JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 5, No. 3, September 2011, pp. 169-182

Limiting Attribute Disclosure in Randomization Based Microdata Release

Ling Guo, Xiaowei Ying, and Xintao Wu*
University of North Carolina at Charlotte, NC 28223, USA
lguo2@uncc.edu, xying@uncc.edu, xwu@uncc.edu

Abstract

Privacy preserving microdata publication has received wide attention. In this paper, we investigate the randomization approach and focus on attribute disclosure under linking attacks. We give 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.

Categories: Smart and intelligent computing

Keywords: Algorithms; Privacy preservation; Randomization; Attribute disclosure; Linking atack

I. INTRODUCTION

Privacy preserving microdata publication has received wide attention [1-5]. Each record corresponding to one individual has a number of attributes, which can be divided into the following three categories: 1) *identity* attributes (e.g., Name and SSN) whose values can uniquely identify an individual; 2) *quasi-identifier (QI)* attributes (e.g., demographic attributes such as ZIP code, age, and gender) whose values when taken together can potentially identify an individual; 3) *sensitive* attributes (e.g., disease and income) that indicate confidential information of individuals.

Before microdata are released, identity attributes are often directly removed to preserve privacy of individuals whose data are in the released table. However, the *QI* information may be used by attackers to link to other public data sets to get the private information of individuals. This is recognized as *linking attacks* in microdata publishing. Two types of information disclosures have been identified under linking attacks: *identity disclosure* and *attribute disclosure* [6]. 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. Samarati and Sweeney [1] proposed the *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 to counter linking attacks based on quasi-identifiers. It was identified that *k*-anonymity is vulnerable to homogeneity and background knowledge attacks when data in the *QI*-group lacks diversity in the sensitive attributes [2]. *l*-diversity [2], as well as following models (e.g., *t*-closeness [7]), were proposed to protect attribute disclosure. *l*-diversity requires that the sensitive attribute has at least one *well-represented* values for each *QI*-group in the generalized dataset.

The randomization approach has also been adopted to publish microdata [8-12]. Instead of generalizing *QI* attribute values, randomization approach distorts the original value to another domain value according to some distortion probabilities. The application of the randomization technique was stud-

Open Access http://dx.doi.org/10.5626/JCSE.2011.5.3.169 http://jcse.kiise.org

Received 01 February 2011, **Accepted** 20 March 2011

*Corresponding Author

pISSN: 1976-4677 eISSN: 2093-8020

ied to prevent identity disclosure under linking attacks in data publishing [12]. 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. Thus, we can maximize data utility, while satisfying privacy preservation requirements for sensitive attributes. We compare our randomization approach with other anonymization approaches within the framework (e.g., two representative approaches *l*-diversity [2] and anatomy [5] are used in this paper). Our result shows the randomization approach can better recover the distribution of the 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.

Our contributions are summarized as follows. We present a systematic study of the randomization method in preventing attribute disclosure under linking attacks. We propose the use of a specific randomization model and present an efficient solution to derive distortion parameters to satisfy requirements for privacy preservation, while maximizing data utilities. We propose a general framework and present a uniform definition for attribute disclosure which is compatible for both randomization and generalization models. 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.

The remainder of this paper is organized as follows. In Section II, we discuss closely related work on group based anonymization approaches and randomization approaches in privacy preservation data publishing. In Section III, we present background on randomization based distortions, including analysis and attacks on the randomized data. In Section IV, we first quantify attribute disclosure under linking attacks and then show our theoretical results on maximizing utility with privacy constraints. In Section V, we conduct empirical evaluations and compare the randomization based distortion with two representative group based anonymization approaches (*l*-diversity [2] and anatomy [5]). We conclude our work in Section VI.

II. RELATED WORK

A. Group Based Anonymization

k-anonymity was proposed [1] to counter linking attacks for data publishing. The method generalizes the values of quasiidentifier attributes to less-specific ones, so that each individual cannot be distinguished from at least $k - 1$ other individuals based on quasi-identifier information. There has been much study in designing efficient algorithms for *k*-anonymity using generalization and suppression techniques [2, 3, 5, 13, 14].

k-anonymity provides protection against identity disclosure, while it contributes little to attribute disclosure. *l*-diversity [2] is proposed to counter attribute disclosure. A table is *l*-diverse if, in each *QI*-group, at most $1/l$ of the tuples possesses the most frequent sensitive value. Thus, an individual can be linked to his sensitive value correctly with probability at most $1/l$. Similarly, the notation of *t*-closeness [7] requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold *t*). The anatomy method was proposed to protect attribute disclosure by breaking the link between quasi-identifiers and sensitive attributes [5]. Anatomy releases all the quasi-identifier and sensitive values directly in two separate tables: a *quasi-identifier table* (QIT) and a *sensitive table* (ST). They presented an algorithm to compute anatomized tables that satisfy the *l*-diversity requirement, while minimizing the error of reconstructing the data.

Kifer and Gehrke [15] investigated the problem of injecting additional information into *k*-anonymous and *l*-diverse tables to improve the utility of published data. Koudas et al. [16] studied the problem of preserving privacy through sensitive attribute permutation and generalization with the addition of fake sensitive attribute values. Narayanan and Shmatikov [17] presented a de-anonymization methodology for sparse multi-dimensional microdata, such as individual transactions, preferences, and so on. The de-anonymization algorithm could be used by attackers against data sets containing anonymous high-dimensional data. Brickell and Shmatikov [18] compared the privacy loss (defined by certain kinds of information learned by attackers) with the utility gain (defined as the same kinds of information learned by analysts) caused by data anonymization and concluded that “even modest privacy gains require almost complete destruction of the data mining utility” in generalization methods. Li and Li [19] proposed the use of the *worst-case* privacy loss rather than the *average* privacy loss among all individuals adopted in [18] to measure privacy loss. They further presented a methodology to evaluate the privacy-utility tradeoff, where privacy loss is quantified by the adversary’s knowledge gain about the sensitive values of specific individuals with the trivially-anonymized data as the baseline and utility loss is measured by the information loss about the sensitive values of large populations with the original data as the baseline. In our paper, we use $1/l$ to bound the attribute disclosure for all records and use the difference of distributions between the original data and the reconstructed data to quantify utility loss. We also examine the utility loss from the perspective of answering aggregate queries and preserving correlations between the sensitive attribute and *QI* attributes.

In this paper, we focus on the randomization model in resisting attribute disclosure of data publishing. The proposed randomization model is compared to traditional *l*-diversity [2] and anatomy [5] models. Recently, Ghinita et al. [14] proposed a framework to solve the privacy-constrained and accuracy constrained data anonymization problems under the *k*-anonymity and *l*-diversity models. They developed efficient heuristics to solve the one-dimensional problems in linear time and generalized solutions to multi-dimensional attributes using space-mapping techniques. Their empirical evaluation results showed that privacy/accuracy-constrained methods outperform existing generalization approaches in terms of the execution time and information loss. We would also point out that researchers have developed various anonymity models, including p-sensitive k-anonymity [20], (α, k)-anonymity [21], (k, e)-anonymity [4], (c, k)-safety [22], m-confidentiality [23], privacy skyline [24], δ-presence [25], (ϵ, m)-anonymity [26], and (t, λ)-uniqueness [27]. Detailed comparisons to these works are beyond the scope of this paper.

B. Randomization

The first randomization model, called Randomized Response (RR), was proposed by Warner [28] in 1965. The model deals with one dichotomous attribute, i.e., every person in the population belongs either to a sensitive group A, or to its complement $\overline{A}$. The problem is to estimate π_A, the unknown proportion of population members in group A. Each respondent is provided with a randomization device by which the respondent chooses one of the following two questions *Do you belong to A*? or *Do you belong to* $\overline{A}$? with respective probabilities p and $1 - p$ and then replies *yes* or *no* to the question chosen. The technique provides response confidentiality and increases respondents' willingness to answer sensitive questions, because no one but the respondent knows to which question the answer pertains. It has been extended to the field of privacy preserving data mining [29].

Aggarwal and Yu [30] provided a survey of randomization models for privacy preservation data mining, including various reconstruction methods for randomization, adversarial attacks, optimality and utility of randomization. Rizvi and Haritsa [9] developed the MASK technique to preserve privacy for frequent item set mining. Agrawal and Haritsa [31] presented a general framework of random perturbation in privacy preservation data mining. Du and Zhan [8] studied the use of a randomized response technique to build decision tree classifiers. Guo et al. [10] investigated data utility in terms of the accuracy of reconstructed measures in privacy preservation market basket data analysis. Aggarwal [32] proposed adding noise from a specified distribution to produce a probabilistic model of k-anonymity. Zhu and Liu [33] investigated the construction of optimal randomization schemes for privacy preservation density estimation and proposed a general framework for randomization using mixture models. Rebollo-Monedero et al. [34] defined the privacy measure similar to t-closeness [7] and formalized the privacy-utility tradeoff from the information theory perspective. The resulting solution turns out to be the postrandomization (PRAM) method [35] in the discrete case and a form of random noise addition in the general case. They also proved that the optimal perturbation is in general randomized, rather than deterministic in the continuous case.

Huang and Du [36] studied the search of optimal distortion parameters to balance privacy and utility. They developed an evolutionary multi-objective optimization method to find optimal distortion matrices when applying the RR technique on a single attribute. The complexity of the algorithm is $O(((N_Q + N_V)^3 + n^2)\ L)$ where N_Q is the population size, N_V is the archive size, n is the number of columns in the RR matrix, and L is the maximum number of iterations. However, there is no guarantee that the derived matrices are optimal in the entire search space and the performance tends to be sacrificed for multiple attributes due to the intractable size of the search space. Similarly, Xiao et al. [37] investigated the optimal random perturbation at multiple privacy levels.

In our paper, we focus on a specific randomization model (i.e., perturbation retains the original value of an attribute A_i with probability p_i and replaces the original value with a random value from the domain of A_i, see Equation 1) and derive an efficient solution to determine the optimal randomization parameters under linking attacks. Chaytor and Wang [38] also applied the use of this simple randomization model to randomize a sensitive value only within a subset of the entire domain. The developed algorithm, called *Small Domain Randomization*, can effectively derive optimal partitions, such that randomizing the sensitive attribute values of records (with small domains) in each partition separately can retain more data utility with the same level of privacy than randomizing the whole data set over the entire domain. Our work focuses on attribute disclosure when applying full domain randomization on both QI attributes and sensitive attribute S. The derived optimal randomization parameter values under privacy constraints can be incorporated in the small domain randomization.

III. 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 that 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 paper are also explained in the single sensitive attribute setting and can be generalized to multiple sensitive attributes.

Attribute A_i has d_i categories denoted by $0,1,\ldots, d_i - 1$. We use Ω_i 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. Similarly, attribute S has d_s categories denoted by $0,1,\ldots, d_s - 1$. We use $\Omega_s = \{0, 1, \ldots, d_s - 1\}$ to denote the domain of S. 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 $D = d_s \Pi_{i=1}^{m} d_i$ 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 π be the column vector with D elements $\pi_{i_1, \cdots, i_m, i_s}$ arranged in a fixed order. Table 1 shows one contingency table example for a data set with one QI attribute (Gender, $d_1 = 2$) and one sensitive attribute S (Disease, $d_s = 3$). Table 1a shows the original contingency table where $\pi = (\pi_{00}, \pi_{01}, \pi_{02}, \pi_{10}, \pi_{11}, \pi_{12})'$ corresponds to a fixed order of cell entries π_{ij} in the 2 × 3 contingency table. π_{10} denotes the proportion of records with *Female* and *Cancer*. The column sum π_{+0} represents the proportion of records with *Cancer* across both genders. Note that contingency tables are widely used in statistics to record and analyze the relationship between

Table 1. 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
M	8	16	48
F	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_{++}

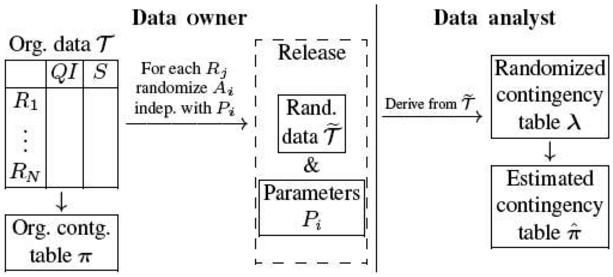

Data owner
Data analyst
Org. data $\mathcal{T}$
QI
S
R_1
R_N
For each R_j randomize A_i indep. with P_i
Release
Rand. data $\tilde{\mathcal{T}}$
&
Parameters P_i
Org. contg. table π
Derive from $\tilde{\mathcal{T}}$
Randomized contingency table λ
Estimated contingency table $\hat{\pi}$

Fig. 1. Randomization based privacy-preserving data publishing.

categorical attributes. Results of most data mining tasks (e.g., clustering, decision tree learning, and association rule mining), as well as aggregate queries, are solely determined by the contingency table. That is, analysis on contingency tables is equivalent to analysis on the original categorical data. Table 1b 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.

A. Distortion Procedure

We use the left part of Fig. 1 to illustrate the process of privacy preserving data publishing. For each record 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 category ($v = 0, \ldots, d_i - 1$) to the u-th category with probability $p_{uv}^{(i)}$: $\Pr(\tilde{A}_i = u|A_i = v) = p_{uv}^{(i)}$. Let $p_i = [p_{uv}^{(i)}]_{d_i \times d_i}$, 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 $\tilde{\mathcal{T}}$, and then both the randomized data set and the distortion matrices are published. The randomization matrices indicate the magnitude of the randomization; this can help data analysts estimate the original data distribution.

Let λ denote the contingency table of the randomized data $\tilde{\mathcal{T}}$. We arrange λ into the column vector with the same order of π. Table 1c shows one example of the randomized contingency table. The randomized contingency table has a close relationship to the original contingency table and the randomization matrices: $E(\lambda) = P\ \pi$ where $P = P_1 \otimes \cdots \otimes P_m \otimes P_s$, and $\otimes$ stands for the Kronecker product. The Kronecker product 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. How to find optimal distortion parameters with privacy or utility constraints has remained a challenging problem [39]. In this paper, we limit the choice of randomization parameters for each QI attribute A_i (and sensitive attribute S) as:

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

That is, 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 IV-B).

B. Analysis on the 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 right part of Fig. 1 shows how data analysts estimate the original data distribution. With the randomized data $\tilde{\mathcal{T}}$ and its contingency table λ, the unbiased estimate of π is given by $\hat{\pi} = P^{-1}\lambda$ and the covariance matrix of the estimator is

$$\Sigma = \mathrm{Cov}(\hat{\pi}) = \frac{1}{N}[P^{-1}(P\pi)^{\delta}(P^T)^{-1} - \pi\pi^T] \tag{2}$$

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

Thus, data analysts derive the estimation for the distribution of original data (in terms of contingency table) without disclosing the individual information of each record. We choose the accuracy of reconstructed distribution as the target utility function in Section IV-B, because most data mining applications are based on the probability distribution of the data.

C. Attacks on Randomized Data

Let X be an attribute or a subset of attributes in the data set $\mathcal{T}$ with domain Ω_X, and $\tilde{X}$ is the randomized value of X in $\tilde{\mathcal{T}}$. It is not reasonable for attackers to regard the observed value as the true value of X, with the randomization process and parameters. Instead, attackers can try to estimate the original value based on the observed data and the released randomization parameters. Let $\hat{X}$ denote the attackers' estimation on the original value of X. Any value in Ω_X is possible due to the randomization procedure. We assume that the attacker is able to calculate the posterior probability of content in the data set and takes the following probabilistic strategy: for any μ, $\nu \in \Omega_X$,

$$\hat{X} = \mu \text{ with prob. } \Pr(X = \mu|\tilde{X} = \nu), \tag{3}$$

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

$$\Pr(X = \mu | \tilde{X} = \nu) = \frac{\pi_\mu \Pr(\tilde{X} = \nu | X = \mu)}{\sum_{\omega \in \Omega_X} \pi_\omega \Pr(\tilde{X} = \nu | X = \omega)}. \quad (4)$$

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

LEMMA 1. *Suppose attackers adopt the probabilistic strategy specified in* (3) *to estimate the data. The probability that attackers accurately estimate the original value of X is given by:*

$$\Pr(\hat{X} = X = \mu) = \sum_{\nu \in \Omega_X} \Pr(\tilde{X} = \nu | X = \mu) \Pr(X = \mu | \tilde{X} = \nu). \quad (5)$$

PROOF. For a particular observed value $\tilde{X} = \nu \in \Omega_X$, $\Pr(\hat{X} = X = \mu, \tilde{X} = \nu) = \Pr(\tilde{X} = \nu | X = \mu) \Pr(X = \mu | \tilde{X} = \nu)$. Then, with the law of total probability, we have

$$\Pr(\hat{X} = X = \mu) = \sum_{\nu \in \Omega_X} \Pr(\hat{X} = X = \mu, \tilde{X} = \nu)$$
$$= \sum_{\nu \in \Omega_X} \Pr(\tilde{X} = \nu | X = \mu) \Pr(X = \mu | \tilde{X} = \nu).$$

The probability of attackers' correct estimation is also defined as the reconstruction probability, $\Pr(X = \mu \rightarrow X = \mu)$, in [9, 12]. It captures the round-trip of going from the original content to the distorted one and then returning to estimate the value of the original content. That is, it indicates how much information of the original data is preserved after randomization. To make the notation concise, we adopt the notation $\Pr(\hat{X} = X = \mu)$ to denote the reconstruction probability to evaluate the risk of the sensitive attribute disclosure.

IV. ATTRIBUTE DISCLOSURE UNDER LINKING ATTACKS

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

DEFINITION 1. *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* $\Pr(S_r | QI_r)$.

We need to quantify the background knowledge of attackers to derive the disclosure probability $\Pr(S_r | QI_r)$. We have the following standard assumptions for background knowledge of attackers in this paper. We assume that the attacker has access to the published data set $\tilde{\mathcal{T}}$ and he knows that $\tilde{\mathcal{T}}$ 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 of 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. The algorithm of *l*-diversity in [2] preserves privacy by generalizing the *QI* attributes to form *QI*-groups. Individuals in the group are linked to any sensitive attributes with probability at most $1/l$, i.e., $\Pr(S_r | QI_r) \leq 1/l$. However, the randomization based approach achieves the privacy protection probabilistically. In the following subsection, we show how to quantify the attribute disclosure risk in the randomization settings.

A. Quantifying Attribute Disclosure

When there is no randomization applied, for those records with their quasi-identifiers equal to QI_r, the attacker simply regards every record as having the same probability to represent individual *r*. The risk of sensitive attribute disclosure is equal to $\frac{\pi_{QI_r, S_r}}{\pi_{QI_r}}$, because there are 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. 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.

1) 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 (4). According to (5), the probability of a correct estimation is $\Pr(\hat{S}_r = S_r | QI_r)$, then the risk of sensitive attribute disclosure is $\frac{\pi_{QI_r, S_r}}{\pi_{QI_r}} \Pr(\hat{S}_r = S_r | QI_r)$.

2) 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 $\Pr(\widehat{QI}_r = QI_r)$, and hence the risk of sensitive attribute disclosure is $\frac{\pi_{QI_r, S_r}}{\pi_{QI_r}} \Pr(\widehat{QI}_r = QI_r)$.

3) Randomize QI and S (RR-Both): When data owners apply randomization to both *QI* and *S*, the attacker first needs to ensure the values of identifier attributes are correctly reconstructed. The probability is given by $\Pr(\widehat{QI}_r = QI_r)$. Second, the attacker needs to ensure the value of the sensitive attribute, given the correctly reconstructed identifier attribute values are correctly reconstructed. We summarize the risk of sensitive attribute disclosure in *RR-Both* , as well as *RR-S* and *RR-QI*, in the following result and give the general calculation of the attribute disclosure risk in the randomization settings.

RESULT 1. *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 successfully predicting the sensitive attribute value S_r of the target individual r given his quasi-identifier values QI_r is:*

$$\Pr(S_r | QI_r) = \frac{\pi_{QI_r, S_r}}{\pi_{QI_r}} \Pr(\widehat{QI}_r = QI_r) \Pr(\hat{S}_r = S_r | QI_r). \quad (6)$$

We give the formal expressions of the two reconstruction probabilities needed in calculating $\Pr(S_r | QI_r)$ in (6). The reconstruction probability of the quasi-identifier is given by:

$$\begin{aligned}
&\Pr(\widehat{QI}_r = QI_r = \alpha) \\
&= \sum_{\beta \in \Omega_{QI}} \Pr(\widetilde{QI}_r = \beta | QI_r = \alpha) \Pr(QI_r = \alpha | \widetilde{QI}_r = \beta) \\
&= \sum_{\beta \in \Omega_{QI}} \frac{\pi_\alpha [\Pr(\widetilde{QI}_r = \beta | QI_r = \alpha)]^2}{\Sigma_{\gamma \in \Omega_{QI}} \pi_\gamma \Pr(\widetilde{QI}_r = \beta | QI_r = \gamma)},
\end{aligned} \tag{7}$$

where $\Pr(\widetilde{QI}_r = \beta | QI_r = \alpha) = \Pi_{k=1}^{m} p_{j_k i_k}^{(k)}$ 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:

$$\begin{aligned}
&\Pr(\hat{S}_r = S_r = u | QI_r) \\
&= \sum_{v \in \Omega_s} \Pr(\tilde{S} = v | S = u, QI_r) \Pr(S = u | \tilde{S} = v, QI_r) \\
&= \sum_{v \in \Omega_s} p_{vu}^{(s)} \Pr(S = u | \tilde{S} = v, QI_r) \\
&= \sum_{v \in \Omega_s} \frac{[p_{vu}^{(s)}]^2 \pi_{QI_r, S=u}}{\Sigma_{t \in \Omega_s} p_{vt}^{(s)} \pi_{QI_r, S=t}}.
\end{aligned} \tag{8}$$

We are interested in when the attribute disclosure reaches the minimum. We show our results in the following property and include our proof in the 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 $\Pr(S_r|QI_r) = \frac{\pi_{QI_r, S_r}^2}{\pi_{QI_r}}$.

Example. We use the instance shown in Table 1b 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. Fig. 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. Fig. 2a shows the scenario when we only randomize S. We can see that min $\Pr(S_r|QI_r) = \left(\frac{\pi_{10}}{\pi_{1+}}\right)^2 = 0.18$ when $p_s = \frac{1}{d_s} = \frac{1}{3}$. Fig. 2b shows the scenario when we only randomize QI. We can see that min $\Pr(S_r|QI_r) = \pi_{10} = 0.12$ when $p_G = \frac{1}{d_G} = \frac{1}{2}$. Fig. 2c 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 $\Pr(S_r|QI_r) = \frac{\pi_{10}^2}{\pi_{1+}} = 0.05$.

Computational Cost. The main computation cost in (6) comes from calculating $\Pr(\widehat{QI}_r = QI_r)$. Let P_{QI} be the distortion matrix on quasi-identifiers: $P_{QI} = P_1 \otimes \cdots \otimes P_m$, π_{QI} be the contingency table on all quasi-identifiers arranged in a column vector, and λ_{QI} denote the expected QI contingency table of the randomized data: $\lambda_{QI} = P_{QI}\, \pi_{QI}$. Then, the denominator in (7) is exactly the cell of λ_{QI} corresponding to β. Let η denote the column vector of the reconstruction probabilities of quasi-identifiers, arranged in the same order of π_{QI}. We can further express (7) in matrix form:

$$\begin{aligned}
\eta &= \pi_{QI} \dot{\times} [(P_{QI}^{\dot{2}})^T (\lambda_{QI}^{\dot{-1}})] \\
&= \pi_{QI} \dot{\times} \{[\otimes_{i=1}^{m} (P_i^{\dot{2}})^T][(\otimes_{i=1}^{m} P_i)\pi_{QI}]^{\dot{-1}}\}
\end{aligned} \tag{9}$$

where $\dot{\times}$ denotes the component-wise multiplication, $P_i^{\dot{2}}$ is the component-wise square of P_i, and $\lambda_{QI}^{\dot{-1}}$ is the component-wise inverse of λ_{QI}.

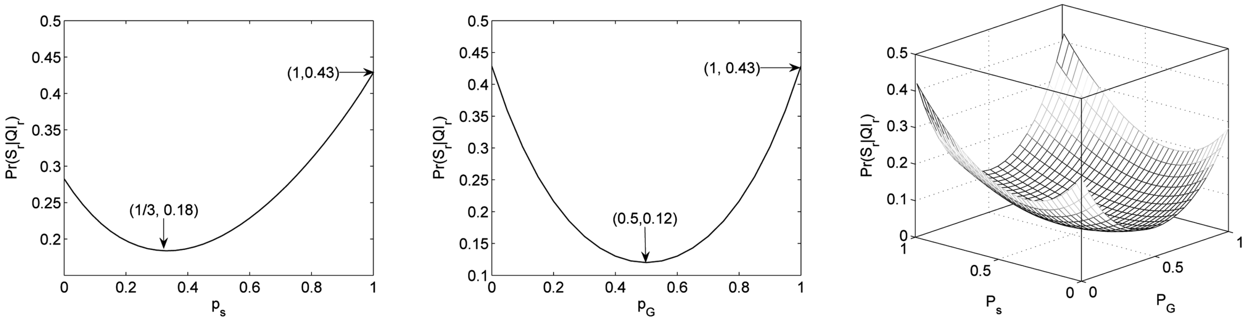
In (9), we need to repeatedly calculate $(P_1 \otimes \cdots \otimes P_m)\boldsymbol{x}$, where $\boldsymbol{x}$ denotes a column vector. Assume we use the naive algorithm in all matrix multiplications. Calculating $(P_1 \otimes \cdots \otimes P_m)\boldsymbol{x}$ directly results in the time and storage complexity of $O(\Pi_i d_i^2)$. The main storage complexity is from storing matrix P_{QI}. The following lemma allows us to reduce the cost of such computation:

LEMMA 2. *Let A, B and X be the matrices of size n × n, m × m, and m × n. Then*

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

where vec(X) denotes the vectorization of the matrix X formed by stacking the columns of X into a single column vector.

Applying Lemma 2 recursively, we can reduce the time complexity to $O([\Sigma_i d_i]\ \Pi_i d_i)$. The storage complexity is also reduced to $O(\Pi_i d_i + \Sigma_i d_i^2)$; this is mainly used to store the



(a) Randomize S only (b) Randomize QI only (c) Randomize QI and S

Fig. 2. Pr(S_r|QI_r) vs. randomization parameters. QI: quasiidentifier.

contingency table.

B. Maximizing Utility with Privacy Constraints

The ultimate goal of publishing data is to maximize utility, while the minimizing risk of attribute disclosure simultaneously. The utility of any data set, 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 data set is useful or not. This motivates us to focus on the data distribution when evaluating the utility of a database, since many data mining applications are based on the probability distribution of the data.

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

$$\text{minimize } E[d(\hat{\pi}, \pi)] \text{ s.t. } \max_r \Pr(S_r|QI_r) \leq \tau, p_i \in (1/d_i, 1], \qquad (10)$$

where $E[d(.)]$ *denotes the expectation of* $d(.)$ *and* $d(\hat{\pi}, \pi)$ *denotes a certain distance between* $\hat{\pi}$ *and* π.

We can set the privacy threshold formalize as the same privacy requirement of l-diversity (i.e., $\tau = 1/l$) to compare the performance of different disguised schemes. That is, 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 data set 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 $\Pr(S_r|QI_r)$.

Problem 1 is a nonlinear optimization problem. In general, we can solve it using optimization packages (e.g., trust region algorithm [40]). In Section IV-A, we 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 $E[\|\hat{\pi}-\pi\|_2^2]$, the expected Euclidean distance difference between the original data and the estimated one.

RESULT 2. *When* $d(\hat{\pi}, \pi)$ *is the squared Euclidean distance, Problem 1 is equivalent to: determine* p_i, $i = 1, \cdots, m$, *and* p_s *to*

$$\text{minimize } \Pi_I \|P_i^{-1}\|_F^2 \text{ s.t. } \max_r \Pr(S_r|QI_r) \leq \tau, p_i \in (1/d_i, 1].(11)$$

We briefly explain how we can derive this result and give its detailed proof in the Appendix. When the distance is the squared Euclidean distance, with Lemma (3) shown in the Appendix, to minimize d($\hat{\pi}$, π) is equivalent to minimize trace(Σ) where Σ is the covariance matrix of the estimator of cell values in the contingency table (shown in Equation 2). Calculating trace(Σ) still involves high computational cost. However, when P_i has the specific form shown in (1), we can further reduce the problem to minimizing $\Pi_i \|P_i^{-1}\|_F^2$, and with Lemma 4, we have

Table 2. Description of the Adult data set used in the evaluation

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

QI: quasiidentifier.

$$P_i^{-1} = \frac{1}{p_i - q_i}(I - q_i \mathbf{1}\mathbf{1}^T), \quad \|P_i^{-1}\|_F^2 = \frac{(d_i-1)^3}{(d_i p_i - 1)^2} + 1.$$

where **1** is the column vector whose cells are all equal to **1**.

V. EMPIRICAL EVALUATION

We ran our experiments on the Adult Database from the UCI data mining repository [41] in our evaluations. The same database has been used in previous work on k-anonymity, l-diversity, t-closeness, and anatomy [2-5]. The Adult Database contains 45,222 tuples from US census data and 14 attributes. Table 2 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. We adopt the following standard distance measures to compare the difference of distributions between the original and reconstructed data to quantify the utility. Given two distributions $P = (p_1, p_2, ..., p_m)$, $Q = (q_1, q_2, ..., q_m)$, **KL** distance is defined as $d_{KL}(P, Q) = \Sigma_{i=1}^{m} p_i \log \frac{p_i}{q_i}$ and χ^2-distance is $d_{\chi^2}(P, Q) = \Sigma_{i=1}^{m} \frac{(p_i - q_i)^2}{p_i}$.

In our evaluation, we first investigate utility vs. privacy of the randomization method in protecting attribute disclosure on two aspects: 1) compare data utility among different scenarios, and 2) the impact of cardinality upon data utility. Then, we compare randomization with l-diversity and anatomy.

Table 3. Randomization parameters p_i for three cases of *RR* (data set EMGRW)

	RR-QI				*RR-S*
l	E	M	G	R	W
2	0.824	0.872	0.920	0.941	0.650
3	0.548	0.812	0.898	0.985	0.267
4	0.382	0.736	0.918	0.961	0.217
5	0.314	0.615	0.873	0.938	0.167
	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

E: education, M: martial-status, G: gender, R: race, W: workclass, RR: randomized response, QI: quasiidentifier.

A. Randomization

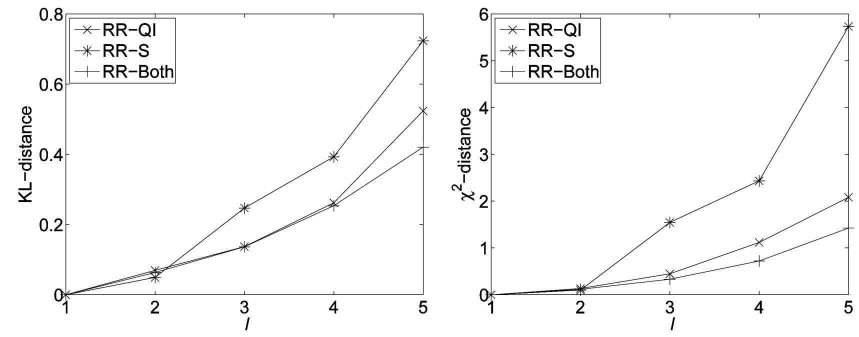
We treated *Education, Martial-status, Gender, and Race* as the quasi-identifier and used *Workclass* as the sensitive attribute. We term this data set EMGRW. We distort only *QI* attributes, or sensitive attribute *S*, or both in different application scenarios for randomization. Table 3 shows the derived randomization parameter *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 to achieve better privacy protections (when l is increased). We can also observe that p_i for *QI* attributes in *RR-Both* is generally closer to 1 than that in *RR-QI*, because *RR-Both* could also distort the sensitive attribute in addition to those *QI* attributes to achieve some given privacy protection. Thus, a small magnitude of distortion is needed for *QI* attributes in *RR-Both*.

Fig. 3 shows results on various distance measures. Naturally, there is a tradeoff between minimizing utility loss and maximizing privacy protection. Fig. 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.

As discussed in Section III, one advantage of the randomization scheme is that the more data we have, the more accurate reconstruction we can achieve. We generate four more data sets with varied sizes by sampling $r * N$ tuples from the Adult Data randomly where N is the size of the original Adult data set and we set $r \in [0.5, 1.5, 2, 2.5]$ to investigate the impact of data size upon the data utility of the randomized data. All four generated data sets have the exact same distribution as the original one. Fig. 4 shows the accuracy of reconstructed data distribution when data size increases. We see that data utility is further improved when more data are available.

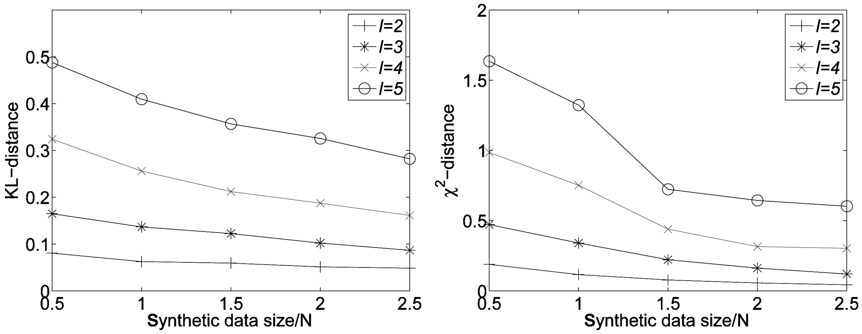
B. Comparison to Other Models

We chose *Education, Salary, Gender, Race* as *QI,* and *Occupation* as the sensitive attribute, similar to the settings of empirical evaluations in [2, 5] to compare randomization scheme with *l*-diversity and anatomy. We term this data set ESGRO. We did not use the previous EMGRW data set, because *l*-diverse partitions cannot be derived by the anatomy algorithm or entropy *l*-diversity. As specified in Machanavajjhala et al. [2] and Xiao and Tao [5], an *l*-diverse partition exists, if and only if at most N/l records are associated with the same sensitive value, where N is the cardinality of the data set. The data distribution of attribute *Work-class* is much skewed and the frequency of one value is much larger than the others. In this section, we focus on the use of *RR-QI* to compare to those two group based models:



(a) KL-distance (b) χ^2-distance

Fig. 3. Distances between $\hat{\pi}$ and π for three scenarios of *RR* (data set EMGRW). RR: randomized response, QI: quasiidentifier.



(a) KL-distance (b) χ^2-distance

Fig. 4. For *RR-Both*, distances between $\hat{\pi}$ and π decrease, as the data set size increases (data set EMGRW). RR: randomized response, QI: quasiidentifier.

l-diversity and anatomy, because the overall distribution of sensitive attribute values is unchanged in both group based methods after anonymization and *RR-QI* after randomization.

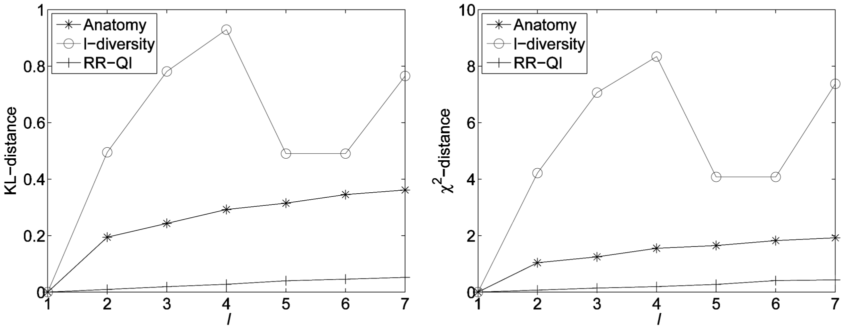
Generalization approaches usually measure the utility syntactically by the number of generalization steps applied to quasi-identifiers [42], average size of quasi-identifier equivalence classes [2], sum of squares of class sizes, or preservation of marginals [3]. We further examine the utility preservation from the perspective of answering aggregate queries, in addition to the previous distribution distance measures, since the randomization scheme is not based on generalization or suppression. We adopt *query answering accuracy*; the same metric is used in Zhang et al. [4]. We also consider the variation of correlations between the sensitive attribute and quasi-identifiers. We used an implementation of the Incognito algorithm [3] to generate the entropy *l*-diverse tables and used the anatomy algorithm in [5] in our experiments.

***1) Distribution Distance*:** We compare data utility of reconstructed probability distributions for different models according to the distance measures. Fig. 5 shows distances between $\hat{\pi}$ and π for anatomy, *l*-diversity and *RR-QI* on the data set ESGRO. We can observe that randomization outperforms both anatomy and *l*-diversity methods, because 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 terms 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 the estimate for the distribution of the original data. However, there is no similar monotonic trend in *l*-diversity, because the generalization algorithm chooses different attributes for generalization with various *l* values; this makes the accuracy of the estimated distribution vary to different extents.

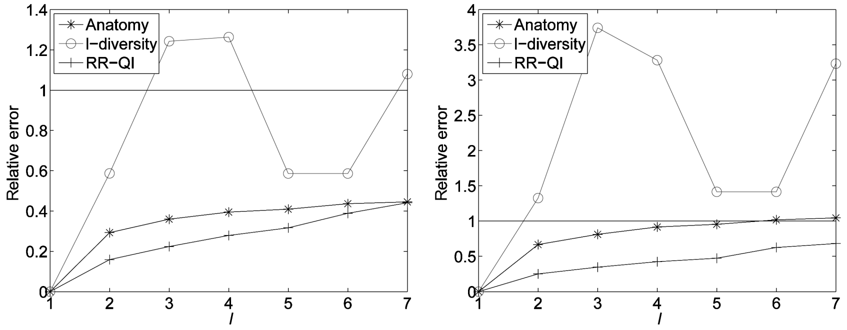
2) Query Answering Accuracy: The accuracy of answering aggregate queries is one of the important aspects to evaluate the utility of distorted data. We compare randomization against two other anonymization approaches [3, 5] using the average relative error in the returned count values. For each count value (corresponding to the number of records in a group), its relative error is defined as $|act - est|/act$, where *act* is the actual count value from the original data, and *est* is the estimate from the reconstructed data for RR approaches or the anonymized data for anonymization approaches. We consider two types of queries in our evaluation.

- Base group-by query with the form:
  SELECT $A_1, \ldots, A_m$, *S*, COUNT(*) FROM data



(a) KL-distance (b) χ^2-distance

Fig. 5. Distances between $\hat{\pi}$ and π for anatomy, *l*-diversity and *RR-QI* (randomized response-quasiidentifier; data set ESGRO).



(a) Base queries (b) Hierarchy-queries

Fig. 6. Relative errors of queries for anatomy, *l*-diversity and *RR-QI* (randomized response-quasiidentifier; data set ESGRO).

WHERE $A_1 = i_1 \ldots A_m = i_m$ AND $S = i_s$ GROUP BY $(A_1, \cdots, A_m, S)$
Where $i_k \in \Omega_k$ and $i_s \in \Omega_s$.
- Cube query with the form:
SELECT $A_1, \cdots, A_m, S$, COUNT(*) FROM data
GROUP BY **CUBE** $(A_1, \cdots, A_m, S)$

The base group-by query returns the number of records in each group of $(A_1 = i_1 \ldots A_m = i_m$ and $S = i_s)$. Note that the total number of groups is equal $D = d_s \prod_{i=1}^{m} d_i)$.

After running the base group-by query on the original data and the reconstructed data or the anonymized data, we get the count values of *act* and *est*, respectively. Fig. 6a shows the calculated average relative errors of three methods (anatomy, *l*-diversity, and *RR-QI*) with varied l ($l = 1, \cdots, 7$). We use the 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 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 Fig. 6b. We can observe from Fig. 6a and 6b that randomization permits significantly more accurate aggregate analysis than both *l*-diversity and anatomy, since it can recover more accurate data distribution. Conversely, *l*-diversity loses the data distribution of records within each generalized QI-group and anatomy loses correlations between the *QI* attributes and the sensitive attribute *S*.

3) 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:

$$U = -\frac{\Sigma_i \Sigma_j \pi_{ij} \log \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}}{\Sigma_j \pi_{+j} \log \pi_{+j}}. \qquad (12)$$

The uncertainty coefficient takes values between -1 and 1; 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.

Tables 4 and 5 show correlation (uncertainty coefficient) val-

Table 4. Variation of correlation (uncertainty coefficient) between pairs of quasiidentifier (*QI*) attributes under different models ($\times 10^{-2}$) (data set ESGRO)

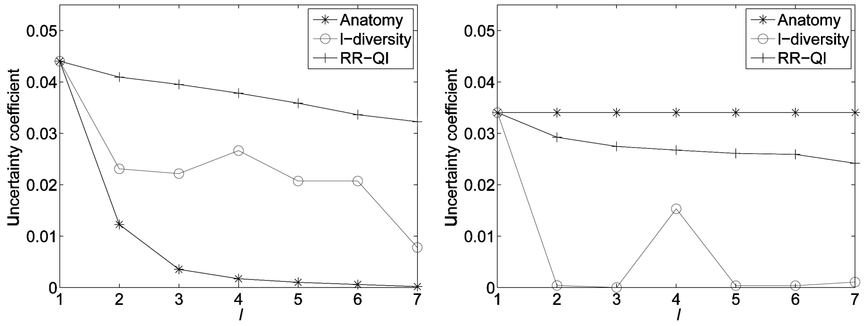
Correlation		*QI* vs. *QI*					
		E vs. S	E vs. G	E vs. R	S vs. G	S vs. R	G vs. R
	Original	11.54	0.68	1.92	4.12	1.07	1.08
l = 3	Anatomy	11.54	0.68	1.92	4.12	1.07	1.08
	l-diversity	0	0	0	0	0	0
	RR-QI	8.57	0.63	1.86	3.65	0.53	1.08
l = 4	Anatomy	11.54	0.68	1.92	4.12	1.07	1.08
	l-diversity	9.18	0	0	0	0	0
	RR-QI	8.05	0.58	1.29	3.37	0.45	1.07
l = 5	Anatomy	11.54	0.68	1.92	4.12	1.07	1.08
	l-diversity	0	0	0.21	0	0	0
	RR-QI	7.50	0.33	1.02	3.15	0.44	1.04

E: education, S: salary, G: gender, R: race, O: occupation.

Table 5. Variation of correlation (uncertainty coefficient) between quasiidentifier (*QI*) and *S* under different models ($\times 10^{-2}$) (data set ESGRO)

Correlation		*QI* vs. *S*			
		E vs. O	**S vs. O**	G vs. O	R vs. O
	Original	9.90	**2.74**	4.40	0.57
l = 3	Anatomy	0.90	**0.20**	0.25	0.06
	l-diversity	8.86	**0**	0	0
	RR-QI	8.72	**2.41**	4.13	0.54
l = 4	Anatomy	0.46	**0.12**	0.08	0.02
	l-diversity	7.91	**2.74**	0	0
	RR-QI	8.33	**2.27**	4.01	0.51
l = 5	Anatomy	0.26	**0.05**	0.07	0.02
	l-diversity	7.91	**0**	0	0.37
	RR-QI	7.77	**2.17**	3.89	0.50

E: education, S: salary, G: gender, R: race, O: occupation.



(a) QI vs.S (b) QI vs.QI

Fig. 7. Average value of uncertainty coefficients among attributes for anatomy, *l*-diversity and *RR-QI* (randomized response-quasiidentifier; data set ESGRO).

ues between each 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. We only include correlation values for *l* = 3, 4, and 5 due to the space limitation. 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) to study how correlations between *QI* and *S* change. 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, and 5 respectively. Conversely, 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, and 5 while it perfectly achieves correlation preservation when *l* = 4, because the Salary attribute is generalized to “All” when *l* = 3, and 5, while it is unchanged across all *QI*-groups when *l* = 4. *l*-diversity in general cannot preserve correlation well, because it is intractable to predict which *QI* attributes will be generalized in *l*-diversity.

Fig. 7 shows the average value of uncertainty coefficients among attributes for anatomy, *l*-diversity and *RR-Both* on the data set ESGRO. As shown in Fig. 7a, randomization achieves the best correlation preservation between the sensitive attribute and quasiidentifiers across all privacy thresholds. Fig. 7b also clearly shows that randomization better preserves correlation among quasi-identifiers than *l*-diversity does. Please note that anatomy can best achieve the correlation among quasi-identifiers, since it does not change values of quasi-identifiers.

C. Evaluation Summary

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 are available. The evaluation also showed that the randomization approach can further improve the accuracy of the reconstructed distribution (and hence utility preservation) when more data are available, while generalization and anatomy approaches do not have this property.

VI. CONCLUSION

In this paper, we investigated attribute disclosure in the case of linking attacks. We compared randomization to other anonymization schemes (*l*-diversity and anatomy) in terms of utility preservation, with the same privacy protection requirements. Our experimental evaluations showed randomization significantly outperforms generalization, i.e., achieving better utility preservation, while yielding the same privacy protection.

There are several other avenues for future work. We aim to extend our research to handle multiple sensitive attributes. In this paper, 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 insensitive attributes in the population. Different forms of background knowledge have been studied in privacy preservation data publishing recently. For example, the formal language [22, 24] and maximum estimation constraints [43] are proposed to express background knowledge of attackers. We will investigate privacy disclosure under those background knowledge attacks. We should note that randomization may not outperform generalization in all cases, especially when some specific background knowledge is utilized in the adversary model, because generalization is truthful and the ambiguity is between true tuples (e.g., individual a and b are indistinguishable), whereas randomization can be regarded as adding noise, and noise can be removed when some background knowledge is known.

We will continue our study of further reducing computation overhead of the randomization approach. The execution time of randomization is usually slower than *l*-diversity and anatomy. For example, the execution time of our *RR-Both* approach on the Adult data sets was 10 times slower than for generalization approaches due to the heavy cost of determining the optimal distortion parameters with attribute disclosure constraints, especially when the attribute domains are large. One advantage of randomization is that the reconstruction accuracy increases when more data are available. Guo et al. [10] preliminarily conducted theoretical analysis on how the randomization process affects the accuracy of various measures (e.g., support, confidence, and lift) in market basket data analysis. In our future work, we will study the accuracy of reconstruction in terms of bias and variance of estimates in randomizing general microdata.

Finally, we are interested in comparing our randomization approaches to most recently developed generalization approaches (e.g., the accuracy-constrained *l*-diversification method in [14]). We are also interested in combining the derived optimal randomization schemes with the small domain randomization [38] to further improve utility preservation. It is our belief that extensive studies on comparing different privacy preservation data publishing approaches are crucial.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. National Science Foundation IIS-0546027 and is an extension of the conference version [44]. Authors thank reviewers for their insightful comments; this greatly helped improve the quality of this work.

REFERENCES

1. P. Samarati and L. Sweeney, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,” *Proceedings of the IEEE Symposium on Research in Security and Privacy*, Oakland, CA, 1998.
2. A. Machanavajjhala, J. Gehrke, D. Kiefer, and M. Venkitasubra-

manian, “L-diversity: privacy beyond k-anonymity,” *Proceedings of the 22nd International Conference on Data Engineering*, Atlanta, GA, 2006.

3. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Incognito: efficient full-domain K-anonymity,” *ACM SIGMOD International Conference on Management of Data*, Baltimore, MD, 2005, pp. 49-60.
4. Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, “Aggregate query answering on anonymized tables,” *Proceedings of the 23rd International Conference on Data Engineering*, Istanbul, Turkey, 2007, pp. 116-125.
5. X. Xiao and Y. Tao, “Anatomy: simple and effective privacy preservation,” *Proceedings of the 32nd International Conference on Very Large Data Bases*, Seoul, Korea, 2006, pp. 139-150.
6. D. Lambert, “Measures of disclosure risk and harm,” *Journal of Official Statistics*, vol. 9, no. 2, pp. 313-331, 1993.
7. N. Li, T. Li, and S. Venkatasubramanian, “T-closeness: privacy beyond k-anonymity and l-diversity,” *Proceedings of the 23rd International Conference on Data Engineering*, Istanbul, Turkey, 2007, pp. 106-115.
8. W. Du and Z. Zhan, “Using randomized response techniques for privacy-preserving data mining,” *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, 2003, pp. 505-510.
9. S. J. Rizvi and J. R. Haritsa, “Maintaining data privacy in association rule mining,” *Proceedings of the 28th International Conference on Very Large Data Bases*, Hong Kong, China, 2002, pp. 682-693.
10. L. Guo, S. Guo, and X. Wu, “Privacy preserving market basket data analysis,” *The 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Warsaw, Poland, 2007, pp. 103-114.
11. L. Guo and X. Wu, “Privacy preserving categorical data analysis with unknown distortion parameters,” *Transactions on Data Privacy*, vol. 2, no. 3, pp. 185-205, 2009.
12. Z. Teng and W. Du, “Comparisons of k-anonymization and randomization schemes under linking attacks,” *Proceedings of the 6th International Conference on Data Mining*, Hong Kong, China, 2006, pp. 1091-1096.
13. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, “Approximation algorithms for k-anonymity,” *Journal of Privacy Technology*, 20051120001, Nov. 2005.
14. G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, “A framework for efficient data anonymization under privacy and accuracy constraints,” *ACM Transactions on Database Systems*, vol. 34, no. 2, pp. 9:1-9:47, 2009.
15. D. Kifer and J. Gehrke, “Injecting utility into anonymized datasets,” *ACM SIGMOD International Conference on Management of Data*, Chicago, IL, 2006, pp. 217-228.
16. N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, “Distribution-based microdata anonymization,” *Proceedings of the 35th International Conference on Very Large Data Bases*, Lyon, France, 2009.
17. A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” *IEEE Symposium on Security and Privacy*, Oakland, CA, 2008, pp. 111-125.
18. J. Brickell and V. Shmatikov, “The cost of privacy: destruction of data-mining utility in anonymized data publishing,” *The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, NV, 2008, pp. 70-78.
19. T. Li and N. Li, “On the tradeoff between privacy and utility in data publishing,” *The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, 2009, pp. 517-525.
20. T. M. Truta and B. Vinay, “Privacy protection: p-sensitive k-anonymity property,” *Proceedings of the 22nd IEEE Internationl Conference on Data Engineering*, Atlanta, GA, 2006, p. 94.
21. R. C. W. Wong, J. Li, A. W. C. Fu, and K. Wang, “(α, k)-anonymity: an enhanced k-anonymity model for privacy-preserving data publishing,” *The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, 2006, pp. 754-759.
22. D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, “Worst-case background knowledge for privacy-preserving data publishing,” *Proceedings of the 23rd International Conference on Data Engineering*, Istanbul, Turkey, 2007, pp. 126-135.
23. R. C. W. Wong, A. W. C. Fu, K. Wang, and J. Pei, “Minimality attack in privacy preserving data publishing,” *Proceedings of the 33nd International Conference on Very Large Data Bases*, Vienna, Austria, 2007, pp. 543-554.
24. B. Chen, K. Lefevre, and R. Ramakrishnan, “Privacy skyline: privacy with multidimensional adversarial knowledge,” *Proceedings of the 33nd International Conference on Very Large Data Bases*, Vienna, Austria, 2007, pp. 770-781.
25. M. E. Nergiz, M. Atzori, and C. Clifton, “Hiding the presence of individuals from shared databases,” *ACM SIGMOD International Conference on Management of Data*, Beijing, China, 2007, pp. 665-676.
26. J. Li, Y. Tao, and K. Xiao, “Preservation of proximity privacy in publishing numerical sensitive data,” *ACM SIGMOD International Conference on Management of Data*, Vancouver, BC, 2008, pp. 473-485.
27. Q. Wei, Y. Lu, and Q. Lou, “(t, λ)-uniqueness: anonymity management for data publication,” *The 7th IEEE/ACIS International Conference on Computer and Information Science*, Portland, OR, 2008, pp. 107-112.
28. S. L. Warner, “Randomized response: a survey technique for eliminating evasive answer bias,” *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63-66, 1965.
29. R. Agrawal and R. Srikant, “Privacy-preserving data mining,” *ACM SIGMOD International Conference on Management of Data*, Dallas, TX, 2000, pp. 439-450.
30. C. C. Aggarwal and P. S. Yu, “A survey of randomization methods for privacy-preserving data mining,” *Privacy-Preserving Data Mining: Models and Algorithms*, C. C. Aggarwal and P. S. Yu, Eds., New York, NY: Springer US, pp. 137-156, 2008.
31. S. Agrawal and J. R. Haritsa, “A framework for high-accuracy privacy-preserving mining,” *Proceedings of the 21st International Conference on Data Engineering*, Tokyo, Japan, 2005, pp. 193-204.
32. C. C. Aggarwal, “On unifying privacy and uncertain data models,” *Proceedings of the 24th International Conference on Data Engineering*, Cancun, Mexico, 2008, pp. 386-395.
33. Y. Zhu and L. Liu, “Optimal randomization for privacy preserving data mining,” *The 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, 2004, pp. 761-766.
34. D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, “From t-

closeness-like privacy to postrandomization via information theory," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 11, pp. 1623-1636, 2010.

35. J. M. Gouweleeuw, P. Kooiman, L. C. R. J. Willenborg, and P. P. de Wolf, "Post randomisation for statistical disclosure control: theory and implementation," *Journal of Official Statistics*, vol. 14, no. 4, pp. 463-478, 1998.
36. Z. Huang and W. Du, "OptRR: optimizing randomized response schemes for privacy-preserving data mining," *Proceedings of the 24th International Conference on Data Engineering*, Cancun, Mexico, 2008, pp. 705-714.
37. X. Xiao, Y. Tao, and M. Chen, "Optimal random perturbation at multiple privacy levels," *Proceedings of the 35th International Conference on Very Large Data Bases*, Lyon, France, 2009, pp. 814-825.
38. R. Chaytor and K. Wang, "Small domain randomization: same privacy, more utility," *Proceedings of the 36th International Conference on Very Large Data Bases*, Singapore, 2010, pp. 608-618.
39. A. Chaudhuri and R. Mukerjee, Randomized Response: Theory and Techniques, New York: Marcel Dekker, 1988.
40. T. F. Coleman, J. Liu, and W. Yuan, "A new trust-region algorithm for equality constrained optimization," *Computational Optimization and Applications*, vol. 21, no. 2, pp. 177-199, 2002.
41. A. Asuncion and D. J. Newman, "UCI machine learning repository," http://mlearn.ics.uci.edu/MLRepository.html.
42. P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010-1027, 2001.
43. W. Du, Z. Teng, and Z. Zhu, "Privacy-MaxEnt: integrating background knowledge in privacy quantification," *ACM SIGMOD International Conference on Management of Data*, Vancouver, BC, 2008, pp. 459-472.
44. L. Guo, X. Ying, and X. Wu, "On attribute disclosure in randomization based privacy preserving data publishing," *Proceedings of the 10th IEEE International Conference on Data Mining Workshops*, Sydney, Australia, 2010, pp. 466-473.
45. G. Strang, Introduction to Linear Algebra, 3rd ed., Wellesley, MA: Wellesley-Cambridge Press, 2003.

Appendix 1. Proof of Property 1

We start with applying *RR* to the sensitive attribute, $\Pr(\widehat{QI}_r = QI_r) = 1$. 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 $\Pr(\hat{S}_r = S_r = 1|QI_r)$. After the adjustment, the randomization probabilities are given by

$$\Pr(\tilde{S}_r = 1|S_r = 1) = p_s, \quad \Pr(\tilde{S}_r = 1|S_r = 0) = q_s. \tag{13}$$

By definition, the posterior probabilities are given by

$$\Pr(S_r = 1|\tilde{S}_r = 1) = \frac{p_s\pi_1}{p_s\pi_1 + q_s\pi_0}, \tag{14}$$

$$\Pr(S_r = 1|\tilde{S}_r = 0) = \frac{(1-p_s)\pi_1}{(1-p_s)\pi_1 + (1-q_s)\pi_0}. \tag{15}$$

Combining (13), (14), and (15), we have

$$\begin{aligned} \Pr(S_r = 1|QI_r) &= \pi_1 \Pr(\hat{S}_r = S_r = 1|QI_r) \\ &= \pi_1 \sum_{i=0,1} \Pr(\tilde{S}_r = i|S_r = 1) \Pr(S_r = 1|\tilde{S}_r = i) \\ &= \pi_1^2 \left[\frac{p_s^2}{p_s\pi_1 + q_s\pi_0} + \frac{(1-p_s)^2}{(1-p_s)\pi_1 + (1-q_s)\pi_0}\right] \end{aligned} \tag{16}$$

Taking the derivative with respect to p_s, we have (16) 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.

Appendix 2. Proof of Result 2

LEMMA 3. *If* $d(\hat{\pi}, \pi) = \|\hat{\pi} - \pi\|_2^2$, *we have* $E[d(\hat{\pi}, \pi)] = \text{trace}(\Sigma)$, *where* Σ *is the covariance matrix of* $\hat{\pi}$ *shown in* (2).

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

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

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

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

The last equality is due to the fact that

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

We proved the lemma.

LEMMA 4. *Let* P_i *be the randomization matrix specified in* (1). *When* $p_i \neq \frac{1}{d_i}$, *we have*

$$P_i^{-1} = \frac{1}{p_i - q_i}(I - q_i\mathbf{1}\mathbf{1}^T), \quad \|P_i^{-1}\|_F^2 = \frac{(d_i-1)^3}{(d_ip_i-1)^2} + 1.$$

where **1** *is the column vector, whose cells all equal 1.*

PROOF. We can re-write P_i as follows:

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

With $q_i = \frac{1-p_i}{d_i-1}$, $\mathbf{1}^T\mathbf{1} = d_i$, and the binomial inverse theorem [45], we can immediately get P_i^{-1}, and $\|P_i^{-1}\|_F^2$ can be directly calculated from P_i^{-1}.

LEMMA 5. *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 4, 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} have 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 3, when the distance is the squared Euclidean distance, to minimize E[d($\hat{\pi}$, π)] is equivalent to minimize trace(Σ). With (2), we have

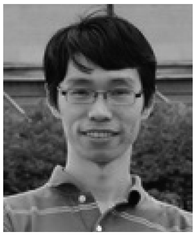
$$\mathrm{trace}(\Sigma) = \mathrm{trace}\left\{\frac{1}{N}[P^{-1}(P\pi)^{\delta}(P^T)^{-1} - \pi\pi^T]\right\}$$

$$\begin{aligned}
&\propto \mathrm{trace}[P^{-1}(P\pi)^{\delta}(P^T)^{-1}] \\
&= \mathrm{trace}[(P\pi)^{\delta}(P^{-1})^T P^{-1}] \\
&= \sum_{i=1}^{D}(P\pi)_i\|(P^{-1})_i\|_F^2 \\
&= \|(P^{-1})_1\|_F^2\sum_{i=1}^{D}(P\pi)_i \quad \text{(with Lemma 5)} \\
&\propto \|P^{-1}\|_F^2 = \prod_i \|P_i^{-1}\|_F^2 .
\end{aligned}$$

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.

Ling Guo

Ling Guo obtained her Ph.D. degree in Information Technology from the University of North Carolina at Charlotte in 2010. She received her BS degree in Electronics Engineering from Shan Dong University of China in 1997. Her major research interests include knowledge discovery and privacy-preservation data mining.



Xiaowei Ying

Xiaowei Ying obtained his Ph.D. degree in Information Technology from the University of North Carolina at Charlotte in 2011. He received his BA degree in Mathematics from Fudan University of China in 2006. His major research interests include privacy-preservation data mining and social network analysis.

Xintao Wu

Xintao Wu is an Associate Professor in the Department of Software and Information Systems at the University of North Carolina at Charlotte, USA. He obtained his Ph.D. degree in Information Technology from George Mason University in 2001. He received his BS degree in Information Science from the University of Science and Technology of China in 1994, an ME degree in Computer Engineering from the Chinese Academy of Space Technology in 1997. His major research interests include data mining and knowledge discovery, data privacy and security.