9-2011

Privacy beyond single sensitive attribute

Yuan FANG
Singapore Management University, yfang@smu.edu.sg

Mafruz Zaman ASHRAFI

See Kiong NG

DOI: https://doi.org/10.1007/978-3-642-23088-2_13

Privacy beyond Single Sensitive Attribute

Yuan Fang[1,2], Mafruz Zaman Ashrafi[2], and See Kiong Ng[2,3]

[1] University of Illinois at Urbana-Champaign, United States
[2] Institute for Infocomm Research, Singapore
[3] Singapore University of Technology and Design, Singapore
fang2@illinois.edu, mafruz@gmail.com, skng@i2r.a-star.edu.sg

Abstract. Publishing individual specific microdata has serious privacy implications. The k-anonymity model has been proposed to prevent identity disclosure from microdata, and the work on ℓ-diversity and t-closeness attempt to address attribute disclosure. However, most current work only deal with publishing microdata with a single sensitive attribute (SA), whereas real life scenarios often involve microdata with *multiple* SAs that may be *multi-valued*. This paper explores the issue of attribute disclosure in such scenarios. We propose a method called CODIP (Complete Disjoint Projections) that outlines a general solution to deal with the shortcomings in a naïve approach. We also introduce two measures, Association Loss Ratio and Information Exposure Ratio, to quantify data quality and privacy, respectively. We further propose a heuristic CODIP* for CODIP, which obtains a good trade-off in data quality and privacy. Finally, initial experiments show that CODIP* is practically useful on varying numbers of SAs.

1 Introduction

Individual specific microdata is essential for advancing empirical research, yet publishing such data can pose serious risks to individual privacy. To minimize the privacy risks, prior methods in k-anonymity [17,16] and its variants [21,6,11], ℓ-diversity [18,13,14,23] and t-closeness [9,10] emphasized on reducing *identity disclosure* and *attribute disclosure* [7]. While these efforts help protect individual privacy to a certain degree, attribute disclosure can still occur if the microdata consist of multiple sensitive attributes (SA). We highlight two shortcomings of the prior methods leading to attribute disclosure in the presence of multiple SAs.

First, prior privacy protection methods is often insufficient when there are multiple SAs, as it is **difficult to ensure good diversity or strong closeness for every SA**.

Example 1. Consider the raw microdata in Table 1(a). Suppose race and sex are quasi-identifiers (QID) [17] and the rest are SAs. We consider all possible 2-anonymized tables as shown in Table 1(b), (c) and (d). If we only want to publish a single SA, say diagnosis, then we can publish either Table 1(c) or (d), as either table each has two distinct diagnoses in every equivalence class (which is a set of tuples that have identical values in QIDs [9]), E_1 and E_2. In addition, both tables satisfy 0.25-closeness [9]. However, if we want to publish all three SAs, each of the 2-anonymized tables has only one distinct value for one of the SAs in some equivalence class (italicized in Table 1(b), (c) and (d)). Consequently, each table only achieves 0.5-closeness for that attribute. □

Table 1. Raw (a) and 2-anonymized tables (b)-(d). E_i are the equivalence classes after 2-anonymization. Abbreviations used: HT (hypertension), DB (diabetes), AS (asthma).

(a) Raw microdata

	race	sex	diagnosis	family_history	job
t_1	white	f	HT	HT	teacher
t_2	white	m	HT	DB	lawyer
t_3	white	m	DB	HT	farmer
t_4	black	m	AS	AS	teacher

(b) Anonymized. $E_1 = \{t_1, t_2\}$, $E_2 = \{t_3, t_4\}$

race	sex	diagnosis	family_history	job
white	*	*HT*	HT	teacher
white	*	*HT*	DB	lawyer
*	m	DB	HT	farmer
*	m	AS	AS	teacher

(c) Anonymized. $E_1 = \{t_1, t_3\}$, $E_2 = \{t_2, t_4\}$

race	sex	diagnosis	family_history	job
white	*	HT	*HT*	teacher
white	*	DB	*HT*	farmer
*	m	HT	DB	lawyer
*	m	AS	AS	teacher

(d) Anonymized. $E_1 = \{t_1, t_4\}$, $E_2 = \{t_2, t_3\}$

race	sex	diagnosis	family_history	job
*	*	HT	HT	*teacher*
*	*	AS	AS	*teacher*
white	m	HT	DB	lawyer
white	m	DB	HT	farmer

Second, when there are multiple SAs, a new type of attack named **background-join attack** emerges. In this new attack, we assume the adversary has some external background knowledge about some individual in the table. By joining the background knowledge and the table, s/he can deduce sensitive information.

Example 2. Suppose Table 1(b) is published. Eve links Bob to equivalence class E_2 based on his QIDs. In E_2 each SA takes two distinct values. Thus, if Eve only focuses on the SA of her interest, say diagnosis, she cannot infer whether Bob has asthma with a probability more than 0.5. However, if Eve has background knowledge that Bob is a teacher, she can deduce that Bob has asthma based on the natural join of "teacher" and the last row of the table. □

Beyond the toy example in Table 1, in real life, microdata that involve multiple SAs are also common. For instance, the dataset "Income Census (KDD)" [1] is extracted from population surveys, which involves many SAs, such as employment_status and wage_per_hour. Publishing such microdata enables useful data mining applications such as classification and association study among different SAs. However, as we have presented, prior methods have two shortcomings in dealing with multiple SAs.

Additionally, an SA can also be *multi-valued* as opposed to *mono-valued*. Given a set of values S, a mono-valued attribute can take only a value v such that $v \in S$.

However, a multi-valued attribute can take any set of values S' such that $S' \subseteq S$. Each value in S is atomic, i.e., there is no nested multi-values within a value. For instance, diagnosis is multi-valued and can take a set of values, say $\{$DB, HT, AS$\}$. In the dataset "Income Census (KDD)" [1], the attribute household_status can be regarded as multi-valued (see Sect. 7). In relational databases, multi-valued attributes are also common, although they are normalized and stored in a separate table. Since normalization is a *lossless* process, normalized tables are thus no different from the original table with multi-valued attributes from an adversary's perspective.

In this paper, we explore privacy methods for publishing microdata with multiple SAs, some of which may be multi-valued. In summary, this paper makes the following contributions:

1. We identified two drawbacks of prior methods on microdata with multiple SAs;
2. We derived a general framework CODIP to address these drawbacks, which can also be applied on multi-valued SAs;

3. We introduced two new measures *Association Loss Ratio* and *Information Exposure Ratio* that quantify data quality and privacy in the new scenario;
4. We proposed a heuristic CODIP* for CODIP, which obtains a good trade-off in data quality and privacy.

2 Related Work

Microdata are usually modelled as a table, where each row corresponds to a tuple for an individual, and each column corresponds to an attribute. It is often assumed that each tuple maps to one individual and no two tuples correspond to the same individual [20].

To prevent identity disclosure, Sweeney proposed k-anonymity [17,16], which introduced the notion of *quasi-identifiers* (QIDs). The set of tuples that have identical values in QIDs are defined as an *equivalence class* [9]. The requirement is each equivalence class must contain at least k tuples. A few variants of k-anonymity also exist, e.g., Anatomy [21] which bucketizes sensitive values instead of QIDs, Micro-aggregation [6,15] and Slicing [11]. While k-anonymity can prevent identity disclosure, it does not prevent attribute disclosure.

Recent extensions of k-anonymity also address attribute disclosure. Their philosophy is to make SA values in each equivalence class more diverse. Ref. [18] proposed p-sensitivity, requiring an SA to take at least p distinct values in every equivalence class. Furthermore, [14] pointed out that the distinct values must be "well represented", and proposed ℓ-diversity based on information entropy and attribute value frequency. In a similar spirit as ℓ-diversity, (k, e)-anonymity [23] can be adopted on continuous values such that each equivalence class must contain sensitive values of a range at least e.

However, according to [9,10], ℓ-diversity is unnecessary and difficult to achieve in some cases, and is prone to skewness and similarity attacks. To address these limitations, Li *et al.* [9] proposed t-closeness. The model requires that the distribution D_j of the SA in each equivalence class E_j is close enough to the overall distribution D in the entire table. Specifically, a table satisfies t-closeness if $\forall_{E_j} : \mathrm{dist}(D_j, D) \leq t$, where $\mathrm{dist}(X, X')$ is the Earth Mover's Distance (EMD) between X and X'. A strong closeness (i.e., a small value of closeness) indicates that the distributions of the SA in each equivalence class are similar to the overall distribution in the entire table, therefore implying less risk for attribute disclosure. Li *et al.* also introduced (n, t)-closeness [10], an extension of the basic t-closeness, which allows more flexibility while retaining closeness.

The above works only deal with a single mono-valued SA. They cannot cope with multiple SAs, with the two drawbacks identified in Sect. 1. This paper extends existing models such as k-anonymity and t-closeness to the new scenario. Currently only a limited number of works deal with multiple mono-valued SAs [12,22,4]. However, they did not deal with the background-join attack (see Sect. 1), a major problem in the presence of multiple SAs– simply because all these works publish the SAs in one table, preserving associations among SAs. Hence an adversary can join the table with his/her background knowledge to reveal other SAs. In addition, they did not address the problem of multi-valued attributes, which we will explore in this paper.

Notation	Representation
$\mathcal{A} = \{A_1, \ldots, A_s\}$	the set of sensitive attributes (SA), $s = \lvert\mathcal{A}\rvert$
$\mathcal{Q} = \{Q_1, \ldots, Q_q\}$	the set of QIDs, $q = \lvert\mathcal{Q}\rvert$
$\mathcal{E} = \{E_1, \ldots, E_c\}$	the set of equivalence classes, $c = \lvert\mathcal{E}\rvert$
D_i	the distribution of A_i in the entire table
D_{ij}	the distribution of A_i in E_j
$\mathrm{dist}(X, X')$	EMD between distributions X and X'

Fig. 1. Notations

3 CODIP: A General Solution

In this section, we first introduce a naïve t-closeness approach, which is a straightforward adaptation of t-closeness in the presence of multiple mono- or multi-valued SAs. Next, we identify the shortcomings in the naïve approach, and propose a general solution CODIP to tackle the shortcomings. Note that although our discussion is based on t-closeness, our approaches also apply to other privacy models such as ℓ-diversity and (n, t)-closeness in a similar fashion. For ease of discussion, we present a list of notations in Fig. 1.

3.1 Naïve t-Closeness Approach

Multiple mono-valued SAs. First consider only multiple mono-valued SAs. Given a k-anonymized table T, suppose all SAs $A_1, \ldots, A_s$ are mono-valued. If two SAs have strong dependency, their joint distribution would be similar to that of a single SA. In this case, we can simply consider the closeness of their distributions individually. If the SAs have weak dependency, their joint values will be very diverse, especially when the number of such SAs are large (the curse of dimensionality). In this case, it is meaningless to require an equivalence class to be "well represented" in terms of the joint values of the SAs.

As such, we define t-closeness of T based on individual SAs instead of their joint distributions. Essentially, in the naïve approach, in order for T to satisfy t-closeness, every SA must satisfy t-closeness for a given k-anonymization.

Definition 1. *A k-anonymized table T, whose SAs are all mono-valued, is said to satisfy* **t-*closeness*** *iff* $\forall_{A_i \in \mathcal{A}} \forall_{E_j \in \mathcal{E}} : \mathrm{dist}(D_{ij}, D_i) \leq t$. □

Multi-valued SAs. Given a raw table with n tuples $t_1, \ldots, t_n$, suppose there is a multi-valued SA B. B can take a subset of values in S, i.e., $\forall_{t_u} : t_u.B \subseteq S$, where $S = \{v_1, \ldots, v_m\}$. Without loss of generality, we assume the values in S are categorical, since continuous values can be discretized. It is easy to transform B into multiple mono-valued attributes.

Definition 2. *$\forall v_i \in S$, define a bit vector $(b_{i1}, b_{i2}, \ldots, b_{in})$, where each $b_{iu} = 1$ if $v_i \in t_u.B$, and $b_{iu} = 0$ otherwise. Attribute B is replaced with m mono-valued attributes $B_1, \ldots, B_m$, such that $\forall_{t_u, B_i} : t_u.B_i = b_{iu}$. We term this process* ***bitmap transformation****, and each B_i the* ***derived attribute*** *of B.* □

Informally, B is transformed into an $m \times n$ bitmap. Note that bitmap transformation is *lossless* and thus does not compromise data quality. In addition, each derived attribute is mono-valued. This allow us to adapt the naïve t-closeness approach on a bitmap transformed table as we have just discussed. We then treat the derived attributes no different from the original mono-valued attributes.

Shortcomings. The naïve approach is a direct adaptation of t-closeness, which suffers the two shortcomings in Sect. 1. We claim that the two shortcomings generally become more severe when there are more SAs.

In the context of t-closeness, the first shortcoming is that we generally have weaker closeness (i.e., a larger value of closeness) when there are more SAs, owing to the *effect of diminishing closeness*. This effect is formalized in Theorem. 1. Its proof is omitted due to space constraint.

Theorem 1. *Given a bitmap transformed table T, let T' be the projection of T on $\mathcal{A}' \cup \mathcal{Q}$, where $\mathcal{A}' \subseteq \mathcal{A}$. Let t_{best} and t'_{best} denote the best closeness that at least one k-anonymized T and T' can satisfy, respectively. The **effect of diminishing closeness** states that $t_{best} \geq t'_{best}$.*

The second shortcoming is that the threat of background-join attacks (abbreviated as "*join-threat*" hereafter) becomes greater as the number of SAs increases. When there are more SAs, an adversary can deduce new information on more SAs in a background-join attack, which increases the join-threat.

Since we are enforcing t-closeness (or other models) on each SA, the threat of traditional background attack on an individual SA is similar to that in the scenario with a single SA. We do not discuss this kind of attack as it has been addressed in previous works involving a single SA. Instead, we focus on the new threat that arises due to the existence of multiple SAs, the so-called "join-threat".

3.2 CODIP: Overcoming the Shortcomings

If we can reduce the number of SAs in a published table, we can alleviate the effect of diminishing closeness and the join-threat. Based on this, we propose a general solution called *Complete Disjoint Projections* or CODIP. In essence, CODIP projects the raw table on subsets of the SAs, and publishes the *projected tables* instead. Each projected table has a smaller number of SAs than the raw table has. Additionally, all of the SAs must be in exactly one of the projection. Formally, we call how CODIP projects the raw table a *projection plan*, or simply a *plan*.

Definition 3. *A **projection plan** projects a bitmap transformed table on its subsets of attributes $\mathcal{A}_1 \cup \mathcal{Q}$, ..., $\mathcal{A}_r \cup \mathcal{Q}$, such that (i) $\cup_{u=1}^{r} \mathcal{A}_u = \mathcal{A}$; (ii) $\forall_u : \mathcal{A}_u \neq \emptyset$; (iii) $\forall_{uw, u \neq w} : \mathcal{A}_u \cap \mathcal{A}_w = \emptyset$. We denote this plan $\Phi(\mathcal{A}_1, \ldots, \mathcal{A}_r)$. The projections are called the **projected tables** of the plan. A plan satisfies t-closeness iff every projected table satisfies t-closeness as in Def. 1.* □

To put in words, a projection plan isolates *disjoint* subsets of SAs in separate tables. Each projected table is then subjected to various anonymity algorithms, and the order of the tuples in each table is randomized. In this paper, we apply k-anonymity [17]

and t-closeness [9]; however, we stress that CODIP is a flexible framework that may adopt any previous privacy models on each projected table. The philosophy is to devise a good projection plan (see Sect. 3.3) so that any algorithm intended for a single SA (e.g., [21,6,13,10]) would also work well on each projected table without suffering significantly from diminishing closeness and the join-threat, while at the same time preserving most of the utility. The SAs of each individual within any projected table is thus protected by such previous algorithms, which offers certain level of protection even in the worst case. Linking SA values across tables is also limited, as will be shown in Sect. 6.1. We will further discuss possible attacks in Sect. 6, which would not succeed on CODIP.

Clearly, the naïve t-closeness approach is a special plan (i.e., $\phi_1 = \Phi(\mathcal{A})$). On the other hand, $\phi_2 = \Phi(\{A_1\}, \ldots, \{A_s\})$ is also a special plan that publishes each SA in a separate table. In this plan, the two shortcomings are completely eliminated, since each table only contains a single SA. Note that all plans except ϕ_1 suffer some information loss. In particular, some of the associations among SA values are lost, as the correspondence of tuples from different projected tables is disturbed. Such loss of associations mitigates the shortcomings of the naïve approach at the cost of data quality. Consider ϕ_2, in which the shortcomings of ϕ_1 are completely eliminated. However, as each projected table only contains a single SA, all associations between any two SAs are lost, making data much less useful. The goal is to overcome the shortcomings as well as to minimize association loss, as we shall discuss next.

3.3 Choosing Better Plans

An optimal plan minimizes the effect of diminishing closeness, association loss, and join-threat. We have made two observations towards such an optimal plan.

Observation 1. *If two SAs have strong dependency, a background-join attack on them reveals less new information beyond the adversary's background knowledge on one of the attribute. In addition, one of them can be closely represented by the other, effectively resulting in fewer than two (independent) attributes. Thus the effect of diminishing closeness on them is less pronounced.* □

Observation 2. *If two SAs are independent or with weak dependency, their joint distribution is insignificant, as no strong associations can be inferred from it. Thus association loss is small if their joint distribution is lost.* □

We use an example to illustrate the intuitions of the two observations.

Example 3. (**Observation 1**) Suppose diagnosis and family_history are two SAs with strong dependency. If they are published in the same table, Eve (who knows Bob has hypertension), learns that Bob has a family history of hypertension by a background-join attack. However, this privacy breach is less serious, as it is quite expected given that Bob has hypertension. Moreover, since the distributions of both attributes would be similar due to their dependency, the effect of diminishing closeness is less pronounced. (**Observation 2**) Suppose job and alcohol are two independent SAs. Their joint distribution appears random– people have different drinking habits regardless of their jobs. The associations between the two provide little information beyond random guessing. Thus, we can afford to lose such associations by publishing them in different tables. □

Algorithm `CODIP` (T, k, t, α, β)
Input: T, a raw table containing microdata.
k, anonymity requirement.
t, closeness requirement.
α, threshold on association loss.
β, threshold on join-threat.
Output: ϕ, a projection plan.
P, the set of projected tables for ϕ.
1) Apply bitmap transformation on T;
2) Partition $\mathcal{A}$ into r disjoint subsets $\mathcal{A}_1, \ldots, \mathcal{A}_r$;
3) $\phi \leftarrow \Phi(\mathcal{A}_1, \ldots, \mathcal{A}_r)$;
4) $P \leftarrow$ `CheckPlan`();
5) **if** $P =$ *null* **then**
6) **return** *failure*;
else
7) **return** (ϕ, P);

Subroutine `CheckPlan` ()
Input: all variables accessible in `CODIP`.
Output: the set of projected tables for ϕ.
8) **if** association loss in $\phi > \alpha$ **then return** *null*;
9) **if** join-threat in $\phi > \beta$ **then return** *null*;
10) **for** $i \leftarrow 1$ **to** r **do**
11) $T_i \leftarrow$ projection of T on $\mathcal{A}_i \cup \mathcal{Q}$;
12) **if** no k-anonymized T_i satisfies t-closeness **then**
13) **return** *null*;
else
14) $T_i \leftarrow$ a k-anonymized T_i satisfying t-closeness;
endfor
15) **return** $\{T_1, T_2, \ldots, T_r\}$;

Fig. 2. General framework for CODIP

To leverage the two observations, we propose a general framework for CODIP as shown in Fig. 2– a high level abstraction assuming an ideal partitioning of SAs (a concrete algorithm is proposed in Sect. 5). It requires the following user inputs: (i) T, the raw microdata table to be published; (ii) k, the anonymity requirement; (iii) t, the closeness requirement; (iv) α, the association loss threshold; (v) β, the join-threat threshold.

For inputs (iv) and (v), we delay the discussion of measuring association loss and join-threat to Sect. 4. For now, assume that they can be quantified. Also, assume that users can specify appropriate values for α and β, following the discussion on their relationships in the experiments (Sect. 7.1), although a more extensive study on this issue is beyond the scope of this paper.

The key operation lies in Step 2, which partitions $\mathcal{A}$ into disjoint subsets. Ideally, the partitioning should be consistent with Observation 1 and 2. In reality it only needs to be consistent to such a degree that a "sufficiently good" plan is obtained, which satisfies user specified thresholds t, α and β. Step 4 invokes the subroutine `CheckPlan`(), which examines if the plan satisfies the thresholds. If so, it returns a set of k-anonymized projected tables; otherwise, it fails. Note that users can optionally impose a quality threshold on QIDs in k-anonymization (Step 12 and 14), e.g., discernibility metric [2].

In this general framework, we do not enforce any specific algorithm to achieve a "sufficiently good" partitioning. A brute force method that enumerates all possible ways of partitioning and then selects one is infeasible since the number of possible ways to partition a set is intractable. We will propose an efficient heuristic CODIP* in Sect. 5 without requiring the costly enumeration.

4 Evaluating Projection Plans

In addition to k-anonymity that measures anonymity and t-closeness that measures closeness, we propose two more measures on a projection plan for CODIP: (1) *Association Loss Ratio* (Γ_α), the degree of association loss due to the lost joint distributions of the SAs; and (2) *Information Exposure Ratio* (Γ_β), the level of join-threat due to background-join attacks. The measures are based on *mutual information* (MI) [5], which can quantify nonlinear dependency between attributes, as opposed to correlation which only measures linear relationships. It means MI can detect dependency caused

by not only positive or negative correlations, but also "mixed" correlations. Hence, it is well-suited for formally capturing the notion of dependency in Observations 1 and 2.

4.1 Association Loss Ratio

We propose a measure to quantify association when SAs are projected onto different tables. Given a bitmap transformed table, for a pair of SAs A_i and A_j, their MI is $I(A_i, A_j) = \sum_{v \in A_i, v' \in A_j} p(v, v') \log\left(\frac{p(v,v')}{p(v)p(v')}\right)$ [5], where $p(x)$ is the pmf of attribute X, and $p(x, y)$ is the joint pmf of X and Y.[1] MI quantifies how much information two attributes share, which also implies the degree of independence between them. In particular $I(A_i, A_j) = 0$ if A_i and A_j are independent. We can use it to quantify how significant the association between the values of A_i and A_j are (Observation 2). Lower MI suggests a higher degree of independence, and thus the association between their values is less significant. This further implies that association loss is smaller if the joint distribution of the two attributes becomes unknown.

By computing the fraction of MI of all pairwise SAs whose joint distributions are unknown, we obtain a [0,1]-normalized measure of association loss– Association Loss Ratio (Γ_α). Given a projection plan $\phi = \Phi(\mathcal{A}_1, \ldots, \mathcal{A}_r)$ such that $\cup_{u=1}^{r} \mathcal{A}_u = \mathcal{A} = \{A_1, \ldots, A_s\}$, the sum of all pairwise MI is $I_\Sigma(\phi) = \frac{1}{2}\sum_{i,j \neq i} I(A_i, A_j)$, and the sum of unknown pairwise MI is $I_\alpha(\phi) = \frac{1}{2}\sum_{i,j \neq i} W_\alpha(A_i, A_j) I(A_i, A_j)$, where $W_\alpha(A_i, A_j)$ assigns a boolean weight— 1 if A_i, A_j are in different projected tables, 0 otherwise (i.e., $I(A_i, A_j)$ is summed in $I_\alpha(\phi)$ only if A_i, A_j are not projected onto the same table). Association Loss Ratio is then defined as a fraction in terms of $I_\Sigma(\phi)$ and $I_\alpha(\phi)$:

$$\Gamma_\alpha(\phi) = \begin{cases} I_\alpha(\phi)/I_\Sigma(\phi) & \text{if } I_\Sigma(\phi) \neq 0; \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Note the special cases that $\Gamma_\alpha(\Phi(\mathcal{A})) = 0$, and $\Gamma_\alpha(\Phi(\{A_1\}, \ldots, \{A_s\})) = 1$. In general, $\Gamma_\alpha(\phi)$ is smaller if the plan ϕ is generated in compliance with Observation 2.

4.2 Information Exposure Ratio

Next, we propose a measure of information exposure resulted from background-join attacks to indicate the level of join-threat. Clearly, when more *new* information is exposed, the threat level is higher. Thus, any background knowledge that is already known to the adversary must be excluded.

Consider any pair of SAs A_i and A_j. Suppose their joint distribution is known to an adversary, i.e., they are published in the same projected table. Assuming that the adversary identifies a tuple and has background knowledge in one of them (say A_i), s/he can then learn the value of the other SA (A_j). Potential new information of the other SA (A_j) could be exposed to the adversary. The other two cases are trivial: (i) if the adversary knows neither A_i nor A_j, no background-join attack can be launched on them; (ii) if the adversary knows both, no new information will be exposed.

The amount of information expressed by an attribute A_i can be represented by its information theoretic entropy [5], which is defined as $H(A_i) = -\sum_{v \in A_i} p(v) \log(p(v))$.

[1] We avoid the notations $p_X(x)$ and $p_{X,Y}(x, y)$ for convenience if no ambiguity arises.

The relationship of the entropies of A_i and A_j is illustrated by the Venn diagram in Fig. 3. For any pair of SAs, if the adversary deduces the value of A_i (or A_j) based on his or her background knowledge of A_j (or A_i), the amount of new information exposed is h_1 (or h_2). Therefore, total amount of new information that can be exposed from this pair is h_1+h_2. Since a larger $h_3 = I(A_i, A_j)$ results in a smaller h_1+h_2, less information can be exposed to an adversary when A_i and A_j have more dependency.

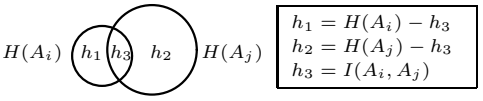

Fig. 3. Relationship of A_i and A_j's entropies

Also, some values in a SA could be non-sensitive depending on the user (e.g., nil value). Hence users should be allowed to define what constitute sensitive values in a SA. Let $\mathrm{Sens}(x)$ denotes the predicate that asserts x is a sensitive value. By default, $\mathrm{Sens}(x)$ is true for all values in a SA; however, users have the flexibility to customize it.

Subsequently we derive $E(A_i, A_j)$, the total amount of new sensitive information that is *exposable* in a pair of SAs A_i and A_j. Taking the sensitivity of values into account, it is computed by summing up exposable information for each joint value, weighted by the joint probability:

$$E(A_i, A_j) = \sum_{v \in A_i, v' \in A_j} p(v, v') \times \begin{cases} 0 & \neg\,\mathrm{Sens}(v) \wedge \neg\,\mathrm{Sens}(v'); \\ H(A_i) - I(A_i, A_j) & \mathrm{Sens}(v) \wedge \neg\,\mathrm{Sens}(v'); \\ H(A_j) - I(A_i, A_j) & \neg\,\mathrm{Sens}(v) \wedge \mathrm{Sens}(v'); \\ H(A_j) + H(A_j) - 2I(A_i, A_j) & \mathrm{Sens}(v) \wedge \mathrm{Sens}(v'). \end{cases}$$

Since a background-join attack is confined within the projected tables that contain the attributes on which the adversary has background knowledge, we compute the fraction of exposable information for each projected table. Given a projection plan $\phi = \Phi(\mathcal{A}_1, \ldots, \mathcal{A}_r)$ such that $\cup_{u=1}^{r} \mathcal{A}_u = \mathcal{A} = \{A_1, \ldots, A_s\}$, the sum of exposable information in all pairwise SAs (i.e., assuming there is only one projected table) is $E_\Sigma(\phi) = \frac{1}{2}\sum_{i,j \neq i} E(A_i, A_j)$, and the sum of *actual exposed* information in all pairwise SAs in the projected table on $\mathcal{A}_u \cup \mathcal{Q}$ can be computed as $E_\beta(\mathcal{A}_u) = \frac{1}{2}\sum_{i,j \neq i} W_\beta(A_i, A_j, \mathcal{A}_u) E(A_i, A_j)$, where $W_\beta(A_i, A_j, \mathcal{A}_u)$ assigns a boolean weight—1 if $A_i \in \mathcal{A}_u$ and $A_j \in \mathcal{A}_u$, and 0 otherwise (i.e., only actual exposed information in the projected table on $\mathcal{A}_u \cup \mathcal{Q}$ is summed). Information Exposure Ratio (Γ_β) is then defined as the sum of fractions in terms of $E_\beta(\mathcal{A}_u)$ and $E_\Sigma(\phi)$ for each projected table, normalized by the number of SAs in that table:

$$\Gamma_\beta(\phi) = \begin{cases} \sum_{u=1}^{r} \left(\frac{E_\beta(\mathcal{A}_u)}{E_\Sigma(\phi)} \cdot \frac{|\mathcal{A}_u|}{|\mathcal{A}|} \right) & \text{if } E_\Sigma(\phi) \neq 0; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Note the special cases that $\Gamma_\beta(\Phi(\{A_1\}, \ldots, \{A_s\})) = 0$, and $\Gamma_\beta(\Phi(\mathcal{A})) = 1$. In general, $\Gamma_\beta(\phi)$ is smaller if the plan ϕ is generated in compliance with Observation 1.

Algorithm CODIP* (T, k, t, α, β)
Input/Output: same as CODIP.
1) Apply bitmap transformation on T;
2) **for** $i = 1$ **to** s **do** $\mathcal{A}_i \leftarrow \{A_i\}$;
3) $\phi \leftarrow \Phi(\mathcal{A}_1, \ldots, \mathcal{A}_s)$;
4) $P \leftarrow$ CheckPlan*();
5) **if** $P = null$ **then return** *failure*;
6) **repeat**
7) $(\mathcal{A}_u, \mathcal{A}_w) \leftarrow \text{argmax}_{u,w:u\neq w} \text{AvgI}(\mathcal{A}_u \cup \mathcal{A}_w)$;
8) $\phi' \leftarrow \phi$; /* temp. placeholder */

(Continued)
9) $P' \leftarrow P$; /* temp. placeholder */
10) Remove $\mathcal{A}_u, \mathcal{A}_w$ from ϕ;
11) Add $\mathcal{A}_u \cup \mathcal{A}_w$ to ϕ;
12) $P \leftarrow$ CheckPlan*();
13) **until** $P = null$;
14) **if** $\Gamma_\alpha(\phi') \leq \alpha$ **then**
15) **return** (ϕ', P');
else
16) **return** *failure*;

Fig. 4. Outline of CODIP*

4.3 Evaluation of Plans

We use Association Loss Ratio and Information Exposure Ratio to evaluate the quality of a projection plan for CODIP. Based on their definitions, smaller ratios indicate a better plan. We propose to evaluate our plans against a baseline, the naïve t-closeness approach in Sect. 3.1, i.e., the plan $\Phi(\mathcal{A})$. By Theorem 1, $\Phi(\mathcal{A})$ has the weakest closeness among all plans. Furthermore, $\Gamma_\alpha(\Phi(\mathcal{A})) = 0$ and $\Gamma_\beta(\Phi(\mathcal{A})) = 1$. Given a plan ϕ, suppose $\Gamma_\alpha(\phi) = \alpha$, $\Gamma_\beta(\phi) = \beta$, ϕ satisfies t'-closeness and $\Phi(\mathcal{A})$ satisfies t-closeness. We say ϕ has a $(1-t'/t)\times 100\%$ improvement in closeness, $(1-\beta)\times 100\%$ reduced join-threat, while suffers $\alpha \times 100\%$ association loss, as compared to the naïve t-closeness approach.

5 CODIP*: A Heuristic for CODIP

In the CODIP framework proposed in Sect. 3.2, we have not described a suitable algorithm for generating good plans. A brute force approach to enumerate all possible plans is infeasible on high dimensional data. Thereby we propose a bottom-up greedy heuristic CODIP*, outlined in Fig. 4.

We start bottom-up from the initial plan $\phi = \Phi(\{A_1\}, \ldots, \{A_s\})$ (Steps 2–3). The basic idea is to ignore $\Gamma_\alpha(\phi)$ first, and merge the disjoint subsets of SAs in ϕ as much as possible. In this way, we attempt to reduce $\Gamma_\alpha(\phi)$ below its threshold while avoid exceeding closeness and $\Gamma_\beta(\phi)$ thresholds. The key operations lie in Steps 6–13, which correspond to the partitioning operation in CODIP (Step 2 in Fig. 2). We greedily pick two subsets of SAs $\mathcal{A}_u$ and $\mathcal{A}_w$ from the plan ϕ, such that the average pairwise MI in $\mathcal{A}_u \cup \mathcal{A}_w$ (AvgI in Step 7) is maximized. We then merge the two subsets $\mathcal{A}_u$ and $\mathcal{A}_w$ in ϕ (Steps 10 and 11). Based on Observations 1 and 2, this merging would greatly reduce $\Gamma_\alpha(\phi)$, and result in a small increase in closeness and $\Gamma_\beta(\phi)$ at least locally. The merging process is repeated until the plan ϕ exceeds the thresholds on closeness or $\Gamma_\beta(\phi)$ (Step 6-13). The subroutine CheckPlan*() checks if ϕ satisfies the thresholds on closeness and Γ_β. It is identical to CheckPlan() in CODIP, except that it does not check for Γ_α (i.e., eliminate Step 8 in Fig. 2), as it will be checked later. Subsequently, the plan before the last merger is returned if it satisfies the threshold on Γ_α (Step 14–16).

CODIP* is efficient by avoiding the combinatorial enumeration of attributes. For a dataset with s number of SAs, in the worst case, only $s-1$ mergers are necessary (i.e., the number of repetitions of Step 6–13 is bounded by $O(s)$).

6 Discussion of Possible Attacks on CODIP

6.1 Intersection Attack

Intersection attack occurs when multiple tables are intersected on common attributes [19], potentially re-establishing the links among sensitive values and QIDs across tables. [19] proposed the notion of (X, Y)-linkability– the extent of "linking" between X (QIDs) and Y (SAs). (X, Y)-linkability is satisfied if the confidence of inferring any value on Y from any value on X (can be joint value on X or Y) does not exceed a threshold $\epsilon \in (0, 1]$. We show that releasing multiple tables using CODIP introduces no more linking risk than releasing a single table using k-anonymity and distinct-ℓ-diversity, i.e., each equivalence class must contain at least ℓ distinct values, $\ell \geq 2$.

Theorem 2. *The tables released by CODIP (each table protected by k-anonymity and distinct-ℓ-diversity) satisfies (X, Y)-linkability with a threshold the same as the case of a single table released using k-anonymity and distinct-ℓ-diversity.*

Proof. As the subset of SAs (Y) in each projected table is disjoint, only QIDs (X) can be intersected. Consider a join of m tables by intersecting on some QIDs. By k-anonymity there are at least k tuples in each table with the same QIDs, producing a join with at least k^m tuples for any (joint) value on QIDs. Among the k^m or more joint tuples, we examine how many have the same value on some SAs. By ℓ-diversity there is at most $k - \ell + 1$ instances for any (joint) value on any SAs from one table. This follows that there are at most $k^{m-p}(k - \ell + 1)^p$ instances for any (joint) value on SAs from p tables ($1 \leq p \leq m$). Thus the confidence of infering SAs from QIDs is at most $\frac{k^{m-p}(k-\ell+1)^p}{k^m} = (\frac{k-\ell+1}{k})^p \leq \frac{k-\ell+1}{k}$. The upperbound is the threshold, which is independent of m. That means the same threshold is obtained when $m = 1$, i.e., a single table using k-anonymity and distinct-ℓ-diversity is released. □

Another type of intersection attack is targeted at incremental releases [3], where new tuples for the same schema are included and re-released with old tuples. Sensitive values can be intersected among old and new releases to derive hidden information. This type of attack is inapplicable to CODIP for two reasons: (i) in each projected table, the tuples all refer to the same set of individuals (i.e., no old and new tuples); (ii) given that there are no common SAs across tables, intersection on sensitive values is not possible.

6.2 Minimality Attack

Minimality attack [20], is possible if the adversary knows the privacy algorithm. The attack utilizes the concept of "minimality", as most privacy algorithms attempt to minimize information loss in order to preserve utility.

For CODIP, minimality attack is possible on two levels. *First*, minimality attack can target at each projected table, where k-anonymity is enforced. In this case, the m-confidentiality model [20] can be applied on each projected table to counter minimality attacks. *Second*, minimality attack can potentially target to restore the correspondence of tuples in different tables. Fortunately, CODIP is not vulnerable to this. While CODIP attempts to minimize the Information Loss Ratio Γ_α, its notion of minimization is relative to the MI of all pairwise attributes, and not to all possible correspondence of tuples

from different tables. Even if an adversary has obtained a correspondence of tuples with smallest possible Γ_α, this smallest Γ_α does not indicate a correct correspondence of tuples.

7 Experiments

We performed some initial experiments to study the trade-off between data quality and privacy. We choose the naïve t-closeness approach in Sect. 3.1 as our baseline. Note that the approaches in [12,22,4] publish all SAs in one table, thus they are vulnerable to background-join attacks in the exact same way as the baseline. Therefore it is fair to compare CODIP* with the baseline only, which suffers the same problem as these previous work. Moreover, to achieve k-anonymity, we adopted a full-domain generalization scheme as outlined in Incognito [8].

The "Census-Income (KDD)" training dataset [1] is used. We chose four QIDs– age, race, sex, citizenship, as well as SAs – seven categorical (worker_class, education, industry, employment_status, business_status, salary_class, occupation), four numeric discretized to $\{0, 1\}$ (wage_per_hour, dividend, capital_gain, capital_loss), and one multi-valued (household_status, giving four derived attributes married, 18^-, descendent, subfamily). There are effectively a total of 15 SAs. Additionally, tuples with missing or unknown values are discarded, giving a total of 98839 tuples that remain.

All algorithms were implemented in Java. The experiments were conducted on a 3.0GHz PC with 3GB memory.

7.1 Relationship of Γ_α and Γ_β

Intuitively, given a plan ϕ, a larger $\Gamma_\alpha(\phi)$ implies a smaller $\Gamma_\beta(\phi)$. This experiment studies the relationship between Association Loss Ratio and Information Exposure Ratio. Since the two ratios only depend on the way the raw table is projected, k-anonymity and t-closeness requirements does not affect them.

We run CODIP* with varying thresholds. Starting from $\beta = 1$, which is the threshold on $\Gamma_\beta(\phi)$, we gradually decrease it. For each β value, we record the smallest $\Gamma_\alpha(\phi)$ that has incurred. A plot of $\Gamma_\beta(\phi)$ against $\Gamma_\alpha(\phi)$ is presented in Fig. 5, where ϕ is the plan generated by CODIP* given a threshold β.

In Fig. 5, when no association loss incurs, i.e., $\Gamma_\alpha(\phi) = 0$, the join-threat is maximum at $\Gamma_\beta(\phi) = 1$. However, if we slightly relax $\Gamma_\alpha(\phi)$, we can trade for a significant reduction in $\Gamma_\beta(\phi)$. This is evident from a sharp decrease in $\Gamma_\beta(\phi)$ from 1 to 0.15, when $\Gamma_\alpha(\phi)$ slowly increases from 0 to 0.19. However, to further reduce the join-threat, a small decrease in $\Gamma_\beta(\phi)$ would result in a drastic increase in $\Gamma_\alpha(\phi)$, which is a less desirable trade-off. Generally, we can get a good trade-off plan if we allow some association loss and join-threat, without attempting to eliminate either factor or impose an extremely small threshold.

Next, we study the effects of the number of projected tables (N) on the plans. We evaluate the plans generated by CODIP* against our baseline, the naïve t-closeness approach (i.e., $N = 1$). Fig. 6 shows the results of our experiment.

As expected, when there are fewer projected tables in a plan, privacy is less protected as shown by the lesser reduction in join-threat in Fig. 6. On the other hand, data quality

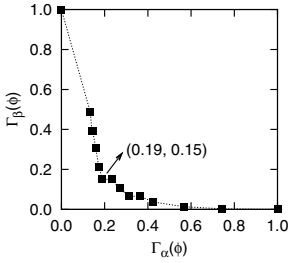

Fig. 5. Relationship of $\Gamma_\beta(\phi)$ and $\Gamma_\alpha(\phi)$

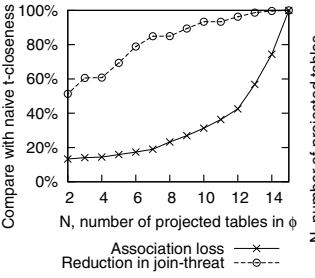

Fig. 6. Effects of no. of tables

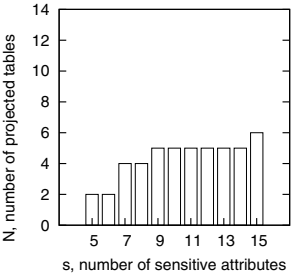

Fig. 7. Effects of no. of SAs ($\alpha = \beta = 0.3$)

improves as reflected in the decreasing association loss. This observation is consistent with CODIP*. In CODIP*, every merging action results in one fewer table causing less association loss while risking more join-threat. Note that as the number of projected tables increases, reduction in join-threat increases in a decreasing rate, whereas association loss increases in an increasing rate. Therefore, a good trade-off plan usually has a smaller number of projected tables (e.g., less than 7 in this experiment), and there are some association loss and join-threat that must be allowed (as we have just discussed based on Fig. 5).

Lastly, to show the scalability of CODIP*, we vary the number of SAs (s). The number of projected tables (N) outputted by CODIP* is shown in Fig. 7. As s increases, N also increases. However, the growth of N is minimal when s is large ($s \geq 9$ in this experiment). This result indicates that CODIP* is effective in protecting privacy while producing a small number of projected tables, even if there are a large number of SAs.

The experiments verified the possibility of greatly enhancing privacy while slightly sacrificing data quality, i.e., a good trade-off can be obtained in practice.

7.2 Closeness and Anonymity

Next, we study the closeness and anonymity requirements t and k, respectively. First, consider $k = 2$. To ensure the quality of QIDs, we also impose a discernibility metric [2] (d_m, in unit of 10^9) threshold on QIDs, such that k-anonymized tables with discernibility metric larger than d_m are not considered. Smaller d_m implies higher quality in QIDs, causing fewer number of valid anonymizations.

Following the analysis in Sect. 7.1, we set thresholds $\alpha = 0.2, \beta = 0.5$. Starting from $\beta = 0.5$, we gradually decrease it, and obtain 6 plans by CODIP*, each with a varying number of projected tables ($N \in [2, 7]$). Fig. 8 shows the best closeness achieved by the plans under different thresholds d_m for $N \in \{2, 4, 6\}$, in addition to the baseline ($N = 1$), and the special plan with each SA published in a separate table ($N = 15$). Specifically, Fig. 8(a) depicts the absolute closeness each plan can achieve at best, whereas Fig. 8(b) compares the plans with the baseline and presents the improvement of each plan.

We observe that smaller N results in weaker closeness. In CODIP*, N becomes smaller when more mergers take place, resulting in a non-decreasing number of SAs in

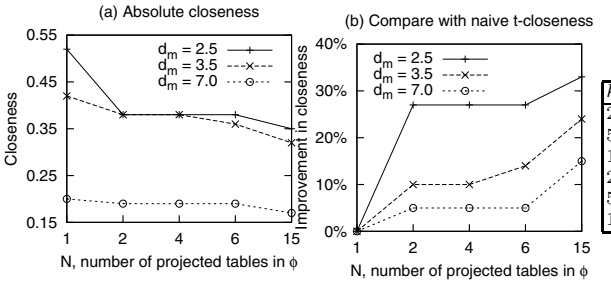

Fig. 8. Best closeness achieved by plans

(a) # of plans

k	$t = 0.4$	$t = 0.37$
2	6	2
5	6	2
10	6	2
20	6	2
50	0	0
100	0	0

(b) # of k-anonymizations

k	$t = 1$
2	46
5	28
10	18
20	14
50	5
100	1

Fig. 9. Effects of k ($\alpha = 0.2, \beta = 0.5, d_m = 3.5$)

each projected table. This result demonstrates the effect of diminishing closeness. Also note that when d_m is smaller, there are fewer valid anonymizations, resulting in weaker closeness. Hence, the improvement in closeness w.r.t. to the naïve t-closeness approach is potentially more significant.

Finally, we study the effects of k on closeness. We count the number of plans that can satisfy the various thresholds in Fig. 9. Fig. 9(a) presents our findings. Note that the baseline can only satisfy 0.42-closeness when $k \in \{2, 5, 10, 20\}$, and 0.52-closeness otherwise. When $k \leq 20$, we have quite a number of plans that can satisfy the requirements on closeness. As expected, a stronger closeness (i.e., a smaller t) results in fewer valid plans. However, when k becomes large ($k \geq 50$), there is apparently no plan that can satisfy the thresholds. The reason is that the number of valid k-anonymizations drops as k increases. Fig. 9(b) shows the number of valid k-anonymizations, assuming no requirement on closeness (i.e., $t = 1$). When k increases from 2 initially, the number of valid plans remains unaffected, as the k-anonymizations that are eliminated due to increased k are expected to have weaker closeness– the eliminated anonymizations contain at least an equivalence class whose cardinality is smaller than k, and smaller equivalence classes are generally less "well represented." When k continues to increase beyond 20, the number of valid k-anonymizations becomes too few. It is likely that none of these few satisfies the given closeness, which is indeed the case in this experiment. Results showed that if k is not too large (e.g., $k < 50$), CODIP* generates plans that satisfy stronger closeness, as compared to the baseline.

8 Conclusion

We studied the privacy issue of attribute disclosure in publishing microdata that have multiple SAs, of some may be multi-valued. We introduced Association Loss Ratio and Information Exposure Ratio to quantify data quality and privacy, respectively. We showed that a direct adaptation of t-closeness is inadequate, and proposed a framework CODIP and a heuristic CODIP*. Experiments showed that CODIP* generates good trade-off plans on a real dataset.

References

1. Asuncion, A., Newman, D.: UCI machine learning repository. Univ. of California, Irvine, ICS (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html
2. Bayardo, R., Agrawal, R.: Data privacy through optimal k-anonymization. In: ICDE, pp. 217–228 (2005)
3. Byun, J., Sohn, Y., Bertino, E., Li, N.: Secure anonymization for incremental datasets. In: Jonker, W., Petković, M. (eds.) SDM 2006. LNCS, vol. 4165, pp. 48–63. Springer, Heidelberg (2006)
4. Chen, Z., Gangopadhyay, A.: A Privacy Protection Model for Patient Data With Multiple Sensitive Attributes. IJISP 2(3), 28–44 (2008)
5. Cover, T., Thomas, J.: Elements of information theory. Wiley, Chichester (1991)
6. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. DMKD 11(2), 195–212 (2005)
7. Lambert, D.: Measures of disclosure risk and harm. JOS 9, 313–331 (1993)
8. LeFevre, K., DeWitt, D., Ramakrishnan, R.: Incognito: Efficient full-domain k-anonymity. In: SIGMOD, p. 60 (2005)
9. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In: ICDE, pp. 106–115 (2007)
10. Li, N., Li, T., Venkatasubramanian, S.: Closeness: A New Privacy Measure for Data Publishing. TKDE (June 2009)
11. Li, T., Li, N., Zhang, J., Molloy, I.: Slicing: a new approach for privacy preserving data publishing. cs.DB, arXiv preprint: 0909.2290v1
12. Li, Z., Ye, X.: Privacy protection on multiple sensitive attributes. In: Qing, S., Imai, H., Wang, G. (eds.) ICICS 2007. LNCS, vol. 4861, pp. 141–152. Springer, Heidelberg (2007)
13. Machanavajjhala, A., Gehrke, J., Kifer, D.: ℓ-diversity: Privacy beyond k-anonymity. In: ICDE, pp. 24–35 (2006)
14. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: ℓ-diversity: Privacy beyond k-anonymity. TKDD 1(1), 3 (2007)
15. Solanas, A., Sebé, F., Domingo-Ferrer, J.: Micro-aggregation-based heuristics for p-sensitive k-anonymity: one step beyond. In: PAIS, pp. 61–69 (2008)
16. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. IJUFKS 10(5), 571–588 (2002)
17. Sweeney, L.: k-anonymity: A model for protecting privacy. IJUFKS 10(5), 557–570 (2002)
18. Truta, T., Vinay, B.: Privacy protection: p-sensitive k-anonymity property. In: ICDE PDM Workshop, p. 94 (2006)
19. Wang, K., Fung, B.: Anonymizing sequential releases. In: SIGKDD, p. 423 (2006)
20. Wong, R., Fu, A., Wang, K., Pei, J.: Minimality attack in privacy preserving data publishing. In: VLDB, pp. 543–554 (2007)
21. Xiao, X., Tao, Y.: Anatomy: Simple and effective privacy preservation. In: VLDB, p. 150 (2006)
22. Ye, Y., Liu, Y., Wang, C., Lv, D., Feng, J.: Decomposition: Privacy preservation for multiple sensitive attributes. In: Zhou, X., Yokota, H., Deng, K., Liu, Q. (eds.) DASFAA 2009. LNCS, vol. 5463, pp. 486–490. Springer, Heidelberg (2009)
23. Zhang, Q., Koudas, N., Srivastava, D., Yu, T.: Aggregate query answering on anonymized tables. In: ICDE, pp. 116–125 (2007)