RANDOMIZATION BASED PRIVACY PRESERVING CATEGORICAL DATA ANALYSIS

by

Ling Guo

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Information Technology

Charlotte

2010

Approved by:

Dr. Xintao Wu

Dr. Zbigniew W. Ras

Dr. Mohamed Shehab

Dr. Zhiyi Zhang

Dr. Ming Dai

ABSTRACT

LING GUO. Randomization based privacy preserving categorical data analysis.
(Under the direction of DR. XINTAO WU)

The success of data mining relies on the availability of high quality data. To ensure quality data mining, effective information sharing between organizations becomes a vital requirement in today's society. Since data mining often involves sensitive information of individuals, the public has expressed a deep concern about their privacy. Privacy-preserving data mining is a study of eliminating privacy threats while, at the same time, preserving useful information in the released data for data mining.

This dissertation investigates data utility and privacy of randomization-based models in privacy preserving data mining for categorical data. For the analysis of data utility in randomization model, we first investigate the accuracy analysis for association rule mining in market basket data. Then we propose a general framework to conduct theoretical analysis on how the randomization process affects the accuracy of various measures adopted in categorical data analysis.

We also examine data utility when randomization mechanisms are not provided to data miners to achieve better privacy. We investigate how various objective association measures between two variables may be affected by randomization. We then extend it to multiple variables by examining the feasibility of hierarchical loglinear modeling. Our results provide 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.

Data privacy and data utility are commonly considered as a pair of conflicting requirements in privacy preserving data mining applications. In this dissertation, we investigate privacy issues in randomization models. In particular, we focus on the attribute disclosure under linking attack in data publishing. We propose efficient solutions to determine optimal distortion parameters such that we can maximize utility

preservation while still satisfying privacy requirements. We compare our randomization approach with l-diversity and anatomy in terms of utility preservation (under the same privacy requirements) from three aspects (reconstructed distributions, accuracy of answering queries, and preservation of correlations). Our empirical results show that randomization incurs significantly smaller utility loss.

ACKNOWLEDGMENTS

First of all, I would like to gratefully and sincerely thank Dr. Xintao Wu for his guidance, understanding, encouragement and patience during my graduate studies. Besides providing me with a terrific environment for doing research, his enthusiasm to work also inspired me to push myself for better. This dissertation would not be possible without his guidance and encouragement.

Thanks are also due to Dr. Zbigniew W. Ras, Dr. Mohamed Shehab, Dr. Zhiyi Zhang and Dr. Ming Dai for serving as my committee members and giving me precious advice to better my work. I thank them for their contribution and their support.

I am thankful to all those at KDD Lab and Data Privacy Lab, past and present. In particular, I thank Songtao Guo, Yong Ye, Xiaowei Ying, Leting Wu and Kai Pan for the friendships and the mind-sparking discussions and suggestions at different phases of this dissertation.

My special gratitude goes to my family. I thank my parents Jianyuan Guo and Siying Jia for their endless love, encouragement, and support throughout my life. I thank my brother Qing Guo and sister-in-law Yuanyuan Zhou, who always be there to share my bitterness and happiness. My special thank goes to my husband Chunlin Liu who loves, supports and encourages me all the time. To them I dedicate this dissertation.

My research was supported by U.S. NSF Grant IIS-0546027. I was also supported by the Department of Software and Information System.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Motivation

With the advance of information technologies, the amount of information collected by different entities is increasing exponentially. Agencies and other organizations often need to publish microdata, e.g., medical data or census data, for research and other purposes. Typically, such data is stored in a table, and each record (row) corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories. (1)*identity* attributes (ID): attributes that identify individuals, e.g., Name and SSN. (2) *quasi-identifier* (QI): attributes which include demographic attributes such as ZIP code, age, gender. (3) *sensitive* attributes: attributes which indicate confidential information of individuals, e.g., disease and salary. Table 1.1 shows one example of financial microdata.

Table 1.1: Personal information of n customers

ID	SSN	Name	Zip	Race	···	Age	Gender	Balance ($1,000)	Income ($1,000)	···	Interest Paid ($1,000)
1	***	***	28223	Asian	···	20	M	10	85	···	2
2	***	***	28223	Asian	···	30	F	15	70	···	18
3	***	***	28262	Black	···	20	M	50	120	···	35
4	***	***	28261	White	···	26	M	45	23	···	134
···	···	···	···	···	···	···	···	···	···	···	···
n	***	***	28223	Asian	···	20	M	80	110	···	15

Identifying attributes are typically removed from microdata records prior to release. But QI attributes may be linked with other public database to disclose the identity of individuals and their sensitive attributes values. A recent study [66] showed that 87% of the population of the United States can be uniquely identified using the seemingly innocuous attributes gender, date of birth, and 5-digit zip code. Authors in [67]

pointed out that these three attributes in anonymized medical data from GIC1 (which included gender, zip code, date of birth and diagnosis) can be linked to Massachusetts voter registration records (which included the name, gender, zip code, and date of birth). Such "linking attack" uniquely identified the medical records of the governor of Massachusetts in the medical data. Privacy is becoming an increasingly important issue in many data mining applications. This has spawned a new research field called privacy preserving data mining (PPDM).

One important issue of PPDM is how to transform data such that a good data mining model can be built on transformed data while preserving privacy at the record level. Data utility and privacy is a trade-off in PPDM. Many methods have been proposed in the literature to transform data for privacy preservation, e.g., k-anonymity [10,42,43,57,58], l-diversity [49], Anatomy [76], randomization [13,48] and so on. In this dissertation, we focus on the data utility and privacy of randomization models.

1.2 Research Statement

My research focus on the analysis of data utility and privacy of randomization-based models in privacy preserving data mining for categorical data. Specifically, my dissertation aims to

1. Analyze the accuracy of estimations and data mining results on transformed data for various perturbation models, especially on randomization-based models in categorical data mining.

2. Analyze data utility of randomized data in privacy preserving data mining without the information of randomization parameters.

3. Investigate the privacy of randomization-based models in privacy preserving categorical data mining.

4. Investigate the attribute disclosure of randomization-based models under linking attacks.

1.3 Dissertation Contributions

This dissertation presents a formal and comprehensive examination of data utility and privacy of randomization-based models in privacy preserving categorical data analysis. The main contributions can be summarized as follows:

- *Accuracy analysis for association rule mining in market basket data.* We investigated the accuracy (in terms of bias and variance of estimates) of both support and confidence estimates of association rules derived from the randomized data. We proposed the novel idea of using interquantile range to bound those estimates derived from the randomized market basket data. We demonstrated that providing confidence on data mining results from randomized data is significant to data miners. They can know how accurate their data mining results are under randomization-based models.

- *A general framework to evaluate the accuracy of estimates of various measures adopted in categorical data analysis .* We presented a general approach to derive variances of estimates of various measures adopted in categorical data analysis. We applied the idea of using interquantile ranges based on Chebyshevs Theorem to bound those estimates derived from the randomized data.

- *Data utility analysis in randomization with unknown distortion parameters.* We investigated whether data mining or statistical analysis tasks can still be conducted on randomized data when distortion parameters are not disclosed to data miners. We examined how various objective association measures between two variables may be affected by randomization. We demonstrated that some measures 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 can be executed on the randomized data directly even without knowing distortion parameters. We then investigated how

the relative order of two association patterns is affected when the same randomization is conducted. We showed that some measures 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. We then extended it to multiple variables by examining the feasibility of hierarchical loglinear modeling. We showed that some classic data mining tasks (e.g., association rule mining, decision tree learning, naive bayes classifier) cannot be applied on the randomized data directly.

- *Analysis of attribute disclosure under linking attacks.* We presented a systematic study of randomization method in preventing attribute disclosure under linking attacks. We proposed a general framework and presented a uniform definition for attribute disclosure which is compatible for both randomization and generalization models.

- *Efficient solution for randomization parameters under linking attacks.* We proposed the use of a specific randomization model. We presented an efficient solution to derive distortion parameters to satisfy requirements for privacy preservation while maximizing data utilities. We compared randomization model with other distortion models, k-anonymity, l-diversity and Anonymity. Our experimental evaluations showed that randomization signicantly outperforms generalization, i.e., achieving better utility preservation while yielding the same privacy protection.

1.4 Dissertation Organization

The remainder of this dissertation is organized as follows:

In Chapter 2, the current research on privacy preserving data mining is briefly reviewed. Various techniques in the privacy preserving data mining, including group

based anonymization, randomization and synthetic microdata generation are introduced. Distortion framework adopted in this dissertation is presented.

In Chapter 3, the accuracy of estimates of various rules derived from the randomized market basket data are investigated. A general framework is presented which can conduct theoretical analysis on how the randomization process affects the accuracy of various measures adopted in categorical data analysis.

In Chapter 4, data utility of randomized data when distortion parameters are not disclosed is investigated. Various objective association measures between two variables are examined and extended to multiple variables by examining the feasibility of hierarchical loglinear modeling.

In Chapter 5, privacy and utility of randomization model under linking attacks are investigated. A uniform definition for attribute disclosure which is compatible for both randomization and generalization models are defined. An efficient solution is proposed to derive distortion parameters to satisfy requirements for privacy preservation while maximizing data utilities.

Chapter 6 concludes this dissertation with a brief summary of the research presented and offers the future directions.

CHAPTER 2: BACKGROUND AND FRAMEWORK

Privacy is becoming an increasingly important issue in many data mining applications. A number of techniques and algorithms have been proposed to modify or transform the data so as to obtain valid data mining results while preserving privacy at different levels. In this chapter, we first overview the existing privacy preserving data mining techniques and outline the important research issues within them in Section 2.1. In Section 2.2, we present the randomized response distortion framework for categorical data that we implemented in this dissertation.

2.1 Privacy Preserving Data Mining Models and Algorithms

We classify representative privacy preserving data mining techniques into three categories, group based anonymization, randomization and synthetic microdata generation.

2.1.1 Group Based Anonymization

In data publishing, uniquely identifying information like names and social security numbers are usually removed before the data are submitted to third party for mining or published to the public. However, this first sanitization does not ensure the privacy of individuals in the data. Attackers may already know the quasi-identifier values of some individuals in the table. This knowledge can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly-available databases (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers. Table 2.1 provides an example of 8 customers' personal health information. Supposing attackers has the personal information (i.e., age 40 and zipcode 13000) of Bob and learn that Bob's health record is included in Table 2.1. Because

Table 2.1: The microdata

ID	Age	Gender	Zipcode	Disease
1	20	M	11000	pneumonia
2	40	M	13000	flu
3	40	M	59000	flu
4	50	M	12000	flu
5	60	F	54000	pneumonia
6	60	F	25000	pneumonia
7	80	F	25000	flu
8	80	F	30000	flu

only tuple 2 matches Bob's QI values, the attackers can assert that Bob contracted flu.

To avoid this problem, group based anonymization methods replace quasi-identifier values with values that are less-specific but semantically consistent. As a result, more records will have the same set of quasi-identifier values. Authors in [57] defined an equivalence class of an anonymized table to be a set of records that have the same values for the quasi-identifiers.

k-anonymity

To counter linking attacks using quasi identifiers, Samarati and Sweeney [58] proposed a definition of privacy called k-anonymity.

Definition 1 *(k-anonymity) A table satisfies k-anonymity if every record in the table is indistinguishable from at least $k-1$ other records with respect to every set of quasi-identifier attributes; such a table is called a k-anonymous table.*

In other words, for every combination of values of the quasi-identifiers in the k-anonymous table, there are at least k records that share those values. This ensures that individuals cannot be uniquely identified by linking attacks. Table 2.2 shows a 4-anonymity for original Table 2.1. *Age* are generalized to the interval $[20, 50], [60, 80]$

Table 2.2: A 4-Anonymity table with 2-Diversity

ID	Age	Gender	Zipcode	Disease
1	[20,50]	M	[10001,60000]	pneumonia
2	[20,50]	M	[10001,60000]	flu
3	[20,50]	M	[10001,60000]	flu
4	[20,50]	M	[10001,60000]	flu
5	[60,80]	F	[10001,60000]	pneumonia
6	[60,80]	F	[10001,60000]	pneumonia
7	[60,80]	F	[10001,60000]	flu
8	[60,80]	F	[10001,60000]	flu

and *Zipcode* to $[10001, 60000]$. Because any one of tuple 1 to tuple 4 can be the record of Bob, Bob can not be identified by attackers from other 3 individuals.

Generalization and suppression algorithms have been well developed to implement k anonymization in categorical data. Generalization involves replacing specific values such as "male" or "female" with a more general one, such as "gender". Suppression is the process of deleting attribute values or entire tuples.

The first algorithm for AG_TS (i.e., generalization over quasi-identifier attributes and tuple suppression) was proposed by Samarati [57]. He introduced the concept of minimal generalization. It captures the property of the generalization process not to distort the data more than needed to achieve k anonymity, and proposed a full-domain generalization algorithm. Bayardo and Agrawal [10] developed an optimal k anonymization algorithm and investigated the impacts of various coding techniques and problem variations on anonymization quality. The proposed algorithm starts with a fully generalized dataset and systematically specializes the dataset into one that is minimally k anonymous. LeFevre et al. [42] proposed Incognito to implement k-anonymous full-domain generalizations using bottom-up aggregation along generalization dimensions and a priori computation. Later in [43], they proposed a new multidimensional model, which provides an additional degree of flexibility not seen in previous (single-dimensional) approaches [10, 42, 57]. Theoretical analysis shows that

the problem of optimal k-anonymity is NP-hard for $k \geq 3$ [1, 10, 51]. A good survey of the algorithms can be found in [15].

However, the generalization and suppression approach proposed in the literature to achieve k-anonymity is not equally suited for all types of attributes, e.g., continuous attributes. Authors in [2] proposed to use data condensation to achieve k-anonymity for the numerical attributes. Authors in [18] proposed to use categorical microaggregation as an alternative to generalization/suppression for nominal and ordinal k-anonymization; they also proposed to use continuous microaggregation to implement continuous k-anonymization.

k-anonymity has been widely adopted in data publishing because of its conceptual simplicity. Nevertheless the technique is susceptible to the following attacks as discussed in [49]:

- **Homogeneity Attack:** In an anonymity table, if there exists an equivalence class in which all tuples share the same value of sensitive attributes, it will be exposed to homogeneity attack, for adversaries can easily infer individuals sensitive values by linking external table. Therefore, although the data is k-anonymized, the value of the sensitive attribute in that equivalence class is disclosed.

- **Background knowledge attack:** An adversary can infer individuals' sensitive information from anonymity table using his/her background knowledge. In [49], the background knowledge that Japanese have an extremely low incidence of heart disease helped attackers narrow down information of what disease the patient might have.

l-diversity

To address the limitations of k-anonymity, Machanavajjhala et al. [49] introduced l-diversity as a stronger notion of privacy.

Definition 2 *(l-diversity) An equivalence class is said to have l-diversity if there are at least l well-represented values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity.*

Table 2.2 shows an example of 2-diversity as each equivalence class has two values for the sensitive attribute *disease*.

Machanavajjhala et al. [49] gave a number of interpretations of the term "well-represented" in this principle:

- **Distinct l-diversity.** A table is said to have distinct l-diversity if for every equivalence class E, there are at least l distinct values for the sensitive attribute. However, an equivalence class may have one value appear much more frequently than other values. It makes the attacker to deduce that an entity in the equivalence class is very likely to have that value. This motivated the development of the following two stronger notions of l-diversity.

- **Entropy l-diversity.** A table is said to have entropy l-diversity if for every equivalence class E, $Entropy(E) \geq \log l$. $Entropy(E)$ is defined to be:

$$Entropy(E) = -\Sigma_{s \in S} p(E, s) \log p(E, s)$$

  Here S is the domain of the sensitive attribute, and $p(E, s)$ is the fraction of records in E that have sensitive value s. The entropy of the entire table must be at least $\log l$ to have entropy l-diversity for each equivalence class.

- **Recursive (c, l)-diversity.** An equivalence class E is said to have recursive (c, l)-diversity if $r_1 < c(r_l + r_{l+1} + ... + r_m)$. Here m is the number of sensitive values and r_i $(1 \leq i \leq m)$ is the frequency of the ith most frequent sensitive value. It makes sure that the most frequent value does not appear too frequently, and the less frequent values do not appear too rarely. A table is said

to have recursive (c, l)-diversity if all of its equivalence classes have recursive (c, l)-diversity.

Authors in [45] implemented l-diversity on top of Incognito and suggested that any k-anonymity technique can be adapted for l-diversity. Although l-diversity can better protect against attribute disclosures than k-anonymity, authors in [45] pointed out that it is vulnerable to skewness attack and similarity attack, and it is difficult and unnecessary to achieve for some specific data set.

t-closeness

Li et al. [45] proposed a novel privacy notion called t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table.

Definition 3 *(t-closeness) An equivalence class is defined to have t-closeness property if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.*

The assumption in [45] is that the distribution Q of the sensitive attribute in the table is known to attackers. Given the anonymized table, the attacker can identified the equivalence class that the individual's record is in and learn the distribution P of the sensitive attribute in that class. The Earth Mover's distance (EMD) [56] is adopted to measure the distance between Q and P, which takes into consideration the semantic closeness of attribute values.

Anatomy

Xiao and Tao [76] proposed anatomy for publishing sensitive data based on the privacy requirement of l-diversity. Anatomy releases all the quasi-identifier and sensitive values directly in two separate tables. Let T be the microdata which contains d

quasi-identifier (QI) attributes $A_1^{qi}, A_2^{qi}, \ldots, A_d^{qi}$ and a sensitive attribute A^s. For any tuple $t \in T$, $t[i]$ is the A_i^{qi} value of t and $t[d+1]$ is its A^s value. Anatomy is defined in [76] as follows:

Definition 4 *(Anatomy) Given an l-diverse partition with m QI-groups, anatomy produces a quasi-identifier table (QIT) and a sensitive table(ST).*

The QIT has schema $(A_1^{qi}, A_2^{qi}, \ldots, A_d^{qi}, Group-ID)$ and has a tuple of the form: $(t[1], t[2], \ldots, t[d], j)$ for each QI-group QI_j $(1 \leq j \leq m)$ and each tuple $t \in QI_j$.

The ST has schema $(Group{-}ID, A^s, Count)$ and has a record of the form: $(j, v, c_j(v))$ for each QI-group QI_j $(1 \leq j \leq m)$ and each distinct A^s value v in QI_j.

They developed a linear-time algorithm for computing anatomized tables that minimize error of reconstructing the mocrodata while satisfying the l-diversity privacy requirement. Their experimental results showed that anatomy provide more accuracy than the generalization based methods. Table 2.3 shows the anatomized tables which also satisfy 2-diversity for each group. Although attackers can derive from QIT Table 2.3(a) that tuple 2 is the record of Bob, they can not deduce which disease he contracted from Table 2.3(b).

Goodness and Weakness

Although group based anonymization preserves privacy and true information of individuals, it often loses considerable information in the microdata, which severely compromises the accuracy of data analysis. Besides, the aonymization process depends on the knowledge of all the records in the data, so it cannot be implemented at data collection time. It requires a trusted server to perform the anonymization on the whole data set.

2.1.2 The Randomization Method

Randomization was initially used in the context of survey which have privacy concerns [46, 72]. It was introduced to preserve privacy data mining by Agrawal and

Table 2.3: The anatomized tables

(a) The quasi-identifier table (QIT)

ID	Age	Gender	Zipcode	Group-ID
1	20	M	11000	1
2	40	M	13000	1
3	40	M	59000	1
4	50	M	12000	1
5	60	F	54000	2
6	60	F	25000	2
7	80	F	25000	2
8	80	F	30000	2

(b) The sensitive table (ST)

Group-ID	Disease	Count
1	flu	3
1	pneumonia	1
2	flu	2
2	pneumonia	2

Srikat [4]. In randomization, noise was added to the data so that the individual values of the records cannot be recovered. However the probability distribution of the aggregate data can be recovered and be used for data mining. Representative randomization methods include the additive-noise-based perturbation, the projection-based perturbation and the Randomized Response.

Additive-noise-based perturbation

The additive-noise-based perturbation model was first proposed by Agrawal and Srikant [4] for building decision-tree classifiers. It can be described as follows:

$$Y = X + E$$

We use X to denote the original data, E to denote the additive noise and Y to represent the perturbed data. Let X_j be the j-th column of the original microdata table corresponding to a sensitive attribute and suppose that there are N tuples.

Each value x_{ij} $(i = 1, \ldots, N)$, is replaced by:

$$y_{ij} = x_{ij} + \epsilon_{ij}$$

Where ϵ_j is a vector of normally distributed errors drawn from a random variable with mean equals to zero. In *uncorrelated noise addition*, it satisfies $\epsilon_j \sim N(0, \sigma^2_{\epsilon_j})$ and $\sigma^2_{\epsilon_j} = \alpha \cdot \sigma^2_{X_j}$ (α is the proportional coefficient). This method can preserves the mean and the co-variance of the original data. In *correlated noise addition*, the covariance matrix of errors is proportional to the covariance matrix of original data ($\epsilon \sim N(0, \alpha\sigma_\epsilon)$), it can preserve the mean and correlation of the original data. Additive-noise-based perturbation is often combined with linear (for continuous attributes [39]) or non linear (for categorical attributes [65]) transformations to provide more protection for the data.

However, this kind of randomization is not secure under some attacks. Kargupta et al. [37] proposed a random matrix-based Spectral Filtering (SF) technique which can recover the original data from the perturbed data. In [36], Huang et al. proposed two data reconstruction methods that are based on data correlations. One method uses the Principal Component Analysis (PCA) technique, while the other one uses the Bayes Estimate (BE) technique. In [31, 32], Guo et al. analyzed the spectral-filtering-based method theoretically and improved the spectrum selecting strategy to achieve the optimal performance. Another contribution is that they provided the bounds of the reconstruction errors, which is meaningful to both the data miner and the attacker. They further proposed IQR attack on additive-noise-based model in [34]. They presented that the individual privacy can be threatened by the estimated distribution according to the defined privacy quantification.

Projection-based perturbation

Table 2.4: Randomized data

ID	Age	Gender	Zipcode	Disease
1	20	M	11000	pneumonia
2	40	*F*	13000	flu
3	40	M	59000	*pneumonia*
4	50	M	12000	flu
5	*40*	F	*25000*	pneumonia
6	60	F	*13000*	*flu*
7	*70*	F	25000	flu
8	80	F	30000	flu

The projection-noise-based perturbation model is defined as follows:

$$Y = RX$$

Where $X \in \mathbb{R}^{p \times n}$ is the original data set consisting of n data records and p attributes. $Y \in \mathbb{R}^{q \times n}$ is the transformed data set consisting of n data records and q attributes. R is a $q \times p$ transformation matrix.

Chen and Liu [13] proposed a rotation based perturbation method, where the transformation matrix R is a $d \times d$ orthogonormal matrix satisfying $R^T R = RR^T = I$. Under this definition of R, the vector length, Euclidean distance and inner product between any pair of points are preserved. Three popular classifiers (kernel method, SVM, and hyperplane-based classifiers) are invariant to such perturbation.

Authors in [47] discussed a Principal Component Analysis(PCA) based attack on the above transformation. They further proposed a random projection-based multiplicative perturbation scheme in [48]. Each entry of R is independent and identically chosen from some normal distribution with mean zero. Authors in [33] proposed an *A-priori-Knowledge* ICA (AK-ICA) reconstruction method, which may be exploited by attackers when a small subset of sample data is available to attackers.

Randomized Response technique

Randomized Response technique was first introduced by Warner [72] in 1965 as a technique to solve the following survey problem: to estimate the percentage of people in a population that has attribute A. Since the attribute A is related to some confidential aspects of human life, respondents may decide not to reply at all or to reply with incorrect answers.

In this model, instead of asking each respondent whether he/she has attribute A, the interviewer asks each respondent two related questions:

- I have the sensitive attribute A.
- I do not have the sensitive attribute A.

Respondents use a randomizing device to decide which question to answer, without letting the interviewer know which question is answered. The randomizing device is designed in such a way that the probability of choosing the first question is θ, and the probability of choosing the second question is $1-\theta$. Although the interviewer learns the responses (e.g. yes or no), he/she does not know which question was answered by the respondents. Thus the respondents privacy is preserved. To estimate the percentage of people who has the attribute A, we can use the following equations:

$$P^*(A = yes) = P(A = yes) * \theta + P(A = no) * (1 - \theta)$$
$$P^*(A = no) = P(A = no) * \theta + P(A = yes) * (1 - \theta)$$

where $P^*(A = yes)$ (resp. $P^*(A = no)$) is the proportion of the yes (resp. no) responses obtained from the survey data, and $P(A = yes)$ (resp. $P(A = no)$) is the estimated proportion of the yes (resp. no) responses to the sensitive questions. Getting P(A = yes) and P(A = no) is the goal of the survey, which can be estimated from the above equations. Further background and more complex randomized response schemes can be found in [11].

Authors in [26] proposed the Post RAndomisation Method (PRAM) as a perturbative method for disclosure protection of categorical variables. PRAM can be seen as applying RR after the data have been collected. They also discussed an invariant PRAM in which the distribution of the distorted data is the same as the original one, so most data mining application can be conducted on randomized data directly. A full comparison between RR and PRAM can be found in [17].

The authors in [54] proposed the MASK scheme, which is based on Randomized Response. They presented strategies of efficiently estimating the original support values of frequent itemsets from the randomized data. Their results empirically showed a high degree of privacy to the user and a high level of accuracy in the mining results can be simultaneously achieved. Agrawal and Haritsa [6] presented a general framework of random perturbation in privacy preserving data mining. Du and Zhan [20] studied the use of randomized response technique to build decision tree classifiers. Zhu and Liu [85] investigated the construction of optimal randomization schemes for privacy preserving density estimation and proposed a general framework for randomization using mixture models. Recently, Huang and Du [35] studied the search of optimal distortion parameters to balance privacy and utility. Similarly, Xiao et. al. [77] investigated the optimal random perturbation at multiple privacy levels.

Guo et. al. [27,28] investigated data utility in terms of the accuracy of reconstructed measures in privacy preserving market basket data analysis. they presented a general method based on the Taylor series to approximate the mean and variance of estimated variables from the randomized data. They also discussed the data utility of randomized data for data miners with unknown distortion parameters in [29]. In [30], they investigated the attribute disclosure under linking attack in privacy preserving data publishing. They presented an efficient solution to derive distortion parameters to satisfy requirements for privacy preservation while maximizing data utilities. They compared randomization model with other distortion models, k-anonymity, l-diversity

and Anonymity. Their experimental evaluations showed that randomization significantly outperforms generalization, i.e., achieving better utility preservation while yielding the same privacy protection.

2.1.3 Synthetic Microdata Generation

Publication of synthetic data is to generate data randomly with the constraint that certain statistics or correlations of the original data is preserved. The idea was first proposed in [55] to generate a synthetic dataset based on the original data and multiple imputation. Later, authors in [59] proposed to use bootstrap method to generate synthetic microdata and it was used for categorical data in [60].

Wu et al. [71, 73–75] proposed a general framework for privacy preserving database application testing by generating synthetic data sets based on some a-priori knowledge about the production databases. Their approach is to fit the general location model using various characteristics (e.g., constraints, statistics, rules) extracted from a production database and then generate synthetic data using model learned. The generated data is valid and similar to real data in terms of statistical distribution, hence it can be used for functional and performance testing.

2.2 Distortion Framework for Categorical Data

In this dissertation, we investigate data utility and data privacy in categorical data under randomization models. In particular, we focus on a simple independent column perturbation, wherein the value of each attribute in the record is perturbed independently. Also, the perturbation is done at the level of individual customer record, without being influenced by the contents of the other records in the database. We will introduce the distortion framework in this section.

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 sen-

sitive 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.

Table 2.5: 2×3 contingency tables for two variables Gender (QI), Disease (sensitive)

(a) Original

	Cancer	Flu	Anemia	
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) Instance

	Cancer	Flu	Anemia
Male	8	16	48
Female	12	14	2

(c) After randomization

	Cancer	Flu	Anemia	
Male	λ_{00}	λ_{01}	λ_{02}	λ_{0+}
Female	λ_{10}	λ_{11}	λ_{12}	λ_{1+}
	λ_{+0}	λ_{+1}	λ_{+2}	λ_{++}

To express the relationship among variables, we can map categorical data sets to contingency tables. Table 2.5 shows one contingency table for a pair of two variables, *Gender* and *Disease* ($d_1 = 2$ 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 *Flu*. The row sum π_{0+} represents the proportion of records with *Male* across all diseases.

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

Table 2.6: 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
$p^{(u)}$	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

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 2.6 summarizes our notations.

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 $p^{(u)}_{j0}, p^{(u)}_{j1}, ..., p^{(u)}_{j\ d_u-1}$, where $\sum_{c=0}^{d_u-1} p^{(u)}_{jc} = 1$. The distortion matrix P_u for A_u is shown as below.

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

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 2.5(c)). Let $P=P_0\times\cdots\times P_{k-1}$, we can obtain

$$\boldsymbol{\lambda}=P\boldsymbol{\pi}=(P_0\times\cdots\times P_{k-1})\boldsymbol{\pi} \tag{2.1}$$

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

$$\hat{\boldsymbol{\pi}}=P^{-1}\hat{\boldsymbol{\lambda}}=(P_0^{-1}\times\cdots\times P_{k-1}^{-1})\hat{\boldsymbol{\lambda}} \tag{2.2}$$

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.

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:

$$P_u=\begin{pmatrix} p_0^{(u)} & 1-p_1^{(u)} \\ 1-p_0^{(u)} & p_1^{(u)} \end{pmatrix} \quad P_v=\begin{pmatrix} p_0^{(v)} & 1-p_1^{(v)} \\ 1-p_0^{(v)} & p_1^{(v)} \end{pmatrix}$$

We have:

$$\lambda_{0+} \;=\; (p_0^{(u)}+p_1^{(u)}-1)\pi_{0+}-p_1^{(u)}+1$$

[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 can see that $\lambda_{0+} - \pi_{0+}$ is a function of π_{0+} , $p_0^{(u)}$, $p_1^{(u)}$, and its value may be greater or less than 0 with varying distortion parameters.

Similarly,

$$\begin{aligned} \lambda_{00} &= p_0^{(u)} p_0^{(v)} \pi_{00} + p_0^{(u)} (1 - p_1^{(v)}) \pi_{01} \\ &+ (1 - p_1^{(u)}) p_0^{(v)} \pi_{10} + (1 - p_1^{(u)})(1 - p_1^{(v)}) \pi_{11} \end{aligned}$$

$\lambda_{00} - \pi_{00}$ is a function of π_{ij} ,$p_0^{(u)}$, $p_1^{(u)}$, $p_0^{(v)}$ and $p_1^{(v)}$, no monotonic relation exists.

We follow the Moment Estimation method as shown in Equation 2.2 to get the unbiased estimate of the distribution for original data. This method has been broadly adopted in the scenarios where RR is used to perturb data for preserving privacy. Although it has good properties as computational simplicity and unbiasedness, some awkward property exists due to random errors [11,17]. That is, the estimate may fall out of the parameter space, which makes the estimate meaningless. This is one reason that Maximum Likelihood Estimation (MLE) is adopted to estimate the distribution in literature [17].

It has been proved in [17] that a good relation holds between these two methods in the scenarios of RR: The moment estimate is equal to the MLE estimate within parameter space. Based on that, we can know that moment estimate from Equation 2.2 achieves the Cramér-Rao bound as MLE does. Therefore, moment estimate is the minimum variance unbiased (MVU) estimator in RR contexts. Our later analysis on accuracy of data mining results is based on such unbiased estimate under the assumption that the estimate is within parameter space.

2.3 Summary

The problem of privacy preserving data mining has become more important recently because of the increasing ability of collecting and storing personal data and the increasing concern of data mining applications to disclose the private information. In this chapter, we provided a overview of existing PPDM techniques present in

the literature. They can be classified into *Group Based Anonymization*, *Randomization* and *Synthetic Microdata Generation*. Besides that, we introduced the distortion model for categorical data that is adopted in this dissertation.

CHAPTER 3: ACCURACY ANALYSIS WITH KNOWN DISTORTION PROBABILITIES P

One of the main challenges of the privacy preserving data mining (PPDM) algorithms is that they need to keep the data utility to an accepted level after the anonymization or randomization process. If the data quality is too degraded, the released database is useless for data mining. Therefore, data quality after perturbation are very important in the evaluation of PPDM techniques.

In this chapter, we investigate the issue of providing accuracy for data in randomization models in PPDM. We propose a general approach to derive confidence range of estimates of various measures adopted in PPDM, which is significant to data miners since they can learn how accurate their reconstructed results are. In Section 3.1, we will discuss some related work in privacy preserving association rule mining. In Section 3.2, we present our approach for accuracy analysis in privacy preserving association rule mining. Extension to general categorical data analysis is discussed in Section 3.3. Some results in this chapter were previously reported in [27, 28].

3.1 Introduction

Denoting the set of transactions in the database D by $\mathcal{T} = \{T_0, \cdots, T_{N-1}\}$ and the set of items in the database by $\mathcal{I} = \{A_0, \cdots, A_{m-1}\}$, 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 the $s(100\%)$ of the transactions in $\mathcal{T}$ 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 $\mathcal{Y}$. The rule is "interesting" if its support and confidence are greater than the user-defined thresholds.

The issue of maintaining privacy in association rule mining has attracted considerable attention in recent years [7, 22, 23, 54]. Most of techniques are based on the

data perturbation or Randomized Response (RR) approach [11], wherein the 0 or 1 (0 denotes absence of an item while 1 denotes presence of an item) in the original user transaction vector is distorted in a probabilistic manner that is disclosed to data miners.

In [54], the authors proposed the MASK technique to preserve privacy for frequent itemset mining and addressed the issue of providing efficiency in calculating the estimated support values. Their results empirically showed a high degree of privacy to users and a high level of accuracy in the mining results can be simultaneously achieved. To evaluate the privacy, they defined a privacy metric and presented an analytical formula for evaluating the privacy obtained under the metric. However, accuracy metric on data mining results was only defined in an aggregate manner as support error and identity error computed over all discovered frequent itemsets.

Authors in [7] addressed the efficiency of MASK technique and proposed a new algorithm that was referred to as EMASK. In EMASK, different distortion parameters are used for 1's and 0's in a transaction. The parameters of distortion are carefully selected beforehand and a variety of optimizations are applied in the mining process to improve the efficiency and effectiveness of the algorithm.

Authors in [22, 23] analyzed the nature of privacy breaches and presented the uniform randomization and select-a-size randomization operators in association rule mining. They derived a formula for an unbiased support estimator and its variance and investigated how to incorporate these formulae into mining algorithms.

Later, Agrawal et al. [6] proposed FRAPP (FRamework for Accuracy in Privacy-Preserving mining), a generalized matrix-theoretic framework for the design of random perturbation schemes for privacy preserving data mining. They argued that the prior techniques differ only in their choices for the perturbation matrix elements. They proposed a novel perturbation mechanism wherein the matrix elements are themselves characterized as random variables, and demonstrated that this feature

provides significant improvements in privacy at only a marginal reduction in accuracy.

3.2 Accuracy in Privacy Preserving Association Rule Mining

3.2.1 Motivation

Our research moves one step further to address the issue of providing accuracy in privacy preserving mining of association rules. We investigate the issue of how the accuracy (i.e., support and confidence) of each association rule mined from randomized data is affected when the randomized response technique is applied.

Specifically, we present an analytical formula for evaluating the accuracy (in terms of bias and variance of estimates) of both support and confidence measures of association rules derived from the randomized data. From the derived bias and variance of estimates, we further derive approximate interquantile ranges. Data miners are ensured that their estimates lie within these ranges with a high confidence, say 95%. We would emphasize that providing confidence on estimated data mining results is significant to data miners since they can learn how accurate their reconstructed results are. We illustrate the importance of those estimated interquantile ranges using an example.

Figure 3.1 shows the original support values, the estimated support values from the randomized data, and their corresponding 95% interquantile ranges of 7 association rules, which are derived from COIL data sets [2]. A distortion parameter $p = 0.65$ and support threshold $sup_{min} = 23\%$ are used in the experiment. The interquantile range of each rule can give data miners confidence about their estimate derived from randomized data. For example, the estimated support of rule 2 is 31.5% and its 95% interquantile range is [23.8%,39.1%], which suggests that the original support value lies in this range with 95% probability. Furthermore, we can observe the 95% interquantile ranges for rules 1-3 are above the support threshold, which guarantees

[2] http://kdd.ics.uci.edu/databases/tic/tic.html

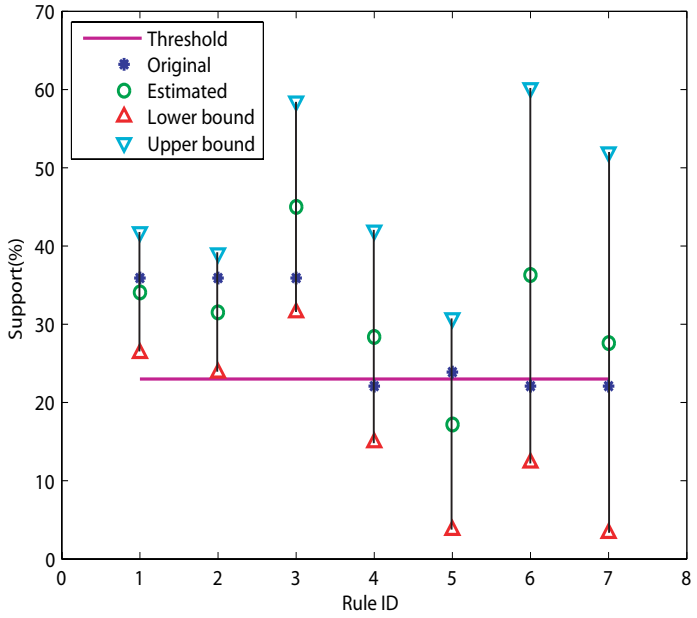


Figure 3.1: Accuracy of the estimated support values of association rules derived from randomized data with p=0.65

those are true frequent itemsets (with at least 95% confidence).

We emphasize providing accuracy of data mining results is important for data miners during data exploration. When the support threshold is set as 23%, we may not only take rule 2 and 6 as frequent sets from the estimated support values, but also conclude rule 6 (35.9%) is more frequent than rule 2 (31.5%). However, rule 2 has the original support as 36.3% while rule 6 has the original support as 22.1%, we mistakenly assign the infrequent itemset 6 as frequent. By using the derived interquantile ranges, we can determine that rule 2 is frequent with high confidence (since its lower bound 23.8% is above the support threshold) and rule 6 may be infrequent (since its lower bound 12.3% is below the support threshold).

3.2.2 Accuracy on Support s

In the scenario of market basket data mining, for each binary variable A_j, which only has two categories (0 = absence, 1 = presence), the distortion parameters can be shown as:

$$P_j = \begin{pmatrix} p_0 & 1-p_1 \\ 1-p_0 & p_1 \end{pmatrix} \tag{3.1}$$

If the original value is in the *absence* (*presence*) category, it will be kept in such category with a probability p_0 (p_1) and changed to *presence* (*absence*) category with a probability $1-p_0$ ($1-p_1$). To make derivations simple, we follow the original Warner Model by setting $p_0 = p_1 = p_j$ and use p_j to denote the distortion parameters. This setting indicates users have the same level of privacy for both 1's and 0's. In general, customers may expect more privacy for their 1's than for their 0's, since the 1's denote specific actions whereas the 0's are the default options.

Denote $\boldsymbol{\pi}^{(j)} = (\pi_0^{(j)}, \pi_1^{(j)})'$ ($\boldsymbol{\lambda}^{(j)} = (\lambda_0^{(j)}, \lambda_1^{(j)})'$) as the vector of marginal proportions corresponding to item A_j in the original (randomized) data set, where $j = 0, \cdots, m-1$. We have

$$\boldsymbol{\lambda}^{(j)} = P_j \boldsymbol{\pi}^{(j)} \tag{3.2}$$

Note that each vector $\boldsymbol{\pi}^{(j)}$ has two values $\pi_0^{(j)}, \pi_1^{(j)}$ and the latter corresponds to the support value of item A_j. For a market data set with N transactions, let $\hat{\boldsymbol{\lambda}}^{(j)}$ be the vector of sample proportions corresponding to $\boldsymbol{\lambda}^{(j)}$. Then an unbiased estimate of $\boldsymbol{\pi}^{(j)}$ is $\hat{\boldsymbol{\pi}}^{(j)} = P_j^{-1} \hat{\boldsymbol{\lambda}}^{(j)}$.

We can easily extend Equation 3.2, which is applicable to one individual item, to compute the support of an arbitrary k-itemset. For simplicity, let us assume that we would compute the support of an itemset which contains the first k items $\{A_0, \cdots, A_{k-1}\}$ (The general case with any k items is quite straightforward but algebraically messy). Let $P = P_0 \times \cdots \times P_{k-1}$, an unbiased estimate of $\boldsymbol{\pi}$ follows as

$$\hat{\boldsymbol{\pi}} = P^{-1} \hat{\boldsymbol{\lambda}} = (P_0^{-1} \times \cdots \times P_{k-1}^{-1}) \hat{\boldsymbol{\lambda}} \tag{3.3}$$

where $\hat{\boldsymbol{\lambda}}$ is the vector of sample proportions corresponding to $\boldsymbol{\lambda}$ and P_j^{-1} denotes the inverse of the matrix P_j. $\hat{\pi}_{1,\cdots,1}$ is the support for the k items. Note that although the distortion matrices $P_0, \cdots, P_{k-1}$ are known, they can only be utilized to estimate the proportions of itemsets of the original data, rather than precisely reconstruct the original 0-1 data.

The whole contingency table is usually modeled as a multinomial distribution in statistics. When we have k items, the number of cells in the contingency table is 2^k. For each cell d, where $d = 1, 2, \cdots, 2^k$, it has a separate binomial distribution with parameters N and η_i. The binomial distribution is the discrete probability distribution of the number of successes in a sequence of N independent 0/1 experiments, each of which yields success with probability η_i. When N is large enough (one rule of thumb is that both $n\eta_i$ and $N(1-\eta_i)$ must be greater than 5), an approximation to $B(N, \eta_i)$ is given by the normal distribution $N(N\eta_i, N\eta_i(1-\eta_i))$.

Result 1 *Since each cell $\pi_{i_1,\cdots,i_k}$ approximately follows normal distribution, its $(1-\alpha)100\%$ interquantile range can be approximated as*

$$[\hat{\pi}_{i_1\cdots i_k} - z_{\alpha/2} * \sqrt{\hat{v}ar(\hat{\pi}_{i_1\cdots i_k})}, \hat{\pi}_{i_1\cdots i_k} + z_{\alpha/2} * \sqrt{\hat{v}ar(\hat{\pi}_{i_1\cdots i_k})}]$$

$z_{\alpha/2}$ is the upper $\alpha/2$ critical value for the standard normal distribution.

$\hat{v}ar(\hat{\pi}_{i_1\cdots i_k})$ can be derived from the covariance matrix [11]:

$$\begin{aligned} \hat{cov}(\hat{\pi}) &= \Sigma_1 + \Sigma_2 \\ &= (N-1)^{-1}(\hat{\pi}^{\delta} - \hat{\pi}\hat{\pi}') + (N-1)^{-1}P^{-1}(\hat{\lambda}^{\delta} - P\hat{\pi}^{\delta}P')P'^{-1} \end{aligned}$$

Note that Σ_1 is the dispersion matrix of the direct estimator of π, which is only related to the data size for estimation. While the data size is usually large in most market basket analysis scenarios, it can be neglected. Σ_2 represents the component of dispersion associated with RR distortion.

We can simply use the derived $\hat{\pi}_{i_1\cdots i_m}$ (from Equation 3.3) as an estimate of μ and

the derived $\sqrt{\hat{v}ar(\hat{\pi}_{i_1 \cdots i_m})}$ as an estimate of σ, where μ and σ are unknown parameters of the normal distribution of each cell. An $(1-\alpha)100\%$ interquantile range, say $\alpha = 0.05$, shows the interval contains the original $\pi_{i_1,\cdots,i_m}$ with 95% probability.

To illustrate this result, we use a simple example $G \Rightarrow H$ (rule 2 in Figure 3.1). The proportion of itemsets of the original data is given as

$$\pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})' = (0.415, 0.043, 0.183, 0.359)'$$

Using the RR scheme presented in the previous section, with the distortion parameters $p_1 = p_2 = 0.9$, we get the randomized responses

$$\hat{\lambda} = (0.368, 0.097, 0.218, 0.316)'$$

By applying Equation 3.3, we derive the unbiased estimate of π as

$$\hat{\pi} = (0.427, 0.031, 0.181, 0.362)'$$

The covariance matrix of $\hat{\pi}$ is unbiasedly estimated as

$$\hat{cov}(\hat{\pi}) = \begin{bmatrix} 7.113 & -1.668 & -3.134 & -2.311 \\ -1.668 & 2.902 & 0.244 & -1.478 \\ -3.134 & 0.244 & 5.667 & -2.777 \\ -2.311 & -1.478 & -2.777 & 6.566 \end{bmatrix} \times 10^{-5}$$

The diagonal elements of the above matrix represent the variances of the estimated $\hat{\pi}$, e.g., $\hat{v}ar(\hat{\pi}_{00}) = 7.113 \times 10^{-5}$ and $\hat{v}ar(\hat{\pi}_{11}) = 6.566 \times 10^{-5}$. Those off-diagonal elements indicate the estimated covariances, e.g., $\hat{cov}(\hat{\pi}_{11}, \hat{\pi}_{10}) = -2.777 \times 10^{-5}$.

From Result 1, we can derive 95% interquantile range of s_{GH} as

$$[\hat{\pi}_{11} - z_{0.025}\sqrt{\hat{v}ar(\hat{\pi}_{11})}, \hat{\pi}_{11} + z_{0.025}\sqrt{\hat{v}ar(\hat{\pi}_{11})}] = [0.346, 0.378]$$

We can also see this derived interquantile range [0.346, 0.378] for rule 2 with $p_1 =$

$p_2 = 0.9$ is shorter than [0.238, 0.391] with $p_1 = p_2 = 0.65$ as shown in Figure 3.1.

3.2.3 Accuracy on Confidence c

We first analyze the accuracy on confidence of a simple association rule $A \Rightarrow B$ where A and B are two single items which have 2 mutually exclusive and exhaustive categories. We denote s_A, s_B, and s_{AB} as the support values of A, B, and AB respectively. Accordingly, we denote $\hat{s}_A$, $\hat{s}_B$, and $\hat{s}_{AB}$ as the estimated support values from randomized data of A, B, and AB respectively.

Result 2 *The confidence (c) of a simple association rule $A \Rightarrow B$ has estimated value as*

$$\hat{c} = \frac{\hat{s}_{AB}}{\hat{s}_A} = \frac{\hat{\pi}_{11}}{\hat{\pi}_{1+}}$$

with the expectation of $\hat{c}$ approximated as

$$\hat{E}(\hat{c}) \approx \frac{\hat{\pi}_{11}}{\hat{\pi}_{1+}} + \frac{\hat{\pi}_{11}}{\hat{\pi}_{1+}^3}\hat{var}(\hat{\pi}_{10}) - \frac{\hat{\pi}_{10}}{\hat{\pi}_{1+}^3}\hat{var}(\hat{\pi}_{11}) + \frac{\hat{\pi}_{11} - \hat{\pi}_{10}}{\hat{\pi}_{1+}^3}\hat{cov}(\hat{\pi}_{11}, \hat{\pi}_{10}) \tag{3.4}$$

and the variance of $\hat{c}$ approximated as

$$\hat{var}(\hat{c}) \approx \frac{\hat{\pi}_{10}^2}{\hat{\pi}_{1+}^4}\hat{var}(\hat{\pi}_{11}) + \frac{\hat{\pi}_{11}^2}{\hat{\pi}_{1+}^4}\hat{var}(\hat{\pi}_{10}) - 2\frac{\hat{\pi}_{10}\hat{\pi}_{11}}{\hat{\pi}_{1+}^4}\hat{cov}(\hat{\pi}_{11}, \hat{\pi}_{10}) \tag{3.5}$$

according to the delta method [38].

Confidence can be regarded as a ratio (W) of two correlated normal random variables (X, Y), $W = X/Y$. However, it is hard to derive the critical value for the distribution of W from its cumulative density function $F(w)$ [63], we provide an approximate interquantile range of confidence based on Chebyshev's Inequality.

Theorem 1 *(Chebyshev's Inequality) For any random variable X with mean μ and variance σ^2*

$$Pr(|X - \mu| \geq k\sigma) \leq 1/k^2 \qquad k > 0$$

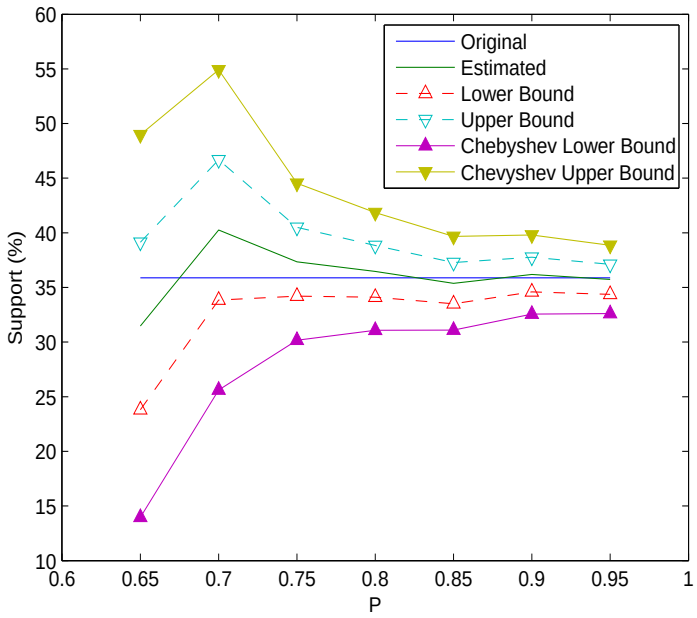


Figure 3.2: Interquantile Range vs. varying p

Chebyshev's Inequality gives a conservative estimate. It provides a lower bound to the proportion of measurements that are within a certain number of standard deviations from the mean.

Result 3 *The loose* $(1-\alpha)100\%$ *interquantile range of confidence (c) of* $A \Rightarrow B$ *can be approximated as*

$$[\hat{E}(\hat{c}) - \frac{1}{\sqrt{\alpha}}\sqrt{\hat{var}(\hat{c})}, \hat{E}(\hat{c}) + \frac{1}{\sqrt{\alpha}}\sqrt{\hat{var}(\hat{c})}]$$

From Chebyshev's Inequality, we know for any sample, at least $(1 - 1/k^2)$ of the observations in the data set fall within k standard deviations of the mean. When we set $\alpha = \frac{1}{k^2}$, we have $Pr(|X - \mu| \geq \frac{1}{\sqrt{\alpha}}\sigma) \leq \alpha$. Hence, $Pr(|X - \mu| \leq \frac{1}{\sqrt{\alpha}}\sigma) \geq 1 - \alpha$. We can simply use the derived $\hat{E}(\hat{c})$ (from Equation 3.4) as an estimate of μ and the derived $\sqrt{\hat{var}(\hat{c})}$ (from Equation 3.5) as an estimate of σ, where μ and σ are unknown parameters of the distribution of confidence. An approximate $(1-\alpha)100\%$ interquantile range of confidence c is then derived.

Note that the interquantile range based on Chebyshev's Theorem is much larger than that based on known distributions such as normal distribution for support estimates. This is because that $\frac{1}{\sqrt{\alpha}} \geq z_{\alpha/2}$ where $z_{\alpha/2}$ is the upper $\alpha/2$ critical value for the standard normal distribution. In Figure 3.2, we show how the 95% interquantile ranges for the estimated support of one particular rule ($G \Rightarrow H$ from COIL data) change with varied distortion p from 0.65 to 0.95. We can see the interquantile range derived based on Chebyshev's theorem is wider than that derived from known normal distribution. As expected, we can also observe that the larger the p, the more accurate the estimate and the tighter the interquantile ranges.

All the above results can be straightforwardly extended to the general association rule $\mathcal{X} \Rightarrow \mathcal{Y}$ and further details can be found in [27].

3.2.4 Empirical Evaluation

In our experiments, 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. The training set consists of 5822 descriptions of customers, including the information of whether or not they have a Caravan insurance policy. Our binary data is formed by collapsing non-binary categorical attributes into binary form (the data can be found at www.cs.uncc.edu/∼xwu/classify/b86.dat), with $n = 5822$ baskets and $m = 86$ binary items. We use ten attributes (denote as A to J) as shown in Table 3.1 to illustrate our results.

1. Accuracy of individual rule vs. varying p

Table 3.2 shows the 7 randomly chosen association rules derived from the randomized COIL data with distortion parameter $p = 0.65$. In this table, s ($\hat{s}$) indicates the original (estimated) support value. $\hat{s}_l$ ($\hat{s}_u$) denotes the lower bound (upper bound) of the 95% interquantile range of the estimated support value. Similarly, c ($\hat{c}$) indicates the original (estimated) confidence value. $\hat{c}_l$ ($\hat{c}_u$) denotes the lower bound (upper

Table 3.1: 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	$> 0 \rightarrow 1$

Table 3.2: Accuracy of the estimated support and confidence for 7 representative rules of COIL

ID	$\mathcal{X}$	$\mathcal{Y}$	s	$\hat{s}$	$\hat{s}_l$	$\hat{s}_u$	c	$\hat{c}$	$\hat{c}_l$	$\hat{c}_u$
1	G	E	35.9	34.1	26.3	41.8	66.2	64.7	31.3	95.3
2	G	H	35.9	31.5	23.8	39.1	66.2	62.2	26.6	90.4
3	EH	G	35.8	45.0	31.5	58.5	89.3	77.5	33.5	100
4	EG	I	22.1	28.4	14.9	42.0	61.7	75.2	0	100
5	HF	I	23.9	17.2	3.7	30.8	100	91.0	0	100
6	EGH	F	22.1	36.3	12.3	60.2	61.7	99.4	0	100
7	FGI	E	22.1	27.6	3.32	52.0	77.9	86.3	0	100

bound) of the 95% estimated confidence value. We have shown how the accuracy of the estimated support values varies in Figure 3.1.

One observation is that interquantile ranges of confidence estimates are usually wider than that of support estimates. For example, the 95% interquantile range of the estimated confidence for rule 2 is $[26.6\%, 90.4\%]$, which is much wider than that of the estimated support $[23.8\%, 39.1\%]$. This is due to three reasons. First, we set the distortion parameter $p = 0.65$ which implies a relatively large noise (the perturbed data will be completely random when $p = 0.5$). Second, the variance of the ratio of two variables is usually larger than the variance of either single variable. Third, the estimated support can be modeled as one approximate normal distribution so we can use the tight interquantile range. On the contrary, we derive the loose interquantile range of confidence using the general Chebyshev's Theorem. We expect that the explicit form of the $F(w)$ distribution can significantly reduce this width. We will investigate the explicit form of the distribution of confidence and all other measures, e.g. correlation, lift, etc. to derive tight bounds in our future work.

Our next experiment shows how the derived estimates (support, confidence, and their corresponding interquantile ranges) of one individual rule vary with the distortion parameter p. We vary the distortion parameter p from 0.65 to 0.95. Figure 3.3(a) (3.3(b)) shows the accuracy of the estimated support (confidence) values with varied

distortion p values for a particular rule $G \Rightarrow H$. As expected, the larger the p, the more accurate the estimate and the tighter the interquantile range is. It was empirically shown in [54] that a distortion probability of $p = 0.9$ (equivalently $p = 0.1$) is ideally suited to provide both privacy and good data mining results for the sparse market basket data. We can observe from Figure 3.3(b) that the 95% interquantile range of the confidence estimate with $p \geq 0.9$ is tight.

2. Accuracy of all rules vs. varying p

The above study of the accuracy of the estimate in terms of each individual rule is based on the variance as criterion. In the case of all rules together, we can evaluate the overall accuracy of data mining results using the average support error, the average confidence error, percentage of false positives, percentage of false negatives etc. as defined in [7].

The metric $\rho = \frac{1}{|R|} \sum_{r \in R} \frac{|\hat{s}_r - s_r|}{s_r} \times 100$ represents the average relative error in the reconstructed support values for those rules that are correctly identified. The identity error σ reflects the percentage error in identifying association rules. $\sigma^+ = \frac{|R-F|}{|F|} \times 100$ indicates the percentage of false positives and $\sigma^- = \frac{|F-R|}{|F|} \times 100$ indicates the percentage of false negatives where R (F) denotes the reconstructed (actual) set of association rules. In addition to the support error (ρ) and the identity error (σ^+, σ^-), we define the following three measures.

- γ: the confidence error $\gamma = \frac{1}{|R|} \sum_{r \in R} \frac{|\hat{c}_r - c_r|}{c_r} \times 100$ represents the average relative error in the reconstructed confidence values for those rules that are correctly identified.

- s-p: the number of pairs of conflict support estimates. We consider $\hat{s}_1, \hat{s}_2$ as a pair of conflict estimates if $\hat{s}_1 < \hat{s}_2$ but $s_1 > \hat{s}_{1l} > s_{min} > s_2$ where $\hat{s}_{1l}$ denotes the lower bound of interqunatile range for s_1.

- c-p: the number of pairs of conflict confidence estimates (similarly defined as

the above s-p).

Errors in support estimation due to the distortion procedure can result in falsely identified frequent itemsets. This becomes especially an issue when the support threshold setting is such that the support of a number of frequent itemsets lie very close to this threshold value (s_{min}). Such border-line itemsets can cause many *false positives* and *false negatives*. Even worse, an error in identifying a frequent itemset correctly in early passes has a ripple effect in terms of causing errors in later passes.

Table 3.3(a) shows how the above measures are varied by changing distortion parameter p from 0.65 to 0.95. We can observe all measures (the support error ρ, the confidence error γ, the false positives σ^+, the false negatives σ^-) decrease when p increases. The number of conflict support pairs (s-p) and conflict confidence pairs (c-p) also have the same trend. Our experiment shows that when $p \geq 0.85$, there are no or very few conflict support (confidence) pairs, which implies the reconstructed set of association rules is close to the original set. However, when $p \leq 0.80$, there are significant number of conflict pairs, which implies the reconstructed set may be quite different from the original one. By incorporating the derived interquantile range for each estimate, we can decrease the error caused by conflict pairs. In Section 3.2.1, we have shown one conflict support pair: rule 2 and rule 6. We can see that $\hat{s}_2 < \hat{s}_6$ (but $s_2 > s_6$). As $\hat{s}_{2l} > s_{min}$ and $\hat{s}_{6l} < s_{min}$, data miners can safely determine rule 2 is frequent but rule 6 may be infrequent. We would emphasize again that providing estimates together with their interquantile ranges (especially for those conflict pairs) through some visualization is very useful for data exploration tasks conducted on the randomized data.

Table 3.3(b) shows the comparison between the identity errors derived using lower bound and upper bound respectively. We define $\sigma_l^+ = \frac{|R_l - F|}{|F|} \times 100$ ($\sigma_u^+ = \frac{|R_u - F|}{|F|} \times 100$) as the false positives calculated from R_l (R_u) where R_l (R_u) denotes the reconstructed set of association rules using lower (upper) bound of interquantile range respectively.

Table 3.3: $sup_{min} = 25\%,\ conf_{min} = 65\%$ for COIL

(a)

p	ρ	σ^-	σ^+	s-p	γ	c-p
0.65	25.6	34.0	53.8	27817	9.90	737
0.70	12.3	21.2	38.1	4803	6.39	393
0.75	7.35	11.8	30.8	729	4.44	85
0.80	3.64	6.82	16.9	0	2.47	28
0.85	2.64	6.67	7.76	0	1.76	0
0.90	1.91	5.18	4.24	0	1.10	0
0.95	0.84	4.63	1.02	0	0.51	0

(b)

p	σ^-	σ_l^-	σ_u^-	σ^+	σ_l^+	σ_u^+
0.65	34.0	98.8	1.25	53.8	0.00	110.7
0.70	21.2	90.9	0.08	38.1	0.08	105.7
0.75	11.8	66.3	0.00	30.8	1.18	96.5
0.80	6.82	50.7	0.31	16.9	0.24	80.9
0.85	6.67	37.7	0.00	7.76	0.55	53.0
0.90	5.18	31.8	0.00	4.24	0.00	35.0
0.95	4.63	26.8	0.00	1.02	0.00	25.7

Similarly we define σ_l^- and σ_u^-. We can observe from Table 3.3(b) that σ_u^- is significantly lower than σ^- while σ_l^+ is significantly lower than σ^+. In other words, using the upper bound of the derived interquantile range can decrease the false negatives while using the lower bound can decrease the false positives. In some scenario, we may emphasize more on decreasing the false positive error. Hence, we can use the lower bound of the derived interquantile range, rather than the estimated value, to determine whether the set is frequent or not (i.e., frequent only if $\hat{s}_l \geq s_{min}$, infrequent otherwise).

3. Other datasets

Since the COIL Challenge data is very sparse (5822 tuples with 86 attributes), we also conducted evaluations on the following representative databases used for association rule mining.

Table 3.4: $sup_{min} = 0.20\%,\ conf_{min} = 20\%$ for BMS-WebView-1

(a)

p	ρ	σ^-	σ^+	s-p	γ	c-p
0.65	362.4	64.1	80.6	632	114.7	11
0.75	72.9	39.9	68.7	418	57.9	2
0.85	19.5	27.9	54.0	67	24.5	0
0.95	5.47	9.66	16.5	56	7.23	0

(b)

p	σ^-	σ_l^-	σ_u^-	σ^+	σ_l^+	σ_u^+
0.65	63.9	100.0	1.34	81.8	0.0	187.6
0.75	40.1	100.0	1.07	69.8	0.0	155.3
0.85	27.9	99.1	0.40	54.0	0.0	152.8
0.95	9.66	70.6	0.00	16.5	0.0	123.8

Table 3.5: $sup_{min} = 0.20\%,\ conf_{min} = 60\%$ for IBM data

(a)

p	ρ	σ^-	σ^+	s-p	γ	c-p
0.65	1234.9	73.4	171.9	971	47.8	7
0.75	99.7	57.8	168.0	11	38.3	0
0.85	19.9	49.7	165.6	3	18.6	0
0.95	5.14	21.3	50.3	0	4.61	0

(b)

p	σ^-	σ_l^-	σ_u^-	σ^+	σ_l^+	σ_u^+
0.65	73.7	100.0	2.99	172.8	0.0	722.5
0.75	57.8	100.0	1.20	167.9	0.0	674.3
0.85	49.7	100.0	0.90	165.6	0.0	673.4
0.95	21.3	99.7	0.00	50.3	0.0	460.8

1. BMS-WebView-1 [3]. Each transaction in the data set is a web session consisting of all the product detail pages viewed in that session. There are about 60,000 transactions with close 500 items.

2. A synthetic database generated from the IBM Almaden market basket data generator with parameters T10.I4.D0.1M.N0.1K., resulting in 10k customer tuples with each customer purchasing about ten items on average.

Tables 3.4 and 3.5 show our results on these two data sets respectively. We can observe similar patterns as shown in COIL data set.

3.3 Extension to General Categorical Data Analysis

Most data mining problems are based on the analysis of associations between variables. No matter how the associations are defined, a suitable measure to evaluate the dependencies between variables is required for such analysis. The problem of analyzing objective measures used by data mining algorithms has attracted much attention in recent years. Many measures have been proposed for different applications in the literature. Depending on the specific properties of it, each measure is useful for some application, but not for others. The objective interestingness measure is usually computed from the contingency table. Table 3.6 shows various measures defined for a pair of binary variables [69].

In this section we conduct theoretical analysis on the accuracy of various measures adopted in categorical data analysis. Our analysis is based on estimating the parameters of derived random variables. The estimated measure (e.g., Interest statistics) is considered as one derived variable. We present a general method in [28], which is based on the Taylor series, for approximating the mean and variance of derived variables. We also derive interquantile ranges of those estimates. Hence, data miners are ensured that their estimates lie within these ranges with a high confidence.

[3] http://www.ecn.purdue.edu/KDDCUP

Table 3.6: Objective association measures for two binary variables

Measure	Expression	Measure	Expression
Pearson's coefficient (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)	$\frac{\frac{\pi_{11}}{\pi_{1+}}-\pi_{+1}}{1-\pi_{+1}}$
Std. residues(e)	$\sqrt{N}\cdot\frac{\pi_{ij}-\pi_{i+}\pi_{+j}}{\sqrt{\pi_{i+}\pi_{+j}}}$	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)	$-\frac{\sum_i\sum_j \pi_{ij} log\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}}{\sum_j \pi_{+j} log\pi_{+j}}$

3.3.1 Variances of Derived Measures

From Table 3.6, we can see that each measure can be expressed as one derived random variable (or function) from the observed variables (π_{ij} or their marginal totals π_{i+}, π_{+j}). Similarly, its estimate from the randomized data can be considered as another derived random variable from the input variables ($\hat{\pi}_{ij}, \hat{\pi}_{i+}, \hat{\pi}_{+j}$). Since we know how to derive variances of the input variables ($\hat{v}ar(\hat{\pi}_{ij})$) from the randomized data, our problem is then how to derive the variance of the derived output variable.

In the following, we first present a general approach based on the Delta method [38] and then discuss how to derive the variance of chi-square statistics (χ^2) as one example.

Let z be a random variable derived from the observed random variables x_i ($i = 1, \cdots, n$): $z = g(x)$. According to the Delta method, a Taylor approximation of the variance of a function with multiple variables can be expanded as

$$var\{g(x)\} = \sum_{i=1}^{k} \{g_i'(\theta)\}^2 var(x_i) + \sum_{i \neq j=1}^{k} \sum g_i'(\theta) g_j'(\theta) cov(x_i, x_j) + o(N^{-r})$$

where θ_i is the mean of x_i, $g(x)$ stands for the function $g(x_1, x_2, \cdots, x_k)$, $g_i'(\theta)$ is the $\frac{\partial g(x)}{\partial x_i}$ evaluated at $\theta_1, \theta_2, \cdots, \theta_k$.

For market basket data with 2 variables, $\hat{\pi} = (\hat{\pi}_{00}, \hat{\pi}_{01}, \hat{\pi}_{10}, \hat{\pi}_{11})'$, the estimated chi-square is shown as

$$\begin{aligned} \hat{\chi}^2 &= N\left(\frac{(\hat{\pi}_{00} - \hat{\pi}_{0+}\hat{\pi}_{+0})^2}{\hat{\pi}_{0+}\hat{\pi}_{+0}} + \frac{(\hat{\pi}_{01} - \hat{\pi}_{0+}\hat{\pi}_{+1})^2}{\hat{\pi}_{0+}\hat{\pi}_{+1}}\right. \\ &+ \left.\frac{(\hat{\pi}_{10} - \hat{\pi}_{1+}\hat{\pi}_{+0})^2}{\hat{\pi}_{1+}\hat{\pi}_{+0}} + \frac{(\hat{\pi}_{11} - \hat{\pi}_{1+}\hat{\pi}_{+1})^2}{\hat{\pi}_{1+}\hat{\pi}_{+1}}\right) \end{aligned}$$

Let $x_1 = \hat{\pi}_{00}$, $x_2 = \hat{\pi}_{01}$, $x_3 = \hat{\pi}_{10}$ and $x_4 = \hat{\pi}_{11}$, we have

$$\begin{aligned} & g(x_1, x_2, x_3, x_4) = \chi^2 \\ & = N[\frac{x_1^2}{(x_1+x_2)(x_1+x_3)} + \frac{x_2^2}{(x_1+x_2)(x_2+x_4)} + \\ & \frac{x_3^2}{(x_3+x_4)(x_3+x_1)} + \frac{x_4^2}{(x_4+x_3)(x_4+x_2)} - 1] \end{aligned}$$

Partial derivatives of the function $g()$ can be calculated respectively. By incorporating estimated expectations, variances and covariances of variables in function $g()$, the variance of function $g()$ can be estimated as

$$\begin{aligned} & \hat{v}ar(g) \approx G_1^2 \hat{v}ar(\hat{\pi}_{00}) + G_2^2 \hat{v}ar(\hat{\pi}_{01}) + G_3^2 \hat{v}ar(\hat{\pi}_{10}) + G_4^2 \hat{v}ar(\hat{\pi}_{11}) \\ & + 2G_1G_2\hat{c}ov(\hat{\pi}_{00}, \hat{\pi}_{01}) + 2G_1G_3\hat{c}ov(\hat{\pi}_{00}, \hat{\pi}_{10}) + 2G_1G_4\hat{c}ov(\hat{\pi}_{00}, \hat{\pi}_{11}) \\ & + 2G_2G_3\hat{c}ov(\hat{\pi}_{01}, \hat{\pi}_{10}) + 2G_2G_4\hat{c}ov(\hat{\pi}_{01}, \hat{\pi}_{11}) + 2G_3G_4\hat{c}ov(\hat{\pi}_{10}, \hat{\pi}_{11}) \end{aligned}$$

where

$$\begin{aligned} G_1 = \frac{\partial g}{\partial x_1} &= N[\tfrac{\hat{\pi}_{00}^2(\hat{\pi}_{01}+\hat{\pi}_{10})+2\hat{\pi}_{00}\hat{\pi}_{01}\hat{\pi}_{10}}{\hat{\pi}_{0+}^2\hat{\pi}_{+0}^2} - \tfrac{\hat{\pi}_{01}^2}{\hat{\pi}_{0+}^2\hat{\pi}_{+1}} - \tfrac{\hat{\pi}_{10}^2}{\hat{\pi}_{+0}^2\hat{\pi}_{1+}}] \\ G_2 = \frac{\partial g}{\partial x_2} &= N[\tfrac{\hat{\pi}_{01}^2(\hat{\pi}_{00}+\hat{\pi}_{11})+2\hat{\pi}_{00}\hat{\pi}_{01}\hat{\pi}_{11}}{\hat{\pi}_{0+}^2\hat{\pi}_{+1}^2} - \tfrac{\hat{\pi}_{00}^2}{\hat{\pi}_{0+}^2\hat{\pi}_{+0}} - \tfrac{\hat{\pi}_{11}^2}{\hat{\pi}_{+1}^2\hat{\pi}_{1+}}] \\ G_3 = \frac{\partial g}{\partial x_3} &= N[\tfrac{\hat{\pi}_{10}^2(\hat{\pi}_{11}+\hat{\pi}_{00})+2\hat{\pi}_{00}\hat{\pi}_{10}\hat{\pi}_{11}}{\hat{\pi}_{1+}^2\hat{\pi}_{+0}^2} - \tfrac{\hat{\pi}_{11}^2}{\hat{\pi}_{1+}^2\hat{\pi}_{+1}} - \tfrac{\hat{\pi}_{00}^2}{\hat{\pi}_{+0}^2\hat{\pi}_{0+}}] \\ G_4 = \frac{\partial g}{\partial x_4} &= N[\tfrac{\hat{\pi}_{11}^2(\hat{\pi}_{01}+\hat{\pi}_{10})+2\hat{\pi}_{11}\hat{\pi}_{01}\hat{\pi}_{10}}{\hat{\pi}_{1+}^2\hat{\pi}_{+1}^2} - \tfrac{\hat{\pi}_{10}^2}{\hat{\pi}_{1+}^2\hat{\pi}_{+0}} - \tfrac{\hat{\pi}_{01}^2}{\hat{\pi}_{+1}^2\hat{\pi}_{0+}}] \end{aligned}$$

Since $\chi^2 = N\phi^2$ where ϕ denotes correlation (A proof is given in Appendix A of [62]), $\phi = \sqrt{\chi^2/N} = \sqrt{g/N}$. As we know, $\frac{\partial \phi}{\partial x_i} = \frac{1}{2\sqrt{gN}} \frac{\partial g}{\partial x_i}$. Following the same procedure above, the variance of correlation ϕ can be approximated as

$$\hat{v}ar(\phi) \approx \frac{\hat{v}ar(g)}{4G_E}$$

where

$$G_E = N^2[\frac{\hat{\pi}_{00}^2}{\hat{\pi}_{0+}\hat{\pi}_{+0}} + \frac{\hat{\pi}_{01}^2}{\hat{\pi}_{0+}\hat{\pi}_{+1}} + \frac{\hat{\pi}_{10}^2}{\hat{\pi}_{1+}\hat{\pi}_{+0}} + \frac{\hat{\pi}_{11}^2}{\hat{\pi}_{1+}\hat{\pi}_{+1}} - 1]$$

Similarly we can derive variances of the estimated values of all measures shown in Table 3.6. Measures such as χ^2, interest factor, IS, PS, and Jaccard coefficient can be extended to more than two variables using the multi-dimensional contingency tables. We show the estimated chi-square statistics for k-itemset as one example.

$$\hat{\chi}^2 = N \sum_{u_1=0}^{1} \cdots \sum_{u_k=0}^{1} \frac{(\hat{\pi}_{u_1 \cdots u_k} - \prod_{j=1}^{k} \hat{\pi}_{u_j}^{(j)})^2}{\prod_{j=1}^{k} \hat{\pi}_{u_j}^{(j)}} \tag{3.6}$$

It is easy to see $\hat{\chi}^2$ can be considered as one derived variable from the observed elements $\hat{\pi}_{u_1 \cdots u_k}$ and the marginal totals $\hat{\pi}_{u_j}^{(j)}$ of the 2^k contingency table. Following the same delta method, we can derive its variance.

3.3.2 Interquantile Ranges of Derived Measures

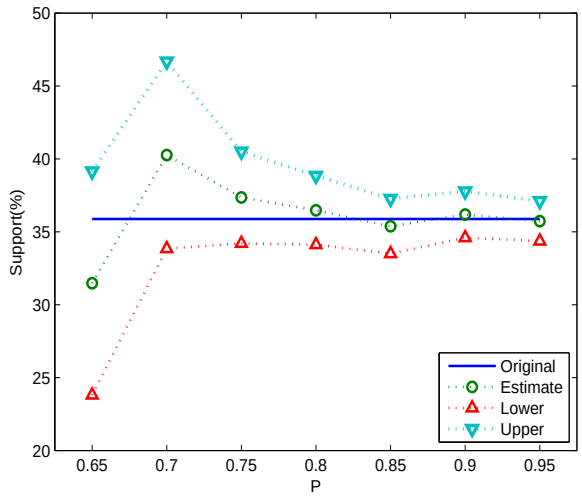
To derive interquantile ranges of estimates, we need to explore the distribution of those derived variables. In [27], the authors have shown the estimate of support follows an approximate normal distribution and the estimate of confidence (i.e., a ratio of two correlated normal variables) follows a very complex $F(w)$ distribution. In general, we can observe that every element (e.g., $\hat{\pi}_{ij}$) in the derived measure expressions (shown in Table 2) has an approximate normal distribution, however, the derived measures usually do not have explicit distribution expressions. Hence we cannot calculate the critical values of distributions to derive the interquantile range. Chebyshev's theorem can be used to give a conservative estimate of it as in Section 3.2.3 for confidence.

3.4 Summary

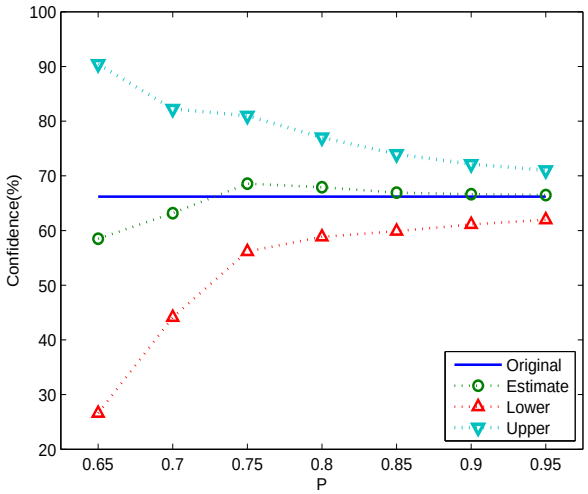
In this chapter, we investigated the issue of providing accuracy in privacy preserving categorical data analysis. We have presented a general approach to derive variances of estimates of various measures adopted in categorical data analysis. We applied the idea of using interquantile ranges based on Chebyshev's Theorem to bound those

estimates derived from the randomized categorical data.

Providing the accuracy of discovered patterns from randomized data is important for data miners. To the best of our knowledge, this has not been previously explored in the context of privacy preserving data mining although defining the significance of discovered patterns in general data mining has been studied (e.g., [25]).



(a) support



(b) confidence

Figure 3.3: Accuracy vs. varying p for rule $G \Rightarrow H$

CHAPTER 4: UTILITY ANALYSIS WITH UNKNOWN DISTORTION PROBABILITIES P

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 as jeopardizing with respect to the original value if the posterior probabilities are significantly greater than the a-priori probabilities. In this chapter, 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.

In Section 4.1, we review some related work in randomization. In Section 4.2, 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 Section 4.3, 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.

In Section 4.4, 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.

Throughout this chapter, we use the COIL Challenge 2000 which provides data from a real insurance business. The description of the data set can be found in Table 3.1 of Section 3.2.4. Some results in this chapter were previously reported in [29].

4.1 Introduction

As we discussed in Chapter 3 , most of previous work on randomization models except [28] investigated the scenario that distortion parameters are fully or partially known by data miners. 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., [6, 28, 54]) or determining the optimal distortion parameters to achieve good performance (e.g., [35]). Data mining tasks were conducted on the reconstructed distribution $\hat{\boldsymbol{\pi}}$ calculated from Equation 3.3. In this Chapter, we investigate the scenario that distortion parameters are not known by data miners. That is, data mining tasks are conducted on the distribution of randomized data λ directly. According to our knowledge, it has not been studied in the literature of privacy preserving data mining.

In [28], the authors very briefly showed that some measures have vertical monotonic property on the market basket data. 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.

4.2 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.

4.2.1 Associations between Two Binary Variables

Table 3.6 shows various association measures for two binary variables (Refer to [69] 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 can be shown as:

$$P_j = \begin{pmatrix} p_0 & 1-p_1 \\ 1-p_0 & p_1 \end{pmatrix}$$

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 2.1. This is because

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

as shown in [11]. $cov(\hat{\boldsymbol{\lambda}})$ 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 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 4 *For any pair of variables A_u, A_v perturbed with any distortion matrix P_u and P_v ($p_0^{(u)}$, $p_1^{(u)}$, $p_0^{(v)}$, $p_1^{(v)} \in [0,1]$) 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:*

$$\chi^2_{ran} \leq \chi^2_{ori}, \qquad G^2_{ran} \leq G^2_{ori}$$

$$M_{ran} \leq M_{ori}, \qquad \tau_{ran} \leq \tau_{ori}$$

$$U_{ran} \leq U_{ori}, \qquad |\phi_{ran}| \leq |\phi_{ori}|$$

$$|D_{ran}| \leq |D_{ori}|, \quad |PS_{ran}| \leq |PS_{ori}|$$

No other measures shown in Table 3.6 holds monotonic property.

For randomization, we know that the distortion is 1) highest with $p = 0.5$ which imparts the maximum randomness to the distorted values; 2) symmetric around $p = 0.5$ and makes no difference, reconstruction-wise, between choosing a value p or its counterpart $1-p$. In practice, the distortion is usually conducted with p greater than 0.5. The following results show the vertical association variations when $p_0^{(u)}$,$p_1^{(u)}$,$p_0^{(v)}$ and $p_1^{(v)}$ are greater than 0.5.

Result 5 *In addition to monotonic relations shown in Result 4, when* $p_0^{(u)}$, $p_1^{(u)}$, $p_0^{(v)}$, $p_1^{(v)} \in [0.5, 1]$, *we have*

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

Proof 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:

$$\begin{aligned}
\pi_{1+} &= \frac{p_1^{(u)} - 1 + (1 + p_0^{(u)} - p_1^{(u)})\lambda_{1+}}{p_0^{(u)} + p_1^{(u)} - 1} \\
\pi_{+1} &= \frac{p_1^{(v)} - 1 + (1 + p_0^{(v)} - p_1^{(v)})\lambda_{+1}}{p_0^{(v)} + p_1^{(v)} - 1} \\
\pi_{11} - \pi_{+1}\pi_{1+} &= \frac{\lambda_{11} - \lambda_{+1}\lambda_{1+}}{(p_0^{(u)} + p_1^{(u)} - 1)(p_0^{(v)} + p_1^{(v)} - 1)}
\end{aligned}$$

Through deduction, AV_{ori} is expressed as:

$$AV_{ori} = \frac{\lambda_{11} - \lambda_{+1}\lambda_{1+}}{(p_0^{(v)} + p_1^{(v)} - 1)[p_1^{(u)} - 1 + (1 + p_0^{(u)} - p_1^{(u)})\lambda_{1+}]}$$

Let $f(p_0^{(u)}, p_1^{(u)}, p_0^{(v)}, p_1^{(v)}, \lambda_{1+}) = |(p_0^{(v)}+p_1^{(v)}-1)[p_1^{(u)}-1+(1+p_0^{(u)}-p_1^{(u)})\lambda_{1+}]|-|\lambda_{1+}|$,

1) When $p_0^{(u)}, p_1^{(u)}, p_0^{(v)}, p_1^{(v)} \in [0.5, 1]$, since

$$\pi_{1+} = \frac{p_1^{(u)} - 1 + (1 + p_0^{(u)} - p_1^{(u)})\lambda_{1+}}{p_0^{(u)} + p_1^{(u)} - 1} \geq 0$$

then

$$p_1^{(u)} - 1 + (1 + p_0^{(u)} - p_1^{(u)})\lambda_{1+} \geq 0$$

we have

$$\begin{aligned}
f(p_0^{(u)}, p_1^{(u)}, p_0^{(v)}, p_1^{(v)}, \lambda_{1+}) &= (p_0^{(v)} + p_1^{(v)} - 1)[p_1^{(u)} - 1 + (1 + p_0^{(u)} - p_1^{(u)})\lambda_{1+}] - \lambda_{1+} \\
&= (p_0^{(v)} + p_1^{(v)} - 1)(p_1^{(u)} - 1)(1 - \lambda_{1+}) \\
&+ [(p_0^{(v)} + p_1^{(v)} - 1)p_0^{(u)} - 1]\lambda_{1+} \\
&\leq 0
\end{aligned}$$

Hence,

$$\begin{aligned}
|AV_{ori}| &= |\frac{\lambda_{11} - \lambda_{+1}\lambda_{1+}}{(p_0^{(v)} + p_1^{(v)} - 1)[p_1^{(u)} - 1 + (1 + p_0^{(u)} - p_1^{(u)})\lambda_{1+}]}| \\
&\geq |\frac{\lambda_{11} - \lambda_{+1}\lambda_{1+}}{\lambda_{1+}}| \\
&\geq |AV_{ran}|
\end{aligned}$$

2) When $p_0^{(u)}, p_1^{(u)}, p_0^{(v)}, p_1^{(v)} \in [0, 0.5]$, since

$$p_1^{(u)} - 1 + (1 + p_0^{(u)} - p_1^{(u)})\lambda_{1+} \geq 0$$

we have

$$\begin{aligned} f(p_0^{(u)}, p_1^{(u)}, p_0^{(v)}, p_1^{(v)}, \lambda_{1+}) &= (p_0^{(v)} + p_1^{(v)} - 1)(p_1^{(u)} - 1)(1 - \lambda_{1+}) \\ &+ [(p_0^{(v)} + p_1^{(v)} - 1)p_0^{(u)} - 1]\lambda_{1+} \end{aligned}$$

when $\lambda_{1+} \geq \frac{(p_0^{(v)}+p_1^{(v)}-1)(p_1^{(u)}-1)}{1-(p_0^{(u)}+p_1^{(v)}-1)(1+p_0^{(u)}-p_1^{(u)})}$

$$f(p_0^{(u)}, p_1^{(u)}, p_0^{(v)}, p_1^{(v)}, \lambda_{1+}) \leq 0, \quad |AV_{ori}| \geq |AV_{ran}|$$

when $\lambda_{1+} < \frac{(p_0^{(v)}+p_1^{(v)}-1)(p_1^{(u)}-1)}{1-(p_0^{(v)}+p_1^{(v)}-1)(1+p_0^{(u)}-p_1^{(u)})}$

$$f(p_0^{(u)}, p_1^{(u)}, p_0^{(v)}, p_1^{(v)}, \lambda_{1+}) > 0, |AV_{ori}| < |AV_{ran}|$$

Similarly, we can prove that $|AV_{ori}| \geq |AV_{ran}|$ is not always held when $p_0^{(u)}, p_1^{(u)}, p_0^{(v)}, p_1^{(v)} \notin [0.5, 1]$.

For all other measures in the above two results, we can prove similarly. 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 4.1(a) and 4.1(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 $p^{(A)}$ and $p^{(D)}$ (In all examples, we follow the original Warner model by setting $p_0^{(u)} = p_1^{(u)} = p^{(u)}$). It can be easily observed that $\chi^2_{ran} \leq \chi^2_{ori}$ for all $p^{(A)}, p^{(D)} \in [0, 1]$ and $IS_{ran} \geq IS_{ori}$ for some $p^{(A)}, p^{(D)}$ 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 [24,68]. 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 [52] proposed three principles that should be satisfied by any good objective measure M for variables X, Y:

- $C1$: $M = 0$ 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 3.6.

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 4.2(a) and 4.2(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 $p^{(u)}$ and $p^{(v)}$. It can be easily observed from Figure 4.2(a) that the blue surface ($PS_{ran}^{A,B}$) is above the brown surface ($PS_{ran}^{I,J}$), which means that $PS_{ran}^{A,B} > PS_{ran}^{I,J}$ for all $p^{(u)}, p^{(v)} \in [0.5, 1]$ with $PS_{ori}^{A,B} > PS_{ori}^{I,J}$ ($PS_{ori}^{A,B}$ and $PS_{ori}^{I,J}$ are the points when $p_u = p_v = 1$). Figure 4.2(b) shows although $\alpha_{ori}^{A,B} < \alpha_{ori}^{I,J}$ ($\alpha_{ori}^{A,B} = 3.23, \alpha_{ori}^{I,J} = 3.94$), $\alpha_{ran}^{A,B} > \alpha_{ran}^{I,J}$ for some distortion parameters $p^{(u)}$ and $p^{(v)}$. For example, $\alpha_{ran}^{A,B} = 1.32$, $\alpha_{ran}^{I,J} = 1.14$ when $p^{(u)} = p^{(v)} = 0.8$.

Result 6 *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 P_v respectively ($p_0^{(u)}$, $p_1^{(u)}$, $p_0^{(v)}$, $p_1^{(v)} \in [0, 1]$) (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 $p_0^{(u)}, p_1^{(u)}, p_0^{(v)}, p_1^{(v)}$.

Proof For any pair of variables, Piatetsky-Shapiro's measure calculated directly from the randomized data without knowing $p_0^{(u)}, p_1^{(u)}, p_0^{(v)}, p_1^{(v)}$ is:

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

The original Piatetsky-Shapiro's measure is:

$$PS_{ori} = \pi_{11} - \pi_{1+}\pi_{+1} = \frac{PS_{ran}}{(p_0^{(u)} + p_1^{(u)} - 1)(p_0^{(v)} + p_1^{(v)} - 1)}$$

$$|PS_{ori}^{u,v}| - |PS_{ori}^{s,t}| = \frac{|PS_{ran}^{u,v}| - |PS_{ran}^{s,t}|}{|(p_0^{(u)} + p_1^{(u)} - 1)(p_0^{(v)} + p_1^{(v)} - 1)|}$$

So $\forall p_0^{(u)}, p_1^{(u)}, p_0^{(v)}, p_1^{(v)} \in [0, 1]$, $\frac{1}{|(p_0^{(u)} + p_1^{(u)} - 1)(p_0^{(v)} + p_1^{(v)} - 1)|} \geq 1$. Result 6 is proved.

Result 7 *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 ($p_0^{(u)}$,$p_1^{(u)} \in [0, 1]$) while B_s and B_t are*

unchanged (Case 2), we have

$$|D^{u,s}_{ori}| \geq |D^{v,t}_{ori}| \Longleftrightarrow |D^{u,s}_{ran}| \geq |D^{v,t}_{ran}|$$
$$|AV^{u,s}_{ori}| \geq |AV^{v,t}_{ori}| \Longleftrightarrow |AV^{u,s}_{ran}| \geq |AV^{v,t}_{ran}|$$

Proof Since

$$D_{ran} = \frac{\lambda_{00}}{\lambda_{+0}} - \frac{\lambda_{01}}{\lambda_{+1}} = \frac{\lambda_{00}\lambda_{11} - \lambda_{01}\lambda_{10}}{\lambda_{+0}\lambda_{+1}}$$
$$D_{ori} = \frac{\pi_{00}\pi_{11} - \pi_{01}\pi_{10}}{\pi_{+0}\pi_{+1}} = \frac{\lambda_{00}\lambda_{11} - \lambda_{01}\lambda_{10}}{(p_0^{(u)} + p_1^{(u)} - 1)\lambda_{+0}\lambda_{+1}}$$

We have $D_{ori} = \frac{1}{(p_0^{(u)} + p_1^{(u)} - 1)} D_{ran}$. Hence,

$$|D^{u,s}_{ori}| - |D^{v,t}_{ori}| = \frac{1}{|p_0^{(u)} + p_1^{(u)} - 1|}(|D^{u,s}_{ran}| - |D^{v,t}_{ran}|)$$

We can show AV also holds. Result 7 is proved.

Through evaluation, no other measure in Table 3.6 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.

4.2.2 Extension to Two Polychotomous Variables

There are five association measures $(\chi^2, G^2, M, \tau, U)$ that can be extended to two variables with multiple categories as shown in Table 4.1.

Vertical Association Variation

Result 8 *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*

Table 4.1: 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)	$-\frac{\sum_i \sum_j \pi_{ij} log \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}}{\sum_j \pi_{+j} log \pi_{+j}}$

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 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 are 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 9 *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 indepen-

dence hypotheses is 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.

Horizontal Association 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.

4.3 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 [8]. 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 4.3.1, we first revisit loglinear modeling and focus on the hierarchical loglinear model fitting. In Section 4.3.2, we present the criterion to determine which hierarchical loglinear models can be preserved after randomization. In Section 4.3.3, we investigate how parameters of loglinear models are affected by randomization.

4.3.1 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 4.1 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.

$$\log \hat{y}_{ijk} = \gamma + \gamma^A_i + \gamma^B_j + \gamma^C_k + \gamma^{AB}_{ij} + \gamma^{AC}_{ik} + \gamma^{BC}_{jk} + \gamma^{ABC}_{ijk} \tag{4.1}$$

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

Table 4.2: 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

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 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 4.2 shows Pearson and p-value of Hypothesis Test for different models. We can see model (AD, G) has the

Table 4.3: Goodness-of-Fit tests for loglinear models on attributes A, D, G after Randomization with different $(p^{(A)}, p^{(D)}, p^{(G)})$

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

smallest χ^2 value (1.60) and the largest p-value (0.66). Hence the best fitted model is (AD, G), i.e.,

$$\log \hat{y}_{ijk} \quad = \quad \gamma + \gamma_i^A + \gamma_j^D + \gamma_k^G + \gamma_{ij}^{AD} \tag{4.2}$$

4.3.2 Equivalent Loglinear Model

Chen [14] first studied equivalent loglinear models under independent misclassification in statistics. Korn [40] extended his work and proposed Theorem 2 as a criterion for obtaining hierarchical loglinear models from misclassified data directly if the misclassification is non-differential and independent.

Theorem 2 *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 2 clearly specifies the criterion of the preserved models, i.e., any randomized variable cannot

Table 4.4: Goodness-of-Fit tests for loglinear models on attributes A, B, E after Randomization with different $(p^{(A)}, p^{(B)}, p^{(E)})$

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

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 4.2 is preserved on all randomized data with different distortion parameters as shown in Table 4.3. We can see that the p-value of model (AD, G) is always prominent no matter how we change the distortion parameters $(p^{(A)}, p^{(D)}, p^{(G)})$. 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 on all the randomized data with different distortion parameters as shown in Table 4.4. We can observe when $p^{(A)} = 0.55$, $p^{(B)} = 0.9$ and $p^{(E)} = 0.9$, 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 [61], the authors proposed the use of the complete independence model (all 1-factor effects and the mean γ) to measure significance of dependence. In [21], 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 4.3 while the all-two-factor model (AB, AC, BC) is shown in Equation 4.4.

$$\log \hat{y}_{ijk} = \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C \tag{4.3}$$

$$\log \hat{y}_{ijk} = \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{jk}^{BC} \tag{4.4}$$

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.

$$pr_{AB.C} = \frac{r_{AB} - r_{AC}r_{BC}}{\sqrt{(1 - r_{AC}^2)(1 - r_{BC}^2)}} \tag{4.5}$$

Equation 4.5 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 10 *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 11 *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.

4.3.3 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.

Result 12 *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.*

Proof The proof is given for three binary variables with the saturated model; the extension to higher dimensions is immediate.

Equation 4.6 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.

$$\begin{aligned}
\gamma &= l_{...} \\
\gamma_i^A &= l_{i..} - \gamma \\
&\cdots \\
\gamma_{ij}^{AB} &= l_{ij.} - \gamma_i^A - \gamma_j^B - \gamma \\
&\cdots \\
\gamma_{ijk}^{ABC} &= l_{ijk} - \gamma_{ij}^{AB} - \gamma_{ik}^{AC} - \gamma_{jk}^{BC} - \gamma_i^A - \gamma_j^B - \gamma_k^C - \gamma
\end{aligned} \tag{4.6}$$

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 12 is proved.

4.4 Effects on Data Mining Applications

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

Association Rule Mining

Association rule learning is a widely used method for discovering interesting relations between items in data mining [3]. 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 4 and Result 5, 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.

Decision Tree Learning

Decision tree learning is a procedure to determine the class of a given instance [53]. 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 13 *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:

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

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

After randomization with distortion parameters $p_0^{(A)} = p_1^{(A)} = 0.6$ and $p_0^{(B)} = p_1^{(B)} = 0.9$ ($\boldsymbol{\lambda}^{ABC}$=(0.2629, 0.1127, 0.1042, 0.0143, 0.2837, 0.0873, 0.1240, 0.0109)$'$), we get:

$$\begin{aligned} gini_{split}(A)_{ran} &= 0.35 \\ gini_{split}(B)_{ran} &= 0.34 \end{aligned}$$

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

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 14 *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 $p_0^{(A)}$=$p_1^{(A)}$=$p_0^{(G)}$=$p_1^{(G)} = 0.6$ ($\boldsymbol{\lambda}^{AGH}$ = (0.1579,0.0848, 0.1351, 0.1163, 0.1643, 0.0845, 0.1408,0.1162)$'$), we get:

$$\begin{aligned} P(\overline{H}|\overline{A}G)_{ran} &= 0.54 \\ P(H|\overline{A}G)_{ran} &= 0.46 \end{aligned}$$

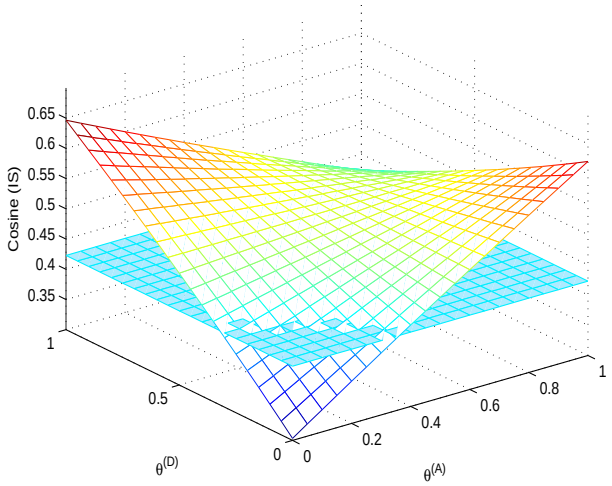
As none of $\pi_{\overline{A}\overline{H}}, \pi_{G\overline{H}}, \pi_{\overline{A}H}, \pi_{GH}$ has monotonic properties after randomization, the relative order of the two probabilities $P(\overline{H}|\overline{A}G)$ and $P(H|\overline{A}G)$ cannot be kept.

4.5 Summary

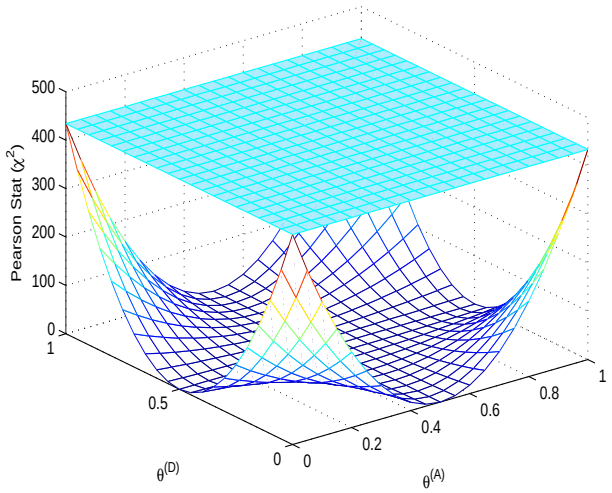
The trade-off between privacy preservation and utility loss has been extensively studied in privacy preserving data mining. However, data owners are still reluctant to release their (perturbed or transformed) data due to privacy concerns. In this chapter, we focus on the scenario where distortion parameters 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.

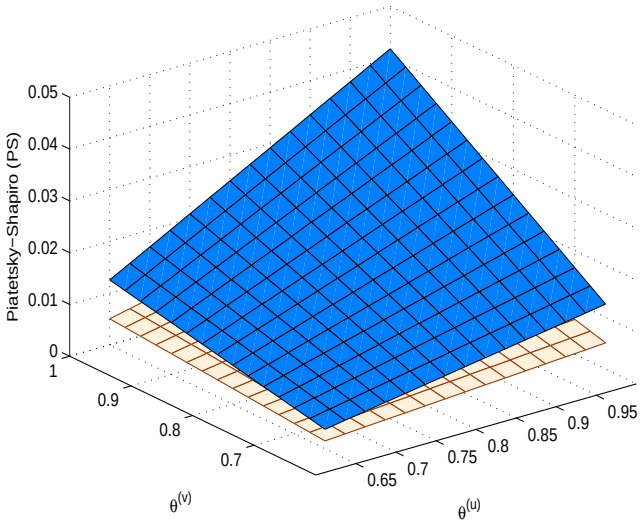


(a) Cosine

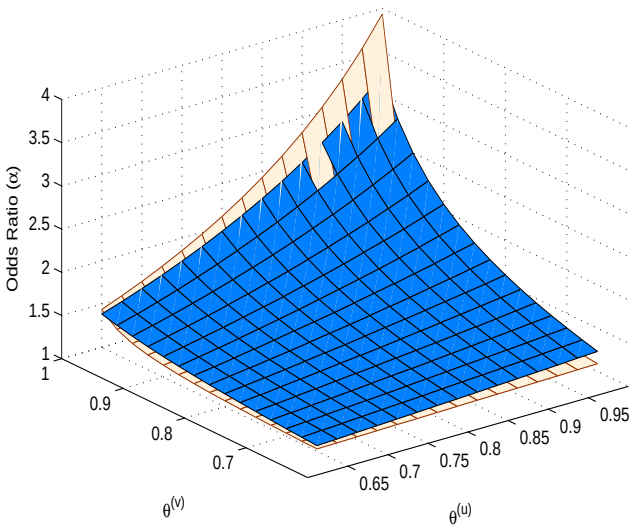


(b) ChiSquare

Figure 4.1: Statistics calculated from original data A, D (flat surface) vs. statistics calculated from randomized data (varied surface) with varying $p^{(A)}$ and $p^{(D)}$



(a) PS



(b) OddsRatio

Figure 4.2: Statistics from randomized data of (A,B) (shown as blue surface) and (I,J) (shown as brown surface) with varying $p^{(u)}$ and $p^{(v)}$

CHAPTER 5: ATTRIBUTE DISCLOSURE UNDER LINKING ATTACKS

Data utility and privacy of individuals are always a trade-off in privacy preserving data mining. Perfect privacy can be achieved without sharing any data, but it offers no utility; perfect utility can be obtained by publishing the data directly, but it disclose privacy of individuals. In this chapter, we will investigate the privacy of Randomization method and focus on attribute disclosure under linking attacks.

In Section 5.1, we will overview different measures and research issues on privacy in randomization models in the literature. In Section 5.2, we quantify attribute disclosure under linking attacks and then show our theoretical results on maximizing utility with privacy constraints. In Section 5.3, we conduct empirical evaluations and compare the randomization based distortion with two representative group based anonymization approaches (l-diversity [49] and anatomy [76]). We conclude our work in Section 5.4. Some results in this chapter were previously reported in [30].

5.1 Introduction

The privacy disclosure of randomization was first discussed in [44] for the traditional Warner model [72]. The posterior probabilities that a respondent belongs to group A and $\bar{A}$, respectively, when he reports R are $P_r(A|R)$ and $P_r(\bar{A}|R)$.

$$\begin{aligned} P_r(A|R) &= \frac{\pi_A P_r(R|A)}{\pi_A P_r(R|A) + (1-\pi_A) P_r(R|\bar{A})} \\ P_r(\bar{A}|R) &= 1 - P_r(A|R) \end{aligned}$$

The response R is regarded as jeopardizing with respect to A or $\bar{A}$ if:

$$P_r(A|R) > \pi_A \quad or \quad P_r(\bar{A}|R) > 1 - \pi_A$$

The more $P_r(A|R)$ and $P_r(\bar{A}|R)$ depart from original distribution π_A and $\pi_{\bar{A}}$, the more privacy will be disclosed. Since:

$$\frac{P_r(A|R)}{P_r(\bar{A}|R)}\frac{1-\pi_A}{\pi_A} = \frac{P_r(R|A)}{P_r(R|\bar{A})}$$

They proposed the following measures of jeopardy carried by R about A and $\bar{A}$, respectively:

$$g(R|A) = \frac{P_r(R|A)}{P_r(R|\bar{A})} \quad g(R|\bar{A}) = \frac{1}{g(R|A)}$$

The response R is nonjeopardizing if and only if $g(R|A) = 1$.

As an unbiased estimate of π_A is:

$$\hat{\pi}_A = \frac{\hat{\lambda} - P_r(R|\bar{A})}{P_r(R|A) - P_r(R|\bar{A})}$$

which is defined if and only if $P_r(R|A) - P_r(R|\bar{A}) \neq 0$, it necessarily makes a response jeopardic with respect to either A or $\bar{A}$. The problem becomes one of constrained optimization which is maximizing data utility with fixing maximal allowable levels of $g(R|A)$ and $g(R|\bar{A})$. Extensions to polychotomous models can refer [11].

Evfimievski et. al. [23] extended the concept of posterior probability and defined the following privacy breach in privacy preserving association rule mining.

Definition 5 *An itemset A causes a privacy breach of level ρ if for some item $a \in A$ and some $i \in 1 \ldots N$ we have $P_r[a \in t_i | A \subseteq t_i'] \geq \rho$.*

The limitation of this measure is it doesn't consider the prior probability of $a \in t_i$. It's not necessarily breach the privacy if $P_r[a \in t_i]$ is high in the original data set. Authors in [54] proposed to use *Reconstruction Probability* to quantify the privacy in MASK scheme. The proposed metric is to answer the question: "With what probability can a given value in the original data be reconstructed?".

Given that the customer indeed did buy item i, the probability of her original $'1'$

can be reconstructed from the distorted entry was expressed as:

$$\begin{aligned} R_1(p, s_i) &= P_r\{Y_i = 1 | X_i = 1\} \times P_r\{X_i = 1 | Y_i = 1\} \\ &+ P_r\{Y_i = 0 | X_i = 1\} \times P_r\{X_i = 1 | Y_i = 0\} \end{aligned}$$

X_i denotes the original entry and Y_i denotes the distorted entry. This expression captures the "round-trip" of going from the true data to the distorted data and then returning to guess the contents of the true one. They further define the total reconstruction probability $R(p)$ as:

$$R_p = aR_1(p) + (1 - a)R_0(p)$$

The privacy measure is defined as:

$$P_p = (1 - R(p)) * 100$$

In [22], the privacy disclosure is measured by estimating the change in probability of a property from original to randomized data. The authors defined a $\rho_1 - to - \rho_2$ privacy breach as following:

Definition 6 *There is a $\rho_1 - to - \rho_2$ privacy breach with respect to property $Q(x)$ if for some $y \in V_Y$, we have:*

$$P[Q(x)] \leq \rho_1 \quad and \quad P[Q(x)|Y = y] \geq \rho_2.$$

Here $0 < \rho_1 < \rho_2 < 1$ and $P[Y = y] > 0$.

They also discussed the sufficient condition for guaranteeing no (ρ_1, ρ_2) privacy breach is:

$$\frac{p[x_1 \longrightarrow y]}{p[x_2 \longrightarrow y]} \leq \gamma$$

This means if we look back from randomized value y, it's difficult to tell whether it was distorted from x_1 or x_2 in the original data. Authors in [6] expressed the notation

of a $\rho_1 - to - \rho_2$ privacy breach in terms of properties of the Markov transition matrix. They further expressed the utility of the scheme in terms of the condition number of the transition matrix.

5.2 Attribute Disclosure under Linking Attacks

5.2.1 Motivation

In privacy preserving data publishing, identity attributes are often directly removed in order to preserve privacy of individuals whose data are in the released table. However, the QI information may be used by attackers to link with other public datasets to get the private information of individuals, which is recognized as *linking attacks* in micro data publishing.

Two types of information disclosures have been identified under linking attacks: *identity disclosure* and *attribute disclosure* [41]. Identity disclosure occurs if attackers can identify an individual from the released data. Attribute disclosure occurs when confidential information about an individual is revealed and can be attributed to the individual. Attribute disclosure may occur when confidential information is revealed exactly or when it can be closely estimated. Thus, attribute disclosure can both incur identification of the individual and comprise her confidential information.

To counter linking attacks based on quasi-identifiers, Samarati and Sweeney [58] proposed k-anonymity model and presented a generalization approach that divides tuples into QI-group by transforming their QI-values into less specific forms such that tuples in the same QI-group cannot be uniquely identified by attackers. It was identified by Machanavajjhala et. al. [49] that k-anonymity is vulnerable to homogeneity and background knowledge attacks when data in QI-group lacks diversity in the sensitive attributes. To protect attribute disclosure, l-diversity [49] as well as other following models (e.g., t-closeness [45]) were proposed recently. l-diversity requires that the sensitive attribute has at least l *well-represented* values for each equivalence class in the generalized dataset. Later, Xiao and Tao [76] proposed the

anatomy method to improve the data utility of anonymized data while achieving the same privacy preservation as l-diversity.

Instead of generalizing QI attribute values to high hierarchies, randomization approach distorts the original value to other domain value according to some distortion probabilities. Teng and Du [70] studied the application of randomization technique to prevent identity disclosure under linking attacks in data publishing. They focused on evaluating the risk of successfully linking a target individual to the index of his record given values of QI attributes.

Our research moves one step further to investigate attribute disclosure under linking attacks. We focus on evaluating the risk of successfully predicting the sensitive attribute value of a target individual given his QI attribute values. We present a general randomization framework and give efficient solutions to determine optimal randomization parameters for both QI and sensitive attributes. As a result, we can maximize data utility while satisfying privacy preservation requirements for sensitive attributes.

Within the framework, we compare our randomization approach with other anonymization approaches (e.g., two representative approaches l-diversity [49] and anatomy [76] are used in this section). Our result shows that randomization approach can better recover the distribution of original data set from the disguised one. Thus intuitively, it might yield a disguised database with higher data utility than l-diversity generalization and anatomy.

The remainder of this section is organized as follows. In Section 5.2.2, we present backgrounds on randomization based distortions including analysis and attacks on the randomized data. In Section 5.2.3, we quantify attribute disclosure under linking attacks. In Section 5.2.4, we give efficient solutions to determine optimal randomization parameters for both QI and sensitive attributes.

5.2.2 Preliminaries

Dataset $\mathcal{T}$ contains N records and $m+1$ categorical attributes: $A_1, A_2, \ldots, A_m$ and S. We use $QI = \{A_1, \ldots, A_m\}$ to denote the set of quasi-identifier attributes (e.g., demographic) whose values may be known to the attacker for a given individual and use S to denote one sensitive attribute whose value should not be associated with an individual by attackers. Generally, $\mathcal{T}$ may also contain other attributes, which are neither sensitive nor quasi-identifying. Those attributes are usually kept unchanged in the released data. We exclude them from our setting since they do not incur privacy disclosure risk or utility loss. All of the discussions in this section are also explained in the single sensitive attribute setting and can be generalized to multiple sensitive attributes.

Attribute $A_i(S)$ has $d_i(d_s)$ categories denoted by 0,1,..., $d_i - 1(d_s - 1)$. We use $\Omega_i(\Omega_s)$ to denote the domain of $A_i(S)$, $\Omega_i = \{0, 1, \ldots, d_i - 1\}$, and $\Omega_{QI} = \Omega_1 \times \cdots \times \Omega_m$ is the domain of quasi-identifiers. The r-th record R_r is denoted by $(A_{1r}, A_{2r}, \ldots, A_{mr}, S_r)$ or simply (QI_r, S_r). Let $D = d_s \prod_{i=1}^{m} d_i$ be the total number of cells in the contingency table.

Let $\pi_{i_1,\cdots,i_m,i_s}$ denote the true proportion corresponding to the categorical combination $(A_1 = i_1, \cdots, A_m = i_m, S = i_s)$. Let $\boldsymbol{\pi}$ be the column vector with D elements $\pi_{i_1,\cdots,i_m,i_s}$ arranged in a fixed order. Table 2.5(b) shows one contingency table instance derived from the data set with 100 tuples. We will use this contingency table as an example to illustrate properties of link disclosure.

Distortion Procedure

We use the upper part of Figure 5.1 to illustrate the process of privacy preserving data publishing. For each records R_j, the data owner independently randomizes attribute A_i using the distortion matrix P_i. Specifically, for attribute A_i (or S) with d_i categories, the randomization process is to change a record belonging to the v-th

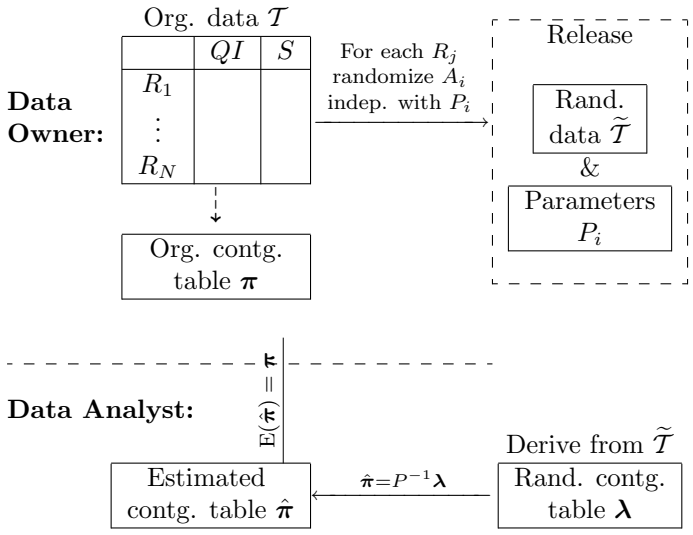


Figure 5.1: Randomization based privacy-preserving data publishing

category ($v = 0, \ldots, d_i - 1$) to the u-th category with probability $p_{uv}^{(i)}$:

$$\Pr(\widetilde{A}_i = u | A_i = v) = p_{uv}^{(i)}.$$

Let $P_i = \left[p_{uv}^{(i)}\right]_{d_i \times d_i}$, and we call P_i (or P_s) the distortion matrix for A_i(or S). Naturally, the column sums of P_i are equal to 1. The original database $\mathcal{T}$ is changed to $\widetilde{\mathcal{T}}$, and then both the randomized data set and the distortion matrices are published. The randomization matrices indicate the magnitude of the randomization, which can help data analysts to estimate the original data distribution.

Let $\boldsymbol{\lambda}$ denote the contingency table of the randomized data $\widetilde{\mathcal{T}}$. We arrange $\boldsymbol{\lambda}$ into the column vector with the same order of $\boldsymbol{\pi}$. Table 2.5(c) shows one example of the randomized contingency table. The randomized contingency table has a close relationship with the original contingency table and the randomization matrices:

$$\mathrm{E}(\boldsymbol{\lambda}) = P\,\boldsymbol{\pi}, \tag{5.1}$$

where $P = P_1 \otimes \cdots \otimes P_m \otimes P_s$, and $\otimes$ stands for the Kronecker product[4].

[4]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

Distortion matrices determine the privacy and utility of the randomized data. Several specific distortion matrices have been investigated in the literature [5,6,11]. How to find optimal distortion parameters with privacy or utility constraints has been remained as a challenging problem [11]. Huang and Du [35] applied an evolutionary multi-objective optimization method to search for optimal distortion matrices from the entire search space for a single attribute. However, the method is difficult to handle multiple attributes due to its high complexity.

In this section, we limit the choice of randomization parameters for each QI attribute A_i (and sensitive attribute S) as:

$$\Pr(\widetilde{A}_i = u | A_i = v) = p_{uv}^{(i)} = \begin{cases} p_i, & u = v, \\ q_i = \frac{1-p_i}{d_i - 1}, & u \neq v. \end{cases} \tag{5.2}$$

In other words, for each attribute A_i, all categories have the same probability p_i to remain unchanged, and have the same probability q_i to be distorted to a different category. With this choice limit, we can derive an efficient algorithm (with explicit formula) to determine the optimal randomization parameters (as shown in Section 5.2.4).

Attacks on Randomized Data

Let X be a content in the data set $\mathcal{T}$ with domain Ω_X, and $\widetilde{X}$ is the randomized value of X in $\widetilde{\mathcal{T}}$. With the randomization process and parameters, it is not reasonable for attackers to regard the observed value as the true value of X. Instead, attackers can try to estimate the original value based on the observed data and the released randomization parameters. Let $\widehat{X}$ denote attackers' estimation on the original value of X. Due to the randomization procedure, any value in Ω_X is possible. We assume that the attacker is able to calculate the posterior probability of a content in the data

set and takes the following probabilistic strategy: for any $\mu, \nu \in \Omega_X$,

$$\widehat{X} = \mu \text{ with prob. } \Pr(X = \mu | \widetilde{X} = \nu), \tag{5.3}$$

where $\Pr(X = \mu | \widetilde{X} = \nu)$ denotes the attacker's posterior belief on the original value $X = \mu$ when he observes $\widetilde{X} = \nu$. With the Bayes' theorem, it can be calculated by:

$$\Pr(X = \mu | \widetilde{X} = \nu) = \frac{\pi_\mu \Pr(\widetilde{X} = \nu | X = \mu)}{\sum_{\omega \in \Omega_X} \pi_\omega \Pr(\widetilde{X} = \nu | X = \omega)}. \tag{5.4}$$

The following lemma gives the accuracy of the attacker's estimation.

Lemma 2 *Suppose attackers adopt the probabilistic strategy specified in* (5.3) *to estimate the data. The probability that attackers accurately estimate the original value of X is given by:*

$$\begin{aligned} &\Pr(\widehat{X} = X = \mu) \\ &= \sum_{\nu \in \Omega_X} \Pr(\widetilde{X} = \nu | X = \mu) \Pr(X = \mu | \widetilde{X} = \nu). \end{aligned} \tag{5.5}$$

Proof For a particular observed value $\widetilde{X} = \nu \in \Omega_X$,

$$\begin{aligned} &\Pr(\widehat{X} = X = \mu, \widetilde{X} = \nu) \\ &= \Pr(\widetilde{X} = \nu | X = \mu) \Pr(X = \mu | \widetilde{X} = \nu). \end{aligned}$$

Then, with the law of total probability, we have

$$\begin{aligned} &\Pr(\widehat{X} = X = \mu) \\ &= \sum_{\nu \in \Omega_X} \Pr(\widehat{X} = X = \mu, \widetilde{X} = \nu) \\ &= \sum_{\nu \in \Omega_X} \Pr(\widetilde{X} = \nu | X = \mu) \Pr(X = \mu | \widetilde{X} = \nu). \end{aligned}$$

The probability of attackers' correct estimation is also defined as the reconstruction probability, $\Pr(X = \mu \rightarrow X = \mu)$, in [54, 70]. Rizvi and Haritsa [54] used

the reconstruction probability to measure privacy protection in randomization based privacy preserving association rule mining. Teng and Du [70] adopted the reconstruction probability to evaluate identity disclosure under linking attacks in micro data publishing. They focused on the problem of successful identification of the record t given its quasi-identifier values. To make the notation concise and consistent with the previous work, we adopt the same notation presented in [70], i.e.,

$$\Pr(X = \mu \rightarrow X = \mu) = \Pr(\widehat{X} = X = \mu),$$

to evaluate the risk of the sensitive attribute disclosure.

5.2.3 Quantification of Attribute Disclosure

We measure privacy in terms of the attribute disclosure risk, whose formal definition is given as follows:

Definition 7 *The attribute disclosure risk under linking attacks is defined to be the probability that the attacker predicts S_r successfully given QI_r of a target individual r, denoted as* $\Pr(S_r|QI_r)$.

To derive the disclosure probability $\Pr(S_r|QI_r)$, we need to quantify the background knowledge of attackers. We have the following standard assumptions for background knowledge of attackers in this dissertation. We assume that the attacker has access to the published data set $\widetilde{\mathcal{T}}$ and he knows that $\widetilde{\mathcal{T}}$ is a randomized version of some base table $\mathcal{T}$. The attacker knows the domain of each attribute of $\mathcal{T}$. We also assume that the attacker can obtain the QI-values of the target individual (e.g., Alice) from some public database or background knowledge and knows that the target individual is definitely contained in the published data. However, he has no knowledge on which record in the published data belongs to the target individual. Finally, we assume that the distortion matrices P_i are available to the attacker, because they are necessary for data miners to conduct analysis.

l-diversity preserves privacy by generalizing the QI attributes to form QI-groups. Each QI-groups contains at least l well-presented sensitive values. Individuals in the group are linked to any sensitive attributes with probability at most $1/l$, i.e., $\Pr(S_r|QI_r) \leq 1/l$. Anatomy achieves the same preservation by publishing QIT and ST tables separately for each l-diverse partitioned group. Xiao and Tao [76] proved that, given a pair of QIT and ST, attackers can correctly infer the sensitive value of any individual with probability at most $1/l$. However, randomization based approach achieves the privacy protection in a probabilistic manner. In the following subsection, we show how to quantify the attribute disclosure risk in the randomization settings.

Quantifying Attribute Disclosure

When there is no randomization applied, for those records with their quasi-identifiers equal to QI_r, the attacker simply regards that every record has the same probability to be individual r. Because there are totally π_{QI_r} records within the group of QI_r and π_{QI_r,S_r} of them have the sensitive value equal to S_r, the risk of sensitive attribute disclosure is equal to $\frac{\pi_{QI_r,S_r}}{\pi_{QI_r}}$. This case corresponds to the worst case of the attribute disclosure risk. When randomization is applied, the attribute disclosure risk will be reduced, because the randomization increases the attacker's uncertainty.

Randomize S only (*RR-S*). When data owners only apply randomization to the sensitive attribute, for each record within the group of QI_r, the attacker takes a guess on its sensitive value using the observed sensitive value and the posterior probability in (5.4). According to (5.5), the probability of a correct estimation is $\Pr(S_r \to S_r|QI_r)$, then the risk of sensitive attribute disclosure is $\frac{\pi_{QI_r,S_r}}{\pi_{QI_r}} \Pr(S_r \to S_r|QI_r)$.

Randomize QI only (*RR-QI*). Similarly as RR-S, when data owners only apply randomization to the quasi-identifiers, the probability of correctly reconstructing QI_r is given by $\Pr(QI_r \to QI_r)$, and hence the risk of sensitive attribute disclosure is $\frac{\pi_{QI_r,S_r}}{\pi_{QI_r}} \Pr(QI_r \to QI_r)$.

Randomize QI and S (*RR-Both*). When data owners apply randomization to both QI and S, the attacker first need to make sure the values of identifier attributes are correctly reconstructed. The probability is given by $\Pr(QI_r \to QI_r)$. Second, the attacker needs to make sure the value of the sensitive attribute given the correctly reconstructed identifier attribute values is correctly reconstructed.

We summarize the risk of sensitive attribute disclosure in *RR-Both* as well as *RR-S* and *RR-QI* into the following result and give the general calculation of the attribute disclosure risk in the randomization settings:

Result 15 *Assume an individual r has quasi-identifier $QI_r = \alpha = \{i_1, i_2, \ldots, i_m\}$, $i_k \in \Omega_k$, and his sensitive attribute $S_r = u$, $u \in \Omega_s$. The probability of successful predicting the sensitive attribute value S_r of the target individual r given his quasi-identifier values QI_r is:*

$$\Pr(S_r|QI_r) = \frac{\pi_{QI_r,S_r}}{\pi_{QI_r}} \Pr(QI_r \to QI_r) \Pr(S_r \to S_r|QI_r). \tag{5.6}$$

We give the formal expressions of the two reconstruction probabilities needed in calculating $\Pr(S_r|QI_r)$ in (5.6). The reconstruction probability of the quasi-identifier is given by:

$$\begin{aligned} &\Pr(QI_r = \alpha \to QI_r = \alpha) \\ &= \sum_{\beta \in \Omega_{QI}} \Pr(\widetilde{QI}_r = \beta|QI_r = \alpha) \Pr(QI_r = \alpha|\widetilde{QI}_r = \beta) \\ &= \sum_{\beta \in \Omega_{QI}} \frac{\pi_\alpha \big[\Pr(\widetilde{QI}_r = \beta|QI_r = \alpha)\big]^2}{\sum_{\gamma \in \Omega_{QI}} \pi_\gamma \Pr(\widetilde{QI}_r = \beta|QI_r = \gamma)}, \end{aligned} \tag{5.7}$$

where $\Pr(\widetilde{QI}_r = \beta|QI_r = \alpha) = \prod_{k=1}^m p_{j_k i_k}^{(k)}$ is the probability that $\alpha = \{i_1, i_2, \ldots, i_m\}$ is distorted to $\beta = \{j_1, j_2, \ldots, j_m\}$.

The reconstruction probability of the sensitive attribute S for the target individual with the quasi-identifier QI_r is given by:

$$\begin{aligned}
&\Pr(S_r = u \to S_r = u|QI_r) \\
&= \sum_{v\in\Omega_s} \Pr(\widetilde{S} = v|S = u, QI_r) \Pr(S = u|\widetilde{S} = v, QI_r) \\
&= \sum_{v\in\Omega_s} p_{vu}^{(s)} \Pr(S = u|\widetilde{S} = v, QI_r) \\
&= \sum_{v\in\Omega_s} \frac{\left[p_{vu}^{(s)}\right]^2 \pi_{QI_r,S=u}}{\sum_{t\in\Omega_s} p_{vt}^{(s)} \pi_{QI_r,S=t}}. \qquad (5.8)
\end{aligned}$$

We are interested in when the attribute disclosure reaches the minimum. We show our results in the following property and include our proof in Appendix.

Property 1 *Given QI_r of individual r,*

- *for RR-S,* $\Pr(S_r|QI_r)$ *is minimized when* $p_s = \frac{1}{d_s}$, $\min \Pr(S_r|QI_r) = \left(\frac{\pi_{QI_r,S_r}}{\pi_{QI_r}}\right)^2$;
- *for RR-QI,* $\Pr(S_r|QI_r)$ *is minimized when* $p_i = \frac{1}{d_i}$ $(i = 1, 2, \ldots, m)$, $\min \Pr(S_r|QI_r) = \pi_{QI_r,S_r}$;
- *for RR-Both,* $\Pr(S_r|QI_r)$ *is minimized when* $p_i = \frac{1}{d_i}$ $(i = 1, 2, \ldots, m,$ *and* $s)$, $\min \Pr(S_r|QI_r) = \frac{\pi^2_{QI_r,S_r}}{\pi_{QI_r}}$.

Proof We start with applying *RR* to the sensitive attribute, $\Pr(QI_r \to QI_r) = 1$. Without loss of generality, we assume that $S_r = 1$. Combine other categories $0, 2, \ldots, d_s - 1$ into a new categories, and still use 0 to denote the new category. To make the notation simple, we simply write $\frac{\pi_{QI_r,1}}{\pi_{QI_r}}$ as π_1 and $\frac{\pi_{QI_r,0}}{\pi_{QI_r}}$ as π_0 in this proof, then $\pi_1 = 1 - \pi_0$. Such adjustment does not change the reconstruction probability $\Pr(S_r = 1 \to S_r = 1|QI_r)$. After the adjustment, the randomization probabilities are

given by

$$\Pr(\tilde{S}_r = 1|S_r = 1) = p_s \tag{5.9}$$

$$\Pr(\tilde{S}_r = 1|S_r = 0) = q_s. \tag{5.10}$$

By definition, the posterior probabilities are given by

$$\Pr(S_r = 1|\tilde{S}_r = 1) = \frac{p_s\pi_1}{p_s\pi_1 + q_s\pi_0}, \tag{5.11}$$

$$\Pr(S_r = 1|\tilde{S}_r = 0) = \frac{(1-p_s)\pi_1}{(1-p_s)\pi_1 + (1-q_s)\pi_0}. \tag{5.12}$$

Combining (5.9), (5.10), (5.11) and (5.12), we have

$$\begin{aligned}
&\Pr(S_r = 1|QI_r) \\
=&\pi_1 \Pr(S_r = 1 \rightarrow S_r = 1|QI_r) \\
=&\pi_1 \sum_{i=0,1} \Pr(\tilde{S}_r = i|S_r = 1) \Pr(S_r = 1|\tilde{S}_r = i) \\
=&\pi_1^2 \left[\frac{p_s^2}{p_s\pi_1 + q_s\pi_0} + \frac{(1-p_s)^2}{(1-p_s)\pi_1 + (1-q_s)\pi_0}\right]
\end{aligned} \tag{5.13}$$

Taking derivative with respect to p_s, we have (5.13) is minimized when $p_s = \frac{1}{d_s}$, and the minimal value is

$$\pi_1^2 = \left(\frac{\pi_{QI_r,S_r}}{\pi_{QI_r}}\right)^2.$$

Following similar strategies, we can prove the general case when we randomize both QI and S.

Example. We use the instance shown in Table 2.5(b) to illustrate this property. For an individual r with ($QI_r = Female$, $S_r = Cancer$), we randomize S (Disease) with p_s and QI (Gender) with p_G independently. Figure 5.2 shows how the attribute disclosure is varied when we apply different randomization parameters. We can see that $\Pr(S_r|QI_r)$ reaches the maximum (i.e., $\frac{\pi_{10}}{\pi_{1+}} = 0.43$) when no randomization is introduced. Figure 5.2(a) shows the scenario when we only randomize S. We can

see that $\min \Pr(S_r|QI_r) = (\frac{\pi_{10}}{\pi_{1+}})^2 = 0.18$ when $p_s = \frac{1}{d_s} = \frac{1}{3}$. Figure 5.2(b) shows the scenario when we only randomize QI. We can see that $\min \Pr(S_r|QI_r) = \pi_{10} = 0.12$ when $p_G = \frac{1}{d_G} = \frac{1}{2}$. Figure 5.2(c) shows the case where randomization is applied to both QI and S. $\Pr(S_r|QI_r)$ reaches the minimum only when both $p_s = \frac{1}{3}$ and $p_G = \frac{1}{2}$, and $\min \Pr(S_r|QI_r) = \frac{\pi_{10}^2}{\pi_{1+}} = 0.05$.

Reducing Computational Cost

The main computation cost in (5.6) comes from calculating $\Pr(QI_r \rightarrow QI_r)$. Let P_{QI} be the distortion matrix on quasi-identifers: $P_{QI} = P_1 \otimes \cdots \otimes P_m$, $\boldsymbol{\pi}_{QI}$ be the contingency table on all quasi-identifiers arranged into a column vector, and $\boldsymbol{\lambda}_{QI}$ denote the expected QI contingency table of the randomized data: $\boldsymbol{\lambda}_{QI} = P_{QI}\, \boldsymbol{\pi}_{QI}$. Then, the denominator in (5.7) is exactly the cell of $\boldsymbol{\lambda}_{QI}$ corresponding to β. Let $\boldsymbol{\eta}$ denote the column vector of the reconstruction probabilities of quasi-identifiers, arranged in the same order of $\boldsymbol{\pi}_{QI}$. We can further express (5.7) in matrix form:

$$\begin{aligned} \boldsymbol{\eta} &= \boldsymbol{\pi}_{QI} \dot{\times} \left[\left(P_{QI}^{\dot{2}} \right)^T \left(\boldsymbol{\lambda}_{QI}^{\dot{-1}} \right) \right] \\ &= \boldsymbol{\pi}_{QI} \dot{\times} \left\{ \left[\otimes_{i=1}^{m} \left(P_i^{\dot{2}} \right)^T \right] \left[\left(\otimes_{i=1}^{m} P_i \right) \boldsymbol{\pi}_{QI} \right]^{\dot{-1}} \right\} \end{aligned} \tag{5.14}$$

where $\dot{\times}$ denotes the componentwise multiplication, $P_i^{\dot{2}}$ is componentwise square of P_i, and $\boldsymbol{\lambda}_{QI}^{\dot{-1}}$ is componentwise inverse of $\boldsymbol{\lambda}_{QI}$.

In (5.14), we need repeatedly calculate $(P_1 \otimes \cdots \otimes P_m)\boldsymbol{x}$, where $\boldsymbol{x}$ denotes a column vector. Assume we use the naive algorithm in all matrix multiplications. Calculating $(P_1 \otimes \cdots \otimes P_m)\boldsymbol{x}$ directly results in the time and storage complexity of $O(\prod_i d_i^2)$. The main storage complexity is from storing matrix P_{QI}. The following lemma allows us to reduce the cost of such computation:

Lemma 3 *Let A, B and X be the matrices of size $n \times n$, $m \times m$, and $m \times n$. Then*

$$(A \otimes B)\,\mathrm{vec}(X) = \mathrm{vec}(BXA^T),$$

where $\mathrm{vec}(X)$ *denotes the vectorization of the matrix* X *formed by stacking the columns of* X *into a single column vector.*

Applying Lemma 3 recursively, we can reduce the time complexity to $O([\sum_i d_i] \prod_i d_i)$. The storage complexity is also reduced to $O(\prod_i d_i + \sum_i d_i^2)$, which is mainly used to store the contingency table.

5.2.4 Maximizing Utility with Privacy Constraints

Analysis on Randomized Data

One advantage of randomization procedure is that the pure random process allows data analysts to estimate the original data distribution based on the released data and the randomization parameters. The bottom part of Figure 5.1 shows how data analysts estimate the original data distribution. With the randomized data $\widetilde{\mathcal{T}}$ and its contingency table $\boldsymbol{\lambda}$, according to (5.1), the unbiased estimate of $\boldsymbol{\pi}$ is given by

$$\hat{\boldsymbol{\pi}} = P^{-1}\boldsymbol{\lambda}, \tag{5.15}$$

and the covariance matrix of the estimator in (5.15) is given by

$$\begin{aligned} \Sigma &= \mathrm{Cov}(\hat{\boldsymbol{\pi}}) \\ &= \frac{1}{N}\left[P^{-1}(P\,\boldsymbol{\pi})^{\delta}(P^T)^{-1} - \boldsymbol{\pi}\,\boldsymbol{\pi}^T\right], \end{aligned} \tag{5.16}$$

where $(P\,\boldsymbol{\pi})^{\delta}$ stands for the diagonal matrix whose diagonal values are $P\,\boldsymbol{\pi}$.

In this way, data analysts derive the estimation for the distribution of original data (in terms of contingency table) without disclosing the individual information of each records. Because most of data mining applications are based on the probability distribution of the data, we choose the accuracy of reconstructed distribution as target utility function in the following efficient solution.

Efficient Solution

The ultimate goal of publishing data is to maximize utility while minimizing risk of attribute disclosure at the same time. Utility of any dataset, whether randomized or not, is innately dependent on the tasks that one may perform on it. Without a workload context, it is difficult to say whether a dataset is useful or not. Since many data mining applications are based on the probability distribution of the data, it motivates us to focus on the data distribution when evaluating the utility of a database.

Problem 1 *Determine* p_i, $i = 1, \cdots, m$, *and* p_s *to*

$$\begin{aligned} &\text{minimize} \ \ \mathrm{E}[\mathrm{d}(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi})] \\ &\text{s.t.} \ \ \max_r \Pr(S_r|QI_r) \leq \tau, p_i \in (1/d_i, 1]\,. \end{aligned} \tag{5.17}$$

where $\mathrm{d}(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi})$ *denotes a certain distance between* $\hat{\boldsymbol{\pi}}$ *and* $\boldsymbol{\pi}$.

To compare the performance of different disguised schemes, we can set the privacy threshold formalize as the same privacy requirement of l-diversity (i.e., $\tau = \frac{1}{l}$). In other words, we would determine the optimal randomization parameters ($p_1, p_2, \ldots, p_m$ and p_s) to maximize the utility while ensuring that the sensitive value of any individual involved in the dataset cannot be correctly inferred by an attacker with probability more than $1/l$. A larger l leads to stronger privacy protection. In general, privacy constraints may be flexible. For example, different individuals may have different concerns about their privacy so we can set different thresholds for $\Pr(S_r|QI_r)$. We can even adopt the *Uninformative Principle* [49] (i.e., there should not be a large difference between the prior and posterior beliefs due to the released data) and apply the relative privacy disclosure measures (e.g., (ρ_1, ρ_2)-privacy breach [22]).

Problem 1 is a nonlinear optimization problem. In general, we can solve it by using optimization packages (e.g., trust region algorithm [16]). In Section 5.2.3, we have discussed how to efficiently calculate the attribute disclosure of target individuals (shown as constraints in Problem 1). Next, we show how we can efficiently calculate

$E[\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2^2]$, the expected Euclidean distance difference between the original data and the estimated one.

Result 16 *When* $\mathrm{d}(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi})$ *is the squared Euclidean distance, Problem 1 is equivalent to: determine* p_i, $i = 1, \cdots, m$, *and* p_s *to*

$$\begin{aligned} &\text{minimize } \textstyle\prod_i \|P_i^{-1}\|_F^2 \\ &\text{s.t. } \max_r \Pr(S_r|QI_r) \leq \tau, p_i \in (1/d_i, 1]\,. \end{aligned} \tag{5.18}$$

Proof Lemma 4 *If* $\mathrm{d}(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}) = \|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2^2$, *we have* $\mathrm{E}[\mathrm{d}(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi})] = \mathrm{trace}(\Sigma)$, *where* Σ *is the covariance matrix of* $\hat{\boldsymbol{\pi}}$ *shown in* (5.16).

Proof We know that $\hat{\boldsymbol{\pi}}$ asymptotically follows the normal distribution $N(\boldsymbol{\pi}, \Sigma)$. Let $\Sigma = X\Lambda X^T$ be the eigen-decomposition of Σ, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $X^T X = I$. Let $\eta = X^T(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$, then η is normally distributed with $\mathrm{E}(\eta) = 0$ and

$$\mathrm{Cov}(\eta) = X^T \mathrm{Cov}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})X = X^T \Sigma X = \Lambda.$$

Notice that Λ is a diagonal matrix, and hence $\mathrm{Cov}(\eta_i \eta_j) = 0$ if $i \neq j$, and $\mathrm{Var}(\eta_i) = \lambda_i$, i.e., η_i independently follows the normal distribution $N(0, \lambda_i)$. Therefore, we have:

$$\begin{aligned} &\mathrm{E}\left[\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2^2\right] = \mathrm{E}\left[(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})^T(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})\right] \\ =& \mathrm{E}\left[(X\eta)^T(X\eta)\right] = \mathrm{E}(\eta^T\eta) = \mathrm{E}(\textstyle\sum_i \eta_i^2) \\ =& \textstyle\sum_i \mathrm{E}(\eta_i^2) = \sum_i \{\mathrm{Var}(\eta_i) + [\mathrm{E}(\eta_i)]^2\} \\ =& \textstyle\sum_i \lambda_i = \mathrm{trace}(\Lambda) = \mathrm{trace}(\Sigma). \end{aligned}$$

The last equality is due to the fact that

$$\mathrm{trace}(\Sigma) = \mathrm{trace}(X\Lambda X^T) = \mathrm{trace}(\Lambda X^T X) = \mathrm{trace}(\Lambda).$$

We proved the lemma.

Lemma 5 *Let P_i be the randomization matrix specified in* (5.2)*. When $p_i \neq \frac{1}{d_i}$, we have*

$$P_i^{-1} = \frac{1}{p_i - q_i}\left(I - q_i \mathbf{1}\mathbf{1}^T\right),$$

where $\mathbf{1}$ *is the column vector whose cells are all equal to 1. Moreover,*

$$\|P_i^{-1}\|_F^2 = \frac{(d_i - 1)^3}{(d_i p_i - 1)^2} + 1.$$

Proof We can re-write P_i as follows:

$$P_i = (p_i - q_i)I + q_i \mathbf{1}\mathbf{1}^T = (p_i - q_i)\left(I + \frac{q_i}{p_i - q_i}\mathbf{1}\mathbf{1}^T\right)$$

With $q_i = \frac{1-p_i}{d_i - 1}$, $\mathbf{1}^T\mathbf{1} = d_i$, and binomial inverse theorem [64], we can immediately get P_i^{-1}, and $\|P_i^{-1}\|_F^2$ can be directly calculated from P_i^{-1}.

Lemma 6 *Let $(P^{-1})_i$ denote the i-th column (or row) of P^{-1}. Then, for any $i = 1, 2, \ldots, D$, we have*

$$\|(P^{-1})_i\|_F^2 = \frac{1}{D}\|P^{-1}\|_F^2.$$

Proof With Lemma 5, we observe that every row (or column) of P_i^{-1} has the same components except for a change of order. Since $P^{-1} = P_i^{-1} \otimes \cdots \otimes P_m^{-1} \otimes P_s^{-1}$, we can also conclude that all rows (or columns) of P^{-1} has the same components except for a change of order. Therefore, for all i's,

$$\|(P^{-1})_i\|_F^2 = \frac{1}{D}\|P^{-1}\|_F^2.$$

Next, we prove the main result. With Lemma 4, when the distance is the squared Euclidean distance, to minimize E[d($\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}$)] is equivalent to minimize trace(Σ). With

(5.16), we have

$$\begin{aligned}
\text{trace}(\Sigma) &= \text{trace}\left\{\frac{1}{N}\left[P^{-1}(P\,\boldsymbol{\pi})^{\delta}(P^{T})^{-1} - \boldsymbol{\pi}\,\boldsymbol{\pi}^{T}\right]\right\} \\
&\propto \text{trace}\left[P^{-1}(P\,\boldsymbol{\pi})^{\delta}(P^{T})^{-1}\right] \\
&= \text{trace}\left[(P\,\boldsymbol{\pi})^{\delta}(P^{-1})^{T}P^{-1}\right]. \\
&= \sum_{i=1}^{D}(P\,\boldsymbol{\pi})_i\|(P^{-1})_i\|_F^2 \\
&= \|(P^{-1})_1\|_F^2\sum_{i=1}^{D}(P\,\boldsymbol{\pi})_i \quad \text{(with Lemma 6)} \\
&\propto \|P^{-1}\|_F^2 = \prod_i \|P_i^{-1}\|_F^2
\end{aligned}$$

Notice that the constraint function is an increasing function of p_i, the optimal solution must occur when the equality stands, and we have proved the result.

When the distance is the squared Euclidean distance, with Lemma (4) shown in Appendix, to minimize $\mathrm{d}(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi})$ is equivalent to minimize $\text{trace}(\Sigma)$ where Σ is the covariance matrix of the estimator of cell values in the contingency table (shown in Equation 5.16). Calculating $\text{trace}(\Sigma)$ still involves high computational cost. However, when P_i has the specific form shown in (5.2), we can further reduce the problem to minimizing $\prod_i \|P_i^{-1}\|_F^2$, and with Lemma 5,

$$P_i^{-1} = \frac{1}{p_i - q_i}\left(I - q_i\mathbf{1}\mathbf{1}^{T}\right),$$

where $\mathbf{1}$ is the column vector whose cells are all equal to 1. We have

$$\|P_i^{-1}\|_F^2 = \frac{(d_i - 1)^3}{(d_i p_i - 1)^2} + 1,$$

which can be directly calculated.

5.3 Empirical Evaluation

We ran our experiments on the Adult Database from the UCI data mining repository [9] in our evaluations. The same database has been used in previous work on

Table 5.1: Description of the *Adult* dataset used in the evaluation

Attribute	Type	Categories
Gender (G)	QI	2
Race (R)	QI	5
Education (E)	QI	16
Marital-status (M)	QI	7
Workclass (W)	Sensitive	7
Occupation (O)	Sensitive	14

k-anonymity, l-diversity, t-closeness, and anatomy [42,49,76,84]. The Adult Database contains 45,222 tuples from US Census data and 14 attributes. Table 5.1 is a summary description of the data including the attributes we used, the number of distinct values for each attribute, and the types of the attributes adopted in the evaluation.

It is expected that a good publication method should preserve both privacy and data utility. We set different l values as privacy disclosure thresholds. To quantify the utility, we adopt the following standard distance measures to compare the difference of distributions between the original and reconstructed data. Given two distributions $P = (p_1, p_2, ..., p_m)$,$Q = (q_1, q_2, ..., q_m)$,

- Variational distance: $\mathrm{d}_V(P,Q) = \sum_{i=1}^{m} \frac{1}{2}|p_i - q_i|$
- L_2 norm distance: $\mathrm{d}_{L_2}(P,Q) = (\sum_{i=1}^{m} |p_i - q_i|^2)^{\frac{1}{2}}$
- KL distance: $\mathrm{d}_{KL}(P,Q) = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i}$
- χ^2-distance: $\mathrm{d}_{\chi^2}(P,Q) = \sum_{i=1}^{m} \frac{(p_i - q_i)^2}{p_i}$

5.3.1 Randomization

We treated *Education, Martial-status, Gender, Race* as the quasi-identifier and used *Work-class* as the sensitive attribute. We call this dataset as EMGRW. For randomization, we distort only QI attributes, or sensitive attribute S, or both in different application scenarios. Table 5.2 shows the derived randomization parameter

Table 5.2: Randomization parameters p_i for three cases of RR (data set EMGRW)

(a) *RR-QI*

l	E	M	G	R
2	0.824	0.872	0.920	0.941
3	0.548	0.812	0.898	0.985
4	0.382	0.736	0.918	0.961
5	0.314	0.615	0.873	0.938

(b) *RR-S*

l	2	3	4	5
W	0.650	0.267	0.217	0.167

(c) *RR-Both*

l	E	M	G	R	W
2	0.824	0.872	0.920	0.941	1
3	0.573	0.821	0.913	0.973	0.955
4	0.428	0.780	0.926	0.953	0.871
5	0.353	0.688	0.902	0.953	0.813

p for three scenarios (*RR-QI*, *RR-S*, and *RR-Both*). We set $l = 2, 3, 4, 5$. We can observe that more distortions (p is away from 1) are needed in order to achieve better privacy protections (l is increased).

Results on various distance measures are shown in Figure 5.3. Naturally, there is a trade-off between minimizing utility loss and maximizing privacy protection. Figure 5.3 indicates that the smaller the distance values, the smaller the difference between the distorted and the original databases, and the better the utility of the distorted database. We can observe that the utility loss (in terms of distance differences) increases approximately linearly with the increasing privacy protection across all three randomization scenarios. *RR-Both* achieves the best in terms of utility preservation, because we use the optimal randomization parameters for both QI and the sensitive attribute.

5.3.2 Effect of Data Sizes in Randomization

As discussed in Section 5.2.4, one advantage of randomization scheme is that the more data we have, the more accurate reconstruction we can achieve. To investigate

the impact of data size upon the data utility of the randomized data, we generate four more datasets with varied sizes by sampling $r*N$ tuples from the Adult Data randomly where N is the size of the original Adult dataset and we set $r \in [0.5, 1.5, 2, 2.5]$. All the four generated datasets have the exact same distribution as the original one. Figure 5.4 shows the accuracy of reconstructed data distribution when data size increases. We can see that data utility is further improved when more data is available. This is a promising avenue for future work because the accuracy (in terms of bias and variance of estimates) of mining results derived from the reconstructed data needs more attentions. It is worth pointing out that the data size does not affect the accuracy of the estimated distribution of original data for generalization approaches (l-diversity and anatomy).

5.3.3 Comparison with Other Models

To compare randomization scheme with l-diversity and anatomy, we chose *Education, Salary, Gender, Race* as QI and *Occupation* as the sensitive attribute, similar to the settings of empirical evaluations in [49, 76]. We call this dataset as ES-GRO. Because the overall distribution of sensitive attribute values is unchanged after anonymization in both l-diversity and anatomy, we use RR-QI to compare with these two group based models.

Generalization approaches usually measure the utility syntactically by the number of generalization steps applied to quasi-identifiers [57], average size of quasi-identifier equivalence classes [49], sum of squares of class sizes, or preservation of marginals [42]. Since the randomization scheme is not based on generalization or suppression, in addition to the previous distribution distance measures, we further examine the utility preservation from the perspective of answering aggregate queries. We adopt *query answering accuracy*, the same metric as the one used in [84]. We also consider the variation of correlations between the sensitive attribute and quasi-identifiers. In our experiments, we used an implementation of Incognito algorithm [42] to generate the

entropy l-diverse tables and used the anatomy algorithm in [76].

Distribution Distance

We compare data utility of reconstructed probability distributions for different models according to the four distance measures. Figure 5.5 shows distances between $\hat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$ for anatomy, l-diversity and *RR-QI* on the data set ESGRO. We can observe that randomization outperforms both anatomy and l-diversity methods. This is because that we can partially recover the original data distribution in the randomization scheme, whereas data distribution within each generalized equivalence class is lost in l-diversity generalization and the relations between the *quasi-identifier table* (QIT) and the *sensitive table* (ST) are also lost in anatomy.

Another observation is that data utility (in term of distance between original and reconstructed distributions) monotonically decreases with the increment of the privacy thresholds (l) for randomization and anatomy. This is naturally expected because more randomness needs to be introduced with the increment of the privacy requirements for randomization and larger l in anatomy means that more tuples are included in each group, which decreases the accuracy of estimate for the distribution of original data. However, there is no similar monotonic trend in l-diversity. This is because the generalization algorithm choose different attributes for generalization with various l values, which makes the accuracy of the estimated distribution vary to different extents.

Query Answering Accuracy

The accuracy of answering aggregate queries is one of the important aspects to evaluate the utility of distorted data. We compare randomization against other anonymization approaches using the average relative error in the returned values. For each value, its relative error equals $|act - est|/act$, where act is the actual result

from the original data, and *est* the estimate derived from the underlying distortion approach. We consider two types of queries in our evaluation.

- Base queries with the form:

  SELECT A_1,..., A_m, S, COUNT(*) FROM data
  WHERE $A_1 = i_1 \ \ldots \ A_m = i_m$ AND $S = i_s$

  Where $i_k \in \Omega_k$ and $i_s \in \Omega_s$. The average relative errors on all the base queries (with size $D = d_s \prod_{i=1}^{m} d_i$) for $l = 1, \cdots, 7$ are shown in Figure 5.6(a).

- Cube query with the form:

  SELECT A_1,$\cdots$, A_m, S, COUNT(*) FROM data
  GROUP BY **CUBE** $(A_1, \cdots, A_m, S)$

  We use the above CUBE query to describe all the possible hierarchical aggregate queries. In multidimensional jargon, a cube is a cross-tabulated summary of detail rows. CUBE enables a SELECT statement to calculate subtotals for all possible combinations of a group of dimensions. It also calculates a grand total. The CUBE query here returns aggregate values of all possible combinations of QI attributes. We calculate the average relative error for each $l = 1, \cdots, 7$ and show the results in Figure 5.6(b).

Figure 5.6 shows relative errors of queries for anatomy, l-diversity and *RR-QI* on data set ESGRO. We can see that randomization permits significantly more accurate aggregate analysis than both l-diversity and anatomy since it can recover more accurate data distribution.

Correlation Between Attributes

A good publishing method should also preserve data correlation (especially between QI and sensitive attributes). We use Uncertainty Coefficient (U) to evaluate the

correlation between two multi-category variables. The expression of U is shown in (5.19).

$$U = -\frac{\sum_i \sum_j \pi_{ij} \log \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}}{\sum_j \pi_{+j} \log \pi_{+j}}. \tag{5.19}$$

The uncertainty coefficient takes values between -1 and 1 and larger values represent a strong association between variables. When the response variable has several possible categorizations, these measures tend to take smaller values as the number of categories increases.

Table 5.3 shows correlation (uncertainty coefficient) values between every pair of attributes vary under three models (anatomy, l-diversity and *RR-QI*) on the data set ESGRO. We vary l from 2 to 7. Due to space limit, we only include correlation values for $l = 3, 4, 5$ in Table 5.3. To study how correlations between QI and S changes, we use the attribute pair of Salary (S, one QI attribute) and Occupation (O, the sensitive attribute) as an example (the column with bold fonts in Table 5.3). The original uncertainty coefficient is 2.74×10^{-2}. *RR-QI* well achieves correlation preservation, i.e., 2.41, 2.27, and 2.17 ($\times 10^{-2}$) for $l = 3, 4, 5$ respectively. On the contrary, the uncertainty coefficient value under anatomy is only 0.20, 0.12, and 0.05 ($\times 10^{-2}$) correspondingly. For l-diversity, it achieves zero correlation preservation when $l = 3, 5$ while it perfectly achieves correlation preservation when $l = 4$. This is because the Salary attribute is generalized to "All" when $l = 3, 5$ while it is unchanged across all QI-groups when $l = 4$. Because it is intractable to predict which QI attributes will be generalized in l-diversity, l-diversity in general cannot well preserve correlation.

To have a clear understanding of correlation preservation for anatomy, l-diversity, and *RR-QI*, we show average values of uncertainty coefficient among attributes on the data set ESGRO in Figure 5.7. Specifically, Figure 5.7(a) (Figure 5.7(b)) corresponds to the case of QI vs S (QI vs QI) while Figure 5.7(c) corresponds to the

case of all attribute pairs. In summary, randomization achieves the best correlation preservation between the sensitive attribute and quasi-identifiers across all privacy thresholds. It is also clearly shown in Figure 5.7(b) that randomization better preserves correlation among quasi-identifiers than l-diversity. Please note that anatomy can best achieve the correlation among quasi-identifiers since it does not change values of quasi-identifiers in its released QIT table.

5.3.4 Summary of Evaluation

In summary, the evaluation shows that randomization can better preserve utility (in terms of distribution reconstruction, accuracy of aggregate query answering, and correlation among attributes) under the same privacy requirements. Utility loss is significantly smaller than that of generalization or anatomy approaches. Furthermore, the effectiveness of randomization can be further improved when more data is available. The evaluation also showed that randomization approach can further improve the accuracy of reconstructed distribution (and hence utility preservation) when more data is available while generalization and anatomy approaches do not have this property.

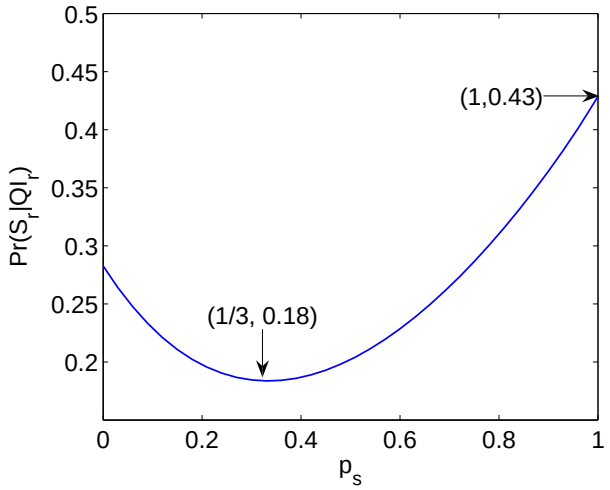
In the sequel, we compared randomization against l-diversity and anatomy in terms of computation overhead. The execution time of randomization was usually 10 times slower than l-diversity and anatomy. While we continue our study of reducing computational cost of randomization, it is our belief that effectiveness is more important than efficiency since randomization only incurs one-time cost.

5.4 Summary

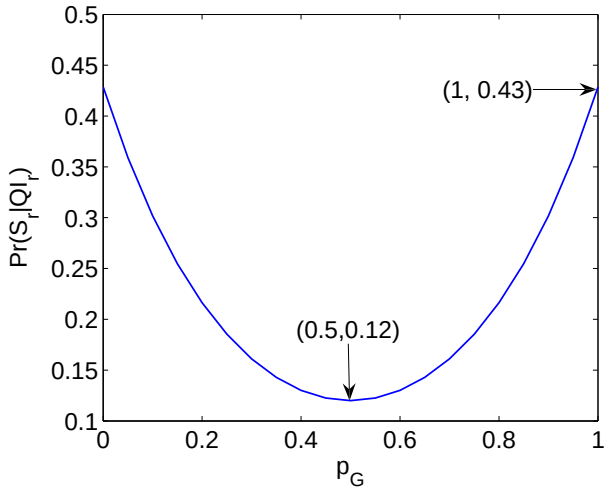
In this chapter, we presented a systematic study of randomization method in preventing attribute disclosure under linking attacks. We proposed a general framework and presented a uniform definition for attribute disclosure which is compatible for both randomization and generalization models. We proposed the use of a specific

randomization model. We presented an efficient solution to derive distortion parameters to satisfy requirements for privacy preservation while maximizing data utilities. We compared randomization model with other distortion models, k-anonymity, l-diversity and Anonymity. Our experimental evaluations showed that randomization significantly outperforms generalization, i.e., achieving better utility preservation while yielding the same privacy protection.

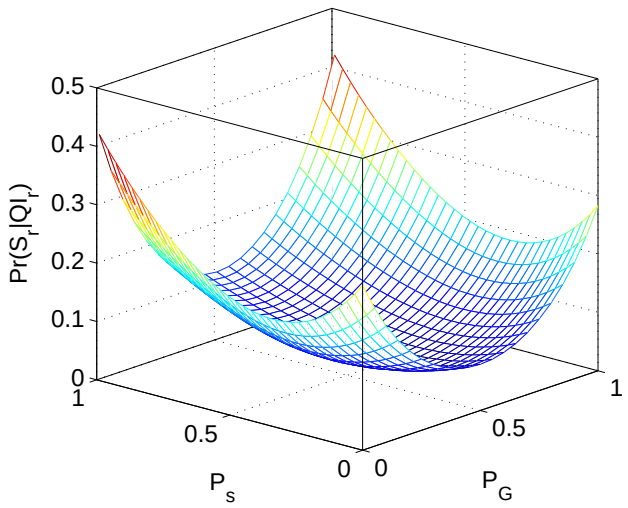
There are several avenues for future work. We aim to extend our research to handle multiple sensitive attributes, which usually happens in practice. In this dissertation, we limit our scope as attackers have no knowledge about the sensitive attribute of specific individuals in the population and/or the table. In practice, this may not be true since the attacker may have *instance-level background knowledge* (e.g., the attacker might know that the target individual does not have cancer; or the attacker might know complete information about some people in the table other than the target individual.) or *partial demographic background knowledge* about the distribution of sensitive and nonsensitive attributes in the population. Different forms of background knowledge have been studied in privacy preserving data publishing recently [12,19,50]. They proposed different approaches (a formal language [12,50] or ME constraints [19]) to express background knowledge of attackers and analyze the privacy risk in data publishing. We will investigate privacy disclosure under those background knowledge attacks [19,50]. We will continue our study of further reducing computation overhead of randomization approach and derive new algorithms to improve the efficiency of the process.



(a) Randomize S only

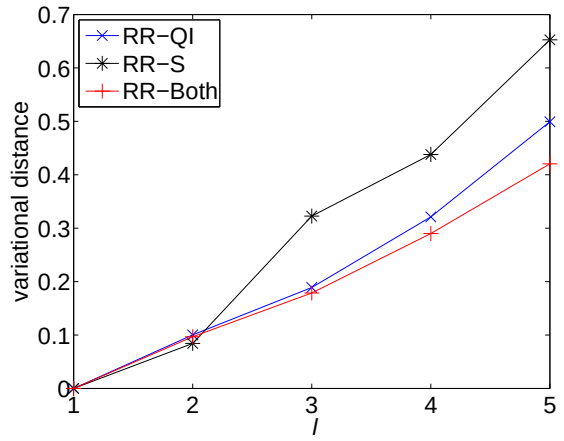


(b) Randomize QI only

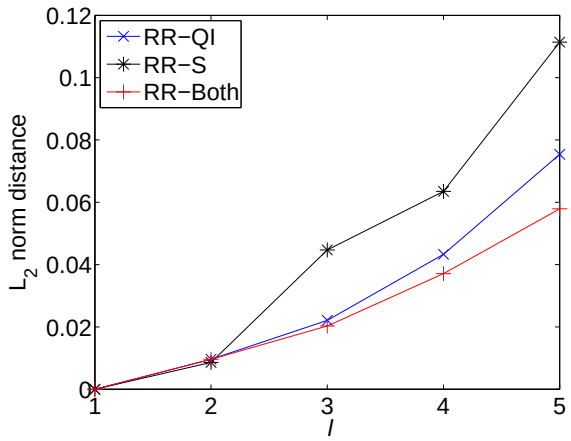


(c) Randomize QI and S

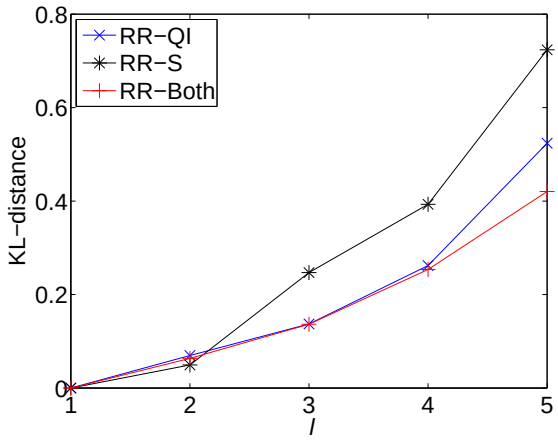
Figure 5.2: $\Pr(S_r|QI_r)$ vs. randomization parameters



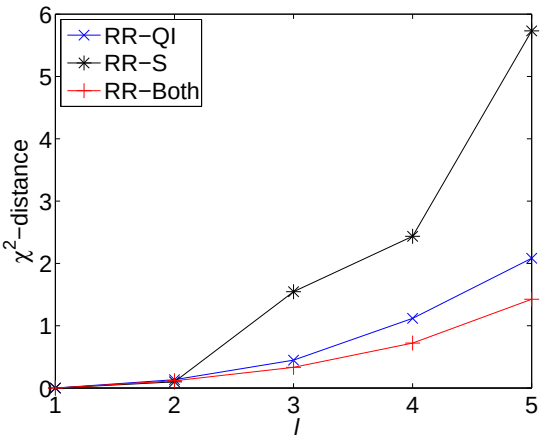
(a) variational distance



(b) L_2 norm distance

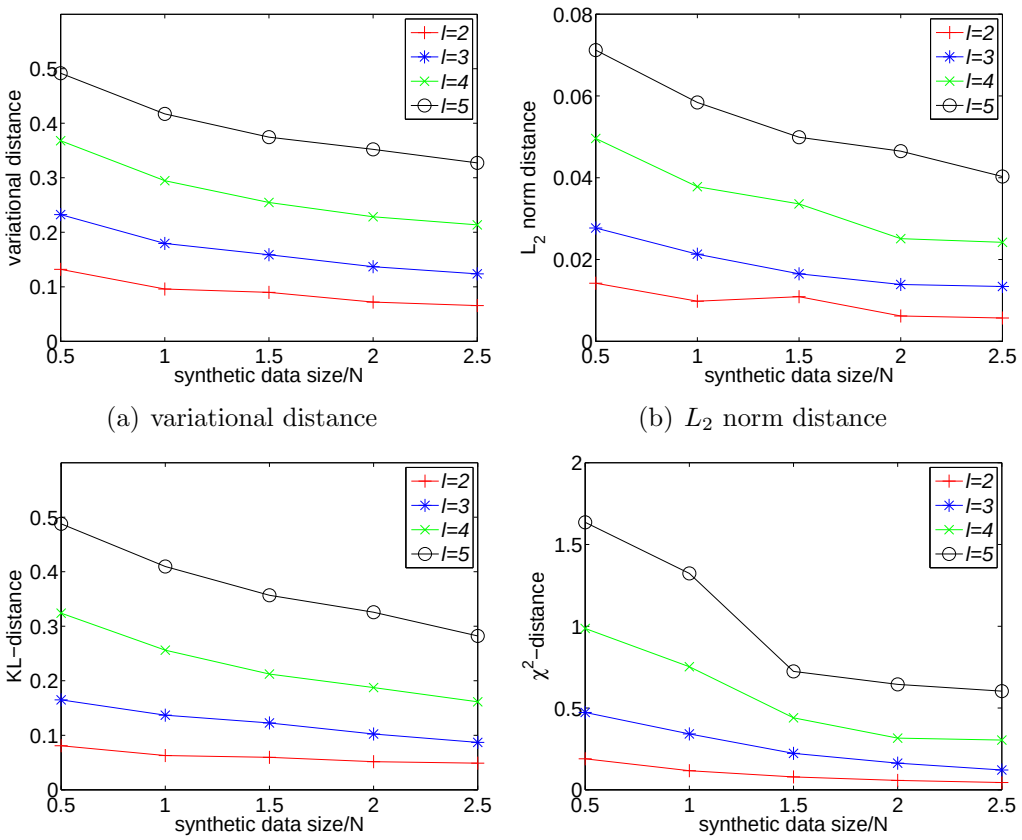


(c) KL-distance



(d) χ^2-distance

Figure 5.3: Distances between $\hat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$ for three scenarios of *RR* (data set EMGRW)



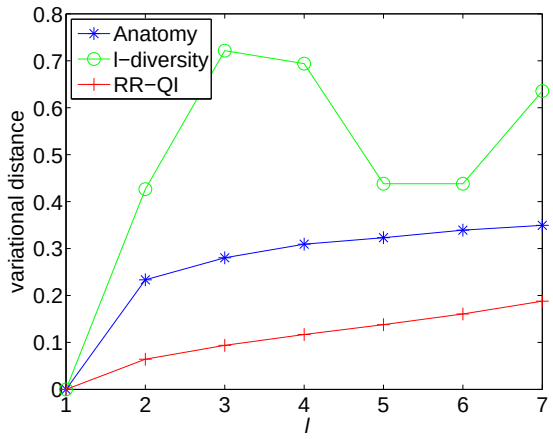
(a) variational distance

(b) L_2 norm distance

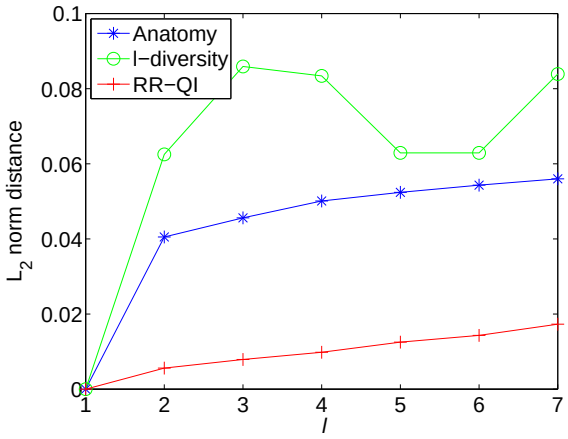
(c) KL-distance

(d) χ^2-distance

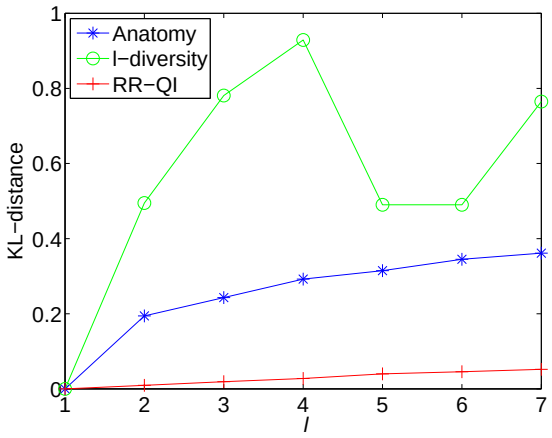
Figure 5.4: For *RR-Both*, distances between $\hat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$ decreases as the data set size increases (data set EMGRW)



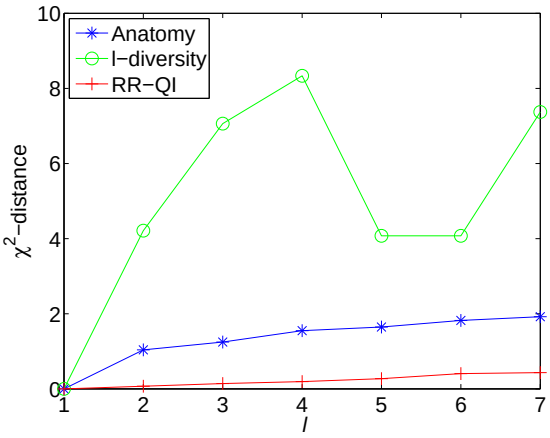
(a) variational distance



(b) L_2 norm distance

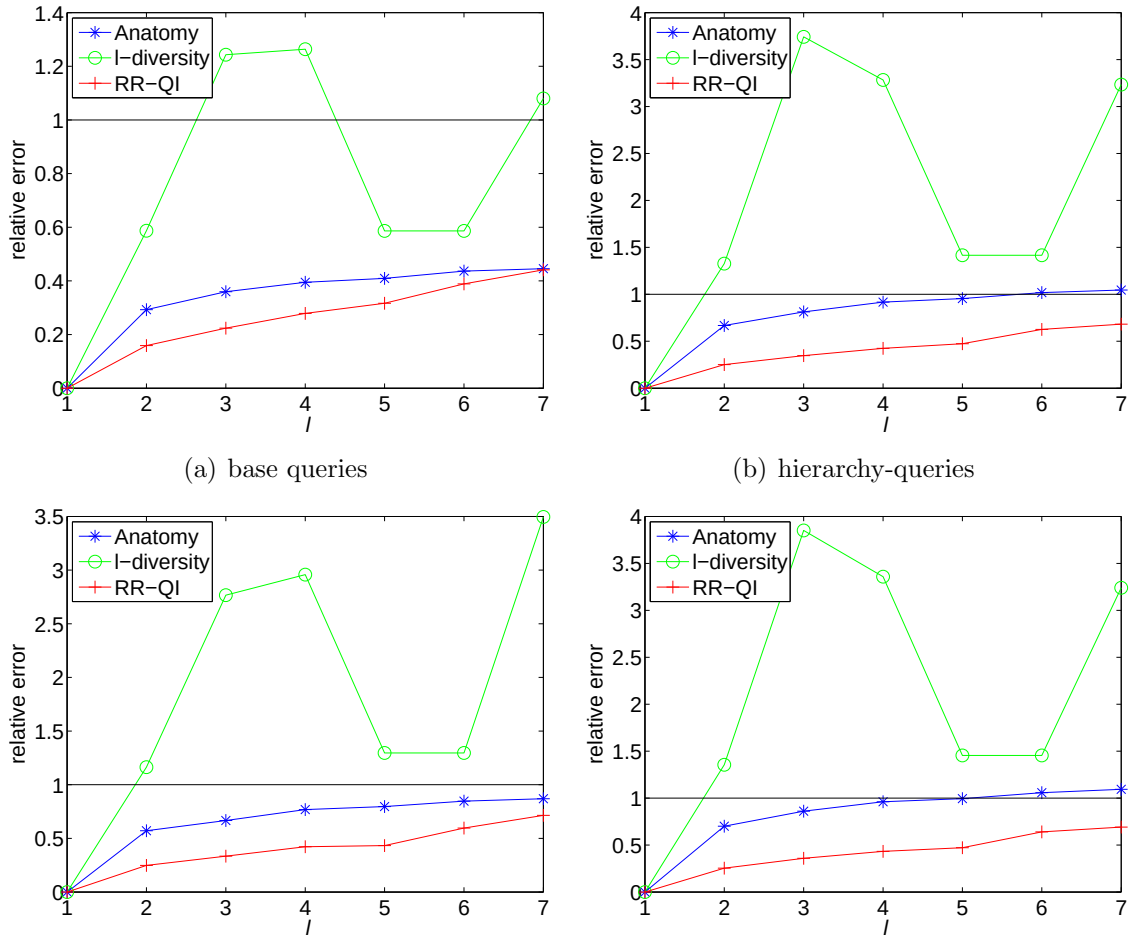


(c) KL-distance



(d) χ^2-distance

Figure 5.5: Distances between $\hat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$ for anatomy, l-diversity and *RR-QI* (data set ESGRO)



(a) base queries

(b) hierarchy-queries

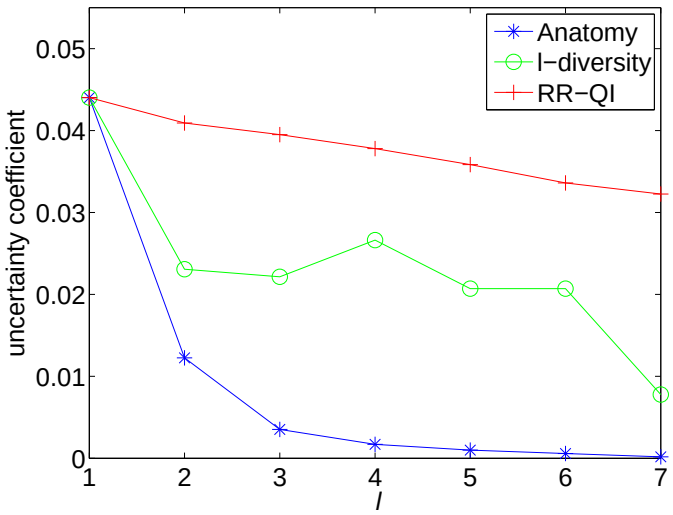
(c) 1000 random queries

(d) 10000 random queries

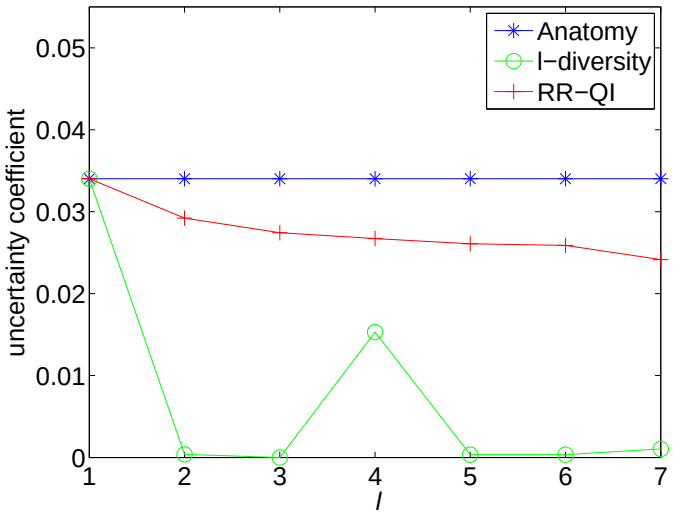
Figure 5.6: Relative errors of queries for anatomy, l-diversity and *RR-QI* (data set ESGRO)

Table 5.3: Variation of correlation (uncertainty coefficient) between attributes under different models ($\times 10^{-2}$) (data set ESGRO)

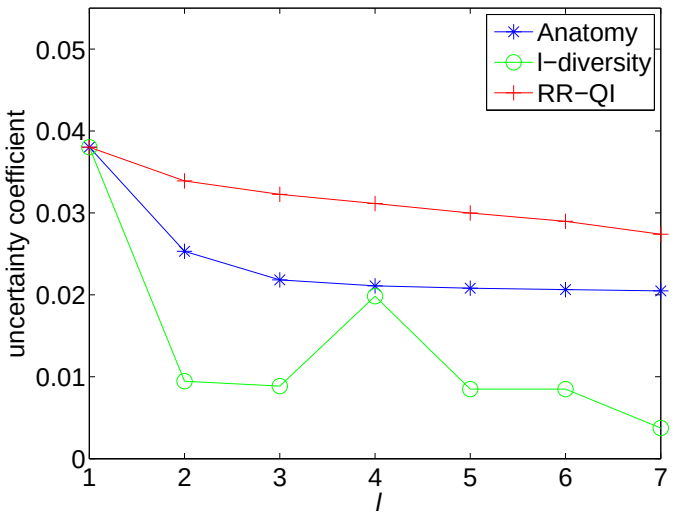
correlation		QI vs QI						QI vs S			
		E vs S	E vs G	E vs R	S vs G	S vs R	G vs R	E vs O	**S vs O**	G vs O	R vs O
	Original	11.54	0.68	1.92	4.12	1.07	1.08	9.90	**2.74**	4.40	0.57
l=3	Anatomy	11.54	0.68	1.92	4.12	1.07	1.08	0.90	**0.20**	0.25	0.06
	l-diversity	0	0	0	0	0	0	8.86	**0**	0	0
	RR-QI	8.57	0.63	1.86	3.65	0.53	1.08	8.72	**2.41**	4.13	0.54
l=4	Anatomy	11.54	0.68	1.92	4.12	1.07	1.08	0.46	**0.12**	0.08	0.02
	l-diversity	9.18	0	0	0	0	0	7.91	**2.74**	0	0
	RR-QI	8.05	0.58	1.29	3.37	0.45	1.07	8.33	**2.27**	4.01	0.51
l=5	Anatomy	11.54	0.68	1.92	4.12	1.07	1.08	0.26	**0.05**	0.07	0.02
	l-diversity	0	0	0.21	0	0	0	7.91	**0**	0	0.37
	RR-QI	7.50	0.33	1.02	3.15	0.44	1.04	7.77	**2.17**	3.89	0.50



(a) QI vs S



(b) QI vs QI



(c) All

Figure 5.7: Average value of uncertainty coefficients among attributes for anatomy, l-diversity and RR-QI (data set ESGRO)

CHAPTER 6: SUMMARY AND FUTURE WORK

6.1 Summary

The problem of privacy in data mining has become more important in recent years. A number of techniques such as randomization and group based generalization have been suggested to perform privacy preserving data mining. Privacy preserving data mining aims at providing a trade-off between sharing information for data mining analysis and protecting confidential information to preserve privacy.

This dissertation presented a formal and comprehensive examination of data utility and privacy of randomization models in privacy preserving categorical data analysis. The main contributions can be summarized as follows:

I: Accuracy analysis in randomization model for categorical data analysis.

- *Accuracy analysis for association rule mining in market basket data.* We investigated the accuracy (in terms of bias and variance of estimates) of both support and confidence estimates of association rules derived from the randomized data. We proposed the novel idea of using interquantile range to bound those estimates derived from the randomized market basket data. We demonstrated that providing confidence on data mining results from randomized data is significant to data miners. They can know how accurate their data mining results are under randomization-based models.

- *A general framework to evaluate the accuracy of estimates of various measures adopted in categorical data analysis .* We presented a general approach to derive variances of estimates of various measures adopted in categorical data analysis.

We applied the idea of using interquantile ranges based on Chebyshevs Theorem to bound those estimates derived from the randomized data.

II: Data utility analysis in randomization with unknown distortion parameters.

- *Data utility analysis in randomization with unknown distortion parameters.* We investigated whether data mining or statistical analysis tasks can still be conducted on randomized data when distortion parameters are not disclosed to data miners. We examined how various objective association measures between two variables may be affected by randomization. We demonstrated that some measures 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 can be executed on the randomized data directly even without knowing distortion parameters. We then investigated how the relative order of two association patterns is affected when the same randomization is conducted. We showed that some measures 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. We then extended it to multiple variables by examining the feasibility of hierarchical loglinear modeling. We showed that some classic data mining tasks (e.g., association rule mining, decision tree learning, naive bayes classifier) cannot be applied on the randomized data directly.

III: Privacy disclosure analysis in randomization models.

- *Analysis of attribute disclosure under linking attacks.* We presented a systematic study of randomization method in preventing attribute disclosure under linking attacks. We proposed a general framework and presented a uniform definition for attribute disclosure which is compatible for both randomization and generalization models.

- *Efficient solution for randomization parameters under linking attacks.* We proposed the use of a specific randomization model. We presented an efficient solution to derive distortion parameters to satisfy requirements for privacy preservation while maximizing data utilities. We compared randomization model with other distortion models, k-anonymity, l-diversity and Anonymity. Our experimental evaluations showed that randomization signicantly outperforms generalization, i.e., achieving better utility preservation while yielding the same privacy protection.

6.2 Future Work

Several directions can be exploited as a continuation of this research. We discuss some extensions we are going to make and technical challenges we would like to address in this section.

I: Investigate randomization models in categorical data analysis under different application scenarios.

In privacy preserving data publishing, most of my previous work focused on the general scenarios that all quasi-identifying attributes are randomized to preserve privacy of published data. In practice, there are situations where some sensitive attributes are randomized while others are released directly, or, some records are randomized while the remaining is unchanged. In the future, we will investigate data utility and privacy under such scenarios and plan to develop randomization models with enhanced flexibility and data utility for these applications.

II: Investigate randomization model under specific background knowledge.

For the attribute disclosure analysis in Chapter 5, we assumed that the attacker has access to the published data set, and knows values of quasi-identify attributes of some individuals in the published data set. In practice, however, the attacker may have instance-level background knowledge or partial demographic background knowledge about the distribution of sensitive and non-sensitive attributes in the population.

Different types of background knowledge have been studied in privacy preserving data publishing recently. We will investigate privacy disclosure of randomization models under those background knowledge attacks in the future.

III: Extend the results to privacy preserving data mining of numerical data.

Randomization model in my current work is based on the Randomized Response (RR) method for categorical data. The RR method has been investigated by statistical researchers to estimate the distribution of sensitive quantitative data. However, it's not well studied in PPDM area. In the future, we will investigate the RR method in preserving privacy for numerical data and extend our theoretical results onto it.

IV: Extend the results to privacy and spectral analysis of social network randomization.

Social networks are of significant importance in various application domains. Most previous studies are focused on revealing interesting properties of networks and discovering efficient and effective analysis methods. However, there has been little work dedicated to privacy preserving social network analysis. Some work has been discussed in [78–83]. In [78, 79], we investigated the application of graph randomization techniques to protect privacy of individual nodes and their sensitive link relationships. We conducted theoretical study and empirical evaluation on the trade-off between utility and privacy of various graph randomization techniques.

V: Extend the empirical evaluation to real data sets on various data mining applications.

We will evaluate the accuracy and privacy analysis of randomization models on other data sets, especially real data sets. We plan to investigate other data mining application in our future work.

REFERENCES

[1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *31st International Conference on Very Large Data Bases*, 2005.

[2] C. C. Aggarwal and P. S. Yu. A condensation approach to privacy preserving data mining. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*, 2004.

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

[4] 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.

[5] R. Agrawal, R. Srikant, and D. Thomas. Privacy preserving olap. In *Proceedings of SIGMOD*, 2005.

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

[7] S. Agrawal, V. Krishnan, and J. Haritsa. On addressing efficiency concerns in privacy-preserving mining. *Proc. of 9th Intl. Conf. on Database Systems for Advanced Applications (DASFAA)*, pages 113–124, 2004.

[8] A. Agresti. *Categorical Data Analysis*. Wiley, 2002.

[9] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007.

[10] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *International Conference on Data Engineering*, 2005.

[11] A. Chaudhuri and R. Mukerjee. *Randomized Response Theory and Techniques*. Marcel Dekker, 1988.

[12] B.-C. Chen, K. LeFevre, and R. Ramakrishnan. Privacy skyline: Privacy with multidimensional adversarial knowledge. In *VLDB Endowment*, 2007.

[13] K. Chen and L. Liu. Privacy preserving data classification with rotation perturbation. In *Proceedings of the 5th IEEE International Conference on Data Mining*. Houston,TX, Nov 2005.

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

[15] V. Ciriani, S. D. C. di Vimercati, S. Foresti, and P.Samarati. k-Anonymity. *Security in Decentralized Data Management*, 2006.

[16] T. F. Coleman, J. Liu, and W. Yuan. A new trust-region algorithm for equality constrained optimization. *Comput. Optim. Appl.*, 21(2):177–199, 2002.

[17] V. den Hout A. and V. der Heijden P.G.M. Randomized response, statistical disclosure control and misclassification: A review. *International Statistical Review*, 70(2):269–288, 2002.

[18] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.

[19] W. Du, Z. Teng, and Z. Zhu. Privacy-MaxEnt: integrating background knowledge in privacy quantification. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 459–472, 2008.

[20] W. Du and Z. Zhan. Using randomized response techniques for privacy-preserving data mining. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 505–510, 2003.

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

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

[23] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–228, 2002.

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

[25] A. Gionis, H. Mannila, T. Mielikainen, and P. Tsaparas. Assessing data mining results via swap randomization. In *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining*, 2006.

[26] J. Gouweleeuw, P. Kooiman, L. Willenborg, and P. de Wolf. Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14(4):463–478, 1998.

[27] 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 2007.

[28] 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.

[29] L. Guo and X. Wu. Privacy preserving categorical data analysis with unknown distortion parameters. *Transaction on Data Privacy*, 2(3):185–205, 2009.

[30] L. Guo, X. Ying, and X. Wu. On attribute disclosure in randomization based privacy preserving data publishing. In *IEEE International Conference on Data Mining Workshops*, 2010.

[31] S. Guo and X. Wu. On the lower bound of reconstruction error for spectral filtering based privacy preserving data mining. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 520–527. Berlin, Germany, Sept. 2006.

[32] S. Guo and X. Wu. On the use of spectral filtering for privacy preserving data mining. In *Proceedings of the 21st ACM Symposium on Applied Computing*, pages 622–626, April 2006.

[33] S. Guo and X. Wu. Deriving private information from arbitraraily projected data. In *Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 84–95. Nanjing, China, May 2007.

[34] S. Guo, X. Wu, and Y. Li. Deriving private information from perturbed data using iqr based approach. In *2nd International Workshop on Privacy Data Management*, 2006.

[35] Z. Huang and W. Du. Optrr: Optimizing randomized response schemes for privacy-preserving data mining. In *ICDE*, pages 705–714, 2008.

[36] 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.

[37] 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.

[38] M. G. Kendall and A. Stuart. *The advanced theory of statistics*, volume 1. New York, Hafner Pub. Co., 1969.

[39] J. Kim. A method for limiting disclosure in microdata based on random noise and transformation. In *In Proc. of the Section on Survey Research Methods*, 1986.

[40] E. L. Korn. Hierarchical Log-Linear Models Not Preserved by Classification Error. *Journal of the American Statistical Association*, 76:110–113, 1981.

[41] D. Lambert. Measures of disclosure risk and harm. *J. Official Stat.*, 9:313, 1993.

[42] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the ACM SIGMOD*, 2005.

[43] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian Multidimensional K-Anonymity. In *ICDE*, 2006.

[44] F. Leysieffer and S. Warner. Respondent jeopardy and optimal designs in randomized response models. *Amer. Statist. Assoc*, 71, 1976.

[45] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the International Conference on Data Engineering*, 2007.

[46] C. K. Liew, U. J. Choi, and C.J.Liew. A data distortion by probability distribution. In *ACM TODS, 10(3):395-411*, 1985.

[47] K. Liu, C. Giannella, and H. Kargupta. An attacker's view of distance preserving maps for privacy preserving data mining. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Berlin,Germany, September 2006.

[48] K. Liu, H. Kargupta, and J. Ryan. Random projection based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transaction on Knowledge and Data Engineering*, 18(1):92–106, 2006.

[49] A. Machanavajjhala, J. Gehrke, D. Keifer, and M. Venkitasubramanian. l-diversity: privacy beyond k-anonymity. In *Proceedings of the IEEE ICDE Conference*, 2006.

[50] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern. Worst-Case Background Knowledge for Privacy-Preserving Data Publishing. In *ICDE*, 2007.

[51] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *23 ACM SIGACT-SIGMOD-SIGART Symposium on principles of Database Systems*, 2004.

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

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

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

[55] D. B. Rubin. Discussion of statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468, 1993.

[56] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth movers distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2), 2000.

[57] P. Samarati. Protecting respondents identities in microdata release. In *IEEE-Trans. Knowl. Data Eng.*, 2001.

[58] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymty and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*, 1998.

[59] S.E.Fienberg. A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical report, Carnegie Mellon University Department of Statistics, 1994.

[60] S.E.Fienberg, U.E.Makov, and R.J.Steele. Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14(4):485–502, 1998.

[61] 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.

[62] C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. In *Proceedings of the 24th VLDB Conference*. New York, 1998.

[63] M. D. Springer. *The Algebra of Random Variables*. John Wiley and Sons, New York, 1979.

[64] G. Strang. *Introduction to Linear Algebra, 3rd edition*. Wellesley-Cambridge Press: Wellesley, MA, 2003.

[65] G. Sullivan. *The use of added error to avoid disclosure in microdata releases*. PhD thesis, Iowa State University, 1989.

[66] L. Sweeney. Uniqueness of simple demographics in the u.s. population. Technical report, Carnegie Mellon University, 2000.

[67] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.

[68] P. Tan, V. Kumar, and S. J. 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.

[69] P. Tan, M. Steinbach, and K. Kumar. *Introuduction to Data Mining*. Addison Wesley, 2006.

[70] Z. Teng and W. Du. Comparisons of k-anonymization and randomization schemes under linking attacks. In *Proceedings of the 6th International Conference on Data Mining*, pages 1091–1096, 2006.

[71] Y. Wang, X. Wu, and Y. Zheng. Privacy preserving data generation for database application performance testing. In *Proceedings of 1st International Conference on Trust and Privacy in Digital Business (TrustBus04)*, pages 142–151, 2004.

[72] S. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *J.Amer.Statist.Assoc.*, 60:63–69, 1965.

[73] X. Wu, C. Sanghvi, Y. Wang, and Y. Zheng. Privacy aware data generation for testing database applications. In *Proceedings of the 9th International Database Engineering and Application Symposium*, pages 317–326, 2005.

[74] X. Wu, Y. Wang, and Y. Zheng. Privacy preserving database application testing. In *Proceedings of the ACM Workshop on Privacy in Electronic Society*, pages 118–128, 2003.

[75] X. Wu, Y. Wu, Y. Wang, and Y. Li. Privacy-aware market basket data set generation: A feasible approach for inverse frequent set mining. In *Proceedings of the 5th SIAM International Conference on Data Mining*, 2005.

[76] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, pages 139–150. VLDB Endowment, September 2006.

[77] X. Xiao, Y. Tao, and M. Chen. Optimal random perturbation at multiple privacy levels. In *Proceedings of VLDB*, 2009.

[78] X. Ying, K. Pan, X. Wu, and L. Guo. Comparisons of randomization and k-degree anonymization schemes for privacy preserving social network publishing. In *SNA-KDD '09: Proceedings of the 3rd SIGKDD Workshop on Social Network Mining and Analysis (SNA-KDD)*, 2009.

[79] X. Ying, K. Pan, X. Wu, and L. Guo. *On the Quantification of Identity and Link Disclosures in Randomizing Social Networks*. Advances in Information & Intelligent Systems, 2009.

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

[81] X. Ying and X. Wu. Graph generation with prescribed feature constraints. In *Proc. of the 9th SIAM Conference on Data Mining*, 2009.

[82] X. Ying and X. Wu. On link privacy in randomizing social networks. In *PAKDD*, 2009.

[83] X. Ying and X. Wu. On randomness measures for social networks. In *SDM*, 2009.

[84] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. In *Proc. 23rd ICDE*, 2007.

[85] Y. Zhu and L. Liu. Optimal randomization for privacy preserving data mining. In *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2008.