Fine-Grain Perturbation for Privacy Preserving Data Publishing

Rhonda Chaytor, Ke Wang
School of Computing Science
Simon Fraser University, BC, Canada
{rlc8, wangk}@cs.sfu.ca

Patricia Brantingham
School of Criminology
Simon Fraser University, BC, Canada
pbrantin@sfu.ca

***Abstract*— Recent work [12] shows that conventional privacy preserving publishing techniques based on anonymity-groups are susceptible to corruption attacks. In a corruption attack, if the sensitive information of any anonymity-group member is uncovered, then the remaining group members are at risk. In this study, we abandon anonymity-groups and hide sensitive information through perturbation on the sensitive attribute. With each record being perturbed independently, corruption attacks cannot be effectively carried out. Previous anti-corruption work did not minimize information loss. This paper proposes to address this issue by allowing fine-grain privacy specification. We demonstrate the power of our approach through experiments on real medical and synthetic datasets.**

I. INTRODUCTION

In the field of *privacy preserving data publishing (PPDP)*, a trusted publisher has collected raw personal data, called *microdata*, and wants to publish it for research purposes. Suppose that a hospital wants to publish the medical microdata in Fig. 1 (a) to researchers at a medical school. As *Disease* is a *sensitive attribute* (SA), the publication must prevent an adversary from inferring the disease of any patient. At the same time, the publication should retain its usefulness for data analysis.

To prepare microdata for publication, unique identifiers like names are first removed; however, this *de-identification* is not enough to safeguard against privacy attacks. *Linking attacks* [10][11] may still occur when an adversary knows a patient's unique combination of public attribute values, called *quasi-identifier* (QI) attributes. To illustrate, in Fig. 1 (a), *Name* has been removed, and public attributes *Age* and *Sex* will be published. Suppose that *Name*, *Age*, and *Sex* appear publicly in other external sources (*e.g.*, voter registration lists), such as the one in Fig. 1 (b). In this case, an adversary can join the two tables in Fig. 1 (a) and (b) to reveal for example that the only 32 year old female, Eva, has swine flu virus "H1N1".

A common technique for preventing linking attacks is hiding an individual in an *anonymity-group*. Consider the generalized publication in Fig. 1 (c), where "*" represents "any sex". It is *k-anonymous* [10][11], $k = 2$, because for each patient, there are at least k-1 other patients having identical values on the QI attributes. Therefore, even if an adversary knows that Eva is a 32 year old female, he would only be $1/k = 50\%$ certain that Eva's disease is "H1N1", since it could equally as likely be "cancer".

(a) Medical microdata

	Age	*Sex*	*Disease*
1	21	M	SARS
2	25	F	HIV
3	26	F	SARS
4	28	M	HIV
5	32	F	H1N1
6	34	F	cancer
7	36	M	H1N1
8	39	M	cancer

(b) External table

Name	*Age*	*Sex*
Andy	21	M
Beth	25	F
Carla	26	F
Doug	28	M
Eva	32	F
Fiona	34	F
Gord	36	M
Hank	39	M

(c) Generalization

	Age	*Sex*	*Disease*
1	[21-25]	*	SARS
2			HIV
3	[26-30]	*	SARS
4			HIV
5	[31-35]	F	H1N1
6			cancer
7	[36-40]	M	H1N1
8			cancer

(d) Perturbed generalization

	Age	*Sex*	*Disease*
1	[21-25]	*	HIV
2			
3	[26-30]	*	H1N1
4			
5	[31-35]	F	SARS
6			
7	[36-40]	M	cancer
8			

Figure 1. Privacy preserving data publication.

A. Motivation

Recent work [12] shows that anonymized tables are susceptible to *corruption attacks*. For example, suppose the adversary learns that 34 year old Fiona has "cancer" from background knowledge [6]. Now from the third group in Fig. 1 (c), the adversary can deduce with 100% certainty that Eva's disease is "H1N1." The authors of [12] propose *Perturbed Generalization (PG)* to prevent corruption attacks. First, they retain a percentage p of SA-values and perturb the rest to other values in the domain. Then, they create anonymity-groups of size k by generalizing QI attributes. Finally, they sample one perturbed record from each group. Fig. 1 (d) shows an example of the final result assuming $p = 1/3$ and $k = 2$. In the third group in Fig. 1 (d), there is a 1/2 chance that Eva's record gets sampled. Now even if Fiona is corrupted, the probability that Eva has "H1N1" is only $1/3 \times 1/2 = 1/6 \approx 17\%$.

Given their pioneering work to combat corruption attacks, the focus in [12] was not on optimizing utility. Specifically, each of perturbation, generalization, and sampling introduces distortion to the data. Generalization loses significant information for aggregate queries [14] and record sampling makes it difficult to reconstruct the distribution on

SA, which depends on a sufficient number of perturbed records.

B. Contributions

The root of corruption attacks is the anonymity-group, which is used to hide an individual. Once a group member's SA-value is corrupted, the remaining group members are at a higher risk. In this work, we abandon the anonymity-group approach. Instead, we disguise a record's SA-value by perturbing it. Unlike PG [12], we publish all QI attributes without modification and all perturbed records without sampling. Since each record's SA-value is perturbed *independently*, the knowledge of one individual's SA-value provides no clue about another individual's SA-value. In this way, corruption attacks are prevented. There is a detailed discussion in Section III. The focus of this paper is on the utility of perturbed records. A key observation motivates our work.

Fine-Grain Privacy SA-values are not equally sensitive and should be perturbed according to a probability distribution that matches their sensitivity. We extend (ρ_1, ρ_2)-privacy in the literature [3] to allow *fine-grain* (ρ_{1i}, ρ_{2i})-privacy for each SA-value x_i. Informally, this privacy notion limits the posterior probability of inferring the original SA-value x_i (after seeing the perturbed record) to ρ_{2i} whenever prior probability is no more than ρ_{1i}. We identify the optimal fine-grain perturbation operator for a given (ρ_{1i}, ρ_{2i})-privacy requirement for all SA-values x_i, where p_i is the retention probability for x_i. In real life applications, the publisher will set the (ρ_{1i}, ρ_{2i})-privacy parameters for each SA-value x_i based on the perceived sensitivity of x_i. We set (ρ_{1i}, ρ_{2i}) parameters based on the intuition that "less frequent values are more sensitive", which holds in many practical cases.

Example 1: Assume the (ρ_{1i}, ρ_{2i})-privacy requirement for "SARS", "HIV", "H1N1", and "cancer" is (1/10, 1/7), (1/10, 1/4), (1/9, 19/35), and (1/8, 18/25), respectively. Given the data in Fig. 1 (a), the uniform and the optimal fine-grain perturbation operators are shown in Fig. 2 (a) and (b) (more details later). Each entry (j, i) contains the probability that value x_i is perturbed to value x_j. As shown by the larger *retention probabilities* along the diagonal, the fine-grain operator retains more of the less sensitive values than uniform operator.

Not only does it make sense from a privacy point of view to give the highly-sensitive SA-values more protection, as our results in Section V demonstrate, this strategy also increases utility for ad-hoc privacy preserving data publishing tasks.

(a) Uniform

SARS	1/3	2/9	2/9	2/9
HIV	2/9	1/3	2/9	2/9
H1N1	2/9	2/9	1/3	2/9
cancer	2/9	2/9	2/9	1/3

(b) Fine-grain

SARS	1/4	1/6	1/6	1/6
HIV	1/4	1/2	1/6	1/6
H1N1	1/4	1/6	1/2	1/6
cancer	1/4	1/6	1/6	1/2

Figure 2. Different perturbation operators.

In the rest of the paper, Section II reviews related work, Section III defines the problem we study, Section IV presents our optimal fine-grain perturbation algorithm, Section V evaluates these solutions, and Section VI concludes our paper.

II. Related Work

Since Samarati and Sweeney's work [10][11] on privacy preserving data publishing, there have been several proposals of privacy principles, (*e.g.*, *l*-diversity [7], *t*-closeness [5]). Tao et al. [12] showed solutions based on these principles are susceptible to corruption attacks because they rely on anonymity-groups to hide information. Parallel to Tao et al. [12], Rastogi et al. [8] also adapted perturbation to data publishing, but did not optimize utility in this setting.

Using perturbation to disguise sensitive information was first studied in a classical surveying technique called *randomized responses* [13], and more recently used for *privacy preserving data mining*, where individuals perturb their data *before* sending it to the publisher, and the publication is intended for a *specific* data mining task, such as classification [2] or association rule mining [3][9]. We consider privacy preserving data publishing, where *unmodified* data has been collected by a publisher and is published for unknown *ad-hoc* tasks.

Our work shares a similarity with Huang and Du [4] on finding optimal perturbation operators. In contrast to our approach, they used a notion of optimality that corresponds to non-dominance in the privacy-utility space and they searched for all non-dominated perturbation operators using a genetic algorithm; we find a single optimal perturbation operator for a given privacy guarantee.

III. Problem

Given the microdata T containing one sensitive attribute (SA) and several quasi-identifier (QI) attributes, we want to find a perturbed version of T for publication, while guaranteeing a notion of privacy and maximizing some utility metric. We assume all attributes in T are categorical. This section provides basic definitions and assumptions necessary to formally define this problem.

SA has a domain $\{x_1,\ldots,x_m\}$ of size m. Let $|x_i|$ denote the number of records in T having the value x_i in SA and $|T|$ denote the number of records in T. The *frequency* f_i of x_i in T is equal to $|x_i| / |T|$. To disguise the sensitive values in SA, the publisher releases a perturbed table T^*, in which SA-values are *perturbed* to other values according to a fixed probability distribution. We use the following convention: x_i denotes a SA-value in T and y_i denotes the same value in T^*. Note that x_i and y_i are the same (*e.g.*, both are "H1N1"), but x_i is an *original value* in T and y_i is a *perturbed value* in T^*. X denotes a random variable for x_i in a record in T and Y denotes a random variable for y_i in a record in T^*.

A. Perturbation operator

For a record in T with SA-value x_i, we retain x_i with probability p_{ii} and perturb x_i to value y_j with probability p_{ji}, $j \neq i$. These probabilities are also denoted by $\Pr[x_i \rightarrow y_i] = p_{ii}$ and $\Pr[x_i \rightarrow y_j] = p_{ji}$. A *perturbation operator* is a perturbation matrix P having the entry (j, i) equal to p_{ji}, $\forall i, j = 1,\ldots,m$. Note $\sum_j p_{ji} = 1$.

We say that P is a *fine-grain* perturbation matrix if $p_{ii} = p_i + q_i$ and $p_{ji} = q_i$, $j \neq i$, where $q_i = (1 - p_i)/m$. In other words, for each record with SA-value x_i, x_i is retained with probability p_i and is perturbed to y_j with probability q_i, where y_j is chosen from the domain of SA. A fine-grain perturbation matrix looks like

$$P^{(1)} = \begin{bmatrix} p_1+q_1 & q_2 & \cdots & q_m \\ q_1 & p_2+q_2 & \cdots & q_m \\ \vdots & \vdots & \ddots & \vdots \\ q_1 & q_2 & \cdots & p_m+q_m \end{bmatrix} \quad (1)$$

Instead of a uniform retention probability p, each SA-value x_i has its own retention probability $p_{ji} = p_i + q_i$.

Like previous work (*e.g.*, [8]), we assume that the adversary is *record-independent*, that is, there is no correlation among records. Each perturbed SA-value y_j is chosen *independently at random* according to the probability distribution given in a perturbation matrix P. Let r be an original record with x_i on SA and let r^* be the perturbed record of r with y_j on SA. By receiving r_i^*, the adversary learns something about the original SA-value x in r_i; however, the independence assumption implies that all r_j^* and any knowledge about r_j, $j \neq i$, disclose nothing about x of r_i and can be ignored in the privacy analysis of r_i. Therefore, the corruption attacks discussed earlier are not effective.

B. Privacy model

Let r be an original record with SA-value x_i and let r^* be the perturbed version of r with SA-value y_j. $\Pr[X = x_i]$ denotes the prior probability of $X = x_i$ in the absence of any knowledge about r's SA, and $\Pr[X = x_i \mid Y = y_j]$ denotes the posterior probability of $X = x_i$ given the perturbed value $Y = y_j$. Since r^* and r agree on QI and since QI is public, the change in belief comes from revealing the perturbed value $Y = y_j$. We discuss the impact of QI on privacy in Section IV.

If the prior probability is too high, the adversary already knows $X = x_i$ before seeing the published data. For this purpose we borrow *(ρ_1, ρ_2)-privacy* from [3]: the posterior probability is not more than ρ_2 whenever the prior probability is not more than ρ_1, where $\rho_1 < \rho_2$. For example, a (1/10, 1/2)-privacy breach occurs when an adversary's prior probability is less than 10% and posterior probability is more than 50%. If the prior probability is more than 10%, there is no (1/10, 1/2)-privacy breach.

Our observation is that ρ_1 and ρ_2 depend on the *sensitivity* of a SA-value. For example, suppose the less sensitive "cancer" requires (1/2, 2/3)-privacy and "HIV" requires (1/10, 1/7)-privacy. Now it is not even possible to have one (ρ_1, ρ_2) setting: setting $\rho_1 \geq 1/2$ would not enforce (1/10, 1/7)-privacy for "HIV"; on the other hand, setting $\rho_1 < 1/2$ would not enforce (1/2, 2/3)-privacy for "cancer" if prior probability falls into $(\rho_1, 1/2]$. To address this issue, the fine-grain privacy below extends the (ρ_1, ρ_2)-privacy [3] to allow a different (ρ_1, ρ_2) for each value in SA.

Definition 1 (Fine-grain (ρ_{1i}, ρ_{2i})-privacy): Let (ρ_{1i}, ρ_{2i}) be the privacy requirement for SA-value x_i. There is an *upward (ρ_{1i}, ρ_{2i})-privacy breach* with respect to $x_i \in$ SA if for some $y_j \in$ SA, $\Pr[X = x_i] \leq \rho_{1i}$ and $\Pr[X = x_i \mid Y = y_j] > \rho_{2i}$. There is a *downward (ρ_{2i}, ρ_{1i})-privacy breach* with respect to $x_i \in$ SA if for some $y_j \in$ SA, $\Pr[X = x_i] \geq \rho_{2i}$ and $\Pr[X = x_i \mid Y = y_j] < \rho_{1i}$. Here, $0 < \rho_{1i} < \rho_{2i} < 1$ and $\Pr[Y = y_j] > 0$.

We say that *(ρ_{1i}, ρ_{2i})-privacy holds* if the above upward and downward privacy breaches are eliminated.

C. Data utility

We consider two types of utility in this paper. The first is the standard utility metric used in evaluating perturbation algorithms in data mining, which we call *aggregate utility*. In many aggregate data mining applications, the distribution of SA-values, instead of the exact SA-value in a record, is the research target.

Given a perturbed table T^*, let $O = (o_1,\ldots,o_m)$ be the vector of the observed frequencies of y_i in T^*, and let $F = (f_1,\ldots,f_m)$ be the vector of the frequencies of x_i in T. $\mathrm{E}(O) = P \times F$. We can estimate F by $F' = P^{-1} \times O$, if P is invertible. F' is an unbiased estimate of F in the sense that $\mathrm{E}(F') = P^{-1} \times \mathrm{E}(O) = F$. To compute F', we can use *inverse* [4] or *Iterative Bayesian* [1] *reconstruction*.

Definition 2 (Aggregate utility): Let $F = (f_1,\ldots,f_m)$ be the actual SA-value frequencies from an original table T and let $F' = (f_1',\ldots,f_m')$ be the frequencies estimated from the perturbed table T^*, discussed above. The *aggregate utility* of F' with respect to F is defined as $1 - 1/m \times \sum_{i=1..m} |f_i - f_i'|$.

We propose a new utility metric, called *record utility*, which measures the expected percentage of records in T whose SA-values are unchanged in T^*. This utility is useful when the truthfulness of data *at the record level* is important. For example, records may be published for human reading, where a published value differing from the original value is considered an error.

Definition 3 (Record utility): Let $F = (f_1,\ldots,f_m)$ be the actual SA-value frequencies from an original table T. The *record utility* is defined as $\sum_i f_i \times p_{ii}$, $\forall i=1,\ldots,m$, where p_{ii} is the retention probability of an SA-value, located along the main diagonal in P.

Record utility is a better utility metric for privacy preserving data publishing because (a) it is *not instance specific*, so it is useful for measuring the quality of M, not just a specific instance T^*, (b) it measures *utility for an ad-*

hoc task, not just an aggregate data mining task and (c) we expect it to have a *positive impact on aggregate utility*, since retaining more SA-values helps reconstruct the distribution on SA. We will evaluate this impact experimentally in Section VI. We can now formally define our problem.

Definition 4 (Optimal Fine-Grain Perturbation Problem): Given the microdata T and privacy parameters (ρ_{1i}, ρ_{2i}) for all x_i in SA, $0 < \rho_{1i} < \rho_{2i} < 1$, i=1,…,m, we want to find an *optimal* fine-grain perturbation matrix P such that $\forall$ i=1,…,m, (i) (ρ_{1i}, ρ_{2i})-privacy holds on any T^* generated by P and (ii) record utility is maximized under (i).

IV. Optimal Fine-Grain Perturbation

Fig. 3 shows the algorithm for finding the optimal fine-grain perturbation operator, called Fine-grain. Given the microdata T and privacy parameters (ρ_{1i}, ρ_{2i}) for each value x_i in SA, i = 1,…,m, Fine-grain computes the frequencies f_i of all SA-values x_i in Step 1. In Step 2, the γ_i values are computed. These values represent how sensitive the SA-values are; a high γ_i means x_i is not very sensitive. In Step 3, a linear program is solved, which determines the optimal probabilities p_i for x_i. In Step 4, a fine-grain perturbation operator P is constructed. By rewriting the privacy constraint in Fig. 3 as $(m - 1) \times p_i + \gamma_i \times p_j \le \gamma_i - 1$, we have a linear program with a global maximizer. There always exists one solution satisfying the privacy constraints (*i.e.*, $p_i = 0$, for i = 1,…,m). Since $0 < \rho_{1i} < \rho_{2i} < 1$, $1 < \gamma_i < \infty$, so $(1 - p_j)/m$ can never be zero, *i.e.*, P never has zero entries.

Fine-grain
Input: T, ρ_{1i} **and** ρ_{2i} **for all SA-values** x_i
Output: optimal fine-grain perturbation matrix
1. Determine SA frequencies f_i from T
2. $\gamma_i = (\rho_{2i} \times (1-\rho_{1i})) / (\rho_{1i} \times (1-\rho_{2i}))$, $\forall i=1,\ldots,m$
3. Solve the program for p_i:
   Objective function: max $\sum_i f_i \times (p_i + (1-p_i)/m)$
   Privacy constraints:

$$\frac{p_i + (1-p_i)/m}{(1-p_j)/m} \le \gamma_i, \forall i, \forall j = 1,\ldots,m, i \ne j$$

$$0 \le p_i \le 1, \forall i = 1,\ldots,m$$

4. Construct P following (1)
5. Return P

Figure 3. Optimal fine-grain perturbation algorithm

Uniform perturbation is a special case of a fine-grain perturbation where all probabilities p_i, i = 1,…,m, are equal, say to p. Therefore, the main diagonal entries are $p + q$ and all other entries are $q = (1 - p)/m$. Also, the privacy constraints in Fig. 3 simplify to $(p + (1 - p)/m) / q \le \gamma$, where $\gamma = \min \gamma_i$, which corresponds to the most restrictive constraint in Fig. 3. The objective function is maximized when $p = (\gamma - 1) / (m - 1 + \gamma)$, so, P becomes

$$P^{(2)} = \frac{1}{m - 1 + \gamma} \begin{bmatrix} \gamma & 1 & \cdots & 1 \\ 1 & \gamma & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & \gamma \end{bmatrix} \quad (2)$$

Example 2 (Fine-Grain vs. Uniform): Continuing from Example 1, in Step 2 of Fig. 3, γ_{SARS} = 1½, γ_{HIV} = 3, γ_{H1N1} = 9½, and γ_{cancer} = 18. Let $\gamma = \min(\gamma_{SARS}, \gamma_{HIV}, \gamma_{H1N1}, \gamma_{cancer})$ = 1½. Fig. 2 (a) gives the uniform perturbation operator with record utility = 1/3, and Fig. 2 (b) gives the optimal fine-grain perturbation operator returned by Fine-grain with record utility = 0.4375. About a 24% increase in record utility is due to the increased retention probability on the main diagonal. This is possible because Fine-grain perturbs less of the less-sensitive SA-values. □

We now prove that Fine-grain's privacy constraints in Fig. 3 guarantee (ρ_{1i}, ρ_{2i})-privacy as stated in Definition 1. In [3], a γ-amplification condition is proposed to guarantee (ρ_1, ρ_2)-privacy. The idea is to bound the ratio $\Pr[x_k \to y] / \Pr[x_i \to y]$ by some γ value derived from the privacy parameters ρ_1 and ρ_2. Intuitively, this says that if the probability of any SA-value being perturbed to the same value y differs by a factor not more than γ, (ρ_1, ρ_2)-privacy will hold. We now extend this approach to (ρ_{1i}, ρ_{2i})-privacy. The extension is not trivial; however, the details are omitted due to space constraints.

Definition 5 (γ_j-amplification condition): For $y_j \in$ SA, let γ_j be a (finite) real greater than 1. We say that a perturbation operator P is *at most* γ_j*-amplifying* if

$$\frac{p_{jj}}{p_{ji}} \le \gamma_j, \forall i = 1,\ldots,m, i \ne j \quad (3)$$

Theorem 1: Let P be a fine-grain operator. Let $0 < \rho_{1i} < \rho_{2i} < 1$ be the privacy parameters for x_i, i = 1,…,m. Assume that P is at most γ_j-amplifying for j = 1,…,m. Revealing "$Y = y_j$" causes neither upward (ρ_{1i}, ρ_{2i})-privacy breaches nor downward (ρ_{2i}, ρ_{1i})-privacy breaches if the following condition is satisfied:

$$\frac{\rho_{2i}}{\rho_{1i}} \times \frac{1 - \rho_{1i}}{1 - \rho_{2i}} \ge \gamma_i \quad (4)$$

The proof is similar to Statement 1 in [3] (details omitted due to space constraints). With Theorem 1, the most relaxed γ_i-amplification condition for (ρ_{1i}, ρ_{2i})-privacy is derived by choosing the maximum γ_i satisfying (4), *i.e.*,

$$\gamma_i = \frac{\rho_{2i}}{\rho_{1i}} \times \frac{1 - \rho_{1i}}{1 - \rho_{2i}} \quad (5)$$

So far, we have ignored the QI attributes in the definition of (ρ_{1i}, ρ_{2i})-privacy. One question is whether background knowledge on QI will affect (ρ_{1i}, ρ_{2i})-privacy. Consider (only) two SA values, "lung cancer" and "breast cancer" and suppose the adversary has background knowledge that a male is very unlikely to have "breast cancer". Upon seeing a record in T^* with *Sex* = "M" (a QI attribute), the adversary can immediately tell the original SA value for this record is "lung cancer". However, this inference is *not* due to data

publication; the adversary already knows that any record with *Sex* = "M" has SA = "lung cancer" *before* seeing the perturbed SA value. Thus the impact of QI is on prior probability, not on posterior probability.

In general, we can show that background knowledge on QI does not affect our approach as long as the perturbation operator P is independent of QI, which is true in our case. To see this, we can model any background knowledge on QI by $Z = z$ for an instance z of some random variable Z that depends on the QI value in a record. In Definition 1, the prior and posterior probabilities are now modified to $\Pr[X = x_i \mid Z = z]$ and $\Pr[X = x_i \mid Y = y_j, Z = z]$, respectively. Therefore, all perturbation probabilities $\Pr[x_i \rightarrow y_j]$ remain the same, so the amplification condition remains unaffected and Theorem 1 still applies. Depending on the background knowledge Z = z (such as *Sex* = "M" vs. *Sex* = "F"), each x_i now may have several (ρ_{1i}, ρ_{2i}) parameters. In this case, we will use the minimum γ_i value for x_i over the Z = z related to x_i (computed by Equation (5)).

V. Experimental Evaluation

In this section, we evaluate the effectiveness of our new strategy for improving utility, namely, *Fine-grain Perturbation*. To our knowledge, [12] is the only work on data publishing that addresses corruption attacks. Their least distorted case is when there is no sampling (*i.e.,* $k = 1$) and no generalization on QI. In this case, their perturbation is exactly the same as *Uniform Perturbation*. We compare the following algorithms: **Uniform** – uses Equation (2) and **Fine-grain** – uses Fig. 3.

We set ρ_{1i} to the frequency f_i, $i = 1,\ldots,m$, and we use a *tolerance parameter* $\theta > 1$ to set ρ_{2i} for the posterior probability. If $\rho_{1i} \geq 1/\theta$, we set γ_i to a large value so that there is no privacy concern on x_i. If $\rho_{1i} < 1/\theta$, we set $\rho_{2i} = \theta \times \rho_{1i}$ and compute γ_i using (5). We collect both record utility and aggregate utility (using iterative Bayesian reconstruction 0 over 10 instances T*). Our algorithms are written in C++ and we used Soplex (http://zibopt.zib.de) to solve our linear program. We ran all experiments on a Pentium IV 3.0 GHz PC with 2.0GB of RAM.

A. Dataset descriptions

CADR Our first experiment uses the real-life Canadian Adverse Drug Reaction (CADR) database (http://www.hc-sc.gc.ca/dhp-mps/medeff/databasdon/extract_extrait-eng.php). We join the Reactions and Reports tables from this database to obtain a table J with QI attributes *report date*, *gender*, *age*, *weight*, *height*, and SA *drug reaction*. Each record in J has a unique patient *Report_id*. SA consists of the $m = 40$ most frequent drug reactions. All records for other drug reactions are discarded. In the remaining records, if a patient has more than one record, we keep the record having the most frequent drug reaction. The dataset size is 95946.

Zipf We also experimented with a synthetic dataset that has frequencies f_i following the *Zipfian distribution*. For the i^{th} most frequent SA-value x_i, $i = 1,\ldots,m$, the frequency f_i is given by

$$f(i, \delta, m) = \frac{1/i^{\delta}}{\sum_{j=1..m} 1/j^{\delta}}$$

where δ determines the decreasing rate of f(i, δ, m). As in the classic version of Zipf law, we use $\delta = 1.0$. This distribution is more skewed than that of the CADR dataset. We set m = 40 and records to 95946 to match the CADR dataset settings. The frequency distributions for both datasets are given in Fig. 4.

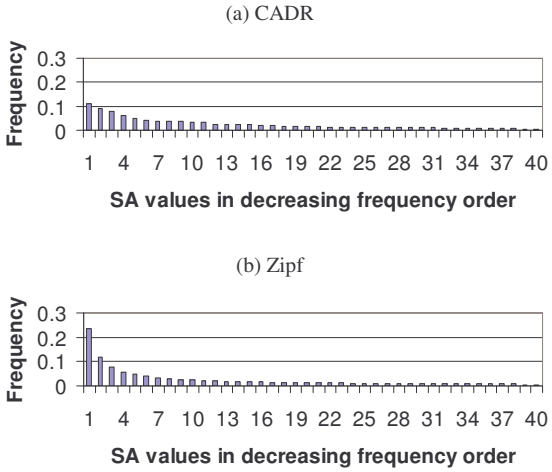

Figure 4. Dataset frequency distributions.

B. Uniform vs. fine-grain perturbation on CADR

Our goal is to evaluate whether fine-grain privacy leads to better retention of data. In this section we consider the real life CADR dataset. The results comparing uniform and fine-grain perturbation are shown in Fig. 5 (a), with record utility on the left and aggregate utility on the right.

Fine-grain perturbation has better record utility than uniform perturbation, with an improvement of about 10% to 20%. However, fine-grain perturbation has almost the same aggregate utility as uniform perturbation. This is because aggregate utility is more robust to data distortion at the record level. As expected, both utilities improve as the privacy tolerance θ increases.

The lower record utility indicates that record utility is more difficult to retain than aggregate utility. This is expected because only unchanged SA-values contribute to record utility. Another interesting point is that, although our algorithms optimize record utility, a better record utility translates into a better aggregate utility.

C. Uinform vs. fine-grain perturbation on Zipf

Again, our goal is to evaluate whether fine-grain privacy leads to better retention of data, but in this section we consider the highly skewed Zipf dataset. The results are shown in Fig. 5 (b). Note that we use larger θ values than in the previous experiment. This is to accommodate privacy

settings for extremely small frequencies (*i.e.*, ρ_1 values) in this skewed dataset. Fine-grain perturbation again has better record utility than uniform perturbation and both utilities improve as the privacy tolerance θ increases. However, we find an unexpected trend: although fine-grain perturbation has a lower aggregate utility than uniform perturbation.

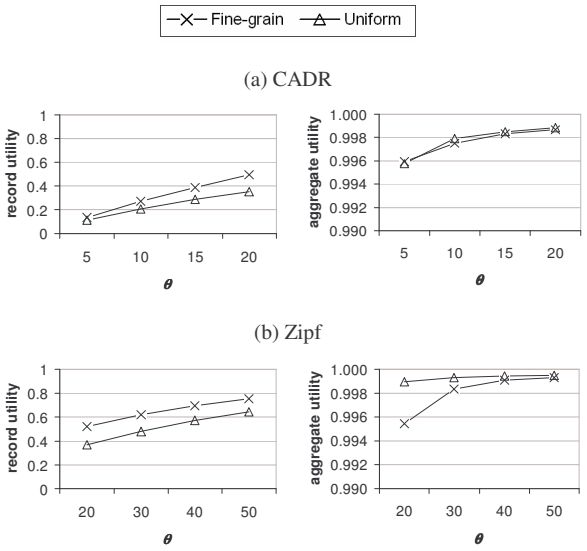

Figure 5. Utility vs. θ for fine-grain and uniform privacy.

On one hand, we are not surprised, since we optimized our algorithm for record utility, not aggregate utility. A high record utility is a sufficient condition for a high aggregate utility, but it is not a necessary condition. For instance, consider an operator which transforms the original dataset T by randomly swapping the SA-values of pairs of records so that every record in T^* ends up with another record's SA-value. Since this operator did not alter the distribution of SA-values, aggregate utility is maximized, while record utility is very low (imagine a researcher trying to use such a scrambled publication for ad-hoc analysis).

On the other hand, it is interesting that a correlation between record and aggregate utility appears to hold for more balanced datasets like CADR, but not for more skewed datasets like Zipf. A key difference in skewed datasets is that low-frequency SA-values only occur in a small number of records, which makes reconstruction unstable. Moreover, our fine-grain algorithm is optimized for record utility, so it prefers the distortion of these low-frequency SA-values (small number of records overall) when it is globally optimal to retain more of the high-frequency SA-values (large number of records overall). Since data mining tasks generally use aggregate information shared by a large number of records, we do not see this as a negative result; rather, it highlights an interesting research problem outside the scope of this privacy preserving data publishing paper: how can we optimize utility for aggregate data mining tasks?

VI. CONCLUSIONS

The conventional anonymity-group approach for privacy preserving data publishing is susceptible to corruption attacks. We presented a novel anti-corruption solution. Fundamentally differing from the anonymity-group approach, we adapted perturbation, previously used for privacy preserving data mining, as the main technique for data publishing. To minimize information loss, we proposed a new perturbation operator called fine-grain perturbation. Our studies showed that while our new approach retains better utility for ad-hoc tasks, more work is needed in optimizing utility for aggregate data mining tasks, especially for highly skewed datasets. In the future, we plan to extend our approach to preserve privacy of individuals in other types of data, such as graphs, spatial data, and query logs.

ACKNOWLEDGMENT

We would like to thank Dr. Tamon Stephen from the Math Department at Simon Fraser University for his advice on mathematical programming software and our reviewers for their feedback. This research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Post Graduate Scholarship and Discovery Grant.

REFERENCES

[1] R. Agrawal, R. Srikant, and D. Thomas, "Privacy Preserving OLAP," Proc. SIGMOD 2005.

[2] W. Du and Z. Zhan, "Using Randomized Response Techniques for Privacy Preserving Data Mining," Proc. KDD 2003.

[3] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting Privacy Breaches in Privacy Preserving Data Mining," Proc. PODS 2003.

[4] Z. Huang and W. Du. "OptRR: Optimizing Randomized Response Schemes for Privacy-Preserving Data Mining," Proc. ICDE 2008.

[5] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy Beyond k-Anonymity and l-diversity," Proc. ICDE 2007.

[6] T. Li and N. Li, "Modeling and Integrating Background Knowledge," Proc. ICDE 2009.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy Beyond k-anonymity," Proc. ICDE 2006.

[8] V. Rastogi, S. Hong, and D. Suciu, "The Boundary Between Privacy and Utility in Data Publishing," Proc. VLDB 2007.

[9] S. Rizvi and J. R. Haritsa, "Maintaining Data Privacy in Association Rule Mining," Proc. VLDB 2002.

[10] P. Samarati, "Protecting Respondents' Identities in Microdata Release," TKDE, vol. 13, no. 6, pp. 1010–1027, 2001.

[11] L. Sweeney, "Achieving k-anonymity Privacy Protection Using Generalization and Suppression," Int'l J. on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 571–588, 2002.

[12] Y. Tao, X. Xiao, J. Li, and D. Zhang, "On Anti-Corruption Privacy Preserving Publication," Proc. ICDE 2008.

[13] S. L. Warner, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. The A. Stat. Assoc., vol. 60, no. 309, pp. 63–69, 1965.

[14] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," Proc. VLDB 2006.