Using Safety Constraint for Transactional Dataset Anonymization

Bechara Bouna, Chris Clifton, Qutaibah Malluhi

Using Safety Constraint for Transactional Dataset Anonymization

Bechara AL Bouna[1], Chris Clifton[2], and Qutaibah Malluhi[1]

[1] Department of Computer Science and Engineering, Qatar University, Qatar
[2] Dept. of Computer Science/CERIAS, Purdue Univ., West Lafayette, Indiana, USA
bechara.albouna@qu.edu.qa
clifton@cs.purdue.edu qmalluhi@qu.edu.qa

Abstract. In this paper, we address privacy breaches in transactional data where individuals have multiple tuples in a dataset. We provide a safe grouping principle to ensure that correlated values are grouped together in unique partitions that enforce l-diversity at the level of individuals. We conduct a set of experiments to evaluate privacy breach and the anonymization cost of safe grouping.

1 Introduction

Data outsourcing is on the rise, and the emergence of cloud computing provides additional benefits to outsourcing. Privacy regulations pose a challenge to outsourcing, as the very flexibility provided makes it difficult to prevent against trans-border data flows, protection and separation of clients, and other constraints that may be required to outsource data. An alternative is encrypting the data [5]; while this protects privacy, it also prevents beneficial use of the data such as value-added services by the cloud provider (e.g., address normalization), or aggregate analysis of the data (and use/sale of the analysis) that can reduce the cost of outsourcing. Generalization-based data anonymization [18, 19, 12, 9] provides a way to protect privacy while allowing aggregate analysis, but doesn't make sense in an outsourcing environment where the client wants to be able to retrieve the original data values.

An alternative is to use *bucketization*, as in the anatomy [23], fragmentation [4], or slicing [11] models. Such a database system has been developed [15, 16]. The key idea is that identifying and sensitive information are stored in separate tables, with the join key encrypted. To support analysis at the server, data items are grouped into buckets; the mapping between buckets (but not between items in the bucket) is exposed to the server. An example is given in Figure 1 where attribute *DrugName* is

Name	Country	Drug Name
Roan (P1)	United States	Mild Exfoliation
Lisa (P4)	Columbia	Azelaic acid
Roan (P1)	United States	Retinoic Acid
Elyse (P2)	United States	Mild Exfoliation
Carl (P3)	France	Azelaic acid
Roan (P1)	United States	Retinoic Acid
Lisa (P4)	Columbia	Cytarabine
Roan (P1)	United States	Azelaic acid
Lisa (P4)	Columbia	Retinoic Acid
Carl (P3)	France	Cytarabine
Carl (P3)	France	Azelaic acid
Roan (P1)	United States	Retinoic Acid
Bob (P5)	Columbia	Esom. Magnesium
Carl (P3)	France	Mild Exfoliation
Alice (P6)	United States	Adapalene

(a) Original Prescription table

Name	Country	GID
Roan (P1)	United States	1
Lisa (P4)	Columbia	1
Roan (P1)	United States	1
Elyse (P2)	United States	2
Carl (P3)	France	2
Roan (P1)	United States	2
Lisa (P4)	Columbia	3
Roan (P1)	United States	3
Lisa (P4)	Columbia	3
Carl (P3)	France	4
Carl (P3)	France	4
Roan (P1)	United States	4
Bob (P5)	Columbia	5
Carl (P3)	France	5
Alice (P6)	United States	5

GID	Drug Name
1	Mild Exfoliation
1	Azelaic acid
1	Retinoic Acid
2	Mild Exfoliation
2	Azelaic acid
2	Retinoic Acid
3	Cytarabine
3	Azelaic acid
3	Retinoic Acid
4	Cytarabine
4	Azelaic acid
4	Retinoic Acid
5	Esom. Magnesium
5	Mild Exfoliation
5	Adapalene

(b) $Prescription_{QIT}$ and $Prescription_{SNT}$

Fig. 1: Table Prescription anonymized

sensitive: Figure 1b is an anatomized version of table prescription with attributes separated into $Prescription_{QIT}$ and $Prescription_{SNT}$.

The bucket size and grouping of tuples into buckets ensures privacy constraints (such as k-anonymity[18, 19] or l-diversity[12]) are satisfied.

Complications arise when extending this approach to transactional datasets. Even with generalization-based approaches, it has been shown that transactions introduce new challenges. While approaches as (X, Y)-privacy [21] and k^m-anonymity [20] include restrictions on the correlation of quasi-identifying values and can be used to model transactional data [3], they still face limitations when applied to bucketization approaches.

We give examples of this based on Figure 1b. The anonymized table satisfies the (X,Y)-privacy and (2,2)-diversity privacy constraints[13]; given the 2-diverse table, an adversary should at best be able to link a patient to a drug with probability 1/2.

Inter-group dependencies occur when an adversary knows certain facts about drug use, e.g., Retinoic Acid is a maintenance drug taken over a long period of time. As P1 is the only individual who appears in all groups where Retinoic Acid appears, it is likely that P1 is taking this drug. Note that this fact can either be background knowledge, or learned from the data.

Intra-group dependencies occur where the number of transactions for a single individual within a group results in an inherent violation of l-diversity (this would most obviously occur if all transactions in a group were for the same individual.) By considering this separately for transactional data, rather than simply looking at all tuples for an individual as a single "data instance", we gain some flexibility.

We present a method to counter such privacy violations while preserving data utility. Our contributions can be summarized as follows:

- An in-depth study of privacy violation due to correlation of individuals' related tuples in *bucketization* techniques.
- A safe grouping technique to eliminate privacy violation. Our safe grouping technique ensures that quasi-identifier and sensitive partitions respect l-diversity privacy constraint.

The approach is based on knowing (or learning) the correlations, and forming buckets with a common antecedent to the correlation. This protects against inter-group dependencies. Identifiers are then suppressed where necessary (in an outsourcing model, this corresponds to encrypting just the portion of the tuple in the identifier table) to ensure the privacy constraint is met (including protection against intra-group correlation.)

In the next section, we present our adversary model. Section 3 gives further background on prior work and its limitations in dealing with this problem. In Section 4, we define the basic notations and key concepts used in the rest of the paper. A definition of correlation-based privacy violation in transactional datasets is given in Section 5. In Section 6, we present our a safe grouping constraint that forms the basis of our anonymization method. Section 7 gives a safe grouping algorithm. A set of experiments to evaluate both the practical efficiency and the loss of data utility (suppression/encryption) is given in Section 8. We conclude with a discussion of next steps to move this work toward practical use.

2 Adversary Model

In our adversary model, we assume that the adversary has knowledge of the transactional nature of the dataset. We also assume that he/she has outside information on correlations between sensitive data items that leads to a high probability that certain sets of items would belong to the same individual. This is illustrated in the introduction (example 1) where the fact that the drug Retinoic Acid is known to be taken for a long period of time makes it possible to link it to patient P1.

We do not care about the source of such background information; it may be public knowledge, or it may be learned from the anatomized data itself. (We view learning such knowledge from the data as beneficial aggregate analysis of the data.)

3 Related Work

In [21], the authors consider that any transaction known by the adversary could reveal additional information that might be used to uncover a sensitive linking between a quasi-identifier and a sensitive value. They define (X,Y)-*privacy* to ensure on one hand that each value of X is linked to at least k different values of Y, and on the other hand, no value of Y can be inferred from a value of X with confidence higher than a designated threshold. A similar approach proposed in [20] in which the authors extend k-anonymity with k^m-anonymity requiring that each combination of at most m items appears in at least k transactions, where m is the maximum number of items per transaction that could be known by the adversary. (Also related is the problem of trail re-identification[14].) As demonstrated in the example in Figure 1b, these techniques are limited when it comes to *bucketization*, as more subtle in *intra* and *intra* group correlations may lead to a breach of l-diversity.

In [11] the authors proposed a slicing technique to provide effective protection against membership disclosure, but it still remains vulnerable to identity disclosure. An adversary with knowledge of the transactional nature of the dataset may still be able to associate an individual identifier to correlated sensitive values. The authors in [6] discuss privacy violations in the anatomy privacy model [23] due to functional dependencies (FDs). In their approach, they propose to create QI-groups on the basis of a FD tree while grouping tuples based on the sensitive attribute to form l-diverse groups. Unfortunately, dealing with FDs' is not sufficient, as less strict dependencies can still pose a threat.

In [22], the authors consider correlation as foreground knowledge that can be mined from anonymized data. They use the possible worlds model to compute the probability of associating an individual to a sensitive value based on a global distribution. In [8], a Naïve Bayesian model is used to compute association probability. They used exchangeability [1] and DeFinetti's theorem [17] to model and compute patterns from the anonymized data. Both papers address correlation in its general form where the authors show *how* an adversary can violate l-diversity privacy constraint through an estimation of such correlations in the anonymized data. As it is a separate matter, we consider that correlations due to transactions where multiple tuples are related to the same individual ensure that particular sensitive values can be linked to a particular individual when correlated in the same group (i.e., bucket). We go beyond this, ad-

dressing any correlation (either learned from the data or otherwise known) that explicitly violates the targeted privacy goal.

4 Formalization

Given a table T with a set of attributes $\{A_1, ..., A_b\}$, $t[A_i]$ refers to the value of attribute A_i for the tuple t. Let U be the set of individuals of a specific population, $\forall u \in U$ we denote by T_u the set of tuples in T related to the individual u. Attributes of a table T that we deal with in this paper are divided as follows;

- *Identifier* (A^{id}) represents an attribute that can be used (possibly with external information available to the adversary) to identify the individual associated with a tuple in a table. We distinguish two types of identifiers; sensitive and nonsensitive. For instance, the attribute *Social Security Number* is a *sensitive identifier*; as such it must be suppressed (encrypted). *Nonsensitive identifiers* are viewed as public information, and include both direct identifiers such as *Patient ID* in Figure 4, and quasi-identifiers that in combination may identify an individual (such as <*Gender, Birthdate, Zipcode*>, which uniquely identifies many individuals.)
- *Sensitive attribute* (A^s) contains sensitive information that must not be linkable to an individual, but does not inherently identify an individual. In our example (Table 1a), the attribute *DrugName* is considered sensitive and should not be linked to an individual.

Definition 1 (Equivalence class / QI-group). *[18] A quasi-identifier group (*QI-group*) is defined as a subset of tuples of* $T = \bigcup_{j=1}^{m} QI_j$ *such that, for any* $1 \leq j_1 \neq j_2 \leq m$, $QI_{j1} \cap QI_{j2} = \phi$.

Note that for our purposes, this can include direct identifiers as well as quasi-identifiers; we stick with the QI-group terminology for compatibility with the broader anonymization literature.

Definition 2 (l-diversity). *[13] a table T is said to be l-diverse if each of the QI-groups $QI_j (1 \leq j \leq m)$ is l-diverse; i.e., QI_j satisfies the condition $c_j(v_s)/|QI_j| \leq 1/l$ where*

- *m is the total number of QI-groups in T*
- *v_s is the most frequent value of A^s*
- *$c_j(v_s)$ is the number of tuples of v_s in QI_j*
- *$|QI_j|$ is the size (number of tuples) of QI_j*

Definition 3 (Anatomy). *Given a table T, we say that T is anatomized if it is separated into a quasi-identifier table (T_{QIT}) and a sensitive table (T_{SNT}) as follows:*

- *T_{QIT} has a schema $(A_1, ..., A_d, GID)$ where A_i $(1 \leq i \leq d)$ is either a nonsensitive identifying or quasi-identifying attribute and GID is the group id of the QI-group.*
- *T_{SNT} has a schema (GID, A^s_{d+1}) where A^s_{d+1} is the sensitive attribute in T.*

Table 1: Notations

T	a table containing individuals related tuples
t_i	a tuple of T
u	an individual described in T
T_u	a set of tuples related to individual u
A	an attribute of T
A^{id}	an identifying attribute of T
A^s	a sensitive attribute of T
QI_j	a quasi-identifier group
T^*	Anonymized version of table T

To express correlation in transactional data we use the following notation $cd^{id} : A^{id}_1, ..., A^{id}_n \dashrightarrow A^s$ where A^{id}_i is a nonsensitive identifying attribute and A^s is a sensitive attribute, and cd^{id} is a correlation dependency between attributes $A^{id}_1, ..., A^{id}_n$ on one hand, and A^s on the other.

Next, we present a formal description of the privacy violation that can be caused due to such correlations.

5 Correlation-Based Privacy Violation

Inter-group correlation occurs when transactions for a single individual are placed in multiple QI-groups (as with P1, P3, and P4 in Figure 1a). The problem arises when the values in different groups are related (as would happen with association rules); this leads to an implication that the values belong to the same individual. Formally:

Definition 1 (Inter QI-group Correlation). *Given a correlation dependency of the form $cd^{id} : A^{id} \dashrightarrow A^s$ over T^*, we say that a privacy*

violation might exists if there are correlated values in a subset $QI_j (1 \leq j \leq m)$ *of* T^* *such that* $v_{id} \in \pi_{A^{id}} QI_1 \cap ... \cap \pi_{A^{id}} QI_m$ *and* $|\pi_{A^s} QI_1 \cap ... \cap \pi_{A^s} QI_m| < l$ *where* $v_{id} \in A^{id}$ *is an individual identifying value,* l *is the privacy constant and an adversary knows of that correlation.*

The example shown in Figure 1, explains how an adversary with prior knowledge of the correlation, in this case that Retinoic Acid must be taken multiple times, is able to associate the drug to the patient Roan (P1) due to their correlation in several QI-groups. (The same would also apply to different drugs that must be taken together.)

An intra-group violation can arise if several correlated values are contained in the same QI-group. Here the problem is that this gives a count of tuples that likely belong to the same individual, which may limit it to a particular individual in the group. Figure 2 is an example of Intra QI-group Correlation, formally defined as follows:

Patient ID	GID
Roan (P1)	1
Roan (P1)	1
Roan (P1)	1
Carl (P3)	1
Carl (P3)	1
Roan (P1)	1

GID	Drug Name
1	Retinoic Acid
1	Retinoic Acid
1	Retinoic Acid
1	Azelaic acid
1	Azelaic acid
1	Azelaic acid

Fig. 2: Intra QI-group correlation

Lemma 1 (Intra QI-group Correlation). *Given a QI-group* QI_j $(1 \leq j \leq m)$ *in* T^* *that is* l*-diverse, we say that a privacy violation might occur if individual's related tuples are correlated in* QI_j *such that* $f(QI_j, u) + c_j(v_s) > |QI_j|$ *where*

- v_s *is the most frequent* A^s *value in* QI_j,
- $c_j(v_s)$ *is the number of tuples* $t \in QI_j$ *with* $t[A^s] = v_s$,
- u *is the individual who has the most frequent tuples in* QI_j,
- $f(QI_j, u)$ *is a function that returns the number of* u *'s related tuples in* QI_j
- $|QI_j|$ *is the size of* QI_j *(number of tuples contained in* QI_j*)*

Proof. Let r be the number of remaining sensitive values in QI_j, $r = |QI_j| - c_j(v_s)$. If $f(QI_j, u) + c_j(v_s) > |QI_j|$, this means that $f(QI_j, u) > |QI_j| - c_j(v_s)$ and therefore $f(QI_j, u) > r$. That is, there are e tuples related to individual u such that $f(QI_j, u) = e$ to be associated to r sensitive values of QI_j where $e > r$. According to the pigeon-hole principle,

at least a tuple t of T_u will be associated to the sensitive value v_s which leads to a privacy violation. □

It would be nice if this lemma was "if and only if", giving criteria where a privacy violation would NOT occur. Unfortunately, this requires making assumptions about the background knowledge available to an adversary (e.g., if an adversary knows that one individual is taking a certain medication, they may be able to narrow the possibilities for other individuals). This is an assumption made by all k-anonymity based approaches, but it becomes harder to state when dealing with transactional data.

Let us go back to Figure 2, an adversary is able to associate both drugs (Retinoic Acid and Azelaic Acid) to patient Roan (P1) due to the correlation of their related tuples in the same QI-group.

In the following, we provide an approach that deals with such privacy violations.

6 Safe Grouping for Transactional Data

As we have shown in the previous section, bucketization techniques do not cope well with correlation due to transactional data where an individual might be represented by several tuples that could lead to identify his/her sensitive values. In order to guarantee safety, we present in this section our safe grouping safety constraint .

Safety Constraint (Safe Grouping)**.** *Given a correlation dependency in the form of* $cd^{id} : A^{id} \dashrightarrow A^s$*, safe grouping is satisfied iff*

1. $\forall u \in U$*, the subset* T_u *of* T *is contained in one and only one quasi identifier group* QI_j *(*$1 \leq j \leq m$*) such that* QI_j *respects* l*-diversity and contains at least* k *subsets* $T_{u_1}, ..., T_{u_k}$ *where* $u_1, ..., u_k$ *are* k *distinct individuals of the population and,*
2. $Pr(u_{i_1}|QI_j) = Pr(u_{i_2}|QI_j) \leq 1/l$ *where* $u_{i_1}, u_{i_2}, i_1 \neq i_2$ *are two distinct individuals in* QI_j *with (*$1 \leq i \leq k$*) and* $Pr(u_i|QI_j)$ *is the probability of* u_i *in* QI_j*.*

Safe grouping ensures that individual tuples are grouped in one and only one QI-group that is at the same time l-diverse, respects a minimum diversity for identity attribute values, and every subset T_u in QI_j are of equal number of tuples.

Figure 3 describes a quasi identifier group (QI_1) that respects safe grouping where on one hand, we assume that there are no other QI-groups containing $P1$ and $P3$ and on the other hand, two tuples from T_{P1} are

anonymized to guarantee that $Pr(P1|QI_1) = Pr(P3|QI_1) \leq 1/2$. Note that we have suppressed some data in order to meet the constraint; this is in keeping with the model in [15] where some data is left encrypted, and only "safe" data is revealed.

Lemma 1. *Let QI_j for ($1 \leq j \leq m$) be a QI-group that includes k individuals, if QI_j satisfies safe grouping then k is at least equal to l*

Proof. Consider an individual u in QI_j, according to the safe grouping, $Pr(u|QI_j) \leq 1/l$. Or $Pr(u|QI_j)$ is equal to $f(QI_j, u)/|QI_j|$ where $f(QI_j, u) = |QI_j|/k$ represents the number of individual's u related tuples in QI_j. Hence, $1/k \leq 1/l$ and $k \geq l$ □

Corollary 1 (Correctness). *Given an anonymized table T^* that respects safe grouping, and a correlation dependency of the form $cd^{id} : A^{id} \dashrightarrow A^s$, an adversary cannot correctly associate an individual u to a sensitive value v_s with a probability $Pr(A^s = v_s, u|T^*)$ greater than $1/l$.*

Proof. Safe grouping guarantees that individual's u related tuples T_u are contained in one and only one QI-group (QI_j), which means that possible association of u to v_s is limited to the set of correlated values that are contained in QI_j. Hence, $Pr(A^s = v_s, u|T^*)$ can be written as $Pr(A^s = v_s, u|QI_j)$. On the other hand,
$Pr(A^s = v_s, u|QI_j) = \frac{Pr(A^s=v_s,u)}{\sum_{i=1}^{k} Pr(A^s=v_s,u_i)}$ where k is the number of individuals in QI_j and $Pr(A^s = v_s, u_i)$ is the probability of associating individual u_i to a sensitive value v_s. Recall that safe grouping guarantees that for a given individual u_i, $Pr(A^s = v_s, u_i)$ is at the most equal to $1/l$. Summarizing, $Pr(A^s = v_s, u|QI_j)$ is at the most equal to $1/k$ where $k \geq l$ according to Lemma 1. □

We can estimate[3], for example, $Pr(A^s = RetinoicAcid, A^{id} = P1|T^*)$ to be 4/5 where it is possible to associate Roan (P1) to Retinoic Acid in 4 of 5 QI-groups as shown in Figure 1b. However, as you can notice from Figure 3, safe grouping guarantees that $Pr(A^s = RetinoicAcid, A^{id} = P1|T^*)$ remains limited to the possible association of values in QI_1 and thus bounded by l-diversity.

The safe grouping constraint is restrictive, but may be necessary. While we do not have a formal proof that it is optimal, we can find

[3] $Pr(A^s = RetinoicAcid, A^{id} = P1|T^*)$ as calculated remains an estimation where a much deeper aspect on how to calculate the exact probability of values correlated across QI-groups can be seen in [22] and [8]

examples where any straightforward relaxation can result in a privacy violation (we do not elaborate due to space constraints.)

We note that using safe grouping, we do not intend to replace anatomy. In fact, we preserve table decomposition as described in the original anatomy model by separating a table T into two subtables (T_{QIT}, T_{SNT}) while providing a safe grouping of tuples on the basis of the attributes related by a correlation dependency.

Presc ID	Patient ID	GID
1	Roan (P1)	1
5	Roan (P1)	1
7	Roan (P1)	1
2	Roan (P1)	1
10	Carl (P3)	1
12	Carl (P3)	1
9	Carl (P3)	1
6	Carl (P3)	1
8	*	2
11	*	2
3	Lisa (P4)	2
4	Elyse (P2)	2
13	Bob (P5)	2

GID	Drug Name
1	Retinoic Acid
1	Retinoic Acid
1	Retinoic Acid
1	Azelaic acid
1	Azelaic acid
1	Azelaic acid
1	Mild Exfoliation
1	Nexium
2	Adapalene
2	Cytarabine
2	Cytarabine
2	Mild Exfoliation
2	Mild Exfoliation

Fig. 3: Table Prescription respecting our safety constraint

7 Safe Grouping Algorithm

In this section, we provide an algorithm to enforce ensure safe grouping for transactional data. The algorithm guaranties the safe grouping of a table T based on an identity attribute correlation dependency $cd^{id} : A^{id} \dashrightarrow A^s$ ($A^{id} \in T_{QIT}$ and $A^s \in T_{SNT}$).

The main idea behind the algorithm is to create k buckets based on the attribute (A^{id}) defined on the left hand side of a correlation dependency in a reasonable time.

The safe grouping algorithm takes a table T, a correlation dependency $A^{id} \dashrightarrow A^s$, a constant l to ensure diversity, and a constant k representing the number of individuals (individuals' related tuples) to be stored in a QI-group. It ensures a safe grouping on the basis of the attribute A^{id}. In Step 2, the algorithm hashes the tuples in T based on their A^{id} values and sorts the resulting buckets. For any individual, all their values will end up in the same bucket. In the group creation process from steps 4-17, the algorithm creates a QI-group with k individuals. If the QI-group respects l-diversity the algorithm adds it to the list of QI-groups and enforces the safety constraint in step 8 by anonymizing tuples in T_{QIT} including values that are frequently correlated in the QI-group. In other terms, it makes sure that individuals' related tuples in the QI-group are of equal number.

Algorithm 1 SafeGrouping

Require: a table T, $cd^{id} : A^{id} \dashrightarrow A^s$, l, k, $minConf$, $maxConf$ and exp
Ensure: safe grouping for T

1: $T_{QIT} = \varnothing$; $T_{SNT} = \varnothing$; $gn = 0$; $i = 0, j = 0$;
2: Hash the tuples in T by their A^{id} values (one bucket per A^{id} value)
3: Sort the set of $Buckets$ based on their number of tuples.
4: **while** there are k groups $QI \in Buckets$ **do**
5: **if** QI is l-diverse **then**
6: $gn = gn + 1$
7: $QI_{gn} = QI$
8: Enforce safety constraint on QI_{gn}
9: Remove QI from $Buckets$
10: **else**
11: Enforce l-diversity over QI
12: **if** size of $QI < k$ **then**
13: Suppress tuples in QI
14: **else**
15: Go to Step 6
16: **end if**
17: **end if**
18: **end while**
19: **for** $j = 1$ to gn **do**
20: **for** each tuple $t \in QI_j$ **do**
21: insert tuple $(t[A_1], ..., t[A], ...t[A_m], j)$ into T_{QIT}
22: **end for**
23: **for** each random value v_s of $A^s \in QI_j$ **do**
24: insert tuple (j, v_s) into T_{SNT}
25: **end for**
26: **end for**

If l-diversity for the QI-group in question is not met, the algorithm enforces it by anonymizing tuples related to the most frequent sensitive value in the QI-group. After the l-diversity enforcement process, the algorithms verifies whether the group contains k buckets, and if not anonymizes (which could mean generalizing, suppressing, or encrypting the values, depending on the target model.)

From steps 19 to 26 the algorithm anatomizes the tables based on the QI-groups created. It stores random sensitive attribute values in the T_{SNT} table.

While safe grouping provides safety, its ability to preserve data utility is limited to the number of distinct values of A^{id} attribute.

8 Experimentation

We now present a set of experiments to evaluate the efficiency of our approach, both in terms of computation and more importantly, loss of data utility. We implemented the safe grouping code in Java based on the Anonymization Toolbox [7], and conducted experiments with an Intel XEON 2.4GHz PC with 2GB RAM.

8.1 Evaluation dataset

In keeping with much work on anonymization, we use the Adult Dataset from the UCI Machine Learning Repository [2]. To simulate real identifiers, we made use of a U.S. state voter registration list containing the attributes *Birthyear*, *Gender* , *Firstname*, and *Lastname*. We combined the adult dataset with the voter's list such that every individual in the voters list is associated with multiple tuples from the adult dataset, simulating a longitudinal dataset from multiple census years. We have constructed this dataset to have a correlation dependency of the following form $Firstname, Lastname \dashrightarrow Occupation$; where *Occupation* is a sensitive attribute, $Firstname, Lastname$ are identifying attributes and remaining attributes are presumed to be quasi-identifiers.

We say that an individual is likely to stay in the same occupation across multiple censuses. Note that this is not an exact longitudinal dataset; n varies between individuals (simulating a dataset where some individuals move into or out of the census area. The generated dataset is of size 48836 tuples with 21201 distinct individuals.

In the next section, we present and discuss results from running our safe grouping algorithm.

8.2 Evaluation Results

We elaborated a set of measurements to evaluate the efficiency of safe grouping. These measurements can be summarized as follows:

- Evaluating privacy breach in a naive anatomization. We note that the same test could be performed on the slicing technique [11] as the authors in their approach do not deal with identity disclosure,
- Determining anonymization cost represented by the loss metric to capture the fraction of tuples that must be (partially or totally) generalized, suppressed, or encrypted in order to satisfy the safe grouping and,
- Comparing the computational cost of our safe grouping algorithm to anatomy [23].

8.2.1 Evaluating Privacy After naïve anatomization over the generated dataset, we have identified 5 explicit violations due to intra QI-group correlations where values of A^{id} are correlated in a QI-group. On the other hand, in order to determine the number of violations due to inter QI-group correlation, we calculate first the possible associations of A^{id} and A^s values across a naïve anatomized table. This is summarized in the following equation for values v_{id} and v_s respectively.

$$\mathcal{G}(v_{id}, v_s) = \frac{\sum_{j=1}^{m} f_j(v_{id}, v_s)}{\sum_{j=1}^{m} g_j(v_{id})}$$

where

$$f_j(v_{id}, v_s) = \begin{cases} 1 \text{ if } v_{id} \text{ and } v_s \text{ are associated in } QI_j \\ 0 \; otherwise \end{cases}$$

and,

$$g_j(v_{id}) = \begin{cases} 1 \text{ if } v_{id} \text{ exists in } QI_j \\ 0 \text{ otherwise} \end{cases}$$

At this point, a violation occurs for significant[4] A^{id} values if;

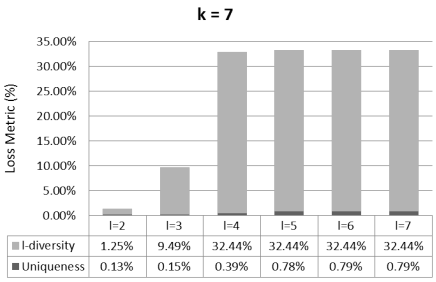
1. $\mathcal{G}(v_{id}, v_s) > 1/l$. This represents a *frequent* association between v_{id} and v_s where v_{id} is more likely to be associated to v_s in the QI-groups to which it belongs and,
2. $|\pi_{A^s} QI_1 \cap ... \cap \pi_{A^s} QI_m| < l$ where $QI_1, ..., QI_m$ are the QI-groups to which v_{id} belongs.

After we applied the above test to the anatomized dataset, we have identified for $l = 2$ and $l = 3$, 167 and 360 inter QI-groups correlation violations. We note that a much deeper study on violations due to data correlation can be found in [22][8][10].

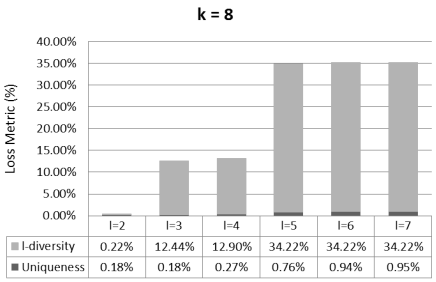
8.2.2 Evaluating Anonymization Cost We evaluate our proposed anonymization algorithms to determine the loss metric ($\mathcal{LM}$) representing the number of tuples in T and T_{QIT} that need to be suppressed in order to achieve the safety constraint. Figure 3 shows a anonymized version of table prescription where grouping is safe and has a loss metric equal to $\mathcal{LM}(Prescription) = 2/13$.

We investigate the anonymization cost for a correlation dependency $cd^{id} : Firstname, Lastname \dashrightarrow Occupation$ using the safe grouping

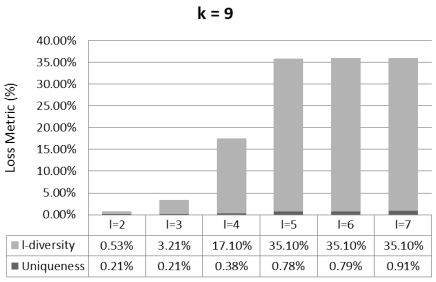
[4] Significance is measured in this case based on the support of A^{id} values and their correlation across QI-groups. For instance, v_{id} is considered significant if it exists in at least α QI-groups where α is a predefined constant greater than 2.



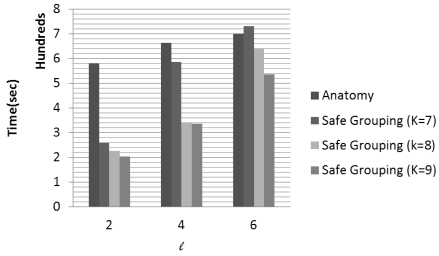
(a) % of tuples anonymized to ensure the safety constraint and l-diversity for $k = 7$



(b) % of tuples anonymized to ensure the safety constraint and l-diversity for $k = 8$



(c) % of tuples anonymized to ensure the safety constraint and l-diversity for $k = 9$



(d) Computational Cost Evaluation

Fig. 4: Safe grouping evaluation in transactional datasets 4a, 4b and 4c

algorithm. We anonymize the dataset with $k = 7, 8, 9$ and $l = 2, 3, 4, 5, 6, 7$ for which the dataset satisfies the eligibility condition (see [13]). At each execution, we compute the $\mathcal{LM}$. The results are shown in Figure 4.

From Figure 4, we can see that the $\mathcal{LM}$ increases with l, and for ($k = 9$, $l = 7$) the computed loss metric $\mathcal{LM}$ is high; notice that the number of tuples to anonymize in order to preserve l-diversity reaches 35%. Nonetheless, for small values of l an acceptable value of $\mathcal{LM}$ is computed. Anonymizing datasets using safe grouping can be seen as a trade-off between cost and privacy where for small values of l, $\mathcal{LM}$ produces values less than 10% leading to a relatively small anonymization cost. Another aspect to consider is how to define k w.r.t l to guarantee a minimum $\mathcal{LM}$. Note that for transactional data, it is possible for k (the number of individuals, not transactions, in a group) to be smaller than l; however, this makes satisfying the privacy criteria difficult, lead-

ing to substantial anonymized data. The experiments show that high data utility can be preserved as long as k is somewhat greater than l.

8.2.3 Evaluating Computation Cost We now give the processing time to perform safe grouping compared to anatomy. Figure 4d shows the computation time of both safe grouping and anatomy over a non-transactional dataset with different k values. Theoretically, a worst case of safe grouping could be much higher; but in practice, for a small values of l the safe grouping has better performance than anatomy. Furthermore, as k increases the safe grouping computation time decreases due to reduced I/O access needed to test QI-groups' l-diversity.

9 Conclusion

In this paper, we proposed a safe grouping method to cope with defects of bucketization techniques in handling correlated values in a transactional dataset. Our safe grouping algorithm creates partitions with an individual's related tuples stored in one and only one group, eliminating these privacy violations. We showed, using a set of experiments, that there is a trade-off to be made between privacy and utility. This trade-off is quantified based on the number of tuples to be anonymized using the safe grouping algorithm. Finally, we investigated the computation time of safe grouping and showed that despite the exponential growth of safe grouping, for a small range of values of l, safe grouping outperforms anatomy while providing stronger privacy guarantees.

10 Acknowledgements

This publication was made possible by NPRP grant 09-256-1-046 from the Qatar National Research Fund. The statements made herein are solely the responsibility of the authors.

References

1. David J. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.*, pages 1–198. Springer, Berlin, 1985.
2. A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.
3. Thorben Burghardt, Klemens Böhm, Achim Guttmann, and Chris Clifton. Anonymous search histories featuring personalized advertisement - balancing privacy with economic interests. *Transactions on Data Privacy*, 4(1):31–50, April 2011.

4. Valentina Ciriani, Sabrina De Capitani Di Vimercati, Sara Foresti, Sushil Jajodia, Stefano Paraboschi, and Pierangela Samarati. Combining fragmentation and encryption to protect privacy in data storage. *ACM Trans. Inf. Syst. Secur.*, 13:22:1–22:33, July 2010.
5. Hakan Hacıgümüş, Balakrishna R. Iyer, and Sharad Mehrotra. Executing SQL over encrypted data in the database-service-provider model. In *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, pages 216–227, Madison, WI, June 4-6 2002.
6. Xiao Jiang, Jun Gao, TengjiaoWang, and Dongqing Yang. Multiple sensitive association protection in the outsourced database. In *Database Systems for Advanced Applications (DASFAA)*, pages 123–137, Tsukuba, Japan, April 1-4 2010.
7. Murat Kantarcioglu, Ali Inan, and Mehmet Kuzu. Anonymization toolbox, 2010.
8. Daniel Kifer. Attacks on privacy and definetti's theorem. In *SIGMOD Conference*, pages 127–138, 2009.
9. Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, pages 106–115, 2007.
10. Tiancheng Li and Ninghui Li. Injector: Mining background knowledge for data anonymization. In *ICDE*, pages 446–455, 2008.
11. Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy. Slicing: A new approach for privacy preserving data publishing. *IEEE Trans. Knowl. Data Eng.*, 24(3):561–574, 2012.
12. Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. *l*-diversity: Privacy beyond *k*-anonymity. In *Proc. 22nd IEEE Intl. Conference on Data Engineering (ICDE 2006)*, Atlanta GA, April 2006.
13. Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. *l*-diversity: Privacy beyond *k*-anonymity. *ACMTransactions on Knowledge Discovery from Data (TKDD)*, 1(1), March 2007.
14. Brad Malin. *Trail Re-identification and Unlinkability in Distributed Databases*. PhD thesis, Carnegie Mellon University, 2006.
15. Ahmet Erhan Nergiz and Chris Clifton. Query processing in private data outsourcing using anonymization. In *The 25th IFIP WG 11.3 Conf. on Data and Applications Security and Privacy (DBSEC-11)*, Richmond, VA, July 11-13 2011.
16. Ahmet Erhan Nergiz, Chris Clifton, and Qutaibah Malluhi. Updating outsourced anatomized private databases. In *16th International Conference on Extending Database Technology (EDBT)*, Genoa, Italy, March 18-22 2013.
17. Paul Ressel. De Finetti-type theorems: an analytical approach. *Ann. Probab.*, 13(3):898–922, 1985.
18. Pierangela Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.*, 13(6):1010–1027, 2001.
19. Latanya Sweeney. k-anonymity: a model for protecting privacy. *Intl. Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.
20. M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy-preserving anonymization of set-valued data. *Proc. VLDB Endow.*, 1(1):115–125, August 2008.
21. Ke Wang and Benjamin C. M. Fung. Anonymizing sequential releases. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 414–423, New York, NY, USA, 2006. ACM.
22. Raymond Chi-Wing Wong, Ada Wai-Chee Fu, Ke Wang, Yabo Xu, and Philip S. Yu. Can the utility of anonymized data be used for privacy breaches? *CoRR*, abs/0905.1755, 2009.
23. Xiaokui Xiao and Yufei Tao. Anatomy: Simple and effective privacy preservation. In *Proceedings of 32nd International Conference on Very Large Data Bases (VLDB 2006)*, Seoul, Korea, September 12-15 2006.