

# On Anti-Corruption Privacy Preserving Publication

Yufei Tao<sup>1</sup>, Xiaokui Xiao<sup>1</sup>, Jiexing Li<sup>1</sup>, Donghui Zhang<sup>2</sup>

<sup>1</sup>*Department of Computer Science and Engineering, Chinese University of Hong Kong*

*Sha Tin, New Territories, Hong Kong*

{taoyf, xkxiao, jxli}@cse.cuhk.edu.hk

<sup>2</sup>*College of Computer and Information Science, Northeastern University*

*360 Huntington Avenue, Boston, MA, USA*

donghui@ccs.neu.edu

**Abstract**— This paper deals with a new type of privacy threat, called “corruption”, in anonymized data publication. Specifically, an adversary is said to have corrupted some individuals, if s/he has already obtained their sensitive values before consulting the released information. Conventional generalization may lead to severe privacy disclosure in the presence of corruption. Motivated by this, we advocate an alternative anonymization technique that integrates generalization with perturbation and stratified sampling. The integration provides strong privacy guarantees, even if an adversary has corrupted any number of individuals. We verify the effectiveness of the proposed technique through experiments with real data.

## I. INTRODUCTION

Anonymized publication has received considerable attention in recent years, due to the awareness of privacy disclosure in data sharing applications. Assume, for example, that a hospital wants to release Table Ia, referred to as the *microdata*. Attribute *Disease* is sensitive, which has two implications. First, the publication must prevent an adversary from inferring accurately the disease of any individual patient. Second, the released content should permit a researcher to understand the correlations between *Disease* and the other attributes, which are statistically significant in a large number of patients.

Obviously, the column *Owner* must not be published. Simply removing that attribute, however, is insufficient, due to the possibility of “linking attacks”. For instance, if an adversary has the voter registration list in Table Ib, s/he can easily obtain the name of any patient, through an equi-join between Tables Ia and Ib. The joining columns *Age*, *Gender*, and *Zipcode* are therefore called the *quasi-identifier* (QI) attributes.

*Generalization* [1], [2], [3], [4], [5] is a popular methodology for preventing linking attacks. The objective is to replace each QI value with a less specific form, so that each tuple is indistinguishable from several others by their QI-values. Table Ic demonstrates a generalized version of Table Ia. The generalization results in four *QI-groups*, each involving a set of tuples with equivalent QI-values. Consider an adversary who aims at inferring the disease of Debbie, knowing her exact QI values {45, F, 20000}. Since the 3rd and 4th rows of Table Ic match Debbie’s QI details, the adversary is not sure whether she contracted *pneumonia* or *breast-cancer*.

### A. Motivation

Generalization provides weak privacy protection when an adversary may corrupt data owners. Consider an adversary who has the QI-values {30, M, 27000} of Calvin. Given Table Ic, s/he sees that the tuple of Calvin is in the first QI-group (consisting of the first two rows). Hence, s/he can only infer that Calvin may have contracted *bronchitis* or *pneumonia*. However, suppose that the adversary has corrupted Bob before, e.g., s/he contacted Bob, and learned that Bob contracted *bronchitis*. As a result, the adversary becomes sure that Calvin must have *pneumonia* (according to Table Ib, only Bob can be in the same QI-group as Calvin). Note that what Bob did is completely conscientious — he is merely giving away his own information.

In the above example, corruption is caused by collusion between a data owner (Bob) and an adversary, whereas, in general, it may happen in many other ways. For instance, an adversary may acquire the diagnostic results of some patients via a friend working in the hospital. As another example, an adversary may be the boss of Bob, who has Bob’s sick-leave application that states explicitly his disease.

### B. Contributions

This paper provides the first study towards eliminating the threat of corruption. First, we formalize anti-corruption anonymization. Following the information-theoretic approach in [6], our formalization aims at achieving *background-sensitive guarantees* (a well-known example is “ $\rho_1$ -to- $\rho_2$  protection” [6]). Such a guarantee models the degree of privacy preservation as a function of an adversary’s background knowledge, and serves as an effective metric for gauging the quality of anonymization.

Second, we elaborate several defects of generalization that have not been revealed in the literature. Our results show that, generalization provides poor background-sensitive guarantees, even in the conventional corruption-free scenarios. Namely, they may allow an adversary to glean considerable new knowledge, even though s/he has almost no knowledge before examining the published data. When corruption is possible, generalization completely fails in guarding privacy.

Third, we overcome the drawbacks of generalization, by integrating it with perturbation [7], [6] and stratified sampling

Owner	Age	Gender	Zipcode	Disease
Bob	25	M	25000	bronchitis
Calvin	30	M	27000	pneumonia
Debbie	45	F	20000	pneumonia
Ellie	50	F	15000	breast cancer
Fiona	55	F	45000	ovarian cancer
Gloria	58	F	32000	hypertension
Henry	65	M	65000	Alzheimer
Isaac	80	M	55000	dementia

(a) Microdata

Name	Age	Gender	Zipcode
Bob	25	M	25000
Calvin	30	M	27000
Debbie	45	F	20000
Ellie	50	F	15000
Emily	52	F	28000
Fiona	55	F	45000
Gloria	58	F	32000
Henry	65	M	65000
Isaac	80	M	55000

(b) A voter registration list

Age	Gender	Zipcode	Disease
[21, 40]	M	[11***, 30***]	bronchitis
[21, 40]	M	[11***, 30***]	pneumonia
[41, 60]	F	[11***, 30***]	pneumonia
[41, 60]	F	[11***, 30***]	breast cancer
[41, 60]	F	[31***, 50***]	ovarian cancer
[41, 60]	F	[31***, 50***]	hypertension
[61, 80]	M	[51***, 70***]	Alzheimer
[61, 80]	M	[51***, 70***]	dementia

(c) A generalized table

TABLE I

PRIVACY PRESERVING PUBLICATION BASED ON GENERALIZATION

[8]. The resulting technique, termed *perturbed generalization*, provides strong background-sensitive guarantees, even if an adversary has corrupted an arbitrary number of individuals.

The rest of the paper is organized as follows. Section II clarifies the objectives of anti-corruption publication. Then, Section III explains why generalization fails to protect privacy in our settings. Section IV presents the proposed anonymization framework. Section V elaborates how an adversary may perform a privacy attack, and Section VI establishes our privacy guarantees. Section VII experimentally evaluates the effectiveness of our solutions. Section VIII briefly reviews the previous work related to ours. Finally, Section IX concludes the paper with directions for future work.

## II. PROBLEM SETTINGS

We consider a microdata table  $\mathcal{D}$ , with  $d$  quasi-identifier (QI) attributes  $A_1^q, \dots, A_d^q$ , and a sensitive attribute  $A^s$ . Each  $A_i^q$  ( $1 \leq i \leq d$ ) can be either discrete or continuous, but  $A^s$  must be discrete. The *domain* of a column  $A$  is the projection of  $\mathcal{D}$  on  $A$ . Let  $U^q$  be the  $d$ -dimensional *QI space*, which is the cartesian product of the domains of  $A_1^q, \dots, A_d^q$ . Use  $U^s$  to denote the domain of  $A^s$ .

For each tuple  $t \in \mathcal{D}$ , define its *QI-vector*  $t.v^q$ , as a  $d$ -dimensional vector containing its QI-values  $t.A_1^q, \dots, t.A_d^q$ . Equivalently,  $t.v^q$  can be regarded as a point in  $U^q$ . Each tuple  $t \in \mathcal{D}$  describes the information of an individual, i.e., the *owner* of  $t$ . All tuples have distinct owners (this is a common assumption in the literature [3], [9]).

Our goal is to publish an anonymized version  $\mathcal{D}^*$ , which satisfies the following requirements:

1. [*Cardinality*]  $\mathcal{D}^*$  has at most  $|\mathcal{D}| \cdot s$  rows, where  $s$  is a real value in  $(0, 1]$ , and a publication parameter.
2. [*Privacy*] Publication of  $\mathcal{D}^*$  ensures strong privacy guarantees, even if an adversary *corrupts* any individuals in  $\mathcal{D}$ .
3. [*Utility*]  $\mathcal{D}^*$  is useful for mining data patterns in  $\mathcal{D}$ .

Next, we discuss each requirement in detail.

### A. Cardinality

This feature is reasonable for several reasons. First, the microdata may be simply too voluminous. For example, a

database in a hospital may be in giga or even tera bytes, rendering transfer in its entirety intractable. Second, given an adequately large subset (of the original dataset), most mining algorithms already return reliable results. Third, the ability of controlling how much percent of a dataset is revealed is an appealing feature for commercial organizations.

### B. Privacy

We aim at preventing linking attacks as exemplified in Section I. Formally, in such an attack, an adversary knows (i) the existence of a victim individual  $o$  in  $\mathcal{D}$ , and (ii) the exact QI values of  $o$ , compactly represented with a QI-vector  $o.v^q$ . The goal of the attack is to infer whether the sensitive value  $o.A^s$  of  $o$  satisfies a predicate  $Q$ , which may be any arbitrarily complex condition. For instance, if  $A^s$  is *Disease*,  $Q$  can be “ $o.A^s$  is a respiratory disease”.

The adversary has access to an *external database*  $\mathcal{E}$ . Given a QI vector  $v^q$ ,  $\mathcal{E}$  returns the identities (e.g., SSNs) of all the people whose QI vectors are equivalent to  $v^q$ . Some of these people may not appear in the microdata, in which case we say that they are *extraneous*, and their sensitive values are  $\emptyset$ . In the example of Section I,  $\mathcal{E}$  is the voter registration list in Table Ib, where Emily is extraneous.

A unique feature of our privacy goal is protection against “corruption”:

*Definition 1 (Corruption):* An adversary is said to have *corrupted* an individual, if s/he learns the exact sensitive value of that individual via resources different from  $\mathcal{D}^*$ .  $\square$

Let  $\mathcal{C}$  be the set of individuals that an adversary is able to corrupt. We model  $\mathcal{C}$  as a subset of  $\mathcal{E}$  (instead of  $\mathcal{D}$ ) to capture the fact that an adversary may be aware of which individuals are extraneous. We allow the size of  $\mathcal{C}$  to be any value from 0 to  $|\mathcal{E}| - 1$ . Obviously, when  $|\mathcal{C}| = 0$ , our scenario degenerates into the traditional assumption that no corruption is possible. The worst case occurs when  $|\mathcal{C}| = |\mathcal{E}| - 1$ ; that is, the adversary has the sensitive values of all the people, except  $o$ .

From her/his own understanding of  $o.A^s$  and the results of corruption, an adversary has developed a certain amount of confidence about how likely  $o.A^s$  would satisfy  $Q$ , even though s/he has not examined  $\mathcal{D}^*$  yet. This is her/his *prior confidence*, denoted as  $P_{prior}(Q)$ . Such confidence results

from the ultimate “background knowledge” of  $\mathcal{D}$  that an adversary can possibly accumulate without  $\mathcal{D}^*$ . It depends on factors that cannot be controlled by the publisher, such as how familiar the adversary is with the victim, her/his expertise on the correlation between the QI and sensitive attributes, her/his corruption power, and so on.

We tackle the challenge that the adversary is an information-theory expert, who is able to combine background knowledge and  $\mathcal{D}^*$  to boost her/his confidence about whether  $o.A^s$  qualifies  $Q$ . We use the term *posterior confidence* to refer to the adversary’s confidence at the end of the whole linking attack, and represent it as  $P_{post}(Q)$ .

The objective of the publisher is to limit the posterior confidence. In particular, we focus on achieving *background-sensitive guarantees*. If one views privacy protection-versus-inference as a game played by the publisher and adversary, a background-sensitive guarantee constrains the adversary’s chance of winning the game, subject to how well s/he can play. The first type of guarantees offered by our technique is:

*Definition 2: [ $\rho_1$ -to- $\rho_2$  Guarantee/Breach [6]]* Let  $\rho_1$  and  $\rho_2$  be values satisfying  $0 \leq \rho_1 < \rho_2 \leq 1$ . A  $\rho_1$ -to- $\rho_2$  guarantee requires that

$$\text{if } P_{prior}(Q) \leq \rho_1, \text{ then } P_{post}(Q) \leq \rho_2.$$

A  $\rho_1$ -to- $\rho_2$  breach occurs, if the guarantee is violated.  $\square$

For instance, it is a 0.3-to-0.5 breach, if an adversary’s posterior confidence exceeds 0.5, when her/his prior confidence is bounded by 0.3. However, once the prior confidence is higher than 0.3, it does not constitute a 0.3-to-0.5 breach, no matter how large the posterior confidence is. Intuitively, in this case the adversary is too powerful, so we cannot constrain her/his chance of winning the game<sup>1</sup>.

We also study another important type of background-sensitive guarantees that have not been analyzed previously.

*Definition 3 ( $\Delta$ -growth):* Let  $\Delta$  be a value in  $(0, 1]$ . A  $\Delta$ -growth guarantee requires

$$P_{post}(Q) - P_{prior}(Q) \leq \Delta.$$

A  $\Delta$ -growth breach occurs, if the guarantee is violated.  $\square$

The  $\Delta$ -growth guarantee is a natural way to control an adversary’s increased knowledge after s/he inspects  $\mathcal{D}^*$ . By setting  $\Delta$  to  $\rho_2 - \rho_1$ , ensuring no  $\Delta$ -growth breach immediately guarantees no  $\rho_1$ -to- $\rho_2$  breach, but the reverse is *not* true. In fact,  $\Delta$ -growth guarantees remedy the deficiency of  $\rho_1$ -to- $\rho_2$  guarantees. Notice that, no 0.3-to-0.5 breach happens, even if an adversary’s prior confidence is (almost) 0, and her/his posterior confidence reaches 0.5. Intuitively, in this case, the deployed privacy preserving approach is not effective, since it allows an adversary to gain considerable new knowledge.

<sup>1</sup>Our formulation is the upward breach defined in [6], which also proposes a downward counterpart. Specifically, a downward  $\rho_1$ -to- $\rho_2$  occurs if the posterior confidence is below  $\rho_2$ , given that the prior confidence is above  $\rho_1$ . We focus on upward breaches, because the absence of  $\rho_1$ -to- $\rho_2$  upward breaches ensures no  $(1 - \rho_1)$ -to- $(1 - \rho_2)$  downward breach.

If a publisher intends to constraint the amount of increased confidence within 0.2, it should enforce a 0.2-growth guarantee instead.

### C. Utility

The utility of an anonymized dataset is typically evaluated by its effectiveness in performing a certain data mining task. Following the previous work [10], [11], we use decision-tree mining as the representative task. In fact,  $\mathcal{D}^*$  can be directly fed into the algorithm in [12] for constructing decision trees, which accurately summarize the data patterns in  $\mathcal{D}$ . The algorithm is *ad-hoc*, since it permits a data analyst to build trees according to her/his own preferences, such as the set of attributes considered, the classification granularity, and so on. Such preferences do not need to be specified at the time of preparing  $\mathcal{D}^*$ . Hence, publication of  $\mathcal{D}^*$  offers significantly more flexibility than releasing only a few trees selected by the publisher.

## III. DEFECTS OF GENERALIZATION

Crucial to generalization is its underlying *generalization principle*, which is a constraint satisfied by every QI-group of  $\mathcal{D}^*$ . The most popular principles involve  $k$ -anonymity [4], [5] and  $l$ -diversity [9]. We will focus on  $l$ -diversity, since  $k$ -anonymity (due to its pioneering role in the literature) has severe vulnerabilities to privacy attacks [9].

$l$ -diversity is most effective when (i) an adversary’s background knowledge about the victim individual  $o$  conforms to a specific type, and (ii) the adversary performs no corruption. In the sequel, we first prove that, even when both conditions are satisfied,  $l$ -diversity can guarantee only weak background-sensitive guarantees. Then, we will show that the guarantees are much worse, when the conditions are violated. Unfortunately, this is true not only for  $l$ -diversity, but for the generalization methodology in general.

### A. Defects of $l$ -diversity When Its Assumptions Are Satisfied

Machanavajjhala et al. [9] give several versions of  $l$ -diversity. Table Ic demonstrates the simplest version, which demands each QI-group to have at least  $l = 2$  different sensitive values. The most powerful and well-adopted version is “ $(c, l)$ -diversity”, where  $c$  is a positive value, and  $l$  an integer. Intuitively, this principle requires that, in every QI group  $QI$  of  $\mathcal{D}^*$ , the most frequent sensitive value should not be too frequent.

Formally, assume that  $QI$  has  $l'$  distinct sensitive values, where  $l'$  can be any integer at least  $l$ . Let  $n_1, n_2, \dots, n_{l'}$  be the numbers of tuples in  $QI$  carrying the most, second most, ..., least frequent sensitive values, respectively (i.e.,  $n_1 \geq n_2 \geq \dots \geq n_{l'}$ ). Then,  $(c, l)$ -diversity requires

$$n_1 \leq c \cdot (n_l + n_{l+1} + \dots + n_{l'}). \quad (1)$$

Figure 1 illustrates an example QI group with size 11 which obeys  $(\frac{1}{2}, 3)$ -diversity. Here, the group has  $l' = 6$  distinct sensitive values, with  $n_1 = 3, n_2 = n_3 = n_4 = 2$ , and  $n_5 =$

owner	QI Attributes	Disease
$o_1$	same	pneumonia
$o_2$		pneumonia
$o_3$		pneumonia
$o_4$		HIV
$o_5$		HIV
$o_6$		bronchitis
$o_7$		bronchitis
$o_8$		lung cancer
$o_9$		lung cancer
$o_{10}$		SARS
$o_{11}$		tuberculosis

Fig. 1. A  $(1/2, 3)$ -diverse QI-group

$n_6 = 1$ . In this case, Inequality 1 becomes  $3 \leq \frac{1}{2}(2+2+1+1)$ , setting  $c$  to  $\frac{1}{2}$  and  $l$  to 3.

Let  $r$  be the real value of  $o.A^s$  in  $\mathcal{D}$ .  $(c, l)$ -diversity aims at preventing an adversary from performing an *exact reconstruction* of  $o.A^s$ . Equivalently, by the terminology of Section II, the predicate  $Q$  has a special form (denoted as  $Q_r$ )

$$Q_r : o.A^s = r.$$

Furthermore, the principle is proposed to tackle adversaries that can identify, without looking at  $\mathcal{D}^*$ , at most  $l - 2$  values in  $U^s$  (the domain of  $A^s$ ) which cannot be the real  $o.A^s$ . In other words, before examining  $U^s$ , the adversary thinks that  $o.A^s$  can be any of the other  $|U^s| - (l - 2)$  values in  $U^s$  with an equal probability, that is, s/he has prior confidence

$$P_{prior}(Q_r) = 1/(|U^s| - l + 2). \quad (2)$$

In this case,  $(c, l)$ -diversity ensures that, after investigating  $\mathcal{D}^*$ , an adversary can figure out  $o.A^s = r$  with probability at most  $\frac{c}{c+1}$  [9]. Namely, her/his posterior confidence

$$P_{post}(Q_r) \leq c/(c + 1). \quad (3)$$

To explain the above derivation with a concrete example, assume that an adversary targets individual  $o = o_1$ , knowing in advance that  $o_1$  does not have *HIV*. Suppose that the *Disease* attribute has a domain size of 100. Thus, before seeing  $\mathcal{D}^*$ , the adversary can guess the real disease *pneumonia* of  $o.A^s$  only with a probability  $1/99$ , as given by Equation 2 ( $l = 3$ ). Now, the adversary studies  $\mathcal{D}^*$ , and finds out that the record of  $o_1$  must be in the QI group in Figure 1. As the adversary can exclude only *HIV* from being the real disease of  $o_1$ , s/he cannot tell which of the 9 tuples not carrying *HIV* belongs to  $o_1$ . Given that the group has 3 *pneumonia* tuples, with a random guess, the adversary infers  $o.A^s = \textit{pneumonia}$  with a probability  $3/9 = 1/3$ , conforming to Inequality 3 ( $c = \frac{1}{2}$ ).

Combining Equation 2 and Inequality 3, when  $Q$  is restricted to  $Q_r$  (i.e., exact reconstruction) and an adversary's background knowledge fulfills the requirement of  $(c, l)$ -diversity, the publisher can ensure a  $\frac{1}{|U^s| - l + 2}$ -to- $\frac{c}{c+1}$  guarantee, and a  $(\frac{c}{c+1} - \frac{1}{|U^s| - l + 2})$ -growth guarantee.

However, recall that our objective is to guard against inference of any predicate  $Q$ , as opposed to merely  $Q_r$ . In other words, the privacy preservation technique should be effective even in the worst case, namely, it must provide good

background-sensitive guarantees which hold for any (even the most adversely-designed)  $Q$ . Unfortunately,  $(c, l)$ -diversity is not worst-case effective, as established in the following lemma.

*Lemma 1:* Let  $u$  be the smallest number of distinct sensitive values in any QI-group of a  $(c, l)$ -diverse  $\mathcal{D}^*$  under the global-recoding scheme [13]. Even if an adversary's background knowledge satisfies the requirement of  $(c, l)$ -diversity and no corruption is performed,  $(c, l)$ -diversity fails to ensure any  $\frac{u-l+2}{|U^s| - l + 2}$ -to- $x$  or  $(x - \frac{u-l+2}{|U^s| - l + 2})$ -growth guarantee, unless  $x = 1$ .

*Proof:* Let  $QI$  be a QI-group in  $\mathcal{D}^*$  with  $u$  distinct sensitive values. There exist at least  $u - l + 2$  sensitive values in  $QI$  that the adversary cannot eliminate from being the real sensitive value of the victim  $o$ . Denote them as  $x_1, x_2, \dots, x_{u-l+2}$ . Consider  $Q = "o.A^s \text{ is any of } \{x_1, \dots, x_{u-l+2}\}"$ . The adversary's prior confidence equals  $\frac{u-l+2}{|U^s| - l + 2}$ . After the attack, s/he will be affirmative that  $Q$  is true, and hence, has posterior confidence 1.  $\square$

In practice,  $u \ll |U^s|$ , rendering  $\frac{u-l+2}{|U^s| - l + 2}$  to be a value by far smaller than 1. Therefore, Lemma 1 indicates that, even if an adversary's prior confidence about  $Q$  is very small, after inspecting a  $(c, l)$ -diverse  $\mathcal{D}^*$ , the adversary may assert that  $o.A^s$  *definitely* satisfies  $Q$ .

Again, we provide the intuition using Figure 1. Suppose that the QI-group in the figure has the smallest number  $u = 6$  of distinct sensitive values, among all the QI-groups in  $\mathcal{D}^*$ . Note that, except *HIV*, the other 5 diseases in the QI-group are respiratory problems. Further assume that they are the only 5 respiratory diseases in the whole domain  $U^s$  of  $A^s$ , which has a size  $|U^s| = 100$ . An adversary intends to pry into the privacy of  $o = o_1$ . However, this time, the goal of privacy attack is the predicate

$$Q : o.A^s \text{ is a respiratory disease.}$$

Conforming to the background knowledge requirement of  $(\frac{1}{2}, 3)$ -diversity, the adversary knows that  $o_1$  does not have *HIV*. Before checking  $\mathcal{D}^*$ , s/he conjectures that  $o.A^s$  is a respiratory disease with probability  $5/99$ , which is her/his prior confidence  $P_{prior}(Q)$ . From  $\mathcal{D}^*$ , the adversary realizes that the record of  $o_1$  must be in the QI-group of Figure 1. After eliminating *HIV*, s/he sees that all the remaining values of the QI-group are respiratory diseases. Hence, s/he becomes affirmative that  $o_1$  definitely has a respiratory problem, that is, her/his posterior confidence  $P_{post}(Q) = 1$ . Therefore, no  $\frac{5}{99}$ -to- $x$  or  $(x - \frac{5}{99})$ -growth guarantee can be claimed for any  $x < 1$ , as stated in Lemma 1.

## B. Failure of Generalization

The above discussion actually "favors"  $l$ -diversity, because it assumes that an adversary's background knowledge follows exactly the requirement of that principle, and the adversary carries out no corruption. As expected, even weaker privacy guarantees can be proved, when these assumptions are invalid. The following lemma holds for any generalized table, no

matter which generalization principle (including those recently developed in [14], [15]) is deployed.

*Lemma 2:* When an adversary can have any background knowledge, and can corrupt any individuals, publication of any generalized  $\mathcal{D}^*$  fails to ensure any

$y$ -to- $x$  or  $(x - y)$ -growth guarantee,

unless  $x = 1$  and  $y = 0$ .

*Proof:* Consider the unfortunate case  $\mathcal{C} = \mathcal{E} - \{o\}$ , namely, the adversary knows the sensitive value of every individual in  $\mathcal{D}$  except  $o$ . Since  $\mathcal{D}^*$  contains all the precise sensitive values, after inspecting it, the adversary will find out the real sensitive value of  $o$ . This means that, no matter how small her/his prior confidence is, her/his posterior confidence is always 1.  $\square$

Lemma 2 theoretically confirms our motivation that generalization provides poor protection against corruption. Specifically, the only provable background-sensitive guarantees of generalization are the useless 0-to-1 and 1-growth guarantees.

The above discussion assumes  $|\mathcal{D}^*| = |\mathcal{D}|$ , namely, the parameter  $s$  of the *Cardinality* constraint equals 1. To modify generalization to support an  $s < 1$ , a trivial solution is to first obtain  $\mathcal{D}^*$  in the same way as  $s = 1$ , and then, publish a random sample set of  $\mathcal{D}^*$  with sampling rate  $s$ . However, the solution does not fulfill our *Privacy* requirement. In particular, Lemma 2 still applies to the random sample set.

#### IV. PERTURBED GENERALIZATION

This section illustrates an alternative anonymization approach that combines generalization with *perturbation* [7], [6] and *stratified sampling* [8]. The framework consists of 3 phases, as detailed in the sequel.

*Phase 1 (Perturbation):* Given a *retention probability*  $p \in [0, 1]$ , we create  $\mathcal{D}^p$ , by independently transforming each tuple  $t \in \mathcal{D}$  to a *perturbed tuple*  $t' \in \mathcal{D}^p$  as follows.

- P1.  $t'.v^q = t.v^q$  (perturbation does not affect QI attributes).
- P2.  $t'.A^s$  is decided by tossing a coin with head probability  $p$ : (i) if the coin heads,  $t'.A^s = t.A^s$ ; (ii) otherwise,  $t'.A^s$  is randomly generated in  $U^s$  following the uniform distribution. In either case, we say that  $t'.A^s$  is a *perturbed value*.  $\square$

In the next phase, we will perform generalization on  $\mathcal{D}^p$ . Before explaining the details, we must clarify the meanings of value-, vector-, and tuple-generalization. Let  $x$  be a value of a QI-attribute  $A_i^q$  ( $1 \leq i \leq d$ ), and  $x'$  a set of values of  $A_i^q$ . We say that  $x'$  *generalizes*  $x$ , if  $x \in x'$ . For example,  $x' = [21, 40]$  generalizes  $x = 25$ ;  $x' = \{M, F\}$  generalizes  $x = M$ . Given  $d$ -dimensional vectors  $v$  and  $v'$ ,  $v'$  *generalizes*  $v$ , if the  $i$ -th component of  $v'$  generalizes the corresponding component of  $v$ , for all  $i \in [1, d]$ . Finally, a tuple  $t'$  *generalizes* another tuple  $t$ , if they share the same sensitive value, and  $t'.v^q$  generalizes  $t.v^q$ .

*Phase 2 (Generalization):* Given  $\mathcal{D}^p$  and an integer  $k \geq 1$ , we obtain  $\mathcal{D}^g$  with these properties:

- G1. Each tuple in  $\mathcal{D}^g$  generalizes a distinct tuple in  $\mathcal{D}^p$ .

- G2. The QI-vector  $t.v^q$  of each tuple  $t \in \mathcal{D}^g$  is identical to the QI-vectors of at least  $k - 1$  other tuples in  $\mathcal{D}^g$ .

- G3. For any two tuples  $t_1, t_2 \in \mathcal{D}^g$ , if  $t_1.v^q \neq t_2.v^q$ , then there does not exist any vector  $v^q \in U^q$  such that  $t_1.v^q$  and  $t_2.v^q$  both generalize  $v^q$ .  $\square$

Property G3 implies that generalization conforms to the *global recoding* scheme [13]. There exist many algorithms [11], [13], [16] that can be used to obtain a  $\mathcal{D}^g$  with all the above properties.

*Phase 3 (Sampling):* Given  $\mathcal{D}^g$ , we produce  $\mathcal{D}^*$  by following these steps:

- S1. Group the tuples of  $\mathcal{D}^g$  by their QI attributes. Each resulting group is called a *QI-group*.
- S2. From each group  $QI$ , randomly sample a tuple  $t$ . We say that  $QI$  is the *source QI-group* of  $t$ .
- S3. Augment  $t$  with an attribute  $t.G$  storing the size of  $QI$ .
- S4. Add  $t$  to  $\mathcal{D}^*$ , and discard the other tuples in  $QI$ .  $\square$

$\mathcal{D}^*$  is a *stratified sample set* [8] of  $\mathcal{D}^g$ . Here, a “stratum” is a QI-group, and a sample is taken from each stratum.

The computation of  $\mathcal{D}^*$  through Phases 1-3 is based on two values  $p$  and  $k$ . We set  $k$  to  $\lceil 1/s \rceil$ , where  $s$  is the parameter of our *Cardinality* constraint. This ensures  $|\mathcal{D}^*|$  to be at most  $|\mathcal{D}| \cdot s$ . The formulation of  $p$ , on the other hand, depends on the degree of privacy control, and will be discussed in Section VI.

We illustrate perturbed generalization by using it to anonymize the microdata  $\mathcal{D}$  in Table Ia, assuming  $p = 0.25$  and  $s = 0.5$  (hence,  $k = 2$ ). Table IIa shows the  $\mathcal{D}^p$  after Phase 1, where all the sensitive values have been altered, except those of Calvin and Gloria. Table IIb illustrates  $\mathcal{D}^g$  at the end of Phase 2. The final  $\mathcal{D}^*$  from Phase 3 is given in Table IIc.  $\mathcal{D}^*$  is augmented with a column  $G$ . All the  $G$ -values are 2, because every QI-group in  $\mathcal{D}^g$  has size 2.

$\mathcal{D}^*$  may involve *absurd tuples*, which contradict common sense, and can never exist in any microdata. For instance, the last tuple in Table IIc is absurd because it associates *ovarian-cancer* with a male. Releasing such tuples is necessary, because they must be present to enable data mining (see [12]). Finally, note that it is not meaningful to judge whether  $\mathcal{D}^*$  captures sufficient information in  $\mathcal{D}$ , when the cardinality of  $\mathcal{D}$  is excessively low. Perturbation-based approaches work well only if  $|\mathcal{D}|$  is large. For instance, some sensitive values in Table IIa disappear in Table IIc; such phenomenon is rather unlikely when  $\mathcal{D}$  is sizable. In Section VII, we will test the utility of  $\mathcal{D}^*$  when  $\mathcal{D}$  is a real dataset.

#### V. MODELING PRIVACY ATTACKS

In the sequel, we provide the mathematical foundation for studying the privacy guarantees offered by perturbed generalization. Since the table  $\mathcal{D}^*$  we release is not a conventional generalized relation, a linking attack is different from that in previous work. Therefore, Section V-A first clarifies the procedural details of an attack. Currently our formulation of an adversary’s knowledge (before and after an attack) has stayed at the conceptual level. In Section V-B, we will make the formulation theoretically specific.

Owner	Age	Gender	Zipcode	Disease	Age	Gender	Zipcode	Disease	Age	Gender	Zipcode	Disease	G
Bob	25	M	25000	hypertension	[21, 40]	M	[11***, 30***]	hypertension	[21, 40]	M	[11***, 30***]	hypertension	2
Calvin	30	M	27000	pneumonia	[21, 40]	M	[11***, 30***]	pneumonia	[41, 60]	F	[11***, 30***]	breast cancer	2
Debbie	45	F	20000	breast cancer	[41, 60]	F	[11***, 30***]	breast cancer	[41, 60]	F	[31***, 50***]	bronchitis	2
Ellie	50	F	15000	bronchitis	[41, 60]	F	[11***, 30***]	bronchitis	[61, 80]	M	[51***, 70***]	ovarian cancer	2
Fiona	55	F	45000	bronchitis	[41, 60]	F	[31***, 50***]	bronchitis					
Gloria	58	F	32000	hypertension	[41, 60]	F	[31***, 50***]	hypertension					
Henry	65	M	65000	dementia	[61, 80]	M	[51***, 70***]	dementia					
Isaac	80	M	55000	ovarian cancer	[61, 80]	M	[51***, 70***]	ovarian cancer					

(a)  $\mathcal{D}^p$  after perturbation(b)  $\mathcal{D}^g$  after generalization(c)  $\mathcal{D}^*$  after sampling

TABLE II

ILLUSTRATION OF OUR PUBLICATION FRAMEWORK ( $p = 0.25, k = 2$ )

### A. Corruption-Aided Linking Attacks

Let us briefly review the basic notations in Section II. We have an adversary who knows the QI-vector  $o.v^q$  of a victim individual  $o$ , and that  $o$  exists in the microdata  $\mathcal{D}$ . S/he aims at inferring how likely the sensitive value  $o.A^s$  of  $o$  satisfies a predicate  $Q$ . Towards this purpose, the adversary may utilize an external database  $\mathcal{E}$  and the precise sensitive values of a set  $\mathcal{C}$  of individuals that s/he has corrupted.

Given a  $\mathcal{D}^*$  released by our solution, the adversary carries out her/his attack in three steps A1, A2, A3.

A1. S/he retrieves the *unique* tuple  $t \in \mathcal{D}^*$  such that  $t.v^q$  generalizes  $o.v^q$ .

The uniqueness of  $t$  is guaranteed by Property G2 and Step S2, as explained in Section IV. We say that  $t$  is the *crucial tuple* of the attack.

A2. S/he collects the set  $\mathcal{O}$  of individuals  $o_1, \dots, o_e$  from  $\mathcal{E}$  that are different from  $o$ , and their QI-vectors  $o_1.v^q, \dots, o_e.v^q$  can be generalized to  $t.v^q$ .

These  $e$  persons, together with  $o$ , are the only *candidates*, who can be the owner of  $t$ . Note that  $e + 1$  is at least  $t.G$ , because  $\{o, o_1, \dots, o_e\}$  must capture the owners of all the tuples in the source QI-group of  $t$ . Since each QI-group has a size at least  $k$ , we have  $e + 1 \geq k$ .

A3. S/he calculates her/his posteriori confidence  $P_{post}(Q)$ , by combining  $\mathcal{D}^*$ ,  $\mathcal{O}$ ,  $\mathcal{C}$  with her/his own expertise.

*Example 1:* Assume that an adversary attempts to derive the probability of Ellie having a respiratory problem, namely, the property  $Q$  of the attack is “ $o.A^s$  is a respiratory disease”, where  $o$  equals Ellie. S/he consults the  $\mathcal{D}^*$  in Table IIc and the voter registration list  $\mathcal{E}$  in Table Ib. Furthermore,  $\mathcal{C} = \{\text{Debbie, Emily}\}$ . That is, the adversary knows that Debbie contracted *pneumonia*, and Emily is extraneous.

At Step A1, the adversary identifies the crucial tuple  $t$  as the second row of  $\mathcal{D}^*$ . At Step A2, s/he retrieves, from  $\mathcal{E}$ ,  $e = 2$  individuals:  $o_1 = \text{Debbie}$  and  $o_2 = \text{Emily}$ , whose QI-vectors can be generalized to  $t.v^q$ . Namely,  $\mathcal{O} = \{\text{Debbie, Emily}\}$ . At Step A3, the adversary analyzes the probability of Ellie’s disease satisfying  $Q$ , from all the information that s/he has acquired. Obviously, as Emily is extraneous, the adversary removes her from further consideration, which leaves only two candidate owners of  $t$ : Debbie and Ellie.

Although the observed *Disease*-value (*breast-cancer*) of  $t$  differs from the real disease *pneumonia* of Debbie, it is wrong to conclude that Debbie does not own  $t$ , because every sensitive value in  $\mathcal{D}^*$  may have been altered in random perturbation. The adversary needs to infer the sensitive value of Ellie through a probabilistic analysis, as detailed in the next subsection.  $\square$

### B. Posterior Confidence Derivation

Before a linking attack, the adversary may already have certain background knowledge about the victim’s sensitive value  $o.A^s$ . We observe that any knowledge essentially permits an adversary to evaluate the probabilities of  $o.A^s$  taking specific values. Thus, we model the background knowledge through a probability density function (pdf):

*Definition 4 (Background Knowledge):* Let  $X$  be a random variable modeling the distribution of  $o.A^s$ . An adversary’s *background knowledge* is a pdf of  $X$ :

$$P[X = x], \quad (4)$$

where  $x$  can be any value in  $U^s$ . If

$$\max_{x \in U^s} P[X = x] \leq \lambda,$$

the background knowledge is  $\lambda$ -skewed.  $\square$

The above definition trivially captures the background knowledge targeted by  $(c, l)$ -diversity (see Section III). Specifically, if an adversary knows that  $o.A^s$  cannot be a value  $x$ , then  $P[X = x] = 0$ .

$\lambda$  limits an adversary’s maximum confidence about the most likely value for  $o.A^s$ . The lower bound of  $\lambda$  equals  $1/|U^s|$ . When  $\lambda$  takes this value, the adversary does not have non-trivial expertise about  $o.A^s$ , and hence, assumes that  $o.A^s$  can be any value in  $U^s$  with the same likelihood. In general, privacy protection is more difficult when  $\lambda$  is higher. In particular, for  $\lambda = 1$ , the adversary is affirmative about the exact  $o.A^s$ ; thus, no protection for  $o$  is possible.

We call the pdf in Definition 4 the *prior pdf* of  $X$ . Let  $Q(X)$  be the set of sensitive values qualifying property  $Q$ . The adversary’s prior confidence can be represented as

$$P_{prior}(Q) = \sum_{x \in Q(X)} P[X = x]. \quad (5)$$

Now we proceed to derive the adversary's posterior confidence. Consider, again, the crucial tuple  $t$  obtained at Step A1 (see Section V-A). This is the only tuple in  $\mathcal{D}^*$  relevant to  $o$ , since the other tuples' (generalized) QI values are inconsistent with  $o.v^q$ . Let  $y$  be the observed sensitive value of  $t$ ; remember that  $y$  may have been perturbed. Our objective is to control the adversary's confidence about  $Q$ , after s/he has observed  $y$ . Therefore, we define her/his posterior confidence as

$$P_{post}(Q) = P[Q|y]. \quad (6)$$

$P_{post}(Q)$  may differ from  $P_{prior}(Q)$  because, after seeing  $y$ , the adversary can derive a *posterior pdf* of  $X$ , which is not necessarily equivalent to the pdf describing her/his background knowledge (Definition 4). Analogous to Formula 4, the new pdf can be represented as:

$$P[X = x|y], \quad (7)$$

where  $x$  is any value in  $U^s$ . Formula 7 can be solved in two steps. First, the adversary figures out the probability  $h$  that the crucial tuple  $t$  indeed belongs to the victim  $o$ , namely:

$$h = P[o \text{ owns } t|y]. \quad (8)$$

Then, the adversary distinguishes two disjoint events:

- *Event 1 (probability  $1 - h$ ):  $o$  is not the owner of  $t$ .* In this case,  $\mathcal{D}^*$  contains no hint about  $o$  at all. Therefore, the adversary's knowledge of  $o.A^s$  remains the same as her/his background knowledge.
- *Event 2 (probability  $h$ ):  $o$  is the owner of  $t$ .* As a result, the adversary obtains a piece of information helpful for calculating Formula 7: the sensitive value of  $o$  has been modified from  $x$  to  $y$  in perturbation. In this case, we use a random variable  $Y$  to capture the perturbed sensitive value of  $o$ .

Combining both events, we have

$$P[X = x|y] = h \cdot P[X = x|Y = y] + (1 - h)P[X = x]. \quad (9)$$

It follows that the adversary's posterior confidence (Equation 6) can be calculated as

$$P_{post}(Q) = \sum_{x \in Q(X)} P[X = x|y]. \quad (10)$$

## VI. FORMAL RESULTS

Following [6], let  $P[a \rightarrow b]$  denote the probability that a sensitive value  $a$  is perturbed to  $b$ . It holds that

$$P[a \rightarrow b] = \begin{cases} p + (1 - p)/|U^s| & \text{if } a = b \\ (1 - p)/|U^s| & \text{otherwise} \end{cases} \quad (11)$$

Let us make several observations about Equation 9. First, the expression  $P[X = x|Y = y]$  can be re-written as:

$$\frac{P[X = x, Y = y]}{P[Y = y]} = \frac{P[X = x] \cdot P[x \rightarrow y]}{p \cdot P[X = y] + (1 - p)/|U^s|}. \quad (12)$$

Next, we study the value of  $h$ . Without loss of generality, assume that  $\mathcal{C} \cap \mathcal{O}$  has  $\alpha$  individuals, among whom  $\beta$  are not extraneous. Let the  $\beta$  people be  $o_1, \dots, o_\beta$ , whose real

sensitive values are  $x_1, \dots, x_\beta$ , respectively. Let  $o_{\beta+1}, \dots, o_\alpha$  be the extraneous persons in  $\mathcal{C}$ .

Since the adversary has confirmed  $\beta + 1$  people (i.e.,  $o, o_1, \dots, o_\beta$ ) in the source QI-group of  $t$ , s/he assumes that each person in  $\mathcal{O} - \mathcal{C} = \{o_{\alpha+1}, \dots, o_e\}$  appears in that group with probability

$$g = (t.G - 1 - \beta)/(e - \alpha). \quad (13)$$

Equation 8 has an equivalent form

$$h = P[o \text{ owns } t, y]/P[y]. \quad (14)$$

The numerator  $P[o \text{ owns } t, y]$  is essentially the probability of two independent events happening simultaneously: (i) the tuple of  $o$  was sampled in Step S2 of Phase 3, and (ii) its sensitive value was perturbed to  $y$ . Hence,

$$\begin{aligned} P[o \text{ owns } t, y] &= \frac{1}{t.G} \sum_{x \in U^s} (P[X = x] \cdot P[x \rightarrow y]) \\ &= \frac{1}{t.G} \left( p \cdot P[X = y] + \frac{1 - p}{|U^s|} \right). \end{aligned} \quad (15)$$

When the adversary's knowledge is  $\lambda$ -skewed,

$$P[o \text{ owns } t, y] \leq \frac{1}{t.G} \left( p \cdot \lambda + \frac{1 - p}{|U^s|} \right). \quad (16)$$

The denominator  $P[y]$  of Equation 14 equals  $P[o \text{ owns } t, y] +$

$$\sum_{i=1}^{\beta} P[o_i \text{ owns } t, y] + \sum_{j=\alpha+1}^e P[o_j \text{ owns } t, y]. \quad (17)$$

For  $i \in [1, \beta]$ ,

$$\begin{aligned} P[o_i \text{ owns } t, y] &= P[x_i \rightarrow y]/t.G \\ &\geq (1 - p)/(t.G \cdot |U^s|). \end{aligned} \quad (18)$$

Given any  $j \in [\alpha + 1, e]$ , we model the sensitive value of  $o_j$  with a random variable  $X_j$ . The derivation of  $P[o_j \text{ owns } t, y]$  is similar to solving  $P[o \text{ owns } t, y]$  into Equation 15, except that we must take into account the fact that  $o_j$  appears in  $\mathcal{D}$  with probability  $g$ , as given in Equation 13. Therefore,

$$\begin{aligned} P[o_j \text{ owns } t, y] &= \frac{g}{t.G} \left( p \cdot P[X_j = y] + \frac{1 - p}{|U^s|} \right) \\ &\geq g \cdot (1 - p)/(t.G \cdot |U^s|). \end{aligned} \quad (19)$$

It follows from Equations 17-19 that

$$P[y] \geq P[o \text{ owns } t, y] + \frac{(t.G - 1)(1 - p)}{t.G \cdot |U^s|}$$

Combining the above formula with Equation 14 and Inequality 16, we have

$$\begin{aligned} h &\leq \frac{p \cdot \lambda + (1 - p)/|U^s|}{p \cdot \lambda + t.G \cdot (1 - p)/|U^s|} \\ &\leq \frac{p \cdot \lambda + (1 - p)/|U^s|}{p \cdot \lambda + k \cdot (1 - p)/|U^s|}. \end{aligned} \quad (20)$$

In the sequel, we use  $h^\top$  to denote the right hand side of the above inequality.

*Theorem 1:* If  $y$  does not satisfy property  $Q$ , no  $\rho_1$ -to- $\rho_2$  and  $\Delta$ -growth breaches can occur, for any  $\rho_1, \rho_2$ , and  $\Delta$ .

*Proof:* When  $x \neq y$ , Equation 12

$$\leq \frac{P[X = x] \cdot (1 - p) / |U^s|}{(1 - p) / |U^s|} = P[X = x]. \quad (21)$$

From Equations 9 and 10, we know  $P_{post}(Q) =$

$$\sum_{x \in Q(X)} (h \cdot P[X = x | Y = y] + (1 - h) \cdot P[X = x]). \quad (22)$$

If  $y \notin Q(X)$ , by Inequality 21, the previous equation

$$\begin{aligned} &\leq \sum_{x \in Q(X)} (h \cdot P[X = x] + (1 - h) \cdot P[X = x]) \\ &= P_{prior}(Q). \end{aligned}$$

Therefore, no privacy breach can occur.  $\square$

For instance, the attack in Example 1 will not incur any breach, because  $y = \text{breast-cancer}$  does not qualify  $Q = \text{“a respiratory disease”}$ . In the subsequent discussion, we focus on a linking attack whose  $Q$  is qualified by  $y$ .

*Theorem 2:* No  $\rho_1$ -to- $\rho_2$  breach can happen if:

$$\frac{\rho'_2(1 - \rho_1)}{\rho_1(1 - \rho'_2)} \geq 1 + \frac{p}{(1 - p) / |U^s|} \quad (23)$$

where  $\rho'_2 = (\rho_2 - \rho_1(1 - h^\top)) / h^\top$ .

*Proof:* Let us rewrite Equation 10 as

$$P_{post}(Q) = h \cdot P[Q | Y = y] + (1 - h) \cdot P_{prior}[Q].$$

Using the same mathematical derivation as in the proof of Statement 1 in [6], we can show that, when Inequality 23 holds,  $P[Q | Y = y] \leq \rho'_2$ . Hence

$$\begin{aligned} P_{post}(Q) &\leq h \cdot \rho'_2 + (1 - h) \cdot \rho_1 \\ &\leq h^\top \cdot \rho'_2 + (1 - h^\top) \cdot \rho_1 = \rho_2 \end{aligned}$$

which completes the proof.  $\square$

*Theorem 3:* Let  $F(w) = (-p \cdot w^2 + p \cdot w) / (p \cdot w + u)$ , where  $u = (1 - p) / |U^s|$ . Let  $w_m = (\sqrt{u^2 + p \cdot u} - u) / p$ . No  $\Delta$ -growth breach occurs if either of the following holds:

- $\lambda \leq w_m$  and  $\Delta \geq h^\top \cdot F(\lambda)$ ;
- $\lambda > w_m$  and  $\Delta \geq h^\top \cdot F(w_m)$ .

*Proof:* Assume that  $Q$  is qualified by  $y$ ; otherwise, no privacy breach can happen according to Theorem 1. By Equations 5, 9, and 10,  $P_{post}(Q) - P_{prior}(Q) =$

$$h \sum_{x \in Q(X)} (P[X = x | Y = y] - P[X = x])$$

According to Inequality 21, the above

$$\leq h^\top (P[X = y | Y = y] - P[X = y])$$

Let  $w = P[X = y] \leq \lambda$ . By Equation 12, the above formula equals

$$h^\top \left( \frac{w \cdot (p + (1 - p) / |U^s|)}{w \cdot p + (1 - p) / |U^s|} - w \right) = h^\top \cdot F(w).$$

Next, we calculate the upper bound of  $F(w)$ , as  $w$  distributes in  $[0, \lambda]$ . We note that  $dF(w)/dw = 0$ , when  $w = w_m = (\sqrt{u^2 + p \cdot u} - u) / p$ . Furthermore,  $F(w)$  is monotonically increasing (or decreasing), if  $w < w_m$  (or  $w > w_m$ ). Hence, in case  $\lambda \leq (\sqrt{u^2 + p \cdot u} - u) / p$ ,  $F(w)$  reaches its maximum at  $w = \lambda$ . Otherwise, the maximum is obtained at  $w = w_m$ .  $\square$

Recall that our publication framework has two parameters  $p$  and  $k$ . As explained in Section II, we always fix  $k$  to  $\lceil 1/s \rceil$  to meet the *Cardinality* requirement. The value of  $p$ , on the other hand, is determined so that an objective *level* of privacy breaches is prevented. Specifically, for  $\rho_1$ -to- $\rho_2$  breaches, a level is specified by a pair of  $\rho_1$  and  $\rho_2$ . Likewise, a level of  $\Delta$ -growth breaches is described by  $\Delta$ . In any case, once the level is fixed,  $p$  is set to the minimum value that guarantees absence of the corresponding breaches, according to Theorems 2 and 3.

## VII. EXPERIMENTS

This section experimentally evaluates the effectiveness of the proposed technique, referred to as *perturbed generalization* (PG) in the sequel. We adapt the algorithm in [11] to implement Phase 2 of PG, and explore its utility in mining decision trees with the algorithm in [12].

### A. Data

We deploy a real database SAL that is widely used in the literature, and downloadable at <http://ipums.org>. SAL contains 700k tuples, each of which describes the personal data of an American. There are 9 discrete attributes: *Age, Gender, Education, Birthplace, Occupation, Race, Work-class, Marital-status, Income*. We treat *Income* as the sensitive attribute and the other columns as *QI*-attributes.

The *Income* domain consists of values 0, 1, 2, ..., 49, where each number  $i$  represents the range (in US dollars) of  $[i \cdot 2000, (i + 1) \cdot 2000)$ . For the purpose of decision tree mining, we divide the domain into  $m$  categories, varying  $m$  between 2 and 3. For  $m = 2$ , the first and second categories cover the ranges  $[0, 24]$  and  $[25, 49]$ , respectively. For  $m = 3$ , the category ranges become  $[0, 24]$ ,  $[25, 36]$ ,  $[37, 49]$ . Notice that, by setting  $m$  to 3, we are refining the “wealthier” category of  $m = 2$ .

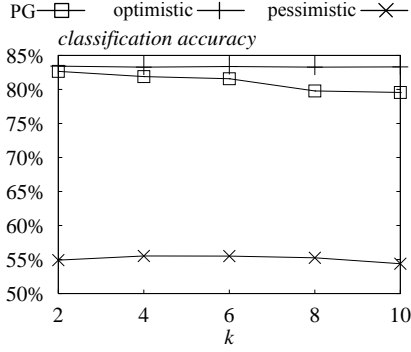
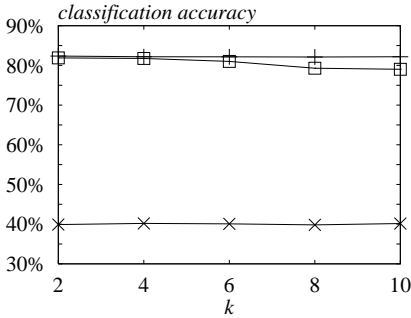
### B. Competitors

Currently no existing solution can provide privacy guarantees matching those offered by PG. Therefore, we compare the utility of PG against those of two yardstick methods: *optimistic* and *pessimistic*. Specifically, both *optimistic* and *pessimistic* build a decision tree from a random subset (of the microdata  $\mathcal{D}$ ) with size  $|\mathcal{D}|/k$  (recall that this is the upper bound of the number of tuples released by PG). For *optimistic*, each tuple

$k$	2	4	6	8	10
$\rho_2$	$\geq 0.69$	$\geq 0.53$	$\geq 0.45$	$\geq 0.40$	$\geq 0.36$
$\Delta$	$\geq 0.47$	$\geq 0.31$	$\geq 0.24$	$\geq 0.19$	$\geq 0.16$

(a)  $p = 0.3$ 

$p$	0.15	0.2	0.25	0.3	0.35	0.4	0.45
$\rho_2$	$\geq 0.34$	$\geq 0.38$	$\geq 0.41$	$\geq 0.45$	$\geq 0.49$	$\geq 0.52$	$\geq 0.56$
$\Delta$	$\geq 0.12$	$\geq 0.16$	$\geq 0.20$	$\geq 0.24$	$\geq 0.28$	$\geq 0.32$	$\geq 0.36$

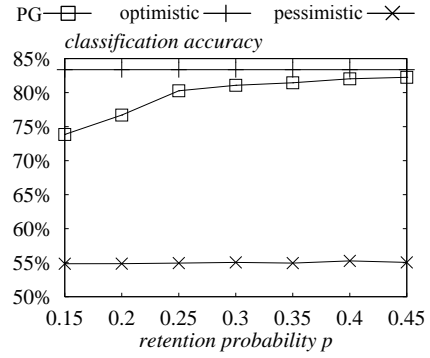
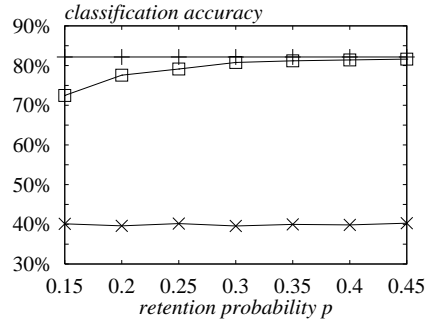
(b)  $k = 6$ TABLE III  
PRIVACY GUARANTEES OF PG(a)  $m = 2$ (b)  $m = 3$ Fig. 2. Utility vs.  $k$  ( $p = 0.3$ )

in the subset is taken directly from  $\mathcal{D}$ , i.e., no perturbation performed. For *pessimistic*, however, all tuples in the subset have their sensitive values randomized, i.e., perturbation with retention probability 0. We employ the tree growing algorithm in [17] to implement *optimistic* and *pessimistic*. We note that *pessimistic* creates a useless decision tree from a randomized dataset that loses all the sensitive information in  $\mathcal{D}$ .

We measure the utility of a method by the *classification accuracy* of its decision tree. Specifically, we use the tree to classify all the tuples in the microdata, and calculate the accuracy as the percentage of the correctly classified tuples. Ideally, the accuracy of PG should be as good as that of *optimistic*, yet significantly better than that of *pessimistic*.

### C. Privacy Guarantees

We aim at privacy protection against 0.1-skewed background knowledge. Furthermore, in preventing  $\rho_1$ -to- $\rho_2$  breaches, we guard against adversaries with prior confidence at most 0.2. Thus,  $\lambda$  and  $\rho_1$  are set to 0.1 and 0.2 respectively in the following experiments.

(a)  $m = 2$ (b)  $m = 3$ Fig. 3. Utility vs.  $p$  ( $k = 6$ )

For PG, the degree of privacy protection is determined by both parameters  $p$  and  $k$ . We will vary  $p$  from 0.15 to 0.45, and  $k$  from 2 to 10. Table III demonstrates the privacy guarantees (derived from Theorems 2 and 3) determined by all pairs of  $p$  and  $k$  to be used together in the subsequent experiments. For example, the second (third) row of Table IIIa indicates that, given  $p = 0.3$ , PG provides a 0.2-to-0.69 (0.47-growth) guarantee for  $k = 2$ , a 0.2-to-0.53 (0.31-growth) guarantee for  $k = 4$ , and so on. As expected, stronger protection is achieved with a lower  $p$  or higher  $k$ .

### D. Utility

The next set of experiments inspects the influence of  $k$  and  $p$  on the utility of PG. In Figure 2a (2b), we use  $m = 2$  ( $=3$ ),  $p = 0.3$ , and measure the classification errors of PG, as  $k$  changes from 2 to 10. We also include the errors of *optimistic* and *pessimistic*, which are not affected by  $k$ , because the two methods do not involve generalization.

The utility of PG stays close to *optimistic*, and degrades very slowly as  $k$  grows. This observation indicates that generalization (Phase 2 of PG) has limited impacts on the utility of the published dataset. In fact, when the microdata  $\mathcal{D}$  has a large cardinality, the QI-vectors of the underlying tuples are dense in  $U^q$ . For  $k \leq 10$ , a generalized QI value is an interval covering a tiny fraction of the QI attribute. This explains why generalization does not affect the quality of data analysis significantly.

In Figure 3, we present the classification errors of alternative methods, by fixing  $k$  to the median value 6, and varying  $p$  from 0.15 to 0.45. The performance of *optimistic* and *pessimistic*

does not change with  $p$ , because the former involves no perturbation, while the latter carries out total perturbation (with  $p = 0$ ). As expected, the utility of PG improves as  $p$  becomes larger. This is a standard characteristic of all the perturbation-based approaches.

### VIII. RELATED WORK

Privacy preserving data publication is first introduced by Sweeney and Samarati [4], [5], who also propose the concept of generalization and  $k$ -anonymity. Since then, numerous generalization principles have been developed. These include  $l$ -diversity [9] (which is discussed in Section IV),  $t$ -closeness [14], *personalization* [15],  $(k, e)$ -anonymity [18],  $\delta$ -presence [19],  $(c, k)$ -safety [20], *privacy skyline* [21], and  $m$ -invariance [22], and so on. Generalization conforming to these principles can be computed by numerous algorithms [23], [1], [11], [24], [2], [25], [3], [13], [16], [26], [27], [28], [29], [30], [31]. Various principles guard anonymity against different types of background knowledge. However, as explained in Section IV, all the above principles succumb to adversaries that have the corruption ability.

*Perturbation* shapes its original form from a classical surveying technique called *randomized responses* [32]. It is renovated for privacy preserving data mining in recent years [7], [6]. Parallel to our work, Rastogi et al. [33] adapt perturbation to data publication. Different from their method, we propose a concrete model for capturing corruption-based privacy attacks, and derive solid privacy guarantees under this model.

### IX. CONCLUSIONS

This paper tackles a new threat of privacy disclosure, called *corruption*, which has not been considered in the literature of privacy preserving publication. The conventional methods may incur severe privacy breaches, when challenged by corruption. Motivated by this, we present a new anonymization technique that integrates generalization, perturbation, and stratified sampling. The integration ensures strong privacy guarantees, even if an adversary has successfully corrupted any number of data owners. Furthermore, the data released by our technique permits a researcher to perform effective data mining about the microdata. Our theoretical findings are confirmed with experimentation.

This work lays down a foundation for further studies of anonymized publication. An exciting topic is re-publication of an anonymized version of the microdata, after it has been updated [34]. This is a difficult problem because we must prevent an adversary from inferring sensitive data by leveraging the correlation among subsequent releases. Another promising direction is to extend the proposed technique to non-relational objects such as spatial data [35]. Privacy guarantees in those scenarios need to be re-derived, since different forms of privacy breaches must be prevented.

### X. ACKNOWLEDGEMENTS

This work is fully supported by Grants CUHK 4161/07 and CUHK 1202/06 from the HKRGC.

### REFERENCES

- [1] R. Bayardo and R. Agrawal, "Data privacy through optimal  $k$ -anonymization," in *ICDE*, 2005, pp. 217–228.
- [2] V. Iyengar, "Transforming data to satisfy privacy constraints," in *SIGKDD*, 2002, pp. 279–288.
- [3] D. Kifer and J. Gehrke, "Injecting utility into anonymized datasets," in *SIGMOD*, 2006, pp. 217–228.
- [4] P. Samarati, "Protecting respondents' identities in microdata release," *TKDE*, vol. 13, no. 6, pp. 1010–1027, 2001.
- [5] L. Sweeney, "Achieving  $k$ -anonymity privacy protection using generalization and suppression," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 571–588, 2002.
- [6] A. V. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *PODS*, 2003, pp. 211–222.
- [7] R. Agrawal, R. Srikant, and D. Thomas, "Privacy preserving olap," in *SIGMOD Conference*, 2005, pp. 251–262.
- [8] S. Chaudhuri, G. Das, and V. Narasayya, "A robust, optimization-based approach for approximate answering of aggregate queries," in *SIGMOD*, 2001, pp. 295–306.
- [9] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian, "l-diversity: Privacy beyond  $k$ -anonymity," in *ICDE*, 2006, p. 24.
- [10] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *SIGMOD*, 2000, pp. 439–450.
- [11] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *ICDE*, 2005, pp. 205–216.
- [12] "Extended version. [Http://www.cse.cuhk.edu.hk/~taoyf](http://www.cse.cuhk.edu.hk/~taoyf)."
- [13] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain  $k$ -anonymity," in *SIGMOD*, 2005, pp. 49–60.
- [14] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond  $k$ -anonymity and  $l$ -diversity," in *ICDE*, 2007, pp. 106–115.
- [15] X. Xiao and Y. Tao, "Personalized privacy preservation," in *SIGMOD*, 2006, pp. 229–240.
- [16] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional  $k$ -anonymity," in *ICDE*, 2006, pp. 277–286.
- [17] M. Mehta, R. Agrawal, and J. Rissanen, "SLIQ: A fast scalable classifier for data mining," in *EDBT*, 1996, pp. 18–32.
- [18] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, "Aggregate query answering on anonymized tables," in *ICDE*, 2007, pp. 116–125.
- [19] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," in *SIGMOD*, 2007, pp. 665–676.
- [20] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern, "Worst-case background knowledge in privacy," in *ICDE*, 2007.
- [21] B.-C. Chen, R. Ramakrishnan, and K. LeFevre, "Privacy skyline: Privacy with multidimensional adversarial knowledge," in *VLDB*, 2007, pp. 770–781.
- [22] X. Xiao and Y. Tao, " $m$ -invariance: towards privacy preserving re-publication of dynamic datasets," in *SIGMOD*, 2007, pp. 689–700.
- [23] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Anonymizing tables," in *ICDT*, 2005, pp. 246–258.
- [24] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," in *VLDB*, 2007, pp. 758–769.
- [25] T. Iwuchukwu and J. F. Naughton, "K-anonymization as spatial indexing: Toward scalable and incremental anonymization," in *VLDB*, 2007, pp. 746–757.
- [26] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Workload-aware anonymization," in *SIGKDD*, 2006.
- [27] A. Meyerson and R. Williams, "On the complexity of optimal  $k$ -anonymity," in *PODS*, 2004, pp. 223–228.
- [28] H. Park and K. Shim, "Approximate algorithms for  $k$ -anonymity," in *SIGMOD*, 2007, pp. 67–78.
- [29] K. Wang and B. C. M. Fung, "Anonymizing sequential releases," in *SIGKDD*, 2006, pp. 414–423.
- [30] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, "Minimality attack in privacy preserving data publishing," in *VLDB*, 2007, pp. 543–554.
- [31] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *VLDB*, 2006, pp. 139–150.
- [32] S. L. Warner, "Randomized response: a survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 6, pp. 63–69, 1965.
- [33] V. Rastogi, S. Hong, and D. Suciu, "The boundary between privacy and utility in data publishing," in *VLDB*, 2007, pp. 531–542.

- [34] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li, "Secure anonymization for incremental datasets," in *SDM*, 2006, pp. 48–63.
- [35] M. F. Mokbel, C.-Y. Chow, and W. G. Aref, "The new casper: query processing for location services without compromising privacy," in *VLDB*, 2006, pp. 763–774.