values. Thus, even the order information used by MASK+ is published in the perturbed table. As such, MASK+ does not violate the conditions set forth by Theorem 5.4. Hence, we have the following theorem:

**Theorem 6.3.** *MASK+  $m$ -confidentiality algorithm is simulatable.*

The time complexity of MASK+ is  $O(n \log n)$ , where  $n$  is the number of tuples in  $T$ . We analyze it as follows. Since MASK+ uses  $k$ -anonymization as a black box, the complexity of the QI perturbation part is  $O(n \log n)$  when a  $k$ -anonymization algorithm such as Mondrian [8] or Hilb [11] is used. The number of iterations (Line 8 - Line 10) is at most  $|V|$  where  $|V|$  is the number of  $k$ -anonymous QI-groups. Line 12 has complexity of  $O(|\mathcal{D}_S|)$ . Since  $n > |V|$  and  $n > |\mathcal{D}_S|$ , the overall time complexity is  $O(n \log n)$ .

## 7. Enhancing Data Utility

Although both tools: global look-ahead and local look-ahead discussed in the previous section suffice to dismiss algorithm-based disclosure, their drawback is that the data utility may be reduced at the cost of enforcing a “more stringent” privacy guarantee. In this section, we introduce stratified pick-up, another tool to enhance utility.

Stratified pick-up takes as input the anonymous QI-groups from any simulatable algorithm and tires to further partition each of these groups greedily based solely on the distinctness of SA values. The design of this phase is principled on two objectives: 1) the algorithm should still be simulatable; and 2) each output QI-group size should be minimized.

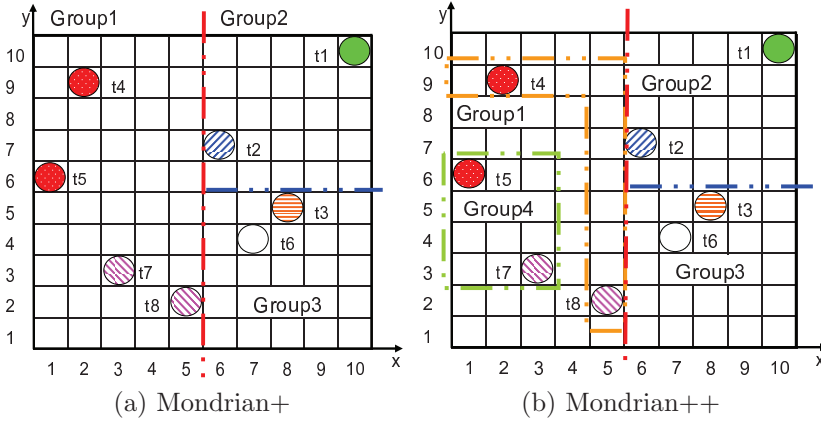


Figure 5: Apply stratified-pick on Mondrian+  $\ell$ -diversity ( $\ell = 2$ )

In particular, a simple solution to achieve these two objectives is to apply Anatomy [10] on each generated QI-group from the simulatable algorithm. Note that as discussed in Section 5.3, Anatomy does not perturb SA values and not use any QI-SA correlation beyond what will be eventually published. Thus, this solution satisfies Theorem 5.3. For example, we apply the stratified pick-up on Group 1 generated from our Mondrian+ algorithm (in Figure 5a). A possible output after performing stratified pick-up on Group 1 is shown in Figure 5b, i.e.,  $\{\{t_4, t_8\}, \{t_5, t_7\}\}$ , which has obviously higher utility than publishing Group 1 as in Figure 5a. Whereas, the generated Group 2 in Figure 5a is already minimized in terms of 2-diversity, and thus stratified pick-up publishes it as is in Figure 5b. Formally, we have the following theorem on the output group size after stratified pick-up.

**Theorem 7.1.** *Each  $\ell$ -diversity QI-group of stratified pick-up contains  $\ell'$  tuples where  $\ell' \in [\ell, 2\ell)$  and each SA value is unique.*

Details of stratified pick-up are shown in Algorithm 4. We test the condition  $\ell \leq \frac{|G|}{2}$  in Line 6 because if a  $\ell$ -diversity group  $G$  has size less than  $2\ell$ ,

---

**Algorithm 4** Stratified Pick-up

---

```
1:  $QIGroup \leftarrow \emptyset$ .
2:  $InputSet \leftarrow$  anonymous groups from any simulatable algorithm.
3: repeat
4:    $G \leftarrow$  the next anonymous group from  $InputSet$ .
5:    $InputSet \leftarrow InputSet \setminus G$ .
6:   if  $\ell \leq \frac{|G|}{2}$  then
7:      $\{g_1, \dots, g_p\} \leftarrow \text{ANATOMY}(G, \ell)$ 
8:      $QIGroup \leftarrow QIGroup \cup \{g_1, \dots, g_p\}$ .
9:   else
10:     $QIGroup \leftarrow QIGroup \cup G$ .
11:  end if
12: until  $InputSet = \emptyset$ .
13: return  $QIGroup$ .
```

---

$G$  must have  $|G|$  distinct SA values and cannot be further partitioned. Such minimized group  $G$  can be directly added to the output (Line 10).

**Theorem 7.2.** *If the algorithm which generates the input to stratified pick-up is simulatable, then the algorithm with stratified pick-up is still simulatable.*

The efficiency of stratified pick-up depends on Anatomy. Following the results from [10], the time complexity of stratified pick-up is  $O(n)$  and the I/O cost is  $O(\lambda)$ , where  $n$  is the total number of tuples and  $\lambda$  is count of distinct SA values.

---

**Algorithm 5** Mondrian++

---

```
1:  $DoneSet \leftarrow \emptyset$ .  $InputSet \leftarrow \{T\}$ .
2:  $DoneSet \leftarrow \text{MONDRIAN+}(InputSet, \ell)$ 
3:  $DoneSet \leftarrow \text{STRATIFIED\_PICKUP}(DoneSet, \ell)$ 
4: return  $DoneSet$ .
```

---

Take Mondrian as an example. Algorithm 5 and Algorithm 6 details a hybrid algorithm of integrating our two tools: (global) look-ahead (Line 2) (i.e., Algorithm 1) and stratified pick-up (Line 3) (i.e., Algorithm 4). The only difference is their usage of generalization and bucketization publishing schemes, respectively. In the same fashion, it is easy to develop a hybrid version: Hilb++ and MASK++. We will test them in the next section.

---

**Algorithm 6** Mondrian++ (in bucketization scheme)

---

- 1:  $DoneSet \leftarrow \emptyset$ .  $InputSet \leftarrow \{T\}$ .
  - 2:  $DoneSet \leftarrow \text{MONDRIAN+}(InputSet, \ell)$
  - 3:  $DoneSet \leftarrow \text{STRATIFIED\_PICKUP}(DoneSet, \ell)$
  - 4: Publish the original QI (without any generalization) for each group in  $DoneSet$
  - 5: **return**  $DoneSet$ .
- 

Table 6: The attributes and its domains in our experiment

(a) Adult dataset				(b) Census dataset			
attribute	domain size	Type	Height	attribute	domain size	Type	Height
age	74	ranges-5, 10, 20	4	age	79	numerical	-
work class	7	taxonomy tree	3	gender	2	suppression	1
marital status	7	taxonomy tree	3	education	17	numerical	-
occupation	14	taxonomy tree	2	marital status	6	taxonomy tree	3
race	5	taxonomy tree	2	race	9	taxonomy tree	2
sex	2	suppression	1	work class	10	taxonomy tree	4
country	41	taxonomy tree	3	country	83	taxonomy tree	3
salary	2	suppression	1	occupation	50	sensitive attr.	-
education	16	sensitive attr.	-				

## 8. Experiment

In this section, we describe our experimental setup, compare the data utility of our simulatable algorithms with the existing  $\ell$ -diversity algorithms, evaluate the impact of our two tools: global/local look-ahead and stratified pick-up, and evaluate the extent of algorithm-based disclosure in MASK and Hilb, respectively.

### 8.1. Experimental Setup

#### 8.1.1. Hardware

All experiments were conducted on a machine with Intel Core 2 Duo 2.6GHz CPU with 2GB RAM and Windows XP OS. All our algorithms were implemented using C++.

#### 8.1.2. Datasets

We conducted the experiments on two datasets: Adult, from UCI Machine Learning Repository and Census, from <http://ipums.org>, which have been extensively used as benchmarks in the literature. For the Adult dataset,

we removed all tuples with missing values to obtain a set of 45,222 tuples. For the Census dataset, we followed the procedure in [10] to sample 300,000 tuples without replacement as our testing bed. Their schemas are summarized in Table 6.

### 8.1.3. Utility Measure

We adopted the same relative error measure proposed in [10]. Consider query workload of the form:

```
SELECT COUNT(*) FROM Dataset
WHERE pred( $Q_1$ ), ..., pred( $Q_{qd}$ ), pred( $S$ )
```

where  $qd$  is the *query dimension* and *pred*( $Q_i$ ) (resp. *pred*( $S$ )) denotes the predicate of  $Q_i$  (resp.  $S$ ) belonging to a range of randomly generated values in its domain. The cardinality of the range is determined by a parameter called *selectivity*. Let *Act* and *Est* be the query result from the microdata table  $T$  and published table  $T^*$ , respectively. The *relative error* is defined as  $|Act - Est|/Act$ . For each set of experiments, we ran a workload of 10000 queries, and calculated the average relative error as the utility measure.

## 8.2. Evaluation of MASK++

We used the same settings as [12]: Adult dataset with attribute *Education* as SA, and all values below high school, i.e., “preschool”, “1st-4th”, “5th-6th” and “7th-8th”, are sensitive. The total number of sensitive tuples is 1566. MASK features two main parameters  $k$  and  $m$  for its two steps:  $k$ -anonymization and SA perturbation, respectively. We tested varying values of  $k \in [5, 8]$  and  $m \in [2, 5]$ .

Before evaluating our MASK++, we first tested the extent of algorithm-based disclosure in MASK [12]. Then, we compared our MASK++ (integrated with both local look-ahead and stratified pick-up) against the original MASK [12]. For the fairness of comparison, we did not compare MASK++ with other data publishing algorithms because the authors’ implementation <http://www.cse.ust.hk/~raywong/code/cred.zip> does not address the case when all the SA values are sensitive.

### 8.2.1. Algorithm-based Disclosure of MASK

Consider a simple attack based on a negative association rule “few (e.g., less than 10%) people under or at the age of 25 have Ph.D. degrees”. This can be drawn either intuitively, or from many educational surveys. Recall from Section 2 that MASK incurs algorithm-based disclosure if a published

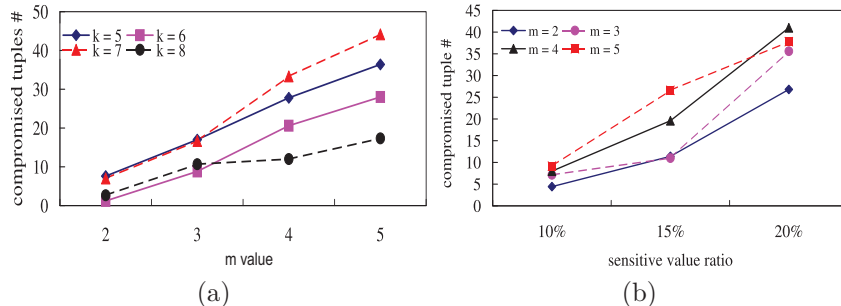


Figure 6: Adult dataset, algorithm-based disclosure of MASK

group violates such rule because, in that case, an adversary can infer that any individual in that group must have probability over  $\frac{1}{m}$  to have a sensitive SA value. As such, our attack can be summarized as follows: if any “Age  $\leq 25$ ” group in the published table has over 10% SA values of “Ph.D.”, we label such a group as problematic group and infer that it originally violates  $m$ -confidentiality. That is, the education level of each individual in that group can be inferred more than  $\frac{1}{m}$  probability to be below high school. We refer to this attack as PhD-25 attack.

In our experiments, we found that the PhD-25 attack never generates any false positive. That is, it never mislabels a group without perturbing SA to be problematic. Thus, the confidence of PhD-25 attack on predicting a violation of  $m$ -confidentiality is 100%. For the recall of this attack, we tested the number of tuples compromised by PhD-25 attack.

First, we tested the extent of algorithm-based disclosure by varying the two parameters of MASK:  $k$  and  $m$ . Figure 6a shows the number of compromised tuples when  $k \in [5, 8]$  and  $m \in [2, 5]$ . One can see that, given  $k$ , the number of compromised tuples increases with  $m$ . This is because a larger  $m$  requires more groups to have SA perturbed after  $k$ -anonymization, thus increasing the probability of “Ph.D.” being added to an “Age  $\leq 25$ ” group. As a result, more tuples are compromised under the PhD-25 attack. On the other hand, given  $m$ , a larger  $k$  does not necessarily increase the number of compromised tuples. This is because the value of  $k$  is generally independent of the probability for a QI-group to violate  $m$ -confidentiality.

Second, we tested the extent of algorithm-based disclosure by varying the percentage of sensitive tuples in the microdata table (through stratified sampling with replacement). Figure 6b depicts the number of compromised

tuples when the percentage ranges from 10% to 20%. We set  $k = 5$  and  $m \in [2, 5]$ . One can see from the figure that more tuples will be compromised when the percentage of sensitive tuples increases. That is because, with more sensitive tuples, more groups are likely to violate  $m$ -confidentiality, leading to more “Ph.D.” being added to an “Age  $\leq 25$ ” group.

### 8.2.2. Utility Comparison with MASK

We illustrated the results in generalization scheme from Figure 7a to 7e, and the results in bucketization scheme from Figure 7f to 7j, respectively. In particular, Figure 7a investigates the tradeoff between the query accuracy and  $k$ -anonymity value when fixing  $m$ -confidential value  $m = 3$ , number of QI  $qi = 8$ , query dimension  $qd = 3$  and selectivity  $s = 5\%$ . Without changing the values of other parameters, Figure 7b sets  $k = 50$  and explores the query accuracy by varying  $m$ . Fixing  $k = 50$  and  $m = 3$ , Figure 7c to 7e studies the query accuracy by varying the number of QI  $qi$ , by varying query dimension  $qd$ , and by varying selectivity  $s$ , respectively. We repeated the same evaluation in bucketization scheme by using the same configuration (Figure 7f to 7j).

All the above figures show that our MASK++ outperforms the original MASK in terms of utility. Nonetheless, when  $k$  is close enough to the  $m$ -confidential value, e.g.,  $k = 3$  or  $5$  in Figure 7a and 7f, the superiority of MASK++ over the original algorithm becomes less significant (or even disappears in the case of bucketization scheme). To explain it, smaller  $k$  generates more small-sized  $k$ -anonymous groups, and thus the effect of our stratified pick-up is weakened. Another interesting point is that the superiority of MASK++ is affected when  $m$  is larger (see Figure 7b and 7g). The reason is that as discussed in Section 6.3, local look-ahead defines a more “stringent” condition with larger  $m$ , and thus our MASK++ tends to perturb SA in more  $k$ -anonymous groups, leading to lower query accuracy.

### 8.2.3. Effects of Local look-ahead and Stratified Pick-up

As we evaluated in Section 8.3.4, Figure 7k demonstrates the individual effect of local look-ahead and stratified pick-up. We set  $qi = 8$ ,  $qd = 3$ ,  $sel = 5\%$  and performed the evaluation on  $k = 10$ , and  $k = 50$ , respectively. As expected, when  $k = 10$ , the query accuracy of MASK+ (i.e., integrated only with local look-ahead) is comparable (no better than) the original algorithm. However, MASK++ (i.e., integrated both local look-ahead and



Figure 7: Adult dataset, MASK++. (a)-(e) generalization scheme; (f)-(j) bucketization scheme; (k) Effect of local look-ahead and stratified pick-up

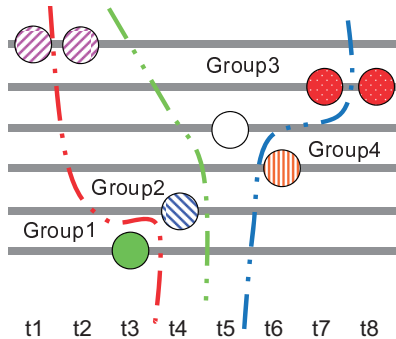


Figure 8: A simple attack on Hilb (2-diversity)

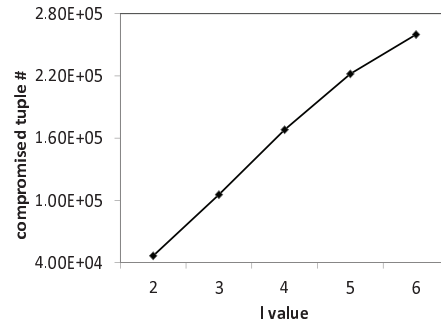


Figure 9: Census data, algorithm-based disclosure of Hilb.

stratified pick-up) achieves better utility than the original MASK. Likewise, we observed the same pattern from the case when  $k = 50$ .

#### 8.2.4. Time performance

Table 7 shows the time performance of our MASK++ against the original MASK, by fixing  $k = 50$ ,  $qi = 8$ ,  $qd = 3$ ,  $sel = 5\%$ . At the cost of eliminating algorithm-based disclosure, MASK++ incorporates local look-ahead, which defines a more “stringent” condition in SA perturbation, leading to SA perturbation in more QI-groups than the original MASK. Therefore, MASK++ runs slower than MASK but their performances are still comparable.

Table 7: Time performance of MASK++

	MASK	MASK++	MASK++ (in bucketization)
time (in seconds)	65.7	66.3	66.3

### 8.3. Evaluation of Mondrian++ and Hilb++

In this set of experiment, we used another popular dataset - Census following the same setting as in [11, 10]. We first tested the extent of algorithm-based disclosure of Hilb, the state-of-the-art algorithm for  $\ell$ -diversity. Then, we evaluated Mondrian++ and Hilb++, our simulatable algorithms (with both global look-ahead and stratified pick-up) against the original Mondrian [9] and Hilb [11] in generalization scheme. For the fairness of comparison, we also considered Mondrian++ and Hilb++ in bucketization scheme (see

Section 7), and compared them with the existing simulatable publishing algorithm Anatomy [10]. We then demonstrated the effect of our two tools: global look-ahead and stratified pick-up individually. Finally, we tested the efficiency.

### 8.3.1. Algorithm-based Disclosure of Hilb

We illustrate a simple attack on Hilb by using the following Figure 8 as an example. Figure 8 illustrates 4 QI-groups produced by a 2-diversity Hilb algorithm.  $t_1$  to  $t_8$  represent a complete list of ascending-ordered tuples for a Hilbert curve defined on the QI-space (see Section 6.2.2 for the algorithm details of Hilb).

Observe from all published QI-groups that the maximum frequency of an SA value in  $\{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8\}$  is 2, i.e.,  $S_{\max}(\{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8\}) = 2$ . Thus, we are able to infer that  $t_1$  and  $t_2$  must have the same SA value because otherwise,  $\{t_1, t_2\}$ , with lower information loss than  $\{t_1, t_3\}$ , would have formed *Group1* because the remaining tuples  $\{t_3, t_4, t_5, t_6, t_7, t_8\}$  would satisfy 2-diversity anyway (as  $S_{\max}(\{t_3, t_4, t_5, t_6, t_7, t_8\}) = S_{\max}(\{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8\}) = 2$ ). With the intersection of the SA values between published *Group1* and *Group2*, we can safely conclude that  $t_1$  and  $t_2$  must have SA value “gastritis”. Similarly, we can also derive that  $t_3$  and  $t_4$  must have “heart disease” and “diabetes”, respectively. Hence, the number of compromised tuples in our example is 4.

A more detailed description of our attack is given in Algorithm 7. The main idea of our attacking is based on the correlation between QI-groups generated by deterministic grouping (i.e., without performing the fall-back step) of Hilb. Line 1 to Line 2 sorts the published QI-groups ascendingly by the respective lowest  $Q_T$  of their members. For example, the published 4 QI-groups in Figure 8 are sorted ascendingly as follows:  $\{t_1, t_3\} \prec_{Q_T} \{t_2, t_4\} \prec_{Q_T} \{t_5, t_7\} \prec_{Q_T} \{t_6, t_8\}$ , where  $\prec_{Q_T}$  denotes the partial order.

Line 5 suffices to guarantee that all the QI-groups attacked by our algorithm are generated by deterministic grouping. As explained in Section 6.2.2, the remaining tuples (i.e.,  $G \setminus G_1$ ) after generating  $G_1$  can achieve  $\ell$ -diversity in the worst-case scenario, and thus randomized grouping would never be initiated by Hilb due to its utility concern. For the same reason, we can easily verify that *Group1* and *Group2* in Figure 8 must be generated by deterministic grouping.

Observe from Hilb that given any deterministic QI-group  $G_1$ , the SA value of a tuple  $t \notin G_1$  must appear in  $G_1$ , if  $t$  was not assigned to any group

---

**Algorithm 7** Algorithm-based Disclosure Attack against Hilb

---

```
1:  $G \leftarrow$  all QI-groups published by Hilb.
2: Calculate  $Q_T$  for each tuple in  $G$ , where  $Q_T$  is 1-D dimensional QI space
   transformed by Hilbert curve. Order QI-groups in  $G$  ascendingly by the
   respective lowest  $Q_T$  in each group.
3: repeat
4:    $G_1 \leftarrow$  select the first QI-group in  $G$ .
5:   if  $(|G| - |G_1|) \geq \ell \times S_{\max}(G)$  then  $\triangleright$  A sufficient condition to find
   deterministic QI-groups.
6:      $\tilde{G} \leftarrow$  all QI-groups  $G_2$  such that  $G_2 \in G \setminus G_1$  and there exists a
   tuple  $t \in G_2$  but  $t \notin G_1$  whose  $Q_T$  falls into the range of  $Q_T$  in  $G_1$ .
7:     if  $\tilde{G} \neq \emptyset$  &&  $|G_1| + 1 < \ell \times 2$ . then
8:       Label  $G_1$  as vulnerable group.
9:     end if
10:  else
11:    Goto Line 14;  $\triangleright$  Early termination.
12:  end if
13:   $G \leftarrow G \setminus G_1$ .
14: until  $G = \emptyset$ .
15: return .
```

---

prior to  $G_1$  and the  $Q_T$  of  $t$  falls exactly into the  $Q_T$  range of  $G_1$ . The reason is that otherwise,  $t$  would have been included when generating  $G_1$ , leading to lower information loss. Therefore, the individual SA values in  $G_1$  may be compromised by the intersection of SA values between  $G_1$  and  $G_2$ . Line 6 to Line 7 describe the above procedure. Note that Line 7 defines the case when  $G_1 \cup \{t\}$  violates  $\ell$ -diversity. Since SA values in each QI-group generated by Hilb are distinct from each other, the number of the most frequent SA values in  $G_1 \cup \{t\}$  is 2. Then, we have:

$$\frac{2}{|G_1| + 1} > \frac{1}{\ell} \Rightarrow |G_1| + 1 < \ell \times 2.$$

when  $G_1 \cup \{t\}$  violates  $\ell$ -diversity.

Let  $m$  be the number of QI-groups generated by Hilb. It is easy to see that the time complexity of Algorithm 7 is  $O(m^2)$ .

Figure 9 shows the extent of algorithm-based disclosure of Hilb on the Census dataset by varying  $\ell$  from 2 to 6. One can see that the number of

compromised tuples increases monotonically with  $\ell$ . The reason is that a larger  $\ell$  significantly increases the probability of finding tuples with the same SA value from different yet correlated groups. It is important to note that the percentage of compromised tuples under our attack can be as high as  $259,937/300,000 = 86.6\%$  (when  $\ell = 6$ ).

### 8.3.2. Utility Comparison with Mondrian and Hilb

We fixed the number of QI  $qi = 7$ , query dimension  $qd = 3$ , selectivity  $s = 5\%$ . Figure 10a illustrates the utility of Mondrian++ and Hilb++ when varying  $\ell$  value. It shows that both Hilb++ and Mondrian++ are not only able to eliminate the algorithm-based disclosure by global look-ahead, but also able to attain via stratified pick-up comparable or even better utility against Hilb, and Mondrian respectively.

Next, we set  $\ell = 4$ , and varied  $qi$  from 3 to 7. Figure 10b shows the impact of  $qi$  on the utility. As we see, Hilb++ provides comparable utility with Hilb in all the cases. Whereas, when  $qi \leq 5$ , Mondrian++ achieves less accuracy than Mondrian. Nonetheless, the accuracy difference is decreasing when  $qi$  increases. This is because Mondrian tends to generate larger groups when there are more QI attributes. This makes the utility improvement by stratified pick-up in Mondrian++ more significant.

Figure 10c examines the utility of Mondrian++ and Hilb++ when query dimension  $qd$  ranges from 2 to 5, and  $\ell = 4$ ,  $qi = 7$ ,  $sel = 5\%$ . Figure 10d investigates the effect of selectivity  $sel$  on the utility when  $\ell = 4$ ,  $qi = 7$ ,  $qd = 3$ . One can see in both two figures that Mondrian++ and Hilb++ maintain comparable or significantly better utility (in the case of Mondrian++).

### 8.3.3. Utility Comparison with Anatomy

We now performed the evaluation on the bucketization publishing scheme, where Mondrian++ and Hilb++ were compared against Anatomy [10]. Recall that since Anatomy is simulatable, the objective of employing Mondrian++ and Hilb++ in bucketization scheme is to provide better utility by taking into account QI-locality information. Using the same parameter settings as the previous generalization case, we conducted experiments as shown from Figures 10e to 10h. As expected, both Mondrian++ and Hilb++ significantly outperform Anatomy in terms of utility.

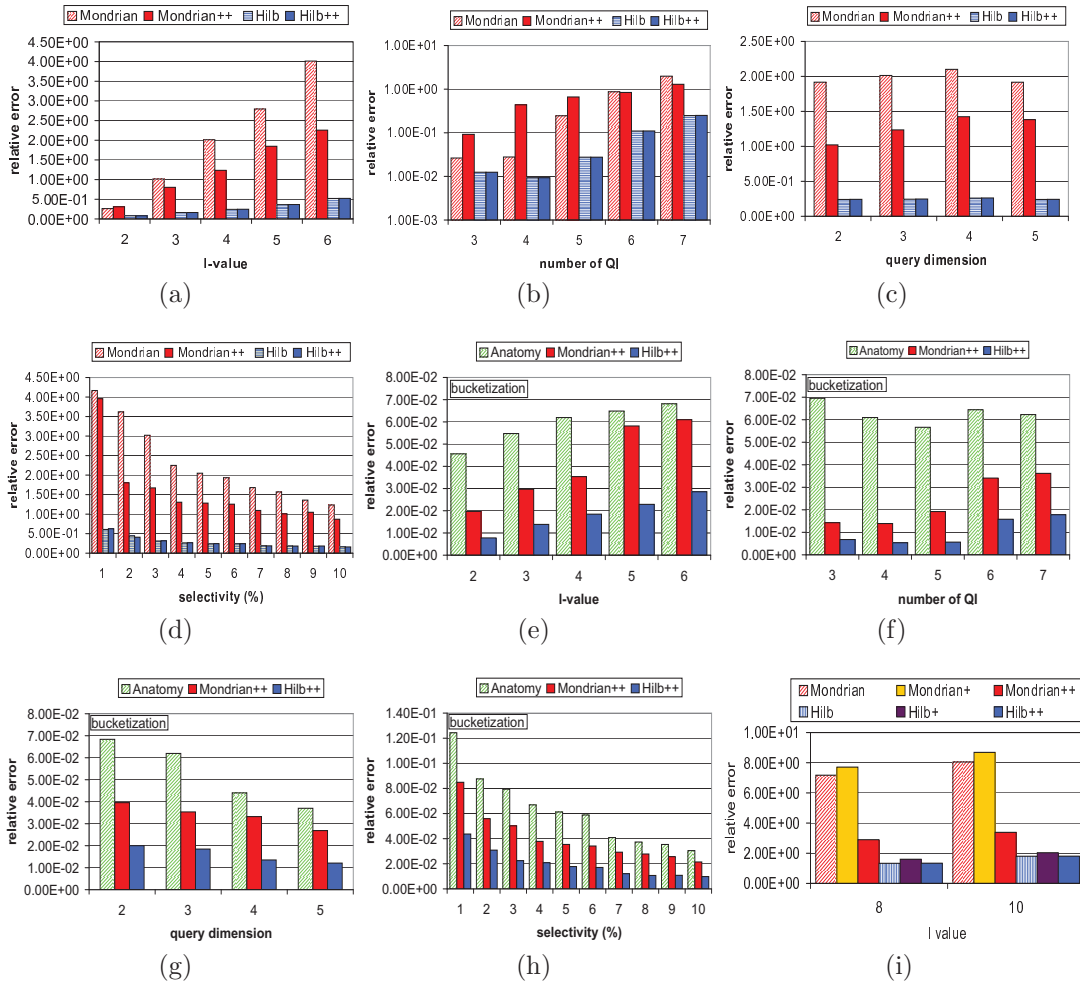


Figure 10: Census dataset, Mondrian++ & Hilb++. (a)-(d) generalization scheme; (e)-(h) bucketization scheme; (i) Effect of global look-ahead and stratified pick-up.

### 8.3.4. Effects of Global look-ahead and Stratified Pick-up

We previously integrated both tools: global look-ahead and stratified pick-up. Now, we demonstrated the effect on utility of each tool separately. We set  $qi = 7$ ,  $qd = 3$ ,  $sel = 5\%$ , and tested the cases when  $\ell = 8$  and  $\ell = 10$ . The reason why we chose higher  $\ell$  values is only for the ease of illustration, because Hilb++ achieves almost the same query accuracy as Hilb when  $\ell \leq 7$ .

To show the effect of our first tool (i.e., global look-ahead), we tested Mondrian+ and Hilb+ in Figure 10i, which were adapted from Mondrian and Hilb via integrating only the first tool (i.e., without stratified pick-up). We compared them against the original Mondrian and Hilb. As expected, both adapted algorithms achieve less utility than Mondrian and Hilb, respectively, as the cost of eliminating algorithm-based disclosure.

To show the effect of our second tool (i.e., stratified pick-up), we compared algorithms integrated with both tools (i.e., global look-ahead and stratified pick-up) against the previously developed algorithms with only the first tool. As we can see from Figure 10i, stratified pick-up improves the utility, leading to comparable or even better utility than Hilb and Mondrian, respectively.

### 8.3.5. Time Performance

We set  $\ell$  to 4. Table 8 depicts the running time of Anatomy, Mondrian, Mondrian++, Hilb and Hilb++. Again, we considered Mondrian++ and Hilb++ in two different schemes, i.e., generalization and bucketization. As expected, different schemes does not affect the time performance of Mondrian++ and Hilb++. Either Mondrian++ or Hilb++ (in both two schemes) runs much faster than the original algorithm because, as discussed in Section 6.2.1, global look-ahead defines a more "stringent" condition, and thus either Mondrian++ or Hilb++ tends to earlier terminate than their original algorithm. As analyzed in Section 6 of the time complexity, Table 8 confirms the running time of Hilb++ to be lower than Mondrian++, but higher than Anatomy.

Table 8: Time performance

	Mondrian	Mondrian++	Mondrian++ (in bucketization)	Hilb	Hilb++	Hilb++ (in bucketization)	Anatomy
time (in seconds)	9.2	2.2	2.2	2.1	0.6	0.6	0.3

## 9. Related Work

Since the introduction of  $k$ -anonymity [1] and  $\ell$ -diversity [2], various privacy models have been proposed including  $(\alpha, k)$ -anonymity [3], personalized privacy [30],  $t$ -closeness [4],  $(k, e)$ -anonymity [5],  $(\epsilon, m)$ -anonymity [31], etc. To achieve these privacy models, researchers studied numerous PPDP algorithms [32, 33, 34, 26, 35, 7, 36, 8, 9, 28, 10, 37, 11, 38, 39, 40]. Orthogonal to the above anonymization techniques at the tuple level, a recent work [41] studied PPDP with multiple privacy rules by focusing on the schema level.

There has been a large body of work on addressing the threats from external knowledge held by adversaries. [23, 42, 43, 44] considered the knowledge about an individual or relationship between individuals. [45] studied the presence of corruption. [21] studied the negative association rule. [22, 46] studied the privacy disclosure from learning whether a certain individual is present in the database or not.

The problem of algorithm-based disclosure was introduced by [12, 14]. [12] provided a new privacy model  $m$ -confidentiality and designed a new algorithm MASK to achieve it. However, we have shown in Section 2.2 that the MASK algorithm in [12] is still vulnerable to algorithm-based disclosure. [14] defined another new privacy model called  $p$ -safety to address the problem, but its efficiency is a problem. Differential privacy proposed in [15] is able to eliminate algorithm-based disclosure by providing a new privacy model and developing a corresponding algorithm. Orthogonal to all the above work, the focus of our paper as mentioned in our introduction part is to develop generic tools to adapt the existing data publishing algorithms, such that these algorithms can be immune from algorithm-based disclosure. [47] proposed a  $k$ -jump strategy to transform a unsafe algorithm to a large family of distinct safe algorithms, but its time complexity is exponential.

Cormode et al. [48] proposed an elegant “symmetric” method to defend against minimality attack, while concluding by theoretical analysis that minimality attack can only render a constant increase on the adversarial belief about individual SA information. While their work is interesting and solid, we argue that our work differs from theirs significantly on the following two key points:

- In terms of the degree of algorithm-based disclosure, we design and evaluate algorithm-based attacks beyond the scope considered in Cormode et al. In particular, an important assumption Cormode et al.

make while quantifying the impact of minimality attack theoretically is that the adversarial knowledge does not involve any SA distribution. As such, their analysis does not apply to algorithm-based disclosures which would occur when an adversary holds such prior knowledge on SA distribution - e.g., the algorithm-based disclosure of SA-perturbation-based algorithms such as MASK (recall from Section 2 that the algorithm-based disclosure of MASK can be exploited by adversaries with external knowledge such as “Japanese have an extremely low incidence of heart disease” - nonetheless, the disclosure itself is still caused by knowledge of the algorithm, not the external knowledge). As such, the conclusions in Cormode et al. do not apply to all algorithm-based disclosures considered in our paper.

Likewise, the attacks we consider in the paper also go beyond another assumption made by Cormode et al. - i.e., each QI-group is generated independently from the others. Our attack actually leverages association between QI-groups which can be learned from knowledge of the publishing algorithm. An example is the attack against Hilb which we discuss in Section 8.3.1. Since we consider attacks beyond the scope defined by Cormode et al., our conclusion is also different from theirs. In particular, as we have shown in Section 8.3.1, our experimental results show that the impact of algorithm-based disclosure can be serious when an adversary exploits QI-group-correlations - e.g., 86.6% tuples out of the Census dataset can be compromised based on knowledge of the Hilb algorithm (in the case of 6-diversity).

- In terms of defending against algorithm-based disclosure (e.g., defense against minimality attacks considered in Cormode et al.), the aforementioned “symmetric” method in Cormode et al. weakens the  $\ell$ -diversity guarantee to  $(\ell - 2/3)$ -diversity (as one can see from Theorem 5 in Cormode et al.). On the other hand, both Local and Global Look-ahead algorithms proposed in our paper can fix the algorithm-based disclosure problem without weakening the privacy guarantee. Meanwhile, we also propose Stratified Pick-up to further improve the utility of the published table.

It is also worth pointing out that we substantially extended our preliminary version [49]. First, we have proposed a brand new tool called local look-ahead which is designed to eliminate algorithm-based disclosure from data

publishing algorithms based on SA-perturbation (e.g., MASK). This stands in contrast with [49] which can only deal with QI-perturbation algorithms that keep SA intact. We have also added the corresponding experiments to illustrate the effectiveness of local look-ahead. Another significant addition is an evaluation of the extent of algorithm-based disclosure in the original MASK algorithm and the state-of-the-art  $\ell$ -diversity algorithm Hilb.

## 10. Conclusion

This paper addressed the problem of algorithm-based disclosure in privacy-preserving data publishing. We proposed Algorithm-Safe Publishing (ASAP), a novel privacy model which defines the space of algorithm-based disclosure. Two necessary conditions and two sufficient conditions of ASAP were given as a toolset to determine whether an existing algorithm is vulnerable to algorithm-based disclosure. To eliminate algorithm-based disclosure, we proposed global and local look-ahead, two generic tools for correcting the design of existing algorithms. To enhance utility, we developed another add-on tool: stratified pick-up. We demonstrated the power of our tools by revising the design of three existing algorithms, Mondrian, Hilb, MASK, to Mondrian++, Hilb++ and MASK++, respectively, for eliminating algorithm-based disclosure. We conducted extensive experiments on real-world datasets to demonstrate the effectiveness, efficiency and utility of our tools.

## References

- [1] P. Samarati, L. Sweeney, Protecting privacy when disclosing information:  $k$ -anonymity and its enforcement through generalization and suppression, Tech. Rep., CMU, SRI, 1998.
- [2] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkatasubramanian,  $\ell$ -diversity: Privacy Beyond  $k$ -anonymity, in: ICDE, 2006.
- [3] R. C. Wong, J. Li, A. W. Fu, K. Wang,  $(\alpha, k)$ -Anonymity: An Enhanced  $k$ -Anonymity Model for Privacy-Preserving Data Publishing, in: KDD, 754–759, 2006.
- [4] N. Li, T. Li, S. Venkatasubramanian,  $t$ -Closeness: Privacy Beyond  $k$ -anonymity and  $\ell$ -diversity, in: ICDE, 106–115, 2007.

- [5] Q. Zhang, N. Koudas, D. Srivastava, T. Yu, Aggregate Query Answering on Anonymized Tables, in: ICDE, 116–125, 2007.
- [6] X. Xiao, Y. Tao,  $m$ -invariance: towards privacy preserving re-publication of dynamic datasets, in: SIGMOD, 689–700, 2007.
- [7] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Incognito: efficient full-domain  $k$ -anonymity, in: SIGMOD, 49–60, 2005.
- [8] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Mondrian Multidimensional  $k$ -anonymity, in: ICDE, 25–35, 2006.
- [9] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Workload-Aware Anonymization, in: KDD, 277–286, 2006.
- [10] X. Xiao, Y. Tao, Anatomy: Simple and Effective Privacy Preservation, in: VLDB, 139–150, 2006.
- [11] G. Ghinita, P. Karras, P. Kalnis, N. Mamoulis, Fast Data Anonymization with Low Information Loss, in: VLDB, 758–769, 2007.
- [12] R. C. Wong, A. W. Fu, K. Wang, J. Pei, Minimality Attack in Privacy-Preserving Data Publishing, in: VLDB, 543–554, 2007.
- [13] A. Kerckhoffs, La cryptographie militaire (Military Cryptography), Journal des sciences militaires IX (1883) 5–83, 161–191.
- [14] L. Zhang, S. Jajodia, A. Brodsky, Information Disclosure under Realistic Assumptions: Privacy versus Optimality, in: CCS, 2007.
- [15] C. Dwork, Differential privacy, in: ICALP, 1–12, 2006.
- [16] A. Machanavajjhala, J. Gehrke, M. Goetz, Data Publishing against Realistic Adversaries, in: VLDB, 2009.
- [17] N. Koudas, D. Srivastava, T. Yu, Q. Zhang, Distribution-based Microdata Anonymization, in: VLDB, 2009.
- [18] C.-Y. C. M. F. Mokbel, W. G. Aref., The New Casper: Query Processing for Location Services without Compromising Privacy, in: VLDB, 2006.
- [19] K. Liu, E. Terzi, Towards Identity Anonymization on Graphs, in: SIGMOD, 2009.

- [20] H. Yeye, J. Naughton, Anonymization of Set-Valued Data via Top-Down, Local Generalization, in: VLDB, 2009.
- [21] T. Li, N. Li, Injector: Mining Background Knowledge for Data Anonymization, in: ICDE, 446–455, 2008.
- [22] M. E. Nergiz, M. Atzori, C. Clifton, Hiding the Presence of Individuals from Shared Databases, in: SIGMOD, 665–676, 2007.
- [23] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, J. Halpern, Worst-Case Background Knowledge for Privacy-Preserving Data Publishing, in: ICDE, 126–135, 2007.
- [24] C. E. Shannon, Communication Theory of Secrecy Systems, Bell System Technical Journal 28 (1949) 656–715.
- [25] T. M. Cover, J. A. Thomas, Elements of Information Theory, Wiley-Interscience, 1991.
- [26] R. J. Bayardo, R. Agrawal, Data privacy through optimal k-anonymization, in: ICDE, 2005.
- [27] V. S. Iyengar, Transforming Data to Satisfy Privacy Constraints, in: KDD, 2002.
- [28] D. Kifer, J. Gehrke, Injecting Utility into Anonymized Datasets, in: SIGMOD, 2006.
- [29] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10 (5) (2002) 571–588.
- [30] X. Xiao, Y. Tao, Personalized Privacy Preservation, in: SIGMOD, 229–240, 2006.
- [31] J. Li, Y. Tao, X. Xiao, Preservation of Proximity Privacy in Publishing Numeric Sensitive Data, in: SIGMOD, 473–486, 2008.
- [32] V. S. Iyengar, Transforming Data to Satisfy Privacy Constraints, in: KDD, 279–288, 2002.

- [33] A. Meyerson, R. Williams, On the Complexity of Optimal  $k$ -anonymity, in: PODS, 223–228, 2004.
- [34] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu, Anonymizing Tables, in: ICDT, 246–258, 2005.
- [35] B. C. M. Fung, K. Wang, P. S. Yu, Top-Down Specialization for Information and Privacy Preservation, in: ICDE, 205–216, 2005.
- [36] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, A. Zhu, Achieving Anonymity via Clustering, in: PODS, 153–162, 2006.
- [37] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A. W. Fu, Utility-based anonymization using local recoding, in: KDD, 785–790, 2006.
- [38] H. Park, K. Shim, Approximate algorithms for  $k$ -anonymity, in: SIGMOD, 67–78, 2007.
- [39] T. Iwuchukwu, J. Naughton,  $k$ -Anonymization as Spatial Indexing: Toward Scalable and Incremental Anonymization, in: VLDB, 746–757, 2007.
- [40] G. Ghinita, Y. Tao, P. Kalnis, On the anonymization of sparse high-dimensional data, in: ICDE, 715–724, 2008.
- [41] X. Jin, M. Zhang, N. Zhang, G. Das, Versatile Publishing for Privacy Preservation, in: KDD, 2010.
- [42] B. Chen, R. Ramakrishnan, K. LeFevre, Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge, in: VLDB, 770–781, 2007.
- [43] W. Du, Z. Teng, Z. Zhu, Privacy-MaxEnt: Integrating Background Knowledge in Privacy Quantification, in: SIGMOD, 459–472, 2008.
- [44] T. Li, N. Li, J. Zhang, Modeling and Integrating Background Knowledge in Data Anonymization, in: ICDE, 2009.
- [45] Y. Tao, X. Xiao, J. Li, D. Zhang, On Anti-Corruption Privacy Preserving Publication, in: ICDE, 725–734, 2008.

- [46] V. Rastogi, S. Hong, D. Suciu, The Boundary Between Privacy and Utility in Data Publishing, in: VLDB, 531–542, 2007.
- [47] W. Liu, L. Wang, L. Zhang, K-Jump Strategy for Preserving Privacy in Micro-Data Disclosure, in: ICDT, 2010.
- [48] G. Cormode, N. Li, T. Li, D. Srivastava, Minimizing Minimality and Maximizing Utility: Analyzing Method-based Attacks on Anonymized Data, in: VLDB, 2010.
- [49] X. Jin, N. Zhang, G. Das, Algorithm-Safe Privacy Preserving Data Publishing, in: EDBT, 2010.