TRANSACTIONS ON DATA PRIVACY 2 (2009) 185–205

Privacy Preserving Categorical Data Analysis with Unknown Distortion Parameters

Ling Guo*, **Xintao Wu***

*Software and Information Systems Department, University of North Carolina at Charlotte, Charlotte, NC 28223, USA. E-mail: {lguo2,xwu}@uncc.edu

Abstract. Randomized Response techniques have been investigated in privacy preserving categorical data analysis. However, the released distortion parameters can be exploited by attackers to breach privacy. In this paper, we investigate whether data mining or statistical analysis tasks can still be conducted on randomized data when distortion parameters are not disclosed to data miners. We first examine how various objective association measures between two variables may be affected by randomization. We then extend to multiple variables by examining the feasibility of hierarchical loglinear modeling. Finally we show some classic data mining tasks that cannot be applied on the randomized data directly.

1 Introduction

Privacy is becoming an increasingly important issue in many data mining applications. A considerable amount of work on randomization based privacy preserving data mining (for numerical data [1, 3, 23, 24], categorical data [4, 22], market basket data [19, 31], and linked data [21, 27, 37]) has been investigated recently.

Randomization still runs certain risk of disclosures. Attackers may exploit the released distortion parameters to calculate the posterior probabilities of the original value based on the distorted data. It is considered to be jeopardizing with respect to the original value if the posterior probabilities are significantly greater than the a-priori probabilities. In this paper, we consider the scenario where the distortion parameters are not released in order to prevent attackers from exploiting those distortion parameters to recover individual data.

In the first part of our paper, we investigate how various objective measures used for association analysis between two variables may be affected by randomization. We demonstrate that some measures (e.g., Correlation, Mutual Information, Likelihood Ratio, Pearson Statistics) have a vertical monotonic property , i.e., the values calculated directly from the randomized data are always less than or equal to those original ones. Hence, some data analysis tasks (e.g., independence testing) can be executed on the randomized data directly even without knowing distortion parameters. We then investigate how the relative order of two association patterns is affected when the same randomization is conducted. We show that some measures (e.g., Piatetsky-Shapiro) have relative horizontal order invariant properties, i.e, if one pattern is stronger than another in the original data, we have that the first one is still stronger than the second one in the randomized data.

In the second part of our paper, we extend association analysis from two variables to multiple variables. We investigate the feasibility of loglinear modeling, which is well adopted to analyze

associations among three or more variables, and examine the criterion on determining which hierarchical loglinear models are preserved in the randomized data. We also show that several multi-variate association measures studied in the data mining community are special cases of loglinear modeling.

Finally, we demonstrate the infeasibility of some classic data mining tasks (e.g., association rule mining, decision tree learning, naïve Bayesian classifier) on randomized data by showing the non-monotonic properties of measures (e.g.,support/confidence, gini) adopted in those data mining tasks. Our motivation is to provide a reference to data miners about what they can do and what they can not do with certainty upon the randomized data directly without distortion parameters. To the best of our knowledge, this is the first such formal analysis of the effects of Randomized Response for privacy preserving categorical data analysis with unknown distortion parameters.

2 Related Work

Privacy is becoming an increasingly important issue in many data mining applications. A considerable amount of work on privacy preserving data mining, such as additive randomization based [1, 3] has been proposed. Recently, a lot of research has focused on the privacy aspect of the above approaches and various point-wise reconstruction methods [23, 24] have been investigated.

The issue of maintaining privacy in association rule mining and categorical data analysis has also attracted considerable studies [4, 11, 14, 15, 31]. Most of techniques are based on a data perturbation or Randomized Response (RR) approach [7]. In [31], the authors proposed the MASK technique to preserve privacy for frequent itemset mining and extended to general categorical attributes in [4]. In [11], the authors studied the use of randomized response technique to build decision tree classifiers. In [19, 20], the authors focused on the issue of providing accuracy in terms of various reconstructed measures (e.g., support, confidence, correlation, lift, etc.) in privacy preserving market basket data analysis when the distortion parameters are available. Recently, the authors in [22] studied the search of optimal distortion parameters to balance privacy and utility.

Most of previous work except [19] investigated the scenario that distortion parameters are fully or partially known by data miners. For example, the authors in [13] focused on measuring privacy from attackers view when the distorted records of individuals and distortion parameters (e.g., f_Y and P) are available. In [19], the authors very briefly showed that some measures have vertical monotonic property on the market basket data. In this paper, we present a complete framework on privacy preserving categorical data analysis without distortion parameters. We extend studies on association measures between two binary variables to those on multiple polychotomous variables. More importantly, we also propose a new type of monotonic property, *horizontal association*, i.e., according to some measures, if the association between one pair of variables is stronger than another in the original data, the same order will still be kept in the randomized data when the same level of randomization is applied.

Randomized Response (RR) techniques have also been extensively investigated in statistics (e.g., see a book [7]). The Post RAndomization Method (PRAM) has been proposed to prevent disclosure in publishing micro data [9, 17, 18, 35, 36]. Specifically, they studied how to choose transition probabilities (a.k.a. distortion parameters) such that certain chosen marginal distributions in the original data are left invariant in expectation of the randomized data. There are some other noise-addition methods have been investigated in the literature, see the excellent survey [6]. Authors in [25] proposed a method by additional transformations that guarantees the covariance matrix of the distorted variables is an unbiased estimate for the one of the original variables. The method works well for numerical variables, but it is difficult to be applied to categorical variables due to the structure of the transformations.

Recently, the role of background knowledge in privacy preserving data mining has been studied

Table 1: COIL significant attributes used in example. The column "Mapping" shows how to map each original variable to a binary variable.

attribute	$i-$th attribute	Name	Description	Mapping
A	18	MOPLLAAG	Lower level education	$> 4 \rightarrow 1$
B	37	MINKM30	Income $<$ 30K	$> 4 \rightarrow 1$
C	42	MINKGEM	Average income	$> 4 \rightarrow 1$
D	43	MKOOPKLA	Purchasing power class	$> 3 \rightarrow 1$
E	44	PWAPART	Contribution private third party insurance	$> 0 \rightarrow 1$
F	47	PPERSAUT	Contribution car policies	$> 0 \rightarrow 1$
G	59	PBRAND	Contribution fire policies	$> 0 \rightarrow 1$
H	65	AWAPART	Number of private third party insurance	$> 0 \rightarrow 1$
I	68	APERSAUT	Number of car policies	$> 0 \rightarrow 1$
J	86	CARAVAN	Number of mobile home policies	$> 0 \rightarrow 1$

[10, 28]. Their focus was on disclosure risk due to the effect of various background knowledge. The focus of our work is on data utility when the distortion parameters are not available. We consider the extreme scenario about what data miners can do and can not do with certainty upon randomized data directly without any other background knowledge. Privacy analysis is beyond the scope of this paper and will be addressed in our future work.

3 Preliminaries

Throughout this paper, we use the COIL Challenge 2000 which provides data from a real insurance business. Information about customers consists of 86 attributes and includes product usage data and socio-demographic data derived from zip area codes. Our binary data is formed by collapsing non-binary categorical attributes into binary form, with 5822 records and 86 binary attributes. We use ten attributes (denote as A to J) as shown in Table 1 to illustrate our results.

3.1 Notations

To be consistent with notations, we denote the set of records in the database $\mathcal{D}$ by $\mathcal{T} = \{T_0, \cdots, T_{N-1}\}$ and the set of variables by $\mathcal{I} = \{A_0, \cdots, A_{m-1}, B_0, \cdots, B_{n-1}\}$. Note that, for ease of presentation, we use the terms "attribute" and "variable" interchangeably. Let there be m sensitive variables $A_0, \cdots, A_{m-1}$ and n non-sensitive variables $B_0, \cdots, B_{n-1}$. Each variable A_u has d_u mutually exclusive and exhaustive categories. We use $i_u = 0, \cdots, d_u - 1$ to denote the index of its categories. For each record, we apply the Randomized Response model independently on each sensitive variable A_u using different settings of distortion, while keeping the non-sensitive ones unchanged.

To express the relationship among variables, we can map categorical data sets to contingency tables. Table 2(a) shows one contingency table for a pair of two variables, *Gender* and *Race* ($d_1 = 2$ and $d_2 = 3$). The vector $\boldsymbol{\pi} = (\pi_{00}, \pi_{01}, \pi_{02}, \pi_{10}, \pi_{11}, \pi_{12})'$ corresponds to a fixed order of cell entries π_{ij} in the 2×3 contingency table. π_{01} denotes the proportion of records with *Male* and *White*. The row sum π_{0+} represents the proportion of records with *Male* across all races.

Table 2: 2×3 contingency tables for two variables Gender, Race

(a) Original

	Black	White	Asian	
Male	π_{00}	π_{01}	π_{02}	π_{0+}
Female	π_{10}	π_{11}	π_{12}	π_{1+}
	π_{+0}	π_{+1}	π_{+2}	π_{++}

(b) After randomization

	Black	White	Asian	
Male	λ_{00}	λ_{01}	λ_{02}	λ_{0+}
Female	λ_{10}	λ_{11}	λ_{12}	λ_{1+}
	λ_{+0}	λ_{+1}	π_{+2}	λ_{++}

Table 3: Notation

Symbol	Definition
A_u	the uth variable which is sensitive
B_l	the lth variable which is not sensitive
P_u	distortion matrix of A_u
$\theta^{(u)}$	distortion parameter of A_u
$\tilde{A}_u$	variable A_u after randomization
χ^2_{ori}	χ^2 calculated from original data
χ^2_{ran}	χ^2 calculated from randomized data
$\pi_{i_0,\cdots,i_{k-1}}$	cell value of original contingency table
$\lambda_{i_0,\cdots,i_{k-1}}$	cell value of randomized contingency table

Formally, let $\pi_{i_0,\cdots,i_{k-1}}$ denotes the true proportion corresponding to the categorical combination of k variables $(A_{0i_0}, \cdots, A_{(k-1)i_{k-1}})$ in the original data, where $i_u = 0, \cdots, d_u - 1$; $u = 0, \cdots, k-1$, and A_{0i_0} denotes the i_0th category of attribute A_0. Let $\boldsymbol{\pi}$ be a vector with elements $\pi_{i_0,\cdots,i_{k-1}}$ arranged in a fixed order. The combination vector corresponds to a fixed order of cell entries in the contingency table formed by these k variables. Similarly, we denote $\lambda_{i_0,\cdots,i_{k-1}}$ as the expected proportion in the randomized data. Table 3 summarizes our notations.

3.2 Distortion Procedure

The first Randomized Response model proposed by Warner in 1965 dealt with one dichotomous attribute, i.e, every person in the population belongs to either a sensitive group A, or to its complement $\bar{A}$. The problem is to estimate the π_A, the unknown proportion of population members in group A. Each respondent is provided with a randomization device by which the respondent chooses one of the following two questions *Do you belong to A?* or *Do you belong to* $\bar{A}$*?* with respective probabilities p and $1 - p$ and then replies *yes* or *no* to the question chosen. Since no one but the respondent knows to which question the answer pertains, the technique provides response confidentiality and increases respondents' willingness to answer sensitive questions. In general, we can consider this dichotomous attribute as one $\{0, 1\}$ variable, e.g., with 0 = absence, 1= presence. Each record is independently randomized using the probability matrix

$$P = \begin{pmatrix} \theta_0 & 1 - \theta_1 \\ 1 - \theta_0 & \theta_1 \end{pmatrix} \tag{1}$$

If the original record is in the *absence*(*presence*) category, it will be kept in such category with a probability θ_0 (θ_1) and changed to *presence*(*absence*) category with a probability $1 - \theta_0$ ($1 - \theta_1$). The original Warner RR model simply sets $\theta_0 = \theta_1 = p$.

We extend RR to the scenario of multi-variables with multi-categories in our distortion framework. For one sensitive variable A_u with d_u categories, the randomization process is such that a record

belong to the jth category ($j = 0, ..., d_u - 1$) is distorted to 0, 1, ... or $d_u - 1$th category with respective probabilities $\theta^{(u)}_{j0}, \theta^{(u)}_{j1}, ..., \theta^{(u)}_{j\ d_u-1}$, where $\sum_{c=0}^{d_u-1} \theta^{(u)}_{jc} = 1$. The distortion matrix P_u for A_u is shown as below.

$$P_u = \begin{pmatrix} \theta^{(u)}_{00} & \theta^{(u)}_{10} & \cdots & \theta^{(u)}_{d_u-1\ 0} \\ \theta^{(u)}_{01} & \theta^{(u)}_{11} & \cdots & \theta^{(u)}_{d_u-1\ 1} \\ & \cdot & & \\ & \cdot & & \\ \theta^{(u)}_{0\ d_u-1} & \theta^{(u)}_{1\ d_u-1} & \cdots & \theta^{(u)}_{d_u-1\ d_u-1} \end{pmatrix}$$

Parameters in each column of P_u sum to 1, but are independent to parameters in other columns. The sum of parameters in each row is not necessarily equal to 1. The true proportion $\boldsymbol{\pi} = (\pi_0, \cdots, \pi_{d_u-1})$ is changed to $\boldsymbol{\lambda} = (\lambda_0, \cdots, \lambda_{d_u-1})$ after randomization. We have

$$\boldsymbol{\lambda} = P_u \boldsymbol{\pi}.$$

For the case of k multi-variables, we denote $\lambda_{\mu_0,\cdots,\mu_{k-1}}$ as the expected probability of getting a response $(A_{0\mu_0}, \cdots, A_{(k-1)\mu_{k-1}})$ and $\boldsymbol{\lambda}$ the vector with elements $\lambda_{\mu_0,\cdots,\mu_{k-1}}$ arranged in a fixed order (e.g., the vector $\boldsymbol{\lambda} = (\lambda_{00}, \lambda_{01}, \lambda_{02}, \lambda_{10}, \lambda_{11}, \lambda_{12})'$ corresponds to cell entries λ_{ij} in the randomized contingency table as shown in Table 2(b)). Let $P = P_0 \times \cdots \times P_{k-1}$, we can obtain

$$\boldsymbol{\lambda} = P\boldsymbol{\pi} = (P_0 \times \cdots \times P_{k-1})\boldsymbol{\pi} \tag{2}$$

where $\times$ stands for the Kronecker product[1].

The original database $\mathcal{D}$ is changed to $\mathcal{D}_{ran}$ after randomization. An unbiased estimate of $\boldsymbol{\pi}$ based on one given realization $\mathcal{D}_{ran}$ follows as

$$\hat{\boldsymbol{\pi}} = P^{-1}\hat{\boldsymbol{\lambda}} = (P_0^{-1} \times \cdots \times P_{k-1}^{-1})\hat{\boldsymbol{\lambda}} \tag{3}$$

where $\hat{\boldsymbol{\lambda}}$ is the vector of proportions calculated from $\mathcal{D}_{ran}$ corresponding to $\boldsymbol{\lambda}$ and P_u^{-1} denotes the inverse of the matrix P_u.

Previous work using RR model either focused on evaluating the trade-off between privacy preservation and utility loss of the reconstructed data with the released distortion parameters (e.g., [4,19,31]) or determining the optimal distortion parameters to achieve good performance (e.g., [22]). Data mining tasks were conducted on the reconstructed distribution $\hat{\boldsymbol{\pi}}$ calculated from Equation 3. In this paper, we investigate the problem whether data mining or statistical analysis tasks can still be conducted with unknown distortion parameters, which has not been studied in the literature.

In Lemma 1, we show that no monotonic relation exists for cell entries of contingency tables due to randomization.

Lemma 1. No monotonic relation exists between $\lambda_{i_0,\cdots,i_{k-1}}$ and $\pi_{i_0,\cdots,i_{k-1}}$.

Proof. We use two binary variables $A_u,\ A_v$ as an example. The proof of multiple variables with multi-categories is immediate. The distortion matrices are defined as:

$$P_u = \begin{pmatrix} \theta^{(u)}_0 & 1-\theta^{(u)}_1 \\ 1-\theta^{(u)}_0 & \theta^{(u)}_1 \end{pmatrix} \quad P_v = \begin{pmatrix} \theta^{(v)}_0 & 1-\theta^{(v)}_1 \\ 1-\theta^{(v)}_0 & \theta^{(v)}_1 \end{pmatrix}$$

[1]It is an operation on two matrices, an m-by-n matrix A and a p-by-q matrix B, resulting in the mp-by-nq block matrix

We have:

$$\lambda_{0+} = (\theta_0^{(u)} + \theta_1^{(u)} - 1)\pi_{0+} - \theta_1^{(u)} + 1$$

We can see that $\lambda_{0+} - \pi_{0+}$ is a function of π_{0+} ,$\theta_0^{(u)}$, $\theta_1^{(u)}$, and its value may be greater or less than 0 with varying distortion parameters. Similarly,

$$\begin{aligned}\lambda_{00} &= \theta_0^{(u)}\theta_0^{(v)}\pi_{00} + \theta_0^{(u)}(1-\theta_1^{(v)})\pi_{01} \\ &+ (1-\theta_1^{(u)})\theta_0^{(v)}\pi_{10} + (1-\theta_1^{(u)})(1-\theta_1^{(v)})\pi_{11}\end{aligned}$$

$\lambda_{00} - \pi_{00}$ is a function of π_{ij} ,$\theta_0^{(u)}$, $\theta_1^{(u)}$, $\theta_0^{(v)}$ and $\theta_1^{(v)}$, no monotonic relation exists.

4 Associations Between Two Variables

In this section, we investigate how associations between two variables are affected by randomization. Specifically, we consider two cases:

- *Case 1*: A_u and A_v, association between two sensitive variables.
- *Case 2*: A_u and B_l, association between a sensitive variable and a non-sensitive variable.

Case 2 is a special case of case 1 while P_l is an identity matrix, so any results for case 1 will satisfy case 2. However, it is not necessarily true vice versa.

4.1 Associations Between Two Binary Variables

Table 4 shows various association measures for two binary variables (Refer to [34] for a survey). We can observe that all measures can be expressed as functions with parameters as cell entries (π_{ij}) and their margin totals (π_{i+} or π_{+j}) in the 2-dimensional contingency table.

Randomization Setting For a binary variable A_u, which only has two categories (0 = absence, 1 = presence), the distortion parameters are the same as those in Equation 1.

In Section 4.1.1, we focus on the problem of vertical association variation, i.e., how association values of one pair of variables based on given measures are changed due to randomization. In Section 4.1.2, we focus on the problem of horizontal association variation, i.e., how the relative order of two association patterns is changed due to randomization.

4.1.1 Vertical Association Variation

We use subscripts ori and ran to denote measures calculated from the original data and randomized data (without knowing the distortion parameters) respectively. For example, χ^2_{ori} denotes the Pearson Statistics calculated from the original data $\mathcal{D}$ while χ^2_{ran} corresponds to the one calculated directly from the randomized data $\mathcal{D}_{ran}$.

There exist many different realizations $\mathcal{D}_{ran}$ for one original data set $\mathcal{D}$. When the data size is large, the distribution $\hat{\boldsymbol{\lambda}}$ calculated from one realization $\mathcal{D}_{ran}$ approaches its expectation $\boldsymbol{\lambda}$, which can be calculated from the distribution $\boldsymbol{\pi}$ of the original data set through Equation 2. This is because

$$cov(\hat{\boldsymbol{\lambda}}) = N^{-1}(\boldsymbol{\lambda}^{\delta} - \boldsymbol{\lambda}\boldsymbol{\lambda}'),$$

as shown in [7]. $cov(\hat{\boldsymbol{\lambda}})$ approaches zero when N is large. Here $\boldsymbol{\lambda}^{\delta}$ is a diagonal matrix with the same diagonal elements as those of $\boldsymbol{\lambda}$ arranged in the same order. All our following results and their

Table 4: Objective association measures for two binary variables

Measure	Expression
Support (s)	π_{11}
Confidence(c)	$\frac{\pi_{11}}{\pi_{1+}}$
Correlation (ϕ)	$\frac{\pi_{11}\pi_{00}-\pi_{01}\pi_{10}}{\sqrt{\pi_{1+}\pi_{+1}\pi_{0+}\pi_{+0}}}$
Cosine (IS)	$\frac{\pi_{11}}{\sqrt{\pi_{1+}\pi_{+1}}}$
Odds ratio (α)	$\frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}$
Interest (I)	$\frac{\pi_{11}}{\pi_{1+}\pi_{+1}}$
Jaccard (ζ)	$\frac{\pi_{11}}{\pi_{1+}+\pi_{+1}-\pi_{11}}$
Piatetsky-Shapiro's(PS)	$\pi_{11}-\pi_{1+}\pi_{+1}$
Mutual Info(M)	$\frac{\sum_i\sum_j \pi_{ij} log\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}}{-\sum_i \pi_{i+} log\pi_{i+}}$
Conviction (V)	$\frac{\pi_{1+}\pi_{+0}}{\pi_{10}}$
J-measure (J)	$\pi_{11} log\frac{\pi_{11}}{\pi_{1+}\pi_{+1}}+\pi_{10} log\frac{\pi_{10}}{\pi_{1+}\pi_{+0}}$
Certainty (F)	$\frac{\frac{\pi_{11}}{\pi_{1+}}-\pi_{+1}}{1-\pi_{+1}}$
Standard residues(e)	$\sqrt{N}\frac{\pi_{ij}-\pi_{i+}\pi_{+j}}{\sqrt{\pi_{i+}\pi_{+j}}}$
Likelihood (G^2)	$2N\sum_i\sum_j \pi_{ij} log\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}$
Pearson (χ^2)	$N\sum_i\sum_j \frac{(\pi_{ij}-\pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}}$
Added Value(AV)	$\frac{\pi_{11}}{\pi_{1+}}-\pi_{+1}$
Risk Difference (D)	$\frac{\pi_{00}}{\pi_{+0}}-\frac{\pi_{01}}{\pi_{+1}}$
Laplace (L)	$\frac{N\pi_{11}+1}{N\pi_{1+}+2}$
Kappa (κ)	$\frac{\pi_{11}+\pi_{00}-\pi_{1+}\pi_{+1}-\pi_{0+}\pi_{+0}}{1-\pi_{1+}\pi_{+1}-\pi_{0+}\pi_{+0}}$
Concentration Coefficient (τ)	$\frac{\sum_i\sum_j \pi_{ij}^2/\pi_{i+}-\sum_j \pi_{+j}^2}{1-\sum_j \pi_{+j}^2}$
Collective Strength (S)	$\frac{\pi_{11}+\pi_{00}}{\pi_{1+}\pi_{+1}+\pi_{0+}\pi_{+0}}\times\frac{1-\pi_{1+}\pi_{+1}-\pi_{0+}\pi_{+0}}{1-\pi_{11}-\pi_{00}}$
Uncertainty Coefficient (U)	$-\frac{\sum_i\sum_j \pi_{ij} log\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}}{\sum_j \pi_{+j} log\pi_{+j}}$

proofs are based on the expectation $\boldsymbol{\lambda}$, rather than a given realization $\hat{\boldsymbol{\lambda}}$. Since data sets are usually large in most data mining scenarios, we do not consider the effect due to small samples. In other words, our results are expected to hold for most realizations of the randomized data.

Result 1. For any pair of variables A_u, A_v perturbed with any distortion matrix P_u and P_v ($\theta_0^{(u)}$, $\theta_1^{(u)}$, $\theta_0^{(v)}$, $\theta_1^{(v)} \in [0,1]$) respectively (Case 1), or any pair of variables A_u, B_l where A_u is perturbed with P_u (Case 2), the $\chi^2, G^2, M, \tau, U, \phi, D, PS$ values calculated from both original and randomized data satisfy:

$$\begin{aligned} \chi^2_{ran} \le \chi^2_{ori}, \quad & G^2_{ran} \le G^2_{ori} \\ M_{ran} \le M_{ori}, \quad & \tau_{ran} \le \tau_{ori} \\ U_{ran} \le U_{ori}, \quad & |\phi_{ran}| \le |\phi_{ori}| \\ |D_{ran}| \le |D_{ori}|, \quad & |PS_{ran}| \le |PS_{ori}| \end{aligned}$$

No other measures shown in Table 4 holds monotonic property.

For randomization, we know that the distortion is 1) highest with $\theta = 0.5$ which imparts the maximum randomness to the distorted values; 2) symmetric around $\theta = 0.5$ and makes no difference, reconstruction-wise, between choosing a value θ or its counterpart $1 - \theta$. In practice, the distortion is usually conducted with θ greater than 0.5. The following results show the vertical association variations when $\theta_0^{(u)}$,$\theta_1^{(u)}$,$\theta_0^{(v)}$ and $\theta_1^{(v)}$ are greater than 0.5.

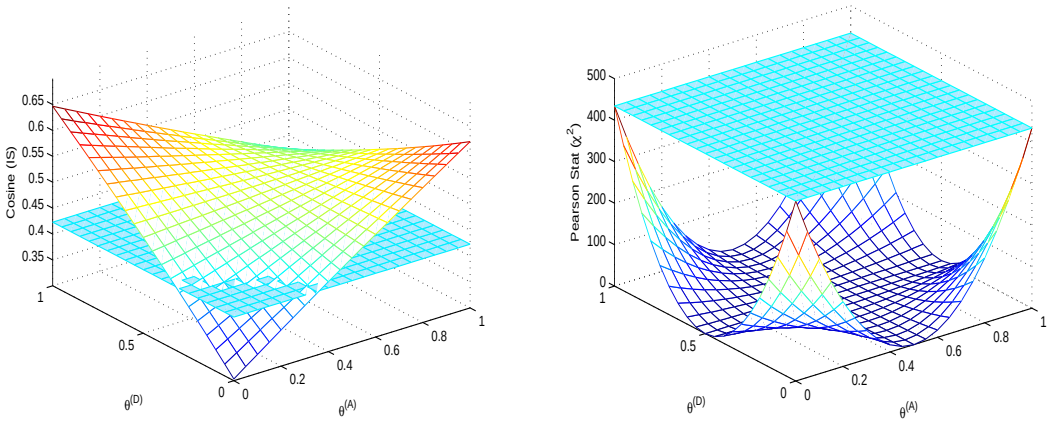
Result 2. In addition to monotonic relations shown in Result 1, when $\theta_0^{(u)}$, $\theta_1^{(u)}$, $\theta_0^{(v)}$, $\theta_1^{(v)} \in [0.5, 1]$, we have

$$\begin{aligned} |F_{ran}| \le |F_{ori}|, \quad & |AV_{ran}| \le |AV_{ori}| \\ |\kappa_{ran}| \le |\kappa_{ori}|, \quad & |\alpha_{ran} - 1| \le |\alpha_{ori} - 1| \\ |I_{ran} - 1| \le |I_{ori} - 1|, \quad & |V_{ran} - 1| \le |V_{ori} - 1| \\ |S_{ran} - 1| \le |S_{ori} - 1| & \end{aligned}$$

We include the proof of Added Value AV in Appendix. For all other measures in the above two results, we can prove similarly. We skip their proofs due to space limits. We can see that four measures (Odds Ratio α, Collective Strength S, Interest PS, and Conviction V) are compared with "1" since values of these measures with "1" indicate the two variables are independent. Next we illustrate this monotonic property using an example.

Example 1. Figure 1(a) and 1(b) show how the Cosine and Pearson Statistics calculated from the randomized data (attributes A and D from COIL data ($\boldsymbol{\pi}^{AD}$=(0.1374, 0.3332, 0.2982, 0.2312)′) vary with distortion parameters $\theta^{(A)}$ and $\theta^{(D)}$ (In all examples, we follow the original Warner model by setting $\theta_0^{(u)} = \theta_1^{(u)} = \theta^{(u)}$). It can be easily observed that $\chi^2_{ran} \le \chi^2_{ori}$ for all $\theta^{(A)}, \theta^{(D)} \in [0,1]$ and $IS_{ran} \ge IS_{ori}$ for some $\theta^{(A)}, \theta^{(D)}$ values.

One interesting question here is how to characterize those measures that have this monotonic property. The problem of analyzing objective measures used by data mining algorithms has attracted much attention in recent years [16, 33]. Depending on the specific properties of it, every measure is meaningful from some perspective and useful for some application, but not for others. Piatetsky-Shapiro [29] proposed three principles that should be satisfied by any good objective measure M for variables X, Y:



(a) Cosine (b) ChiSquare

Figure 1: statistics calculated from original data A, D (flat surface) vs. statistics calculated from randomized data (varied surface) with varying $\theta^{(A)}$ and $\theta^{(D)}$

- $C1$: $M = 0$ if X and Y are statistically independent, that is, $Pr(XY) = Pr(X)Pr(Y)$.
- $C2$: M monotonically increases with $Pr(XY)$ when $Pr(X)$ and $Pr(Y)$ remain the same.
- $C3$: M monotonically decreases with $Pr(X)$ (or $Pr(Y)$) when $Pr(XY)$ and $Pr(Y)$ (or $Pr(X)$) remain the same.

We can observe that all measures which obey $C1$ and $C2$ principles have monotonic properties after randomization by examining measures shown in Table 4.

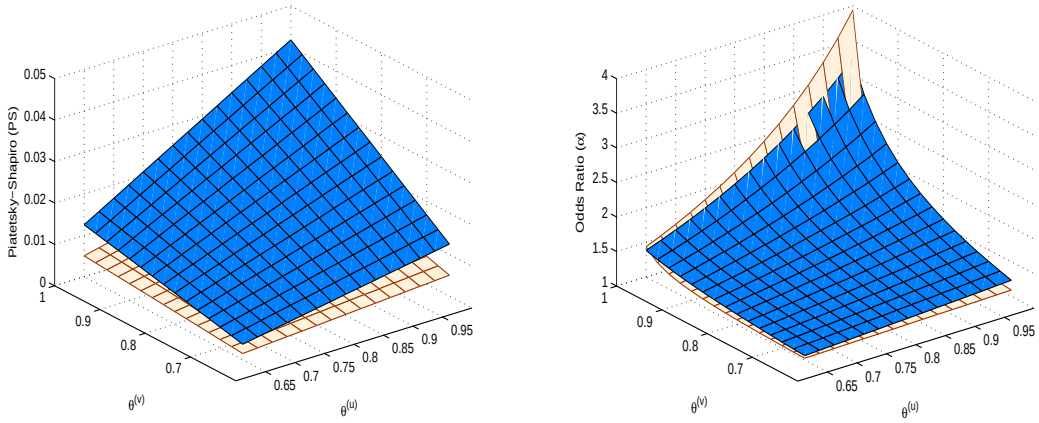
4.1.2 Horizontal Association Variation

In this section, we investigate the horizontal association variation problem, i.e., if the association based on a given association measure between one pair of variables is stronger than another in the original data, whether the same order will still be kept in the randomized data when the same level of randomization is applied.

We first illustrate this horizontal property using an example and then present our results.

Example 2. Figure 2(a) and 2(b) show how the Piatetsky-Shapiro's measure and Odds Ratio (A,B ($\boldsymbol{\pi}^{A,B}$=$(0.4222, 0.0484, 0.3861, 0.1432)'$) and I,J ($\boldsymbol{\pi}^{I,J}$=$(0.4763, 0.0124, 0.4639, 0.0474)'$)) calculated from the randomized data vary with distortion parameters $\theta^{(u)}$ and $\theta^{(v)}$. It can be easily observed from Figure 2(a) that the blue surface ($PS^{A,B}_{ran}$) is above the brown surface ($PS^{I,J}_{ran}$), which means that $PS^{A,B}_{ran} > PS^{I,J}_{ran}$ for all $\theta^{(u)}, \theta^{(v)} \in [0.5, 1]$ with $PS^{A,B}_{ori} > PS^{I,J}_{ori}$ ($PS^{A,B}_{ori}$ and $PS^{I,J}_{ori}$ are the points when $\theta_u = \theta_v = 1$). Figure 2(b) shows although $\alpha^{A,B}_{ori} < \alpha^{I,J}_{ori}$ ($\alpha^{A,B}_{ori} = 3.23, \alpha^{I,J}_{ori} = 3.94$), $\alpha^{A,B}_{ran} > \alpha^{I,J}_{ran}$ for some distortion parameters $\theta^{(u)}$ and $\theta^{(v)}$. For example, $\alpha^{A,B}_{ran} = 1.32$, $\alpha^{I,J}_{ran} = 1.14$ when $\theta^{(u)} = \theta^{(v)} = 0.8$.

Result 3. For any two sets of binary variables $\{A_u, A_v\}$ and $\{A_s, A_t\}$, A_u and A_s are perturbed with the same distortion matrix P_u while A_v and A_t are perturbed with the same distortion matrix



(a) PS (b) OddsRatio

Figure 2: statistics from randomized data of (A,B) (shown as blue surface) and (I,J) (shown as brown surface) with varying $\theta^{(u)}$ and $\theta^{(v)}$

P_v respectively ($\theta_0^{(u)}$, $\theta_1^{(u)}$, $\theta_0^{(v)}$, $\theta_1^{(v)} \in [0,1]$) (Case 1), we have

$$|PS_{ori}^{u,v}| \geq |PS_{ori}^{s,t}| \Longleftrightarrow |PS_{ran}^{u,v}| \geq |PS_{ran}^{s,t}|$$

where $PS_{ori}^{u,v}, PS_{ori}^{s,t}$ denote Piatetsky-Shapiro's measure calculated from the original dataset $\{A_u, A_v\}$ and $\{A_s, A_t\}$ respectively and $PS_{ran}^{u,v}, PS_{ran}^{s,t}$ correspond to measures calculated directly from the randomized data without knowing $\theta_0^{(u)}, \theta_1^{(u)}, \theta_0^{(v)}, \theta_1^{(v)}$.

Result 4. For any two pairs of variables $\{A_u, B_s\}$ and $\{A_v, B_t\}$, A_u and A_v are perturbed with the same distortion matrix P_u ($\theta_0^{(u)}$,$\theta_1^{(u)} \in [0,1]$) while B_s and B_t are unchanged (Case 2), we have

$$|D_{ori}^{u,s}| \geq |D_{ori}^{v,t}| \Longleftrightarrow |D_{ran}^{u,s}| \geq |D_{ran}^{v,t}|$$
$$|AV_{ori}^{u,s}| \geq |AV_{ori}^{v,t}| \Longleftrightarrow |AV_{ran}^{u,s}| \geq |AV_{ran}^{v,t}|$$

We include our proofs in Appendix. Through evaluation, no other measure in Table 4 except Piatetsky-Shapiros, Risk Difference, and Added Values measures has this property. Intuitively, if the same randomness is added to the two pairs of variables separately, the relative order of the association patterns should be kept after randomization. Piatetsky-Shapiro measure can be considered as a better measure than others to preserve such property.

4.2 Extension to Two Polychotomous Variables

There are five association measures (χ^2, G^2, M, τ, U) that can be extended to two variables with multiple categories as shown in Table 5.

Table 5: Objective measures for two polychotomous variables

Measure	Expression
Mutual Info (M)	$\frac{\sum_i \sum_j \pi_{ij} log \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}}{-\sum_i \pi_{i+} log \pi_{i+}}$
Likelihood (G^2)	$2 \sum_i \sum_j \pi_{ij} log \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}$
Pearson (χ^2)	$N \sum_i \sum_j \frac{(\pi_{ij} - \pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}}$
Concentration Coefficient (τ)	$\frac{\sum_i \sum_j \pi_{ij}^2 / \pi_{i+} - \sum_j \pi_{+j}^2}{1 - \sum_j \pi_{+j}^2}$
Uncertainty Coefficient (U)	$-\frac{\sum_i \sum_j \pi_{ij} log \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}}{\sum_j \pi_{+j} log \pi_{+j}}$

4.2.1 Vertical Variation

Result 5. For any pair of variables A_u, A_v perturbed with any distortion matrix P_u and P_v, the χ^2, G^2, M, τ, U values calculated from both original and randomized data satisfy:

$$\chi^2_{ran} \leq \chi^2_{ori}, \quad G^2_{ran} \leq G^2_{ori}$$
$$M_{ran} \leq M_{ori}, \quad \tau_{ran} \leq \tau_{ori}$$
$$U_{ran} \leq U_{ori}$$

We omit the proofs from this paper. We would emphasize that this result is important for data analysis tasks such as hypothesis testing. According to the above result, associations between two sensitive variables or associations between one sensitive variable with non-sensitive one will be attenuated by randomization. An important consequence of the attenuation results is that if there is no association between A_u, A_v or A_u, B_l in the original data, there will also be no association in randomized data.

Result 6. The χ^2 test for independence on the randomized $\tilde{A}_u$ with $\tilde{A}_v$ or on $\tilde{A}_u$ with B_l is a correct α-level test for independence on A_u with A_v or A_u with B_l while with reduced power.

This result shows testing pairwise independence between the original variables is equivalent to testing pairwise independence between the corresponding distorted variables. That is, the test can be conducted on distorted data directly when variables in the original data are independent. However, the testing power to reject the independence hypotheses may be reduced when variables in the original data are not independent. For independence testing, we have two hypotheses:

- H_0: $\pi_{ij} = \pi_{i+}\pi_{+j}$, for $i = 0, ..., d_1 - 1$ and $j = 0, ..., d_2 - 1$.
- H_1: the hypotheses of H_0 is not true.

The test procedure is to reject H_0 with significance level α if $\chi^2 \geq C$. In other words, $Pr(\chi^2 \geq C|H_0) \leq \alpha$. The probability of making Type I error is defined as $Pr(\chi^2 \geq C|H_0)$ while $1 - Pr(\chi^2 \geq C|H_1)$ denotes the probability of making Type II error. To maximize the power of the test, C is set as χ^2_α, i.e., the $1 - \alpha$ quantile of the χ^2 distribution with $(d_1 - 1)(d_2 - 1)$ degrees of freedom.

If two variables are independent in original data, i.e., $\chi^2_{ori} < \chi^2_\alpha$, when testing independence on the randomized data, we have $\chi^2_{ran} < \chi^2_{ori} < \chi^2_\alpha$. We can observe that randomization does not affect the validity of the significance test with level α. The risk of making Type I error is not increased.

If two variables are dependent in original data, i.e., $\chi^2_{ori} \geq \chi^2_\alpha$. The power to reject H_0 ($Pr(\chi^2_{ori} \geq \chi^2_\alpha | H_1)$) will be reduced to $Pr(\chi^2_{ran} \geq \chi^2_\alpha | H_1)$ when testing on randomized data. That is, χ^2_{ran} may be decreased to be less than χ^2_α. Hence we may incorrectly accept H_0. The probability of making Type II error is increased.

4.2.2 Horizontal Variation

Since none of Risk Difference, Added Value, and Piatetsky-Shapiro can be extended to polychotomous variables, no measure has the monotonic property in terms of horizontal association variation for a pair of variables with multi categories.

5 High Order Association based on Loglinear Modeling

Loglinear modeling has been commonly used to evaluate multi-way contingency tables that involve three or more variables [5]. It is an extension of the two-way contingency table where the conditional relationship between two or more categorical variables is analyzed. When applying loglinear modeling on randomized data, we are interested in the following problems. First, is the fitted model learned from the randomized data equivalent to that learned from the original data? Second, do parameters of loglinear models have monotonic properties? In Section 5.1, we first revisit loglinear modeling and focus on the hierarchical loglinear model fitting. In Section 5.2, we present the criterion to determine which hierarchical loglinear models can be preserved after randomization. In Section 5.3, we investigate how parameters of loglinear models are affected by randomization.

5.1 Loglinear Model Revisited

Loglinear modeling is a methodology for approximating discrete multidimensional probability distributions. The multi-way table of joint probabilities is approximated by a product of lower-order tables. For a value $y_{i0i1\cdots i(n-1)}$ at position i_r of the rth dimension d_r ($0 \leq r \leq n-1$), we define the log of anticipated value $\hat{y}_{i0i1\cdots i(n-1)}$ as a linear additive function of contributions from various higher level group-bys as:

$$\hat{l}_{i0i1\cdots i(n-1)} = \log \hat{y}_{i0i1\cdots i(n-1)} = \sum_{\mathcal{G} \subseteq \mathcal{I}} \gamma^{\mathcal{G}}_{(i_r | d_r \in \mathcal{G})}$$

We refer to the γ terms as the coefficients of the model. For instance, in a 3-dimensional table with dimensions A, B, C, Equation 4 shows the saturated loglinear model. It contains the 3-factor effect γ^{ABC}_{ijk}, all the possible 2-factor effects (e.g.,γ^{AB}_{ij}), and so on up to the 1-factor effects (e.g., γ^A_i) and the mean γ.

$$\log \hat{y}_{ijk} = \gamma + \gamma^A_i + \gamma^B_j + \gamma^C_k + \gamma^{AB}_{ij} + \gamma^{AC}_{ik} + \gamma^{BC}_{jk} + \gamma^{ABC}_{ijk} \quad (4)$$

As the saturated model has the same amount of cells in the contingency table as its parameters, the expected cell frequencies will always exactly match the observed ones with no degree of freedom. Thus, in order to find a more parsimonious model that will isolate the effects best demonstrating the data patterns, a non-saturated model must be sought.

Fitting hierarchical loglinear models Hierarchical models are nested models in which when an interaction of d factors is present, all the interactions of lower order between the variables of that interaction are also present. Such a model can be specified in terms of the configuration of highest-order interactions. For example, a hierarchical model denoted as (ABC, DE) for five variables (A-E) has two highest factors (γ^{ABC} and γ^{DE}). The model also includes all the interactions of

Table 6: Goodness-of-Fit tests for loglinear models on A, D, G

Model	χ^2	df	p-Value
A, D, G	435.70	4	<0.001
AD, G	1.60	3	0.66
AG, D	434.40	3	<0.001
DG, A	435.71	3	<0.001

lower order factors such as two factor effects (γ^{AB}, γ^{AC}, γ^{BC}), one factor effects (γ^{A}, γ^{B}, γ^{C}, γ^{D}, γ^{E}) and the mean γ.

To fit a hierarchical loglinear model, we can either start with the saturated model and delete higher order interaction terms or start with the simplest model (independence model) and add more complex interaction terms. The Pearson statistic can be used to test the overall goodness-of-fit of a model by comparing the expected frequencies to the observed cell frequencies for each model. Based on the Pearson statistic value and degree of freedom of each model, the p-value is calculated to denote the probability of observing the results from data assuming the null hypothesis is true. Large p-value means little or no evidence against the null hypothesis.

Example 3. For variables A, D, G in COIL data ($\boldsymbol{\pi}^{ADG}$=(0.0610, 0.0764, 0.1506, 0.1826, 0.1384, 0.1597, 0.1079, 0.1233)′) in COIL data, Table 6 shows Pearson and p-value of Hypothesis Test for different models. We can see model (AD, G) has the smallest χ^2 value (1.60) and the largest p-value (0.66). Hence the best fitted model is (AD, G), i.e.,

$$\log \hat{y}_{ijk} \quad = \quad \gamma + \gamma_i^A + \gamma_j^D + \gamma_k^G + \gamma_{ij}^{AD} \tag{5}$$

5.2 Equivalent Loglinear Model

Chen [8] first studied equivalent loglinear models under independent misclassification in statistics. Korn [26] extended his work and proposed Theorem 1 as a criterion for obtaining hierarchical loglinear models from misclassified data directly if the misclassification is non-differential and independent.

Theorem 1. A hierarchical model is preserved by misclassification if no misclassified variable appears more than once in the specification in terms of the highest order interactions of the model. A model is said to be preserved if the misclassified data fits the same model as the original data (i.e., the misclassification induces no spurious associations between the variables).

Since the Randomized Response in our framework is one kind of such non-differential and independent misclassification, we can apply the same criterion to check whether a hierarchical loglinear model is preserved in the randomized data. Theorem 1 clearly specifies the criterion of the preserved models, i.e., any randomized variable cannot appear more than once in the highest order interactions of the model specification. We first illustrate this criterion using examples and then examine the feasibility of several widely adopted models on the randomized data.

Example 4. The loglinear model (AD, G) as shown in Equation 5 is preserved on all randomized data with different distortion parameters as shown in Table 7. We can see that the p-value of model (AD, G) is always prominent no matter how we change the distortion parameters $(\theta^{(A)}, \theta^{(D)}, \theta^{(G)})$. On the contrary, the loglinear model (AB, AE) that best fits the original data with attributes A,B,E ($\boldsymbol{\pi}^{ABE}$=(0.2429, 0.1793, 0.0258, 0.0227, 0.2391, 0.1470, 0.0903, 0.0529)′) cannot be preserved

Table 7: Goodness-of-Fit tests for loglinear models on attributes A, D, G after Randomization with different $(\theta^{(A)}, \theta^{(D)}, \theta^{(G)})$

Model	Original		(0.9,0.9,0.9)		(0.7,0.7,0.7)		(0.7,0.8,0.9)	
	χ^2	P-value	χ^2	P-value	χ^2	P-value	χ^2	P-value
A, D, G	435.70	<0.001	177.16	<0.001	10.97	0.03	24.82	<0.001
$\boldsymbol{AD, G}$	**1.60**	**0.66**	**0.61**	**0.89**	**0.04**	**0.99**	**0.15**	**0.98**
AG, D	434.40	<0.001	176.60	<0.001	10.93	0.01	24.68	<0.001
DG, A	435.71	<0.001	177.17	<0.001	10.97	0.01	24.83	<0.001

Table 8: Goodness-of-Fit tests for loglinear models on attributes A, B, E after Randomization with different $(\theta^{(A)}, \theta^{(B)}, \theta^{(E)})$

Model	Original		(0.9,0.9,0.9)		(0.7,0.7,0.7)		(0.55,0.9,0.9)	
	χ^2	P-value	χ^2	P-value	χ^2	P-value	χ^2	P-value
A, B, E	280.87	<0.001	95.05	<0.001	4.84	0.30	1.59	0.81
$\boldsymbol{AB, E}$	18.33	<0.001	6.78	0.08	0.40	0.94	**0.21**	**0.98**
AE, B	264.81	<0.001	88.51	<0.001	4.44	0.22	1.49	0.69
BE, A	279.18	<0.001	94.68	<0.001	4.83	0.19	1.48	0.69
$\boldsymbol{AB, AE}$	**2.28**	**0.32**	**0.32**	**0.85**	**0.01**	**0.99**	0.11	0.95
AB, BE	18.03	<0.001	6.67	0.04	0.40	0.82	0.10	0.95
AE, BE	264.07	<0.001	88.35	<0.001	4.44	0.11	1.38	0.50

on all the randomized data with different distortion parameters as shown in Table 8. We can observe when $\theta^{(A)} = 0.55$, $\theta^{(B)} = 0.9$ and $\theta^{(E)} = 0.9$, the p-value of model (AB, E) is greater than that of model (AB, AE). Hence, the fitted model on randomized data is changed to (AB, E).

Independence model and all-two-factor model. In [32], the authors proposed the use of the complete independence model (all 1-factor effects and the mean γ) to measure significance of dependence. In [12], the authors proposed the use of all-two-factor effects model to distinguish between multi-item associations that can be explained by all pairwise associations, and item sets that are significantly more frequent than their pairwise associations would suggest.

For a 3-dimensional table, the complete independence model (A, B, C) is shown in Equation 6 while the all-two-factor model (AB, AC, BC) is shown in Equation 7.

$$\log \hat{y}_{ijk} = \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C \tag{6}$$

$$\log \hat{y}_{ijk} = \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{jk}^{BC} \tag{7}$$

According to the criterion, we can conclude that the independence model can be applied on randomized data to test complete independence among variables of original data. However, we cannot test the all-two-factor model on randomized data directly since the all-two-factor model cannot be preserved after randomization.

Conditional independence testing. For a 3-dimensional case, testing conditional independence of two variables, A and B, given the third variable C is equivalent to the fitting of the loglinear model (AC, BC). Based on the criterion, we can easily derive that the model (AC, BC) is not preserved after randomization when variable C is randomized.

In practice, the *partial correlation* is often adopted to measure the correlation between two variables after the common effects of all other variables in the data set are removed.

$$pr_{AB.C} = \frac{r_{AB} - r_{AC}r_{BC}}{\sqrt{(1 - r_{AC}^2)(1 - r_{BC}^2)}} \tag{8}$$

Equation 8 shows the form for the partial correlation of two variables, A and B, while controlling for a third variable C, where r_{AB} denotes Pearson's correlation coefficient. If there is no difference between $pr_{AB.C}$ and r_{AB}, we can infer that the control variable C has no effect. If the partial correlation approaches zero, the inference is that the original correlation is spurious (i.e., there is no direct causal link between the two original variables because the control variable is either the common anteceding cause, or the intervening variable).

According to the criterion, we have the following results.

Result 7. The χ^2 test of the independence on two randomized variables $\tilde{A}_u$ with $\tilde{A}_v$ (or on $\tilde{A}_u$ with B_l) conditional on a set of variables $\mathcal{G}$ ($\mathcal{G} \subseteq \mathcal{I}$) is a correct α-level test for independence on A_u with A_v (or A_u with B_l) conditional on $\mathcal{G}$ while with reduced power if and only if no distorted sensitive variable is contained in $\mathcal{G}$.

Result 8. The partial correlation of two sensitive variables or the partial correlation of one sensitive variable and one non-sensitive variable conditional on a set of variables $\mathcal{G}$ ($\mathcal{G} \subseteq \mathcal{I}$) has monotonic property $|pr_{ran}| \leq |pr_{ori}|$ if and only if no distorted sensitive variable is contained in $\mathcal{G}$.

Other association measures for multi variables. There are five measures (IS, I, PS, G^2, χ^2) that can be extended to multiple variables. Association measures for multiple variables need an assumed model (usually the complete independence model). We have shown that G^2 and χ^2 on the independence model have monotonic relations. However, we can easily check that IS, I, PS do not have monotonic properties since they are determined by the difference between one cell entry value and its estimate from the assumed model. On the contrary, G^2 and χ^2 are aggregate measures which are determined by differences across all cell entries.

5.3 Variation of Loglinear Model Parameters

Parameters of loglinear models indicate the interactions between variables. For example, the γ_{ij}^{AB} is two-factor effect which shows the dependency within the distributions of the associated variables A, B. We present our result below and leave detailed proof in Appendix.

Result 9. For any k-factor coefficient $\gamma^{\mathcal{G}_k}_{(i_r|d_r \in \mathcal{G}_k)}$ in hierarchical loglinear model, no vertical monotonic property or horizontal relative order invariant property is held after randomization.

6 Effects on Other Data Mining Applications

In this section, we examine whether some classic data mining tasks can be conducted on randomized data directly.

6.1 Association Rule Mining

Association rule learning is a widely used method for discovering interesting relations between items in data mining [2]. An association rule $\mathcal{X} \Rightarrow \mathcal{Y}$, where $\mathcal{X}, \mathcal{Y} \subset \mathcal{I}$ and $\mathcal{X} \cap \mathcal{Y} = \phi$, has two measures: the support s defined as $s(100\%)$ of the transactions in $\mathcal{T}$ that contain $\mathcal{X} \cup \mathcal{Y}$, and the confidence c is defined as $c(100\%)$ of the transactions in $\mathcal{T}$ that contain $\mathcal{X}$ also contain $\mathcal{Y}$. From Result 1 and

Result 2, we can easily learn that neither support nor confidence measures of association rule mining holds monotonic relations. Hence, we cannot conduct association rule mining on randomized data directly since values of support and confidence can become greater or less than the original ones after randomization.

6.2 Decision Tree Learning

Decision tree learning is a procedure to determine the class of a given instance [30]. Several measures have been used in selecting attributes for classification. Among them, gini function measures the *impurity* of an attribute with respect to the classes. If a data set D contains examples from l classes, given the probabilities for each class (p_i), $gini(D)$ is defined as $gini(D) = 1 - \sum_{i=1}^{l} p_i^2$.

When D is split into two subsets D_1 and D_2 with sizes n_1 and n_2 respectively, the gini index of the split data is:

$$gini_{split}(D) = \frac{n_1}{n} gini(D_1) + \frac{n_2}{n} gini(D_2)$$

The attribute with the smallest $gini_{split}(D)$ is chosen to split the data.

Result 10. The relative order of gini values can not be preserved after randomization. That is, there is no guarantee that the same decision tree can be learned from the randomized data.

Example 5. For variables A,B,C ($\boldsymbol{\pi}^{ABC}$=(0.2406, 0.1815, 0.0453, 0.0031, 0.3458, 0.0404, 0.1431, 0.0002)′) in COIL data, we set A,B as two sensitive attributes and C as class attribute. The $gini$ values of A, B before randomization are:

$$\begin{aligned} gini_{split}(A)_{ori} &= \pi_{\overline{A}} gini(A_1) + \pi_A gini(A_2) \\ &= \pi_{\overline{A}}[1 - (\frac{\pi_{\overline{A}C}}{\pi_{\overline{A}}})^2 - (\frac{\pi_{\overline{A}C}}{\pi_{\overline{A}}})^2] + \pi_A[1 - (\frac{\pi_{A\overline{C}}}{\pi_A})^2 - (\frac{\pi_{AC}}{\pi_A})^2] \\ &= 0.30 \end{aligned}$$

Similarly, $gini_{split}(B)_{ori} = 0.33$.

After randomization with distortion parameters $\theta_0^{(A)} = \theta_1^{(A)} = 0.6$ and $\theta_0^{(B)} = \theta_1^{(B)} = 0.9$ ($\boldsymbol{\lambda}^{ABC}$=(0.2629, 0.1127, 0.1042, 0.0143, 0.2837, 0.0873, 0.1240, 0.0109)′), we get:

$$gini_{split}(A)_{ran} = 0.35 \qquad gini_{split}(B)_{ran} = 0.34$$

The relative order of $gini_{split}(A)$ and $gini_{split}(B)$ can not be preserved after randomization.

6.3 Naïve Bayes Classifier

A naïve Bayes classifier is a probabilistic classifier to predict the class label for a given instance with attributes set $\mathcal{X}$. It is based on applying Bayes' theorem (from Bayesian statistics) with strong assumptions that the attributes are conditional independence given class label C .

Given an instance with feature vector $\boldsymbol{x}$, the naïve Bayes classifier to determine its class label C is defined as:

$$h^*(\boldsymbol{x}) = argmax_i \frac{P(\mathcal{X} = \boldsymbol{x}|C = i)P(C = i)}{P(\mathcal{X} = \boldsymbol{x})}$$

It chooses the maximum a posteriori probability (MAP) hypothesis to classify the example.

Result 11. The relative order of posteriori probabilities can not be preserved after randomization. That is, instances can not be classified correctly based on the Naïve Bayes classifier derived from randomized data directly.

Example 6. For variables A,G,H ($\boldsymbol{\pi}^{AGH}$=$(0.1884, 0.0232, 0.0802, 0.1788, 0.2264, 0.0199, 0.1031, 0.1800)'$) in COIL data, we set A,G as two sensitive attributes and H as class attribute. For an instance with attributes $A = 0, G = 1$, the probability of its class $H = 0$ before randomization is:

$$\begin{aligned} P(\overline{H}|\overline{A}G)_{ori} &= P(\overline{A}|\overline{H}) \times P(G|\overline{H}) \times P(\overline{H})/P(\overline{A}G) \\ &= \frac{\pi_{\overline{A}\overline{H}}}{\pi_{\overline{H}}} \times \frac{\pi_{G\overline{H}}}{\pi_{\overline{H}}} \times \pi_{\overline{H}}/\pi_{\overline{A}G} \\ &= \frac{\pi_{\overline{A}\overline{H}}\pi_{G\overline{H}}}{\pi_{\overline{H}}}/\pi_{\overline{A}G} \\ &= 0.31 \end{aligned}$$

Similarly, the probability of its class $H = 1$ is:

$$P(H|\overline{A}G)_{ori} = \frac{\pi_{\overline{A}H}\pi_{GH}}{\pi_H}/\pi_{\overline{A}G} = 0.69$$

After randomization with distortion parameters $\theta_0^{(A)}$=$\theta_1^{(A)}$=$\theta_0^{(G)}$=$\theta_1^{(G)} = 0.6$ ($\boldsymbol{\lambda}^{AGH}$=$(0.1579, 0.0848, 0.1351, 0.1163, 0.1643, 0.0845, 0.1408, 0.1162)'$), we get:

$$P(\overline{H}|\overline{A}G)_{ran} = 0.54 \quad P(H|\overline{A}G)_{ran} = 0.46$$

As none of $\pi_{\overline{A}\overline{H}}, \pi_{G\overline{H}}, \pi_{\overline{A}H}, \pi_{GH}$ has monotonic properties after randomization, the relative order of the two probabilities $P(\overline{H}|\overline{A}G)$ and $P(H|\overline{A}G)$ cannot be kept.

7 Conclusion

The trade-off between privacy preservation and utility loss has been extensively studied in privacy preserving data mining. However, data owners are still reluctant to release their (perturbed or transformed) data due to privacy concerns. In this paper, we focus on the scenario where distortion parameters are not disclosed to data miners and investigate whether data mining or statistical analysis tasks can still be conducted on randomized categorical data. We have examined how various objective association measures between two variables may be affected by randomization. We then extended to multiple variables by examining the feasibility of hierarchical loglinear modeling. We have shown that some classic data mining tasks (e.g., association rule mining, decision tree learning, naïve Bayes classifier) cannot be applied on the randomized data with unknown distortion parameters. We provided a reference to data miners about what they can do and what they can not do with certainty upon randomized data directly without the knowledge about the original distribution of data and distortion information.

In our future work, we will comprehensively examine various data mining tasks (e.g., causal learning) as well as their associated measures in detail. We will conduct experiments on large data sets to evaluate how strong our theoretical results may hold in practice. We are also interested in extending this study to numerical data or networked data.

Acknowledgment

This work was supported in part by U.S. National Science Foundation IIS-0546027.

References

[1] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the 20th Symposium on Principles of Database Systems*, 2001.

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD Conference*, pages 207–216, 1993.

[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 439–450. Dallas, Texas, May 2000.

[4] S. Agrawal and J. R. Haritsa. A framework for high-accuracy privacy-preserving mining. In *Proceedings of the 21st IEEE International Conference on Data Engineering*, pages 193–204, 2005.

[5] A. Agresti. *Categorical data analysis*. Wiley, 2002.

[6] R. Brand. Microdata protection through noise addition. *Lecture Notes in Computer Science*, 2316:97–116, 2002.

[7] A. Chaudhuri and R. Mukerjee. *Randomized response: theory and techniques*. Marcel Dekker, 1988.

[8] T. T. Chen. Analysis of randomized response as purposively misclassified data. *Journal of the American Statistical Association*, pages 158–163, 1979.

[9] J. Domingo-Ferrer, J.M. Mateo-Sanz, and V. Torra. Comparing SDC methods for micro-data on the basis of information loss and disclosure risk. In *Proceedings of NTTS and ETK*, 2001.

[10] W. Du, Z. Teng, and Z. Zhu. Privacy-maxent: integrating background knowledge in privacy quantification. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 459–472, 2008.

[11] W. Du and Z. Zhan. Using randomized response techniques for privacy-preserving data mining. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 505–510, 2003.

[12] W. DuMouchel and D. Pregibon. Empirical bayes screening for multi-item association. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, August 2001.

[13] A. Evfimievski. Randomization in privacy preserving data mining. *ACM SIGKDD Explorations Newsletter*, 4(2):43–48, 2002.

[14] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the 22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 211–222, 2003.

[15] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–228, 2002.

[16] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey . *ACM Computing Surveys*, 38(3):9, 2006.

[17] S. Gomatam and A. F. Karr. Distortion measures for categorical data swapping. Technical Report, Number 131, National Institute of Statistical Sciences, 2003.

[18] J. M. Gouweleeuw, P. Kooiman, L. C. R. J. Willenborg, and P. P. de Wolf. Post randomization for statistical disclosure control: theory and implementation. *Journal of Official Statistics*, 14(4):463–478, 1998.

[19] L. Guo, S. Guo, and X. Wu. Privacy preserving market basket data analysis. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, September 2007.

[20] L. Guo, S. Guo, and X. Wu. On addressing accuracy concerns in privacy and preserving association rule mining. In *Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, May 2008.

[21] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. *Technical Report, University of Massachusetts*, 07-19, 2007.

[22] Z. Huang and W. Du. Optrr: Optimizing randomized response schemes for privacy-preserving data mining. In *Proceedings of the 24th IEEE International Conference on Data Engineering*, pages 705–714, 2008.

[23] Z. Huang, W. Du, and B. Chen. Deriving private information from randomized data. In *Proceedings of the ACM SIGMOD Conference on Management of Data*. Baltimore, MA, 2005.

[24] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *Proceedings of the 3rd International Conference on Data Mining*, pages 99–106, 2003.

[25] J. Kim. A method for limiting disclosure in microdata based on random noise and transformation. In *Proceedings of the American Statistical Association on Survey Research Methods*, 1986.

[26] E. L. Korn. Hierarchical log-linear models not preserved by classification error. *Journal of the American Statistical Association*, 76:110–113, 1981.

[27] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *Proceedings of the ACM SIGMOD Conference*, Vancouver, Canada, 2008. ACM Press.

[28] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge in privacy. Technical Report, Cornell University, 2006.

[29] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*, pages 229–248, 1991.

[30] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[31] S. J. Rizvi and J. R. Haritsa. Maintaining data privacy in association rule mining. In *Proceedings of the 28th International Conference on Very Large Data Bases*, 2002.

[32] C. Silverstein, S. Brin, and R. Motwani. Beyond market baskets: generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2:39–68, 1998.

[33] P. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, pages 32–41, 2002.

[34] P. Tan, M. Steinbach, and K. Kumar. *Introduction to data mining*. Addison Wesley, 2006.

[35] A. Van den Hot. Analyzing misclassified data: randomized response and post randomization. Ph.D. Thesis, University of Utrecht, 2004.

[36] L. Willenborg and T. De Waal. Elements of statistical disclosure control in practice. *Lecture Notes in Statistics*, 155, 2001.

[37] X. Ying and X. Wu. Randomizing social networks: a spectrum preserving approach. In *Proceedings of the 8th SIAM Conference on Data Mining*, April 2008.

A Proof of Results

Proof of Result 1 and Result 2

The Added Value calculated directly from the randomized data without knowing P_u, P_v is

$$AV_{ran} = \frac{\lambda_{11}}{\lambda_{1+}} - \lambda_{+1} = \frac{\lambda_{11} - \lambda_{+1}\lambda_{1+}}{\lambda_{1+}}$$

The original Added Value can be expressed as

$$AV_{ori} = \frac{\pi_{11} - \pi_{+1}\pi_{1+}}{\pi_{1+}}$$

As $\boldsymbol{\pi} = (P_u^{-1} \times P_v^{-1})\boldsymbol{\lambda}$, we have:

$$\begin{aligned}
\pi_{1+} &= \frac{\theta_1^{(u)} - 1 + (1 + \theta_0^{(u)} - \theta_1^{(u)})\lambda_{1+}}{\theta_0^{(u)} + \theta_1^{(u)} - 1} \\
\pi_{+1} &= \frac{\theta_1^{(v)} - 1 + (1 + \theta_0^{(v)} - \theta_1^{(v)})\lambda_{+1}}{\theta_0^{(v)} + \theta_1^{(v)} - 1} \\
\pi_{11} - \pi_{+1}\pi_{1+} &= \frac{\lambda_{11} - \lambda_{+1}\lambda_{1+}}{(\theta_0^{(u)} + \theta_1^{(u)} - 1)(\theta_0^{(v)} + \theta_1^{(v)} - 1)}
\end{aligned}$$

Through deduction, AV_{ori} is expressed as:

$$AV_{ori} = \frac{\lambda_{11} - \lambda_{+1}\lambda_{1+}}{(\theta_0^{(v)} + \theta_1^{(v)} - 1)[\theta_1^{(u)} - 1 + (1 + \theta_0^{(u)} - \theta_1^{(u)})\lambda_{1+}]}$$

Let $f(\theta_0^{(u)}, \theta_1^{(u)}, \theta_0^{(v)}, \theta_1^{(v)}, \lambda_{1+}) = |(\theta_0^{(v)} + \theta_1^{(v)} - 1)[\theta_1^{(u)} - 1 + (1 + \theta_0^{(u)} - \theta_1^{(u)})\lambda_{1+}]| - |\lambda_{1+}|$,
1) When $\theta_0^{(u)}, \theta_1^{(u)}, \theta_0^{(v)}, \theta_1^{(v)} \in [0.5, 1]$, since $\pi_{1+} = \frac{\theta_1^{(u)} - 1 + (1 + \theta_0^{(u)} - \theta_1^{(u)})\lambda_{1+}}{\theta_0^{(u)} + \theta_1^{(u)} - 1} \geq 0$, then $\theta_1^{(u)} - 1 + (1 + \theta_0^{(u)} - \theta_1^{(u)})\lambda_{1+} \geq 0$, we have

$$\begin{aligned}
f(\theta_0^{(u)}, \theta_1^{(u)}, \theta_0^{(v)}, \theta_1^{(v)}, \lambda_{1+}) &= (\theta_0^{(v)} + \theta_1^{(v)} - 1)[\theta_1^{(u)} - 1 + (1 + \theta_0^{(u)} - \theta_1^{(u)})\lambda_{1+}] - \lambda_{1+} \\
&= (\theta_0^{(v)} + \theta_1^{(v)} - 1)(\theta_1^{(u)} - 1)(1 - \lambda_{1+}) + [(\theta_0^{(v)} + \theta_1^{(v)} - 1)\theta_0^{(u)} - 1]\lambda_{1+} \\
&\leq 0
\end{aligned}$$

Hence,

$$\begin{aligned}
|AV_{ori}| &= |\frac{\lambda_{11} - \lambda_{+1}\lambda_{1+}}{(\theta_0^{(v)} + \theta_1^{(v)} - 1)[\theta_1^{(u)} - 1 + (1 + \theta_0^{(u)} - \theta_1^{(u)})\lambda_{1+}]}| \\
&\geq |\frac{\lambda_{11} - \lambda_{+1}\lambda_{1+}}{\lambda_{1+}}| \\
&\geq |AV_{ran}|
\end{aligned}$$

2) When $\theta_0^{(u)}, \theta_1^{(u)}, \theta_0^{(v)}, \theta_1^{(v)} \in [0, 0.5]$, since $\theta_1^{(u)} - 1 + (1 + \theta_0^{(u)} - \theta_1^{(u)})\lambda_{1+} \geq 0$, we have

$$f(\theta_0^{(u)}, \theta_1^{(u)}, \theta_0^{(v)}, \theta_1^{(v)}, \lambda_{1+}) = (\theta_0^{(v)} + \theta_1^{(v)} - 1)(\theta_1^{(u)} - 1)(1 - \lambda_{1+}) + [(\theta_0^{(v)} + \theta_1^{(v)} - 1)\theta_0^{(u)} - 1]\lambda_{1+}$$

when $\lambda_{1+} \geq \frac{(\theta_0^{(v)} + \theta_1^{(v)} - 1)(\theta_1^{(u)} - 1)}{1 - (\theta_0^{(u)} + \theta_1^{(v)} - 1)(1 + \theta_0^{(u)} - \theta_1^{(u)})}$

$$f(\theta_0^{(u)}, \theta_1^{(u)}, \theta_0^{(v)}, \theta_1^{(v)}, \lambda_{1+}) \leq 0, \quad |AV_{ori}| \geq |AV_{ran}|$$

when $\lambda_{1+} < \frac{(\theta_0^{(v)} + \theta_1^{(v)} - 1)(\theta_1^{(u)} - 1)}{1 - (\theta_0^{(v)} + \theta_1^{(v)} - 1)(1 + \theta_0^{(u)} - \theta_1^{(u)})}$

$$f(\theta_0^{(u)}, \theta_1^{(u)}, \theta_0^{(v)}, \theta_1^{(v)}, \lambda_{1+}) > 0, |AV_{ori}| < |AV_{ran}|$$

Similarly, we can prove that $|AV_{ori}| \geq |AV_{ran}|$ is not always held when $\theta_0^{(u)}, \theta_1^{(u)}, \theta_0^{(v)}, \theta_1^{(v)} \notin [0.5, 1]$.

Proof of Result 3 and Result 4

For any pair of variables, Piatetsky-Shapiro's measure calculated directly from the randomized data without knowing $\theta_0^{(u)}, \theta_1^{(u)}, \theta_0^{(v)}, \theta_1^{(v)}$ is:

$$PS_{ran} = \lambda_{11} - \lambda_{1+}\lambda_{+1} = \lambda_{00}\lambda_{11} - \lambda_{01}\lambda_{10}$$

The original Piatetsky-Shapiro's measure is:

$$PS_{ori} = \pi_{11} - \pi_{1+}\pi_{+1} = \frac{PS_{ran}}{(\theta_0^{(u)} + \theta_1^{(u)} - 1)(\theta_0^{(v)} + \theta_1^{(v)} - 1)}$$

$$|PS_{ori}^{u,v}| - |PS_{ori}^{s,t}| = \frac{|PS_{ran}^{u,v}| - |PS_{ran}^{s,t}|}{|(\theta_0^{(u)} + \theta_1^{(u)} - 1)(\theta_0^{(v)} + \theta_1^{(v)} - 1)|}$$

So $\forall \theta_0^{(u)}, \theta_1^{(u)}, \theta_0^{(v)}, \theta_1^{(v)} \in [0,1]$, $\frac{1}{|(\theta_0^{(u)}+\theta_1^{(u)}-1)(\theta_0^{(v)}+\theta_1^{(v)}-1)|} \geq 1$. Result 3 is proved.

Since

$$D_{ran} = \frac{\lambda_{00}}{\lambda_{+0}} - \frac{\lambda_{01}}{\lambda_{+1}} = \frac{\lambda_{00}\lambda_{11} - \lambda_{01}\lambda_{10}}{\lambda_{+0}\lambda_{+1}}$$

$$D_{ori} = \frac{\pi_{00}\pi_{11} - \pi_{01}\pi_{10}}{\pi_{+0}\pi_{+1}} = \frac{\lambda_{00}\lambda_{11} - \lambda_{01}\lambda_{10}}{(\theta_0^{(u)} + \theta_1^{(u)} - 1)\lambda_{+0}\lambda_{+1}}$$

We have $D_{ori} = \frac{1}{(\theta_0^{(u)}+\theta_1^{(u)}-1)} D_{ran}$. Hence,

$$|D_{ori}^{u,s}| - |D_{ori}^{v,t}| = \frac{1}{|\theta_0^{(u)} + \theta_1^{(u)} - 1|}(|D_{ran}^{u,s}| - |D_{ran}^{v,t}|)$$

We can show AV also holds. Result 4 is proved.

Proof of Result 9

The proof is given for three binary variables with the saturated model; the extension to higher dimensions is immediate. Equation 9 shows how to compute the coefficients for the model of variables A, B, C, where a dot "." means that the parameter has been summed over the index.

$$\begin{aligned}
\gamma &= l_{...} \\
\gamma_i^A &= l_{i..} - \gamma \\
&\cdots \\
\gamma_{ij}^{AB} &= l_{ij.} - \gamma_i^A - \gamma_j^B - \gamma \\
&\cdots \\
\gamma_{ijk}^{ABC} &= l_{ijk} - \gamma_{ij}^{AB} - \gamma_{ik}^{AC} - \gamma_{jk}^{BC} - \gamma_i^A - \gamma_j^B - \gamma_k^C - \gamma
\end{aligned} \tag{9}$$

From randomized data we get:

$$\gamma_{0ran}^A = \frac{1}{8}\log\frac{\lambda_{000}\lambda_{001}\lambda_{010}\lambda_{011}}{\lambda_{100}\lambda_{101}\lambda_{110}\lambda_{111}}$$

Similarly, we have:

$$\gamma_{0ori}^A = \frac{1}{8}\log\frac{\pi_{000}\pi_{001}\pi_{010}\pi_{011}}{\pi_{100}\pi_{101}\pi_{110}\pi_{111}}$$

There is no monotonic relation between λ_{ijk} and π_{ijk} $(i, j, k = 0, 1)$. γ^A can be greater or less than the original value after randomization. Same results can be proved for other γ parameters. Result 9 is proved.