Efficient Sanitization of Unsafe Data Correlations

Bechara AL Bouna
Department of Computer Science and Engineering
Qatar University & Antonine University
Doha, Qatar - Baabda, Lebanon
bechara.albouna@upa.edu.lb

Chris Clifton
Department of Computer Science
Purdue University
West Lafayette, Indiana - USA
clifton@cs.purdue.edu

Qutaibah Malluhi
Department of Computer Science and Engineering
Qatar University
Doha, Qatar
qmalluhi@qu.edu.qa

ABSTRACT

In this paper, we present a study to counter privacy violation due to unsafe data correlation. We propose a safe correlation requirement to keep correlated values bounded by l-diversity and evaluate the trade-off to be made for the sake of a strong privacy guarantee. Finally, we present a correlation sanitization algorithm that enforces our safety constraint and demonstrates its efficiency.

1. INTRODUCTION

Preserving privacy in outsourced databases has received considerable attention in the last decade. Several privacy constraints [23, 22, 16, 11] have been defined on datasets to prevent disclosure of sensitive information related to individuals. These constraints are based on generalizations that transform quasi-identifiers values into a general form and create quasi-identifier groups to eliminate possible linking attacks. A second approach is *table decomposition*: Quasi-identifiers and sensitive values are placed in separate tables, and tuples are divided into groups that are linked in a way that provides sufficient uncertainty in the join criteria to meet privacy constraints. This approach has been alternately termed anatomy [29], fragmentation [4] and slicing [14]; we will use the term anatomy to refer to this class of approaches, as it does not have other meanings in the database community.

Anatomy has the advantage that exact data values are maintained, allowing data and actions on individual data values to be outsourced. Only the link between identifying and sensitive values is generalized. We envision this work being used in the context of [20], where the actual links (and in our case, some data values) are encrypted to ensure the server cannot violate privacy, while still enabling some server-side use of the data.

As an example, Figure 1a shows prescription history, where the attribute *DrugName* is sensitive. Figure 1b represents an anatomized version of Table Prescription with attributes separated into $Prescription_{QIT}$ and $Prescription_{SNT}$. The anonymized table satisfies the 2-diversity privacy constraint[16]; given the 2-diverse table, an adversary can at best link a patient to a drug with a probability equal to 1/2.

Despite anatomy's efficiency in preserving privacy and data fidelity, it and other generalization based techniques defects when some types of correlation exist in the data.

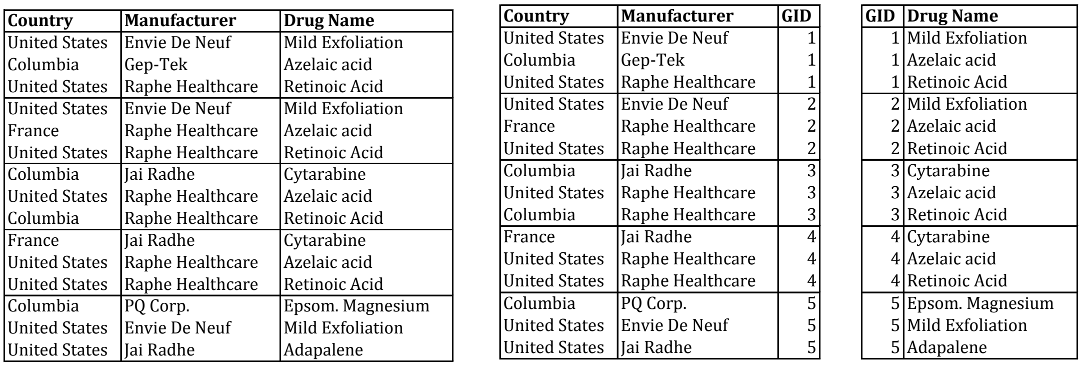
The most obvious problem is when values in identifiers (or quasi-identifiers) are directly correlated with sensitive values as discussed in [27] and [8]. Based on knowledge of such correlation (possibly learned from the data), an adversary may increase the probability that a given individual is linked to a given sensitive value with probability greater than the $1/l$ enforced by the anatomization groups. The inter quasi-identifying group correlation between *United States* and *Retinoic Acid* given in Figure 1 shows that with respect to knowledge mined from the anonymized data, an adversary is able to assert knowledge regarding the global distribution of countries and drugs. Such global distribution increases the probability of linking individuals to sensitive values on the basis of their countries. The authors of [27, 8] demonstrated how correlation can be used to violate privacy constraints. They argued that the (sometimes implicit) assumptions of an i.i.d. model and random worlds model, when tuple independence does not hold in the actual data, allows adversaries to learn and use cases where these assumptions do not hold to violate privacy. In this paper, we shed more light on the threat of data correlation that could be found after a naïve anonymization of a table, and give methods to control that risk. While we use the anatomy model [29] in our examples, this work also applies to other bucketization techniques such as fragmentation [4] and slicing [14].

1.1 Contributions

We present a study to counter privacy violations and at the same time preserve data utility. Our contributions can be summarized as follows:

- We propose a safe correlation requirement to reduce the threat of exposed correlations between quasi-identifier and sensitive attributes. We show that under this requirement, correlations can be bounded by a trade-off between utility and privacy.
- We provide a sanitization algorithm to ensure safety from correlations by solving a linear programming problem in a post-anonymization process.

The key idea is that we do not completely hide correlations



Country	Manufacturer	Drug Name
United States	Envie De Neuf	Mild Exfoliation
Columbia	Gep-Tek	Azelaic acid
United States	Raphe Healthcare	Retinoic Acid
United States	Envie De Neuf	Mild Exfoliation
France	Raphe Healthcare	Azelaic acid
United States	Raphe Healthcare	Retinoic Acid
Columbia	Jai Radhe	Cytarabine
United States	Raphe Healthcare	Azelaic acid
Columbia	Raphe Healthcare	Retinoic Acid
France	Jai Radhe	Cytarabine
United States	Raphe Healthcare	Azelaic acid
United States	Raphe Healthcare	Retinoic Acid
Columbia	PQ Corp.	Epsom. Magnesium
United States	Envie De Neuf	Mild Exfoliation
United States	Jai Radhe	Adapalene

(a) Original Prescription table

Country	Manufacturer	GID
United States	Envie De Neuf	1
Columbia	Gep-Tek	1
United States	Raphe Healthcare	1
United States	Envie De Neuf	2
France	Raphe Healthcare	2
United States	Raphe Healthcare	2
Columbia	Jai Radhe	3
United States	Raphe Healthcare	3
Columbia	Raphe Healthcare	3
France	Jai Radhe	4
United States	Raphe Healthcare	4
United States	Raphe Healthcare	4
Columbia	PQ Corp.	5
United States	Envie De Neuf	5
United States	Jai Radhe	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	Epsom. Magnesium
5	Mild Exfoliation
5	Adapalene

(b) Anonymized Prescription table $Prescription_{QIT}$ and $Prescription_{SNT}$

Figure 1: Example scenario

(we want to support learning from the data), this follows the spirit of t-closeness [11], but building on anatomy allows us greater grouping flexibility without the utility loss from over-generalizing data values.

2. ADVERSARY MODEL

We assume that both the adversary and defender have knowledge of correlations in the data; in the case of the adversary, his/her knowledge is mainly based on what can be learned from the anonymized data. As for the defender, it can include any correlations that can be learned from the original data. We also assume that an adversary has outside information enabling it to link (quasi)-identifying information with individuals. Thus all quasi-identifiers and identifiers are considered individually identifiable.

We assume that the adversary *does not* have prior knowledge of sensitive values for specific individuals. For example, if an adversary knew the prescriptions being taken by all of the individuals in Figure 1b except for a specific individual, then it is clearly possible for the adversary to determine his/her prescriptions. While there are methods to deal with data analysis under such a scenario up to a point (e.g., [5]), they violate our goal of storing and disclosing actual data values. Full protection against other kind of background knowledge is impossible while still maintaining data utility [5].

3. RELATED WORK

The anatomy [29], fragmentation [4] and slicing [14] models have been proposed to provide a technique that ensures privacy and preserves data granularity lost using generalization - based approaches such as k-anonymity [22, 23], l-diversity [16], (α, k)-anonymity [28], and t-closeness [11].

Unfortunately, these models fail to provide the promised privacy because of the dependencies that might exit in the data. In [14] the authors provide a technique that combines both generalization and bucketization to protect datasets against membership disclosure. Despite its originality, this approach remains vulnerable to negative correlations even while grouping attributes that are highly correlated. In [25] [15], disassociation is applied in a way to preserve both, the original terms to leverage utility and the k^m-anonymity privacy constraint. A privacy breach can still occur due to the lack of diversity. Particularly, when ensuring k^m-anonymity without using generalization which makes the technique vulnerable to homogeneity attacks.

An adversary discovering correlations in the data can use these correlations to discover information about individuals [27] [8]. In [27], the authors consider correlations as foreground knowledge that can be mined from anonymized data. They use the possible worlds model to compute the probability of associating an individual with a sensitive value based on a global distribution. In [8], a Naïve Bayesian model is used to compute association probability. They use exchangeability [2] and DeFinetti's theorem [21] to model and compute patterns from the anonymized data.

There are two components to each of these papers. The first is a relatively simple idea - that we can use correlations to link identifying information to sensitive values. A much deeper aspect is that they show *how* an adversary can find such correlations in the anonymized data. Our work addresses the first component directly: We ensure that even given knowledge of the true correlations present in the data, the probability that a particular sensitive value can be linked to a particular individual is below a threshold (e.g., $1/l$ for l-diversity, or α for (α, k)-anonymity.) This ensures that our method prevents not only the attacks in [27, 8], but any other correlation-based attacks that may be developed. Furthermore, we try to preserve and expose correlations where possible, increasing utility of the data. In [13], the authors deal with background knowledge that can be mined from the data. In their paper, they focus mainly on what is known as negative correlations limiting by that the ability to handle positive and exposed correlations.

[5] defines the notion of differential privacy to handle private data publishing efficiently. The technique gained much popularity among computer scientists providing strong assumptions on the way that data should be released. In essence, differential privacy guarantees privacy without making any assumption on the adversary's background knowledge. More accurately, it shows robustness when a certain number of tuples in the dataset are known by the adversary. Despite its originality, differential privacy tends to be less efficient when correlation among the tuples is high [9]. In addition, the appropriate value of ϵ to achieve the needed

real-world privacy is unclear [10].

While there are approaches that bridge differential privacy and generalization for data release [17, 12], they are not applicable in our environment. For example, [17] releases noisy group sizes; if applied in our model, the server would likely be able to use query history to distinguish true vs. fake tuples and thus reduce this noise, violating ϵ-differential privacy. Alternatively, [12] uses sampling to show that at some point k-anonymization techniques can achieve a relaxation of ϵ-differential privacy with a small error probability δ. This, however, significantly decreases the utility of the data which already suffers from constraints imposed by generalization.

4. FORMALIZATION

We first define basic concepts and notations used in the paper (see also Table 1).

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. Attributes of a table are divided as follows:

- *Quasi-identifiers* A^{qi} represent attributes that can be used (possibly with external information available to the adversary) to identify the individual associated with a tuple in a table. *Name, Gender, Age and Zipcode* are examples of quasi-identifiers.
- *Sensitive attributes* A^s contain sensitive information that must not be linkable to an individual. In our example (Table 1), *DrugName* is considered sensitive and should not be linked to an individual.

Definition 1 (Equivalence class / QI-group). *[22] 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* $1 \leq j_1 \neq j_2 \leq m$, $QI_{j1} \cap QI_{j2} = \phi$.

Table Prescription shown in Figure 1a is composed of 6 different quasi-identifier groups identified by their GID attribute's values.

Definition 2 (l-diversity). *[16] 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 in QI_j*
- *$c_j(v_s)$ is the number of tuples of v_s in QI_j*
- *$|QI_j|$ is the size (number of tuples) of QI_j*

For instance, quasi-identifier group QI_1 in Figure 1a is 3-diverse containing 3 distinct sensitive values.

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 or quasi-identifier 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
A	an attribute of T
A^{qi}	a quasi-identifier attribute of T
A^s	a sensitive attribute of T
QI_j	a quasi-identifier group
T^*	Anonymized version of table T
$\mathcal{CD}$	a set of correlation dependencies
$cd : A^{qi} \dashrightarrow A^s$	a correlation dependency between attribute A^{qi} and the sensitive attribute A^s

Figure 1b is an anatomized version of Table Prescription in Figure 1a in which only the links between individuals and their sensitive values are generalized.

To express correlations between attributes of an anonymized table T^*, we use the term correlation dependencies $\mathcal{CD}$ formally defined as follows:

Definition 4 (Correlation Dependency). *Let A^{qi} be an attribute of T^*, and A^s be the sensitive attribute of T^*. A correlation dependency ($cd^{qi} \in \mathcal{CD}$) of the form of $cd^{qi} : A^{qi} \dashrightarrow A^s \in \mathcal{CD}$ exists over T^* if $\exists v_s \in A^s$ and $v_{qi} \in A^{qi}$ s.t. $P(v_s|v_{qi}) >> P(v_s)$.*

We assume that dealing with correlation dependencies is not a straightforward process in which we can assume that every correlation is unsafe. Such assumption contradicts the basic utility of data outsourcing and causes dramatic damage to the utility of aggregate analysis. It is important to specify to what extent correlation is unsafe and define its legitimate boundaries during the anonymization process. For completeness, we define the significance of a sensitive value v_s w.r.t. a quasi-identifier value v_{qi} based on a confidence and support measures to be discussed below.

Definition 5 (Significant Sensitive Value). *Given a correlation dependency of the form $cd^{qi} : A^{qi} \dashrightarrow A^s$ over a table T, we say that a sensitive value v_s is significantly related to v_{qi} iff*

- *$conf(v_{qi}, v_s) = Pr(A^s = v_s, A^{qi} = v_{qi})/Pr(A^{qi} = v_{qi})$ is less than or equal to minConf threshold ($conf(v_{qi}, v_s) \leq minConf$) or greater than or equal to a maxConf threshold ($conf(v_{qi}, v_s) \geq maxConf$) and,*
- *$sup(v_{qi}, v_s) \geq minSup$ where minSup is defined to capture sensitive values that are frequently correlated with the quasi-identifier values.*

We use *confidence* (*conf*), easily mined from the data during anonymization, to determine the strength of a correlation dependency and limit the number of **significantly** related sensitive values. Specifically, a sensitive value related to a quasi-identifier value by a correlation dependency is significant if its *confidence* is at least equal to a maximum confidence (*maxConf*) threshold or at the most equal to a minimum confidence (*minConf*) threshold, and it has a *support* greater than a minimum support (*minSup*) threshold. *minConf*, *maxConf* and *minSup* are set to satisfy safety requirements as shown in the next section.

5. CORRELATION-BASED PRIVACY VIOLATION

High correlation would allow us to use the values of one attribute to predict the values of other attributes. While this is valuable knowledge, it can also violate the privacy constraints. The problems detailed in [27, 8] lie with the ability of an adversary to extract patterns (correlations) from an anonymized table that can be used to violate privacy. Summarizing, we define here the privacy problem as follows:

Definition 6 (Privacy Problem)**.** *A privacy violation occurs if for a given individual u, $Pr(u_s = v_s|T^*) > 1/l$, where v_s is a sensitive value of A^s, and T^* is an l-diverse anonymized version of T.*

Definition 6 provides a general perspective of the privacy breach but yet we cannot assume that every correlation is unsafe. As mentioned earlier, such an assumption contradicts the basic utility of data outsourcing. For this reason, we consider that for a given an anonymized table T^*, if an adversary is able to associate a *significant* sensitive value v_s to an individual u with a probability greater than $1/l$ based on the assumed adversary knowledge, we say that the privacy principle has been violated.

It is essential to enforce proper safety requirements during the anonymization process to keep *significant* correlations bounded and eliminate by that any possible breach of privacy.

We present in the following our safe correlation safety constraint to bound correlation dependencies of the form $cd^{qi} : A^{qi} \dashrightarrow A^s$.

Safety Constraint (Safe Correlation)**.** *Given a correlation dependency of the form ($cd^{qi} : A^{qi} \dashrightarrow A^s$) over T. Let v_{qi} be a value of quasi-identifier attribute A^{qi} and $v_s \in A^s$ be a sensitive value significantly related to v_{qi}. We say a safe correlation constraint is satisfied for T^* iff*

1. ***significant** sensitive values are uniformly distributed such that $Pr(A^s = v_{s_i}, A^{qi} = v_{qi}|T^*) = 1/\lambda_{v_{qi}}$ for $(1 \leq i \leq |\mathcal{S}(v_{qi})|)$ and,*

2. *there are at least l distinct significant sensitive values for v_{qi}, $|S(v_{qi})| \geq l$*

where

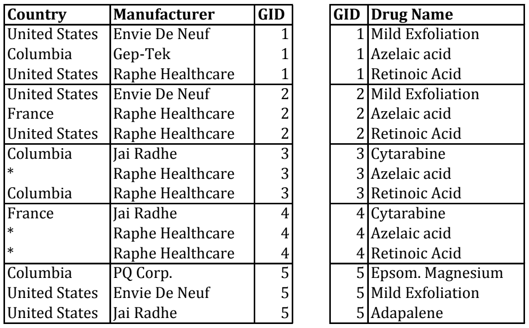
- *$\mathcal{S}(v_{qi})$ is the set of sensitive values **significantly** related to v_{qi} and,*
- *$\lambda_{v_{qi}} \geq l$ is the correlation constant.*

Using this safe correlation requirement we provide boundaries to correlation while making sure that the most frequent correlated value does not appear too frequently, and that the low correlation values do not appear too rarely in T^*. We note that the correlation constant $\lambda_{v_{qi}}$ depends on the actual correlation between a quasi-identifying value v_{qi} and the significant sensitive values. $\lambda_{v_{qi}}$ is determined in a post-anonymization process explained in the next section.

Theorem. *An adversary cannot use his/her previous knowledge of some of the significant correlations to link individuals to sensitive values in the anonymized dataset.*

Proof. Given that $Pr(A^s = v_s, u|T^*)$ can be written as $Pr(A^s = v_s, t_{v_{qi}}|T^*)$ where $t_{v_{qi}}$ is individual u's tuple and $t[A^{qi}] = v_{qi}$. Assuming that v_s is significantly related to v_{qi} meaning that $v_s \in S(v_{qi})$ and thus $Pr(A^s = v_s, A^{qi} = v_{qi}|T^*)$ is equal to $1/\lambda_{qi}$. If a privacy violation occurs as such $Pr(A^s = v_s, t_{v_{qi}}|T^*) > 1/l$, the correlation itself must violate our assumptions. According to the safe correlation constraint, significant correlations between sensitive and quasi-identifying values are bounded by l-diversity. In other terms, there are $l - 1$ other sensitive values such that $Pr(A^s = v_{s_i}, A^{qi} = v_{qi}|T^*) = 1/\lambda_{v_{qi}}$ for $(1 \leq i \leq l - 1)$. □

Figure 2 shows how we can achieve this safety constraint using the correlation sanitization algorithm defined in Section 6. As we can see, several values have been suppressed to make sure that both probabilities remain equal after the anonymization process.



Country	Manufacturer	GID
United States	Envie De Neuf	1
Columbia	Gep-Tek	1
United States	Raphe Healthcare	1
United States	Envie De Neuf	2
France	Raphe Healthcare	2
United States	Raphe Healthcare	2
Columbia	Jai Radhe	3
*	Raphe Healthcare	3
Columbia	Raphe Healthcare	3
France	Jai Radhe	4
*	Raphe Healthcare	4
*	Raphe Healthcare	4
Columbia	PQ Corp.	5
United States	Envie De Neuf	5
United States	Jai Radhe	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	Epsom. Magnesium
5	Mild Exfoliation
5	Adapalene

Figure 2: Safe correlation: a post-anonymization safety constraint.

One subtle remaining issue is multi-dimensional correlations, where several combined attribute values can correlate with a sensitive attribute. Formally, we define a p-dimensional correlation dependency as follows:

Definition 7 (p-Dimensional Correlation Dependency)**.** *Let A^{qi} be quasi-identifying attribute of table T, we say a correlation dependency of the form $cd_p : (A_1^{qi}, ..., A_p^{qi}) \dashrightarrow A^s$ is p-dimensional where A^s is a sensitive attribute of T iff $\exists$ p values $v_1 \in A_1^{qi}, ..., v_p \in A_p^{qi}$ such that for a given $v_s \in A^s$, v_s is significantly related to $(v_{qi_1}, ..., v_{qi_p})$.*

Typically, dealing with p-dimensional correlation dependencies cannot be done while assuming a straightforward extension of the safety constraint. It is essential to consider parameters related to data utility with respect to safety. While this is left for a future work, we assume that safety is guaranteed if and only if any subset of possible attribute combinations of the p-dimensional correlation dependency antecedent is 'safe'.

6. PRIVACY ENFORCEMENT

We now provide the correlation sanitization algorithm, a mechanism to enforce the safe correlation requirement.

6.1 Correlation Sanitization: a Linear Programming Problem

Given an anonymized table T^*, the correlation between a significant sensitive value v_s and a quasi-identifying value v_{qi} can be referred to as $Pr(A^s = v_s, A^{qi} = v_{qi}|T^*)$ and determined as follows:

$$Pr(A^s = v_s, A^{qi} = v_{qi}|T^*) = \frac{\sum_{QI_j \in \mathcal{QI}(v_{qi})} c_j(v_{qi}) \times Pr(A^s = v_s, t_i|QI_j)}{\sum_{QI_j \in \mathcal{QI}(v_{qi})} c_j(v_{qi})} \quad (1)$$

To achieve the safe correlation constraint, we solve the linear programming (LP) problem subject to maximizing the sum of count of QI-values in each QI-group in T^* such that, $\forall v_s \in S(v_{qi})$, $\sum_{j=1}^{m} p_{j,k} x_{i,j} = \frac{1}{\lambda_{v_{qi}}} \times c(v_{qi})$, where

- $x_{i,j}$ is a variable that represents the count of v_{qi} in QI-group QI_j denoted by $c_j(v_{qi_i})$,
- $p_{j,k}$ represents the probability of associating a tuple t_i with the sensitive value v_{s_k} in QI-group QI_j denoted by $Pr(A^s = v_{s_k}, t_i|QI_j)$, and
- $c(v_{qi_i})$ is the total number of tuples with v_{qi_i} in T^*.

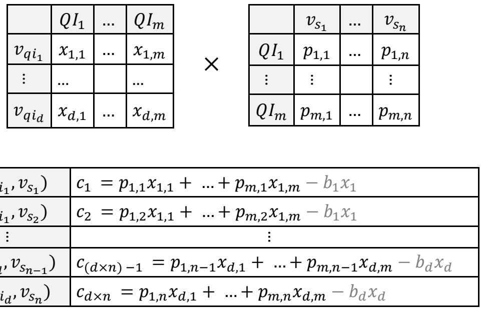
The problem can be viewed as an anonymization problem in which we determine the number of QI-values that should be suppressed in each QI-group in order to guarantee an appropriate correlation constant $(1/\lambda_{v_{qi}})$. To summarize, the linear programming problem can be expressed as follows:

$$\begin{array}{ll} \max & \sum_{i,j} x_{i,j} \\ \text{s.t.} & 0 \leq \sum_j p_{j,k} x_{i,j} - x_i \leq \epsilon, \text{ if } v_{s_k} \in S(v_{qi_i}) \\ & 0 \leq \sum_j p_{j,k} x_{i,j} \leq c(v_{qi_i} : v_{s_k}), \text{ if } v_{s_k} \notin S(v_{qi_i}) \\ & 0 \leq x_{i,j} \leq c_j(v_{qi_i}) \\ & 0 \leq x_i \leq c(v_{qi_i}) \times \frac{1}{l}. \end{array}$$

where,

- x_i is a variable that expresses the correlation constant te be determined during the anonymization process. We note that x_i is equal to $\frac{1}{\lambda_{v_{qi_i}}} \times c(v_{qi_i})$ such that $x_i \leq \frac{1}{l} \times c(v_{qi_i})$. Figure 3 shows the set of constraints of the LP problem including variables x_i.
- ϵ is a user defined error bound.
- $c(v_{qi_i} : v_{s_k})$ is the actual correlation of v_{qi_i} and v_{s_k} determined from the anonymized table T^*

The constraints coefficients matrix is computed based on the set of constraints expressed in Figure 3.



	QI_1	...	QI_m
v_{qi_1}	$x_{1,1}$	...	$x_{1,m}$
⋮	...		...
v_{qi_d}	$x_{d,1}$	...	$x_{d,m}$

$\times$

	v_{s_1}	...	v_{s_n}
QI_1	$p_{1,1}$	...	$p_{1,n}$
⋮	⋮		⋮
QI_m	$p_{m,1}$	...	$p_{m,n}$

(v_{qi_1}, v_{s_1})	$c_1 = p_{1,1}x_{1,1} + \ldots + p_{m,1}x_{1,m} - b_1 x_1$
(v_{qi_1}, v_{s_2})	$c_2 = p_{1,2}x_{1,1} + \ldots + p_{m,2}x_{1,m} - b_1 x_1$
⋮	⋮
$(v_{qi_d}, v_{s_{n-1}})$	$c_{(d\times n)-1} = p_{1,n-1}x_{d,1} + \ldots + p_{m,n-1}x_{d,m} - b_d x_d$
(v_{qi_d}, v_{s_n})	$c_{d\times n} = p_{1,n}x_{d,1} + \ldots + p_{m,n}x_{d,m} - b_d x_d$

Figure 3: Constraints for LP problem formed based on T^*

Algorithm 1 Correlation Sanitization

```
Require: a table $T$, a correlation dependency ($cd^{qi} : A^{qi} \dashrightarrow A^s$), a minimum and maximum confidence thresholds ($minConf$, $maxConf$), a minimum support threshold $minSup$, $l$ the privacy constant and $\epsilon$ the error bound
Ensure: safe correlation for $T^*$
  /**Pre-anonymization: determine significant sensitive values */
  for each distinct $v_{qi}$ in $A^{qi}$ do
      $S(v_{qi})$={$v_s$ | $v_s$ is a sensitive value significantly related to $v_{qi}$ w.r.t $minConf$, $maxConf$ and $minSup$}
      if $|S(v_{qi})| < l$ then
          for each $v_s$ in $S(v_{qi})$ do
              Suppress ($c(v_{qi}, v_s)$) tuples with $t[A^{qi}] = v_{qi}$ and $t[A^s] = v_s$ in $T$
          end for
      end if
  end for
  $T^* = Anonymize(T, l)$
  /**Post-anonymization: formalizing an LP problem */ /** 1 - Determine structural variables $X$ for objective function $z = \sum_{i,j} x_{i,j}$ from $T^*$ */
  $cl = 0$;
  for each distinct $v_{qi_i}$ in $A^{qi}$ do
      for each $QI$ in $T^*$ do
          $X[cl] \leftarrow x_{i,j}$
          Set $0 \leq x_{i,j} \leq c_j(v_{qi_i})$
          $cl = cl + 1$;
      end for
  end for
  for each distinct $v_{qi_i}$ in $A^{qi}$ do
      if $S(v_{qi})$ is not empty then
          $X[cl] \leftarrow x_i$
          Set $0 \leq x_i \leq \frac{1}{l} \times c(v_{qi_i})$
          $cl = cl + 1$;
      end if
  end for
  /** 2- Determine constraints coefficients matrix from $T^*$ */
  $cI = 0, r = 0, C[][] = 0$;
  for each distinct $v_{qi_i}$ in $A^{qi}$ do
      $cI = i * m$;
      for each distinct $v_{s_k}$ in $A^s$ do
          $cl = cI$;
          for each $QI_j$ in $T^*$ do
              $C[r][cl] = p_{j,k}$;
              $cl = cl + 1$;
          end for
          if $v_{s_k} \in S(v_{qi_i})$ then
              $cl = getColFor(v_{qi_i})$;
              $C[r][cl] = -1$;
              $B[r] = \epsilon$;
          else
              $B[r] = getCorrelation(v_{qi_i}, v_{s_k})$;
          end if
          $r = r + 1$;
      end for
  end for
  Solve LP problem {max. $z | CX \leq B$}
  /**Anonymize QI-Values*/
  for each $QI_j$ in $T^*$ do
      Suppress $c_j(v_{qi_i}) - x_{i,j}$ values of $v_{qi_i}$ in $QI$
  end for
```

Now that we have shown how we can guarantee the safe correlation safety constraint, we present our correlation sanitizer algorithm that ensures that the most frequent correlated values do not appear too frequently, and that the less frequent correlated values do not appear too rarely in T^*. The algorithm takes a table T, a quasi-identifier correlation dependency cd^{qi}, minimum and maximum confidence thresholds ($minConf$, $maxConf$), the minimum support threshold $minSup$ and the error bound ϵ. It ensures the safe correlation requirement for T.

The algorithm is composed of two main tasks, pre- anonymization and post-anonymization. In pre-anonymization, from Step 1 to 8, the algorithm retrieves the set of significant sensitive values $S(v_{qi})$ for each distinct value v_{qi} in the quasi-identifier attribute A^{qi}, based on $minConf$, $maxConf$ and $minSup$. Hence, a privacy breach could occur at this level when an adversary is able to determine possible associations with sensitive values based on the size of $S(v_{qi})$. That is why the algorithm from Step 3 to 7 suppresses the tuples related to v_{qi} and v_s if $|\mathcal{S}(v_{qi})|$ is less than l.

In post-anonymization, we ensure that the probability of associating v_{qi} with any of its significantly related sensitive values $v_s \in S(v_{qi})$ is equal to $1/\lambda_{v_{qi}}$ which is achieved by solving the linear programming problem discussed in the previous section. It first retrieves the structural variables from T^* (Step 10 to 24). Each variable $x_{i,j}$ representing the count of v_{qi_i} in QI_j is bounded by $c_j(v_{qi_i})$, variable x_i expressing $\frac{1}{\lambda_{v_{qi_i}}} \times c(v_{qi})$ is determined based on the LP solution. Note that x_i is bounded by $\frac{1}{l} \times c(v_{qi})$.

In the second block of post-anonymization from Step 25 to 43, the algorithm determines the constraints coefficients matrix. In Step 31, we store $p_{j,k}$ corresponding to $Pr(t_i, A^s = v_{s_k}|QI_j)$ and associated with variable $x_{i,j}$ of column cl in the constraint coefficient matrix C. In order to guarantee safe correlation, the algorithm verifies if $v_{s_k} \in S(v_{qi_i})$ where $Pr(A^s = v_{s_k}, A^{qi} = v_{qi_i}|T^*)$ should be equal to $1/\lambda_{v_{qi}}$. In this case, the algorithm stores a -1 coefficient for variable $x_i \leq a$ corresponding to v_{qi_i} for column cl in C and the error bound ϵ in B. On the other hand, if $v_{s_k} \notin S(v_{qi_i})$, the auxiliary variable in this case is bounded by the actual correlation of v_{s_k} and v_{qi_i} as shown in Step 39.

The LP problem is solved in Step 44 such that for each QI-group QI_j, a number of $c_j(v_{qi_i}) - x_{i,j}$ is suppressed from Step 45 to 47.

Framing this as an optimization problem raises concerns of a minimality attack [26]. The safety constraint addresses this: Because of the requirement that all exposed values have equal number, the optimal suppression will always remove the more numerous values. A minimality attack will thus assume that the suppressed values are only the more common values. This would be the (presumably known) correlations; the probability of any given value being suppressed is based on its probability given correlations. In other words, the optimality of the suppression tells us that what we can estimate from the data is exactly what we would expect from just knowing the correlation.

There is still an issue of minimality attacks on the underlying anonymization method. This can be addressed through using a non-deterministic approach in Step 9. This protects against minimality attacks, as described in [3].

Let $|A^{qi}|, |A^s|$ be the number of distinct quasi-identifying and sensitive values in attributes A^{qi} and A^s respectively, the time complexity of the sanitization algorithm can be estimated by $O(m \cdot |A^{qi}| \cdot |A^s|)$ where m is the number of QI-groups in T^*.

In addition, based on the linear programming problem defined in 6.1, we can say that the sanitization algorithm scales. In fact, $\forall i, j$, if $x_{i,j}$ and x_i are equal to zero, we can easily verify that all constraints are satisfied.

7. EXPERIMENTAL EVALUATION

We now present a set of experiments to evaluate the efficacy of our approach. We implemented the correlation sanitization code in Java based on the Anonymization Toolbox [7], running on an Intel XEON 2.4GHz PC with 2GB RAM.

7.1 Evaluation Dataset

In keeping with much work on anonymization, we use the Adult Dataset from UCI Machine Learning Repository [6]. We treat *Occupation* as a sensitive attribute; other attributes are presumed to be (quasi- or actual-) identifiers.

We used cd^{qi} : *Education* ⇢ *Occupation* as a correlation dependency for the adult dataset containing 32561 tuples. We note that using such correlation dependency, an adversary is able to identify the occupation of an individual in the dataset according to education.[1]

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

7.2 Evaluation Results

We conducted a set of measurements to evaluate the efficiency of our correlation sanitization algorithm. These measurements can be summarized as follows:

- Evaluating the correlation threat after anonymization,
- 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 safety constraints, and
- Comparing anonymization cost in two different datasets w.r.t several minimum and maximum confidence values (minConf and maxConf),

7.2.1 *Correlation Evaluation*

We evaluate here the remaining correlation in the dataset after a naïve anonymization using the correlation sanitization algorithm. In fact, we compare the outcome of anonymization techniques, more precisely anatomy and correlation sanitization, using a java-based implementation of Wong's approach [27]. We use in this test $l = 3, 4$ and 5 for several significant sensitive values as shown in Figure 4.

We note that in order to calculate $Pr(A^s = v_s, t_i|QI_j)$ defined in the correlation sanitization algorithm, we used the possible world model with actual correlations as shown in the example of Section 5 for the following significant sensitive values; *Handlers_cleaner*, *Craft_repair*, *Exec_managerial* and *Adm_clerical*.

As expected, the correlation sanitization algorithm bounds the correlations with confidence greater than 0.9 and lower than 0.1 while others eventually remain representing the y-axis in the Figures 4b, 4c and 4d expressing residual non-violating correlations related to non-significant sensitive values that could not be exposed.

7.2.2 *Anonymization Cost Evaluation*

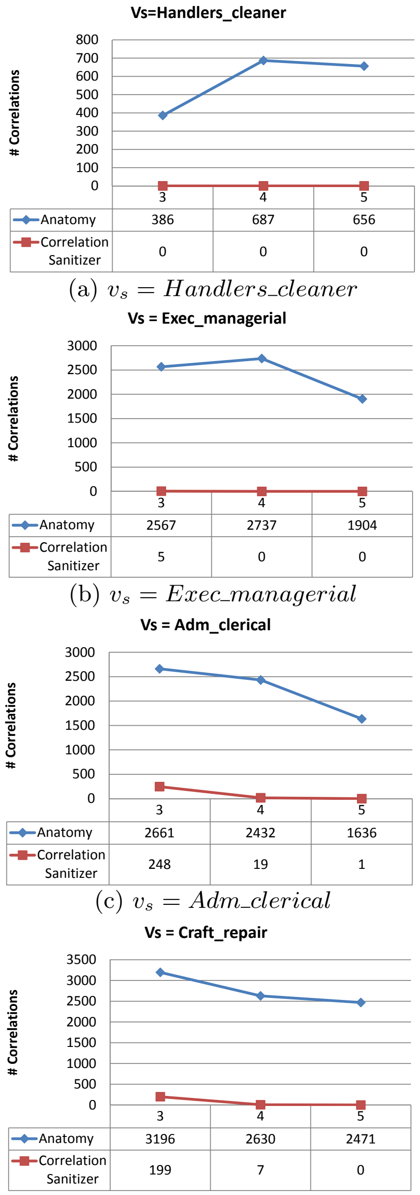
We evaluate our proposed correlation sanitization algorithm to determine the number of tuples and values that are suppressed to achieve the safety constraint. We use the following loss metric to quantify such loss of data fidelity.

[1] We invite the reader to check out [27] for more details on how to compute the global distribution and the privacy breach value for each attribute value.

Definition 8 (Loss Metric ($\mathcal{LM}$)). *Let $g(T^*, v)$ be a function that returns the number of tuples where the value v is suppressed in the anonymization T^* of T. The loss metric ($\mathcal{LM}$) for table T^* and value v is*

$$\mathcal{LM}(T, v) = \frac{g(T^*, v)}{|T|} \tag{2}$$

Figure 2 shows an anonymized version of table prescription where the grouping is safe. The loss metric for this anonymization has a loss metric equal to $\mathcal{LM}(Prescription, UnitedStates) = 1/3$.



(a) $v_s = Handlers\_cleaner$

(b) $v_s = Exec\_managerial$

(c) $v_s = Adm\_clerical$

(d) $v_s = Craft\_repair$

Figure 4: Correlation Sanitization for $l = 3, 4$ and 5

We applied the algorithm on table T to ensure safe correlation for values for cd^{qi} : $Education \dashrightarrow Occupation$ with $minConf$, $maxConf$ and $minSup$ equal to 0.1, 0.9 and 0.2.

Results in Figure 5 show explicitly the trade-off between privacy and utility such that for the sensitive value $Craft\_repair$ and $l = 5$, $\mathcal{LM}$ reaches 56%. At some point, we can see that the result can be dwarfed by the loss of utility. We have not identified any inherent reason why this must hold. Further research into more effective anonymization algorithms may produce techniques that meet privacy requirements while increasing the ability to learn from the data.

7.2.3 *Cost Evaluation in Different Datasets*

We also compared the anonymization costs computed when applying the correlation sanitization algorithm to the Adult dataset and the Bank Marketing Dataset used in [19]. In the latter, we treat *Balance* as a sensitive attribute while the remaining attributes are presumed to be (quasi- or actual-) identifiers. For computational reasons, we generalize the values of attribute *Balance* to 21 intervals to reduce the total number of distinct sensitive values. The results are shown in Figure 6 for $l = 2, 3$ and 4 with 5 different values for minConf and maxConf respectively represented in the X-axis.

Not surprisingly, the results are similar for both datasets showing that the cost increases when anonymizing the correlations. This is only to confirm as in [12] that there is a trade-off to be made at the stake of utility in order to meet strong privacy requirements. While this could be limiting to generalization techniques, it remains debatable in our approach where exact data values are maintained[2].

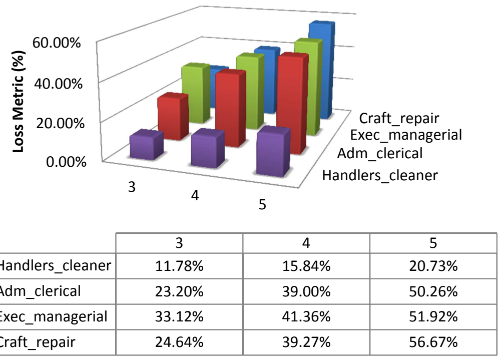


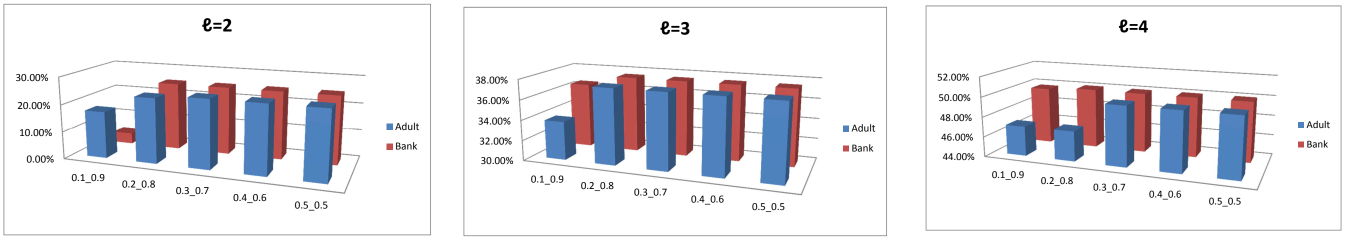
Figure 5: Evaluating loss for Correlation Sanitization

8. CONCLUSION

In this paper, we presented new methods to cope with defects of anonymization techniques resulting from unsafe data correlation. We defined a new safety constraint to deal with correlation between quasi-identifier and sensitive attributes. We provided a sanitization algorithm to ensure the safe correlation in a post-anonymization process. Finally, 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 and values to be anonymized using anonymization algorithms.

A related problem is coping with correlations in transactional datasets where multiple tuples could be related to an

[2] Note that while suppression prevents privacy violations, it does not necessarily prevent discovery of correlations. Preliminary results on a decision tree learning approach customized to anatomized data show comparable classification accuracy to decision trees learned on the original data.



(a) Bank vs. Adult dataset with l=2

(b) Bank vs. Adult dataset with l=3

(c) Bank vs. Adult dataset with l=4

Figure 6: Correlation Sanitization for $l = 2, 3$ and 4

individual [1]. Under such assumption, a straightforward extension of safety constraint could not be achieved leading eventually to more sophisticated privacy violation detection and elimination methods. Achieving sufficient utility in such environments may also need to consider alternative privacy models such as LKC-privacy [18] or (k, m)-anonymity [24].

9. ACKNOWLEDGEMENTS

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

10. REFERENCES

[1] B. al Bouna, C. Clifton, and Q. M. Malluhi. Using safety constraint for transactional dataset anonymization. In *DBSec*, pages 164–178, 2013.

[2] D. 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.

[3] G. Cormode, N. Li, T. Li, and D. Srivastava. Minimizing minimality and maximizing utility: Analyzing method-based attacks on anonymized data. In *Proceedings of the VLDB Endowment*, volume 3, pages 1045–1056, 2010.

[4] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, G. Livraga, S. Paraboschi, and P. Samarati. Extending loose associations to multiple fragments. In *Proceedings of the 27th International Conference on Data and Applications Security and Privacy XXVII*, DBSec'13, pages 1–16, Berlin, Heidelberg, 2013. Springer-Verlag.

[5] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[6] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[7] M. Kantarcioglu, A. Inan, and M. Kuzu. Anonymization toolbox, 2010.

[8] D. Kifer. Attacks on privacy and definetti's theorem. In *SIGMOD Conference*, pages 127–138, 2009.

[9] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *SIGMOD Conference*, pages 193–204, 2011.

[10] J. Lee and C. Clifton. Differential identifiability. In *The 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1041–1049, Beijing, China, Aug. 12-16 2012.

[11] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, pages 106–115, 2007.

[12] N. Li, W. Qardaji, and D. Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, ASIACCS '12, pages 32–33, New York, NY, USA, 2012. ACM.

[13] T. Li and N. Li. Injector: Mining background knowledge for data anonymization. In *ICDE*, pages 446–455, 2008.

[14] T. Li, N. Li, J. Zhang, and I. Molloy. Slicing: A new approach for privacy preserving data publishing. *IEEE Trans. Knowl. Data Eng.*, 24(3):561–574, 2012.

[15] G. Loukides, J. Liagouris, A. Gkoulalas-Divanis, and M. Terrovitis. Disassociation for electronic health record privacy. *Journal of Biomedical Informatics*, 50(0):46 – 61, 2014. Special Issue on Informatics Methods in Medical Privacy.

[16] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACMTransactions on Knowledge Discovery from Data (TKDD)*, 1(1), Mar. 2007.

[17] N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu. Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 493–501, New York, NY, USA, 2011. ACM.

[18] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. kwong Lee. Anonymizing healthcare data: a case study on the blood transfusion service. In *KDD*, pages 1285–1294, 2009.

[19] S. Moro, R. Laureano, and P. Cortez. Using data mining for bank direct marketing: An application of the crisp-dm methodology. In P. N. et al., editor, *Proceedings of the European Simulation and Modelling Conference - ESM'2011*, pages 117–121, Guimaraes, Portugal, Oct. 2011. EUROSIS.

[20] A. E. Nergiz and C. Clifton. Query processing in private data outsourcing using anonymization. In *The 25th IFIP WG 11.3 Conference on Data and Applications Security and Privacy (DBSEC-11)*, Richmond, Virginia, July 11-13 2011.

[21] P. Ressel. De Finetti-type theorems: an analytical approach. *Ann. Probab.*, 13(3):898–922, 1985.

[22] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.*, 13(6):1010–1027, 2001.

[23] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[24] M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy-preserving anonymization of set-valued data. *Proceedings of the VLDB Endowment*, 1(1):115–125, Aug. 2008.

[25] M. Terrovitis, N. Mamoulis, J. Liagouris, and S. Skiadopoulos. Privacy preservation by disassociation. *Proc. VLDB Endow.*, 5(10):944–955, June 2012.

[26] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *VLDB*, pages 543–554, 2007.

[27] R. C.-W. Wong, A. W.-C. Fu, K. Wang, P. S. Yu, and J. Pei. Can the utility of anonymized data be used for privacy breaches? *ACM Trans. Knowl. Discov. Data*, 5(3):16:1–16:24, Aug. 2011.

[28] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang. (alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *KDD*, pages 754–759, 2006.

[29] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *Proceedings of 32nd International Conference on Very Large Data Bases (VLDB 2006)*, Seoul, Korea, Sept. 12-15 2006.