Correlation Hiding by Independence Masking

Yufei Tao[1], Jian Pei[2], Jiexing Li[1], Xiaokui Xiao[3], Ke Yi[4], Zhengzheng Xing[2]

[1]Chinese University of Hong Kong [2]Simon Fraser University
[3]Nanyang Technological University [4]Hong Kong University of Science and Technology

***Abstract*— Extracting useful correlation from a dataset has been extensively studied. In this paper, we deal with the opposite, namely, a problem we call *correlation hiding* (CH), which is fundamental in numerous applications that need to disseminate data containing sensitive information. In this problem, we are given a relational table $\mathcal{T}$ whose attributes can be classified into three disjoint sets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. The objective is to distort some values in $\mathcal{T}$ so that $\mathcal{A}$ becomes independent from $\mathcal{B}$, and yet, their correlation with $\mathcal{C}$ is preserved as much as possible. CH is different from all the problems studied previously in the area of data privacy, in that CH demands *complete* elimination of the correlation between two sets of attributes, whereas the previous research focuses on *partial* elimination up to a certain level. A new operator called *independence masking* is proposed to solve the CH problem. Implementations of the operator with good worst case guarantees are described in the full version of this short note.**

I. Introduction

Correlation elimination is fundamental in applications that need to disseminate data containing sensitive information. Assume, for example, that the census bureau has produced a table $\mathcal{T}$ with attributes $\mathcal{A} = \{Race\}$, $\mathcal{B} = \{Income\}$, and $\mathcal{C} = \{Investment\text{-}expense, Food\text{-}expense, Entertainment\text{-}expense\}$, which needs to be put online to allow the public to study the spending behavior of various ethnic groups and income classes. Doing so, however, will also reveal the correlation between *Race* and *Income*, which should be avoided, because such correlation may lead to sensitive debates such as how much wealthier an ethnic group is than another. In other words, the government would like to release $\mathcal{T}$ in such a way that hides the correlation between *Race* and *Income*, and yet preserves all the other correlations.

The problem cannot be settled by giving away two tables: (i) T_1 that has only $\mathcal{A}$ and $\mathcal{C}$, and (ii) T_2 with only $\mathcal{B}$ and $\mathcal{C}$. This is because their equi-join $T_1 \bowtie T_2$ may restore a significant portion of the original $\mathcal{T}$. The worst case is that no two tuples in $\mathcal{T}$ have the same values on $\mathcal{C}$, allowing the equi-join to rebuild $\mathcal{T}$ precisely.

A good solution in the above scenario should fulfill two requirements. First, it must totally destroy the dependence between *Race* and *Income*. That is, the *Income* distribution of any specific race, such as Caucasian, should look exactly the same as that of any other race, such as Asian. Second, it needs to do so by distorting as little information in T as possible. Otherwise, the resulting table would not allow researchers to perform meaningful data mining, thus defeating the objective of publication.

This paper deals with a general version of the problem described earlier. Assume that we are given a table $\mathcal{T}$ whose attributes have been classified into three disjoint subsets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. The goal is to compute another table $\mathcal{T}^*$ where

- the set $\mathcal{A}$ of attributes is independent from the set $\mathcal{B}$ of attributes;
- a large number of values of $\mathcal{T}$ are retained.

We refer to the above problem as *correlation hiding* (CH), to which we are not aware of any adequate solution. It is opposite to the classic topic of "correlation extraction" in data mining. The closest existing works are found in the areas of *privacy preserving data publication* (PPDP), and *association rule hiding* (ASH). However, as elaborated in Section II, CH is fundamentally different from both PPDP and ASH, in that CH demands *complete* elimination of some designated correlation, whereas PPDP and ASH focus on *partial* elimination up to a certain level. Such a difference renders the solutions of PPDP and ASH inapplicable to CH.

In this short note, we propose an operator called *independence masking* (I-masking) to carry out correlation hiding. The operator works by masking some values in $\mathcal{T}$ in order to make the sets of attributes $\mathcal{A}$ and $\mathcal{B}$ appear mutually independent. The goal is to minimize the number of values masked. In the full version of this short paper, we show that the problem is NP-hard, and describe several fast approximate algorithms with good approximation guarantees. The full paper also contains experiments that verify the practical effectiveness of I-masking in the context of association rule mining.

II. Related Work

In this section, we review two problems that are similar to the problem of correlation hiding (CH) studied in our paper, and explain why the existing solutions are not applicable to CH.

Privacy preserving data publication (PPDP). In recent years, PPDP has received considerable attention from the database and data mining communities (see the recent work [1] and the references therein). Given a table $\mathcal{T}$ containing sensitive information, PPDP aims at computing an anonymized version $\mathcal{T}^*$ that satisfies a privacy constraint, such as k*-anonymity* [3], [4] or ℓ*-diversity* [2], pre-determined by the administrator. PPDP is similar to CH in the sense that both of them need to distort the correlation between two sets of attributes $\mathcal{A}$ and $\mathcal{B}$ in the original table $\mathcal{T}$.

The key difference between PPDP and CH is the *degree* of distortion. Specifically, the goal of PPDP is to distort *just* to the level required by the underlying privacy constraint. Any

additional distortion should be avoided because, in PPDP, the remaining (un-distorted) correlation of $\mathcal{A}$ and $\mathcal{B}$ is meant to be released, as it is the target of research studies. In contrast, CH aims at *complete* removal of the correlation between $\mathcal{A}$ and $\mathcal{B}$, namely, they must appear totally independent from each other in the published $\mathcal{T}^*$.

In fact, CH even differs from PPDP in their fundamental rationales. Specifically, the motivation of PPDP is to conceal the details of a distribution, but preserve its "big picture". The reason is that statistical analysis does not require fine details anyway; hence, even with some details removed, the resulting dataset can still be useful for statistical studies. In contrast, the distribution targeted by CH must be fully destroyed, and is not meant to be analyzed at all.

Association rule hiding (ASH). Let S be a set of transactions, each of which is a set of items from a discrete domain such as products sold by a supermarket. Denote by R the set of association rules that can be mined from S. Assume that there is a subset R' of R that is sensitive, and should not be known by the public. The objective of ASH [5] is to modify S to another dataset S^* such that the (insensitive) association rules in $R - R'$ can still be discovered from S^*, but those in R' cannot.

Unlike ASH that deals with transactions, CH is concerned with relational tables. Even if one regards a relational table as a set of transactions (by treating each cell as an item, and each tuple an item set), ASH solutions still cannot be applied to CH due to two reasons. First, ASH demands two parameters, namely *support* and *confidence*, to define association rules. It is unclear how these parameters would fit in CH. Second, similar to PPDP, an ASH algorithm does not fully eliminate the data correlation, because some (insensitive) association rules must remain discoverable. CH, as mentioned before, demands complete removal of the correlation between two sets of attributes. In fact, it appears rather difficult to even cast CH as an ASH problem. The reason is that association rules are essentially a *specific* type of correlations, whereas the objective of CH is to eliminate any correlation in general.

III. Problem Definition

In this section, we will formalize the problem of *correlation hiding* (CH). Let $\mathcal{T}$ be a relational table whose attributes have been classified into three disjoint subsets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. The correlation between attribute sets $\mathcal{A}$ and $\mathcal{B}$ is sensitive, and must be fully concealed. $\mathcal{C}$, on the other hand, involves all the remaining attributes of $\mathcal{T}$ that do not belong to $\mathcal{A}$ and $\mathcal{B}$. CH aims at converting $\mathcal{T}$ to another dataset $\mathcal{T}^*$ such that

I. $\mathcal{A}$ and $\mathcal{B}$ are independent in $\mathcal{T}^*$, and

II. $\mathcal{T}^*$ preserves the other correlations in $\mathcal{T}$ as much as possible.

We refer to $\mathcal{T}$ as the *source table*, and $\mathcal{T}^*$ as the *sanitized table*.

Note that the two requirements are not equally important: requirement I has a higher priority. Specifically, if $\mathcal{T}^*$ has not destroyed the correlation of $\mathcal{A}$ and $\mathcal{B}$, it is always a poor solution no matter how well it retains the other correlations in $\mathcal{T}$. This is due to the protective nature of the applications modeled by CH. For instance, consider the motivating example in Section I where $\mathcal{A}$ = {*Race*}, $\mathcal{B}$ = {*Income*}, and $\mathcal{C}$ = {*Investment-expense*, *Food-expense*, *Entertainment-expense*}. As long as the government is not convinced that the correlation between *Race* and *Income* has disappeared, it would prefer to keep the data away from public access.

Requirement I can be described in a more rigorous manner as follows. Let a be any value in the domain of $\mathcal{T}^*.\mathcal{A}$, and similarly, b be any value in the domain of $\mathcal{T}^*.\mathcal{B}$. Note that in case $\mathcal{A}$ ($\mathcal{B}$) has multiple attributes, a (b) is a multidimensional vector. Denote by $Pr[a]$ (or $Pr[b]$) the percentage of tuples in $\mathcal{T}^*$ whose values of $\mathcal{A}$ ($\mathcal{B}$) are equal to a (b). Likewise, denote by $Pr[a, b]$ as the percentage of tuples in $\mathcal{T}^*$ carrying both a and b simultaneously. Then, the independence in requirement I can be specified as

$$Pr[a, b] = Pr[a] \cdot Pr[b]. \tag{1}$$

As an immediate corollary, one can easily verify that if the above equation holds, then any subset of $\mathcal{A}$ is also independent from any subset of $\mathcal{B}$.

The "other correlations" in requirement II can be specialized into 5 concrete types:

1. *Correlation* $\mathcal{A}$: the correlation among the attributes in $\mathcal{A}$.
2. *Correlation* $\mathcal{B}$: the correlation among the attributes in $\mathcal{B}$.
3. *Correlation* $\mathcal{C}$: the correlation among the attributes in $\mathcal{C}$.
4. *Correlation* $\mathcal{AC}$: the correlation between the attributes of $\mathcal{A}$ and those of $\mathcal{C}$.
5. *Correlation* $\mathcal{BC}$: the correlation between the attributes of $\mathcal{B}$ and those of $\mathcal{C}$.

Apparently, it is straightforward to preserve correlation $\mathcal{C}$ because the attributes in $\mathcal{C}$ do not need to be touched in correlation hiding at all. Among the other types of correlations, correlations $\mathcal{AC}$ and $\mathcal{BC}$ are of higher importance. This is because in practice $\mathcal{C}$ is typically a set of "measures", such that the goal of scientific studies is to find out how those measures are influenced by the attributes in $\mathcal{A}$ and $\mathcal{B}$, respectively.

CH is a general problem that may be attacked using different methodologies. This is analogous to the opposite problem of "correlation extraction", which can be performed by clustering, association rule mining, classification with decision trees, etc. In the next section, we describe a feasible methodology to perform CH.

IV. Independence Masking

We present an operator called *independence masking* (I-masking) to carry out CH. The operator takes an integer parameter u, whose effect will be explained later. Denote by n the cardinality of the source table $\mathcal{T}$. In the sequel, we will assume that n is a multiple of u. In case it is not, we randomly remove at most $u - 1$ tuples from $\mathcal{T}$ to make the property hold. As will be clear later, u is typically by far smaller than

tuple ID	Age	Occupation	Income
1	30~50	CEO	25k
2	30~50	Salesman	4k
3	30~50	Prof	10k
4	> 50	Prof	20k
5	30~50	Prof	11k
6	< 30	Prof	6k
7	< 30	Manager	18k
8	< 30	Manager	13k
9	< 30	Manager	7k
10	30~50	Prof	15k
11	30~50	Prof	9k
12	30~50	Prof	6k

(a)

tuple ID	Age	Occupation	Income
1	30~50	CEO	3
2	30~50	Salesman	1
3	30~50	Prof	2
4	> 50	Prof	3
5	30~50	Prof	2
6	< 30	Prof	1
7	< 30	Manager	3
8	< 30	Manager	2
9	< 30	Manager	1
10	30~50	Prof	3
11	30~50	Prof	2
12	30~50	Prof	1

(b)

tuple ID	Age	Occupation	Income
1	30~50	*	3
2	30~50	*	1
3	30~50	*	2
4	*	Prof	3
5	*	Prof	2
6	*	Prof	1
7	< 30	Manager	3
8	< 30	Manager	2
9	< 30	Manager	1
10	30~50	Prof	3
11	30~50	Prof	2
12	30~50	Prof	1

(c)

Fig. 1. Example 1 ($\mathcal{B}$ has only one attribute). (a) shows the source table $\mathcal{T}$. (b) shows the intermediate table $\mathcal{T}^{\triangle}$ after replacing each *Income* value with a cluster label. (c) shows the final sanitized table $\mathcal{T}^*$.

n; hence, removing at most $u-1$ tuples will not lose much information.

To facilitate discussion, let us assume, without loss of generality, that $\mathcal{A}$ has d attributes A_1, A_2, ..., A_d. Denote by $dom(A_i)$ the domain of A_i. I-masking outputs a sanitized table $\mathcal{T}^*$ that has as many tuples as $\mathcal{T}$. Furthermore, the attributes of $\mathcal{T}^*$ can also be classified into disjoint subsets $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ such that

- $\mathcal{A}$ has d attributes corresponding to those in T. Specifically, the domain of each attribute in $\mathcal{A}$ augments the domain of the corresponding attribute of $\mathcal{T}$ by a special symbol '$*$'. Formally, $dom(\mathcal{T}^*.A_i) = dom(\mathcal{T}.A_i) \cup \{*\}$ for all $1 \leq i \leq d$.
- $\mathcal{B}$ has a *single* attribute whose domain is the set of integers from 1 to u. It relates to the set $\mathcal{B}$ of attributes in $\mathcal{T}$ such that each integer of $\mathcal{T}^*.\mathcal{B}$ represents a cluster of $\mathcal{T}$ in its subspace $\mathcal{B}$.
- $\mathcal{C}$ is exactly the same as the $\mathcal{C}$ of $\mathcal{T}$. Namely, I-masking touches only $\mathcal{A}$ and $\mathcal{B}$, and leaves $\mathcal{C}$ intact.

More precisely, I-masking has three steps detailed as follows.

1. Partition $\mathcal{T}$ based on its attribute set $\mathcal{B}$ into u clusters such that every cluster has an equal number of tuples. Label the clusters as 1, 2, ..., u, respectively.

   Once this is done, the values of $\mathcal{T}.\mathcal{B}$ are no longer needed: we will be concerned only with the clusters. Hence, let us replace the $\mathcal{B}$ values of each tuple with the label of the cluster it belongs to. This creates an *intermediate table* $\mathcal{T}^{\triangle}$, whose $\mathcal{A}$ and $\mathcal{C}$ are the same as $\mathcal{T}$, but its $\mathcal{B}$ has an integer domain of $[1, u]$.
2. Mask some $\mathcal{A}$ values of $\mathcal{T}^{\triangle}$ as '$*$' so that $\mathcal{A}$ and $\mathcal{B}$ are independent from each other. Let the resulting table be $\mathcal{T}^*$.
3. Return $\mathcal{T}^*$, as well as the u clusters of $\mathcal{B}$ values obtained earlier in the first step.

Next we illustrate the above steps with two examples. Since I-masking does not alter the attributes in $\mathcal{C}$, we will omit them in our examples but their presence should be implicitly understood.

Example 1. Consider that $\mathcal{T}$ is the table in Figure 1a, where $\mathcal{A}$ = {*Age*, *Occupation*} and $\mathcal{B}$ = {*Income*}. Note that "30∼50" is a raw value of *Age* (as opposed to a "generalized" value as one would find in k-anonymization [3], [4]). Assume u equals 3. Step 1 of I-masking creates u equal-sized clusters out of the values in $\mathcal{T}.\mathcal{B}$. Here, $\mathcal{B}$ has a single numeric attribute. So the clustering results in three intervals: [4k, 7k], [9k, 13k], and [15k, 25k], each of which covers the *Income* of exactly 4 tuples. Label these intervals (i.e., clusters) as 1, 2, 3, respectively. Replacing each *Income* value with the label of the interval it falls in gives the intermediate table $\mathcal{T}^{\triangle}$ in Figure 1b.

Step 2 of I-masking hides some values of $\mathcal{A}$ with the symbol '$*$', until $\mathcal{A}$ and $\mathcal{B}$ have become independent. The resulting table $\mathcal{T}^*$ is shown in Figure 1c, where 6 '$*$' are used. To see the independence of $\mathcal{A}$ and $\mathcal{B}$, notice that Equation 1 holds for any values a and b in the domains of $\mathcal{A}$ and $\mathcal{B}$ in Figure 1c, respectively. For example, let a = (*, Prof) and b = 1. Then, we have $Pr(a, b) = 1/12$, which is indeed the product of $Pr(a) = 3/12$ and $Pr(b) = 4/12$.

Finally, I-masking returns the table in Figure 1c, together with the *Income* values in clusters, namely, cluster 1 = {4k, 6k, 6k, 7k}, cluster 2 = {9k, 10k, 11k, 13k}, cluster 3 = {15k, 13k, 20k, 25k}. □

In Example 1, $\mathcal{B}$ involves only a single attribute. Next, we will see another example where $\mathcal{B}$ has multiple attributes.

Example 2. Let $\mathcal{T}$ be the table in Figure 2a, where $\mathcal{A}$ = {*Race*} and $\mathcal{B}$ = {*Income*, *Saving*}. Assume u is 2. For clarity, we denote the $\mathcal{B}$ values of each tuple as a 2D point in Figure 2b. As before, Step 1 of I-masking clusters these points into u clusters, as illustrated in Figure 2b. Figure 2c shows the intermediate table $\mathcal{T}^{\triangle}$ after replacing the $\mathcal{B}$ values with cluster labels. Step 2 of I-masking generates the sanitized

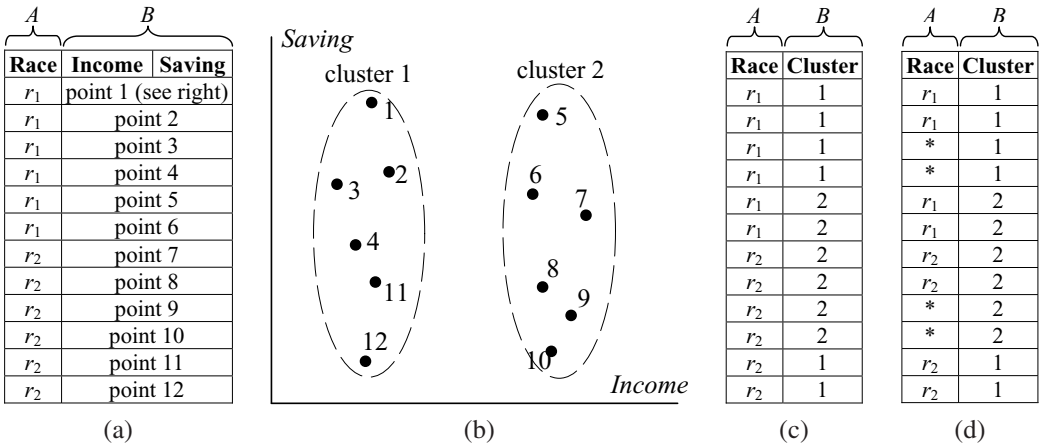

Fig. 2. Example 2 ($\mathcal{B}$ has multiple attributes). (a) shows the source table $\mathcal{T}$. (b) illustrates the clustering on $\mathcal{T}.\mathcal{B}$. (c) gives the intermediate table $\mathcal{T}^{\triangle}$, and (d) the sanitized table $\mathcal{T}^*$.

table $\mathcal{T}^*$ as in Figure 2d, where the two attributes have become independent. This $\mathcal{T}^*$ is returned, together with the points in Figure 2b (i.e., the $\mathcal{B}$ values of $\mathcal{T}$) in their respective clusters. □

Although in the previous examples $\mathcal{B}$ contains only numeric attributes, I-masking also works even if some or all the attributes in $\mathcal{B}$ are categorical. The only thing we require on $\mathcal{B}$ is the ability to cluster its values. This can be achieved by many algorithms, such as k-means (with straightforward adaptation to ensure all clusters are equally large), that perform clustering in the *metric space*. These algorithms are applicable as long as one can supply a distance function to calculate the similarity of two objects, which is easy to formulate in most applications.

Quality of correlation preserving. It is clear that I-masking completely retains Correlations $\mathcal{B}$ and $\mathcal{C}$ (see the classification of correlations in Section III), because the values of those attributes are returned directly. The introduction of '$*$', apparently, loses a part of Correlations $\mathcal{A}$ and $\mathcal{AC}$. Furthermore, the choice of u also affects Correlation $\mathcal{BC}$. To see this, recall that, in $\mathcal{T}^*$, I-masking essentially presents the projection of $\mathcal{T}$ on $\mathcal{B}$ in the form of u clusters; hence, a larger u captures Correlation $\mathcal{BC}$ better.

It would be nice if we could put an upper bound on how much of Correlations $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{AC}$ is lost. This, unfortunately, is not possible because CH has a compulsory goal – eliminating the correlation between $\mathcal{A}$ and $\mathcal{B}$ (see requirement I in Section III). In the worst case $\mathcal{A}$ and $\mathcal{B}$ can be extremely related such that removing their correlation would necessitate masking all values. It thus follows that I-masking is a best-effort process. Namely, we should reduce the information loss as much as possible, on condition that requirement I is satisfied. Put differently, given a specific value of u, we would like to generate $\mathcal{T}^*$ using the smallest number of '$*$'. Following this rationale, I-masking can be cast as the following optimization problem.

PROBLEM 1: Let $\mathcal{T}^{\triangle}$ be a table with $d+1$ attributes. Among these attributes, there is one, denoted as $\mathcal{B}$, whose domain consists of integers from 1 to u. Denote by $\mathcal{A}$ as the set of all other d attributes in $\mathcal{T}^{\triangle}$.

$\mathcal{T}^{\triangle}$ has the property that exactly $1/u$ of its tuples carry 1, 2, ..., u as their $\mathcal{B}$ values, respectively. Let $\mathcal{T}^*$ be a table that is identical to $\mathcal{T}^{\triangle}$ except that some values in the attributes of $\mathcal{A}$ have been masked by '$*$'. $\mathcal{T}^*$ is said to be *independent* if its $\mathcal{A}$ and $\mathcal{B}$ are independent from each other. The goal is to find an independent $\mathcal{T}^*$ containing the smallest number of stars. □

V. WHAT IS IN THE FULL PAPER

The full paper contains several formal results on Problem 1. Specifically, the paper first shows that the problem is NP-hard, and then, gives two fast approximate algorithms with approximation ratios $u-1$ and d, respectively. Furthermore, the full version also includes experimentation to demonstrate the effectiveness of I-masking in association rule mining.

ACKNOWLEDGEMENTS

This work is supported by Grants GRF 4161/07 and GRF 4173/08 from HKRGC.

REFERENCES

[1] Y. Bu, A. W.-C. Fu, R. C.-W. Wong, L. Chen, and J. Li. Privacy preserving serial data publishing by role composition. *Proc. of the VLDB Endowment (PVLDB)*, 1(1):845–856, 2008.

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. In *Proc. of International Conference on Data Engineering (ICDE)*, page 24, 2006.

[3] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 13(6):1010–1027, 2001.

[4] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):571–588, 2002.

[5] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni. Association rule hiding. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 16(4):434–447, 2004.