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 attack

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-
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 quasi-identifier 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 gen-
eralization 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, ε)-anonymity [4], (c, k)-safety [22], m-confidence [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 $\bar{A}$. The problem is to estimate $\pi_A$, the unknown proportion of population members in group A. Each respondent is provided with a randomization device by which the respondent chooses one of the following two questions: Do you belong to A? or Do you belong to $\bar{A}$? with respective probabilities $p$ and $1 − p$ and then replies yes or no to the question chosen. 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].


# III. PRELIMINARIES

Dataset $T$ contains $N$ records and $m + 1$ categorical attributes: $A_1, A_2, \ldots, A_m$, and $S$. We use $QI = \{A_1, A_2, \ldots, A_m\}$ 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, $T$ 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 $\Omega_i$ to denote the domain of $A_i$, $\Omega = \{0, 1, \ldots, d − 1\}$, 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 $(A_{r1}, A_{r2}, \ldots, A_{rm}, S_r)$ or simply $(QI_r, S_r)$. Let $D = d_1 \times \cdots \times d_m$ be the total number of cells in the contingency table.

Let $\pi_{i1} \times \cdots \times \pi_{is}$ denote the true proportion corresponding to the categorical combination $(A_i = i_1, \ldots, A_m = i_m, S = s)$. Let $\pi$ be the column vector with $D$ elements $\pi_{i1} \times \cdots \times \pi_{is}$ arranged in a fixed order. Table 1 shows one contingency table example for a data set with one QI attribute (Gender, $d = 2$) and one sensitive attribute $S$ (Disease, $d = 3$). Table 1a shows the original contingency table where $\pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ corresponds to a fixed order of cell entries $\pi_i$ in the $2 \times 3$ contingency table. $\pi_0$ denotes the proportion of records with Female and Cancer. The column sum $\pi_0$ 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)

<table>
<thead>
<tr>
<th></th>
<th>Cancer</th>
<th>Flu</th>
<th>Anemia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>π₁₀</td>
<td>π₁₁</td>
<td>π₁₂</td>
</tr>
<tr>
<td>Female</td>
<td>π₂₀</td>
<td>π₂₁</td>
<td>π₂₂</td>
</tr>
</tbody>
</table>

(b) Instance

<table>
<thead>
<tr>
<th></th>
<th>Cancer</th>
<th>Flu</th>
<th>Anemia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>14</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

(c) After randomization

<table>
<thead>
<tr>
<th></th>
<th>Cancer</th>
<th>Flu</th>
<th>Anemia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>λ₁₀</td>
<td>λ₁₁</td>
<td>λ₁₂</td>
</tr>
<tr>
<td>Female</td>
<td>λ₂₀</td>
<td>λ₂₁</td>
<td>λ₂₂</td>
</tr>
</tbody>
</table>

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 \( R_i \), 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 \( p_{iu}^{(v)} \): \( \Pr(\tilde{A}_i = u|A_i = v) = p_{iu}^{(v)} \). Let \( p_{iu} = [p_{iu}^{(v)}]_{v=0}^{d_i-1} \), and we call \( p_i \) (or \( P_i \)) the distortion matrix for \( A_i \) (or \( S \)). Naturally, the column sums of \( P_i \) are equal to 1. The original database \( T \) is changed to \( \tilde{T} \), 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 \( \pi \) denote the contingency table of the randomized data \( \tilde{T} \). We arrange \( \pi \) into the column vector with the same order of \( \pi \). 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 \( P = P_1 \otimes \cdots \otimes P_n \otimes P_s \), and \( \otimes \) stands for the Kronecker product. The Kronecker product is an operation on two matrices, an \( m \times n \) matrix \( A \) and a \( p \times q \) matrix \( B \), resulting in the \( mp \times nq \) 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:

\[ \Pr(\tilde{A}_i = u|A_i = v) = \begin{cases} p_i, & u = v, \\ q_i = 1 - p_i, & u \neq v. \end{cases} \] (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 \( \tilde{T} \) and its contingency table \( \lambda \), the unbiased estimate of \( \pi \) is given by \( \tilde{\pi} = P^{−1} \lambda \), and the covariance matrix of the estimator is

\[ \Sigma = \text{Cov}(\tilde{\pi}) = \frac{1}{N} [E(\pi)^{T} (E\pi)^{-1} - \pi E^{T}] \] (2)

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

Thus, data analysts derive the estimate 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 \( T \) with domain \( \Omega_x \), and \( \tilde{X} \) is the randomized value of \( X \) in \( \tilde{T} \). 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 \( \hat{\Omega}_x \) 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 \( \mu, v \in \hat{\Omega}_x \),

\[ \hat{X} = \mu \text{ with prob. } \Pr(X = \mu | \tilde{X} = v), \] (3)
where $\Pr(X = \mu | \bar{X} = \nu)$ denotes the attacker’s posterior belief on the original value $X = \mu$ when he observes $\bar{X} = \nu$. With the Bayes’ theorem, it can be calculated by:

$$\Pr(X = \mu | \bar{X} = \nu) = \sum_{\nu' \in \Omega} \pi_{\nu'} \Pr(\bar{X} = \nu | X = \mu) \Pr(X = \mu | \bar{X} = \nu).$$  \hspace{1cm} (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:

$$\Pr(\bar{X} = X = \mu) = \sum_{\nu \in \Omega} \Pr(\bar{X} = \nu | X = \mu) \Pr(X = \mu | \bar{X} = \nu). \hspace{1cm} (5)$$

**Proof.** For a particular observed value $\bar{X} = \nu \in \Omega$, $\Pr(\bar{X} = X = \mu, \bar{X} = \nu) = \Pr(\bar{X} = \nu | X = \mu) \Pr(X = \mu | \bar{X} = \nu)$.

Then, with the law of total probability, we have

$$\Pr(X = \mu) = \sum_{\nu \in \Omega} \Pr(\bar{X} = \nu | X = \mu) \Pr(X = \mu | \bar{X} = \nu).$$

The probability of attackers’ correct estimation is also defined as the reconstruction probability, $\Pr(\bar{X} = X = \mu) \rightarrow X = \mu$, 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 $\Pr(\bar{X} = X = \mu)$ 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 $\hat{S}_r$ 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 $\bar{T}$ and he knows that $\bar{T}$ is a randomized version of some base table $T$. The attacker knows the domain of each attribute of $T$. 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 $\Pr(S_r | QI_r) \leq 1/l$, because there are $\pi_{QI_r}$ records within the group of $QI_r$, and $\pi_{QI_r}$ 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 $\Pr(S_r | QI_r)$.

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 $QI$ is given by $\Pr(QI_r = QI_r)$, and hence the risk of sensitive attribute disclosure is $\Pr(S_r | QI_r) = QI_r$.

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 $\Pr(QI_r = QI_r)$, and 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 $QI_r = \alpha = \{i_1, i_2, \ldots, i_m\}$, $i_k \in \Omega$, and his sensitive attribute $S_r = u$, $u \in \Omega$. The probability of successfully predicting the sensitive attribute value $S_r$ of the target individual $r$ given his quasi-identifier values $QI_r$ is:

$$\Pr(S_r | QI_r) = \frac{\pi_{QI_r}}{\pi_{QI_r}} \Pr(QI_r = QI_r) \Pr(S_r | QI_r).$$  \hspace{1cm} (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:

```
For an individual $m$, we apply different randomization parameters. We can include our proof in the Appendix.

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

$$
Pr(S|QI) = \frac{\pi_{QI}(S, π_{QI}|S)}{\pi_{QI}(T, π_{QI}|T)}
$$

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 $QI$, of individual $r$:

- for RR-S, $Pr(S|QI)$ is minimized when $p_1 = \frac{1}{d_1}$, min $Pr(S|QI)$ is $\left(\frac{\pi_{QI}|S_1}{\pi_{QI}}\right)$;
- for RR-QI, $Pr(S|QI)$ is minimized when $p_1 = \frac{1}{d_1}$ ($i = 1, 2, \ldots, m$), min $Pr(S|QI) = \frac{\pi_{QI}|S_1}{\pi_{QI}}$;
- for RR-Both, $Pr(S|QI)$ is minimized when $p_1 = \frac{1}{d_1}$ ($i = 1, 2, \ldots, m$ and $s$), min $Pr(S|QI) = \frac{\pi_{QI}|S_1}{\pi_{QI}}$.

**Example.** We use the instance shown in Table 1b to illustrate this property. For an individual $r$ with ($QI =$ Female, $S =$ Cancer), we randomize $S$ (Disease) with $p_1$, and $QI$ (Gender) with $p_2$ independently. Fig. 2 shows how the attribute disclosure is varied when we apply different randomization parameters. We can see that $Pr(S|QI)$ reaches the maximum (i.e., $\frac{\pi_{QI}}{\pi_{QI}} = 0.43$) when no randomization is introduced. Fig. 2a shows the scenario when we only randomize $S$. We can see that min $Pr(S|QI) = \left(\frac{\pi_{QI}}{\pi_{QI}}\right) = 0.18$ when $p_1 = \frac{1}{d_1} = \frac{1}{3}$. Fig. 2b shows the scenario when we only randomize $QI$. We can see that min $Pr(S|QI) = \pi_{QI} = 0.12$ when $p_2 = \frac{1}{d_2} = \frac{1}{2}$. Fig. 2c shows the case where randomization is applied to both $QI$ and $S$. $Pr(S|QI)$ reaches the minimum only when both $p_1 = \frac{1}{3}$ and $p_2 = \frac{1}{2}$, and min $Pr(S|QI) = \frac{\pi_{QI}}{\pi_{QI}} = 0.05$.

**Computational Cost.** The main computation cost in (6) comes from calculating $Pr(QI|QI)$. 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: $λ_{QI} = P_{QI} π_{QI}$. 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:

$$
η = π_{QI} × [(P_{QI} | (λ_{QI}^T)\]\]

$$
= π_{QI} × [(\otimes_{i=1}^m (P_i)^T | \otimes_{i=1}^m (P_i) π_{QI})^T]
$$

where $×$ denotes the component-wise multiplication, $P_i^T$ is the component-wise square of $P_i$, and $λ_{QI}^T$ is the component-wise inverse of $λ_{QI}$.

In (9), we need to repeatedly calculate $(P_i \otimes \cdots \otimes P_m)\vec{x}$, where $\vec{x}$ denotes a column vector. Assume we use the naive algorithm in all matrix multiplications. Calculating $(P_i \otimes \cdots \otimes P_m)\vec{x}$ directly results in the time and storage complexity of $O(\Pi d_i')$. 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 $n \times n$, $m \times m$, and $m \times n$. Then

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

where $\vec{vec}(X)$ 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 $O(\Sigma d_i | Π d)$. The storage complexity is also reduced to $O(Π d + Σ d_i')$; this is mainly used to store the
Table 2. Description of the Adult data set used in the evaluation

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Type</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (G)</td>
<td>QI</td>
<td>2</td>
</tr>
<tr>
<td>Race (R)</td>
<td>QI</td>
<td>5</td>
</tr>
<tr>
<td>Education (E)</td>
<td>QI</td>
<td>16</td>
</tr>
<tr>
<td>Marital-status (M)</td>
<td>QI</td>
<td>7</td>
</tr>
<tr>
<td>Salary class (S)</td>
<td>QI</td>
<td>2</td>
</tr>
<tr>
<td>Work-class (W)</td>
<td>Sensitive</td>
<td>7</td>
</tr>
<tr>
<td>Occupation (O)</td>
<td>Sensitive</td>
<td>14</td>
</tr>
</tbody>
</table>

QI: quasiidentifier.

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 intrinsically 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 \( p \) to minimize \( E[d(\hat{\pi}, \pi)] \) s.t. \( \max \Pr(S_i|Q_i) \leq \tau, p_i \in (1/d_{i1}, 1), \)

(10)

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

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 \( (p_1, p_2, \cdots, p_m, \) and \( p) \) 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 \( \Pr(S_i|Q_i) \).

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 \( E[d(\hat{\pi}, \pi)] \), 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 \( p \) to minimize \( \Pi_i \{d_i^2 \} \) s.t. \( \max \Pr(S_i|Q_i) \leq \tau, p_i \in (1/d_{i1}, 1) \),

(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 \( d(\hat{\pi}, \pi) \) is equivalent to minimize trace(\( \Sigma \)) where \( \Sigma \) is the covariance matrix of the estimator of cell values in the contingency table (shown in Equation 2). Calculating trace(\( \Sigma \)) still involves high computational cost. However, when \( P_\tau \) has the specific form shown in (1), we can further reduce the problem to minimizing \( \Pi_i \{d_i^2 \} \), and with Lemma 4, we have

\[
\begin{align*}
\frac{d^2}{\sum p_i} &= \frac{1}{p_i - q_i} \left( \frac{q_i - q_i^R}{1 - q_i} \right)^2 + 1,
\end{align*}
\]

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 \( P = (p_1, p_2, \cdots, p_m, Q = (q_1, q_2, \cdots, q_m) \), KL distance is defined as \( d_{KL}(P, Q) = \sum_{i=1}^{m} \log \frac{p_i}{q_i} \) and \( \chi^2 \)-distance is \( d_{\chi^2}(P, Q) = \sum_{i=1}^{m} \frac{(p_i - q_i)^2}{p_i} \).

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 \( p \) for three cases of \( RR \) (data set EMGRW)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>RR-QI</th>
<th>RR-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>( l )</td>
<td>E</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>0.824</td>
<td>0.872</td>
</tr>
<tr>
<td>3</td>
<td>0.548</td>
<td>0.812</td>
</tr>
<tr>
<td>4</td>
<td>0.382</td>
<td>0.736</td>
</tr>
<tr>
<td>5</td>
<td>0.314</td>
<td>0.615</td>
</tr>
<tr>
<td>( RR-Both )</td>
<td>I</td>
<td>E</td>
</tr>
<tr>
<td>2</td>
<td>0.824</td>
<td>0.872</td>
</tr>
<tr>
<td>3</td>
<td>0.573</td>
<td>0.821</td>
</tr>
<tr>
<td>4</td>
<td>0.428</td>
<td>0.780</td>
</tr>
<tr>
<td>5</td>
<td>0.353</td>
<td>0.688</td>
</tr>
</tbody>
</table>

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 i = 2, 3, 4, 5. We can observe that more distortions (p is away from 1) are needed to achieve better privacy protections (when i is increased). We can also observe that p, 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 ∈ [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:

![Fig. 3. Distances between \( \hat{\pi} \) and \( \pi \) for three scenarios of RR (data set EMGRW). RR: randomized response, QI: quasiidentifier.](image)

![Fig. 4. For RR-Both, distances between \( \hat{\pi} \) and \( \pi \) decrease, as the data set size increases (data set EMGRW). RR: randomized response, QI: quasiidentifier.](image)
limiting attribute disclosure in randomization based microdata release

Ling Guo et al. 177 http://jcse.kiise.org

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 \( \pi \) for anatomy, \( l \)-diversity and RR-QI on the data set ESGRO. We can observe that randomization outperforms both anatomy and \( l \)-diversity methods, 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 \( \mid \text{act} - \text{est} \mid / \text{act} \), where \( \text{act} \) is the actual count value from the original data, and \( \text{est} \) 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

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

Fig. 6. Relative errors of queries for anatomy, \( l \)-diversity and RR-QI (randomized response-quasiidentifier; data set ESGRO).
Fig. 7. Average value of uncertainty coefficients among attributes for anatomy, \( l \)-diversity and \( RR-QI \) (randomized response-quasiidentifier; data set ESGRO).

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

<table>
<thead>
<tr>
<th>Correlation</th>
<th>( QI ) vs. ( QI )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E vs. S</td>
</tr>
<tr>
<td>Original</td>
<td>11.54</td>
</tr>
<tr>
<td>( l ) = 3</td>
<td>Anatomy</td>
</tr>
<tr>
<td>( l )-diversity</td>
<td>0</td>
</tr>
<tr>
<td>( RR-QI )</td>
<td>8.57</td>
</tr>
<tr>
<td>( l ) = 4</td>
<td>Anatomy</td>
</tr>
<tr>
<td>( l )-diversity</td>
<td>9.18</td>
</tr>
<tr>
<td>( RR-QI )</td>
<td>8.05</td>
</tr>
<tr>
<td>( l ) = 5</td>
<td>Anatomy</td>
</tr>
<tr>
<td>( l )-diversity</td>
<td>0</td>
</tr>
<tr>
<td>( RR-QI )</td>
<td>7.50</td>
</tr>
</tbody>
</table>


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

<table>
<thead>
<tr>
<th>Correlation</th>
<th>( QI ) vs. ( S )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E vs. O</td>
</tr>
<tr>
<td>Original</td>
<td>9.90</td>
</tr>
<tr>
<td>( l ) = 3</td>
<td>Anatomy</td>
</tr>
<tr>
<td>( l )-diversity</td>
<td>8.86</td>
</tr>
<tr>
<td>( RR-QI )</td>
<td>8.72</td>
</tr>
<tr>
<td>( l ) = 4</td>
<td>Anatomy</td>
</tr>
<tr>
<td>( l )-diversity</td>
<td>7.91</td>
</tr>
<tr>
<td>( RR-QI )</td>
<td>8.33</td>
</tr>
<tr>
<td>( l ) = 5</td>
<td>Anatomy</td>
</tr>
<tr>
<td>( l )-diversity</td>
<td>7.91</td>
</tr>
<tr>
<td>( RR-QI )</td>
<td>7.77</td>
</tr>
</tbody>
</table>

ues between each pair of attributes vary under three models (anatomy, 1-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 \( 2.74 \times 10^{-2} \). 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 1-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 \( l = 4 \). 1-diversity in general cannot preserve correlation well, because it is intractable to predict which \( QI \) attributes will be generalized in 1-diversity.

Fig. 7 shows the average value of uncertainty coefficients among attributes for anatomy, 1-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 quasi-identifiers across all privacy thresholds. Fig. 7b also clearly shows that randomization better preserves correlation among quasi-identifiers than 1-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 (1-diversity and anatomy) in terms of privacy 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 1-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 1-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

2. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubra-
34. D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, “From t-


Appendix 1. Proof of Property 1

We start with applying RR to the sensitive attribute, \( \Pr(\bar{Q}_1 = Q_1) = 1 \). Without loss of generality, we assume that \( S_1 = 1 \). Combine other categories \( 0, 2, \ldots, d_i - 1 \) into a new categories, and still use 0 to denote the new category. To make the notation simple, we simply write \( \frac{\pi_{i0}}{\pi_{i1}} \) as \( \pi_i \) and \( \frac{\pi_{i2}}{\pi_{i1}} \) as \( \pi_2 \) in this proof, then \( \pi_i = 1 - \pi_2 \). Such adjustment does not change the reconstruction probability \( \Pr(\bar{S}_1 = S_1 = 1|Q_1) \). After the adjustment, the randomization probabilities are given by

\[
\Pr(\bar{S}_1 = 1|\bar{S}_0 = 1) = p_1, \quad \Pr(\bar{S}_1 = 1|\bar{S}_0 = 0) = q_1.
\]

By definition, the posterior probabilities are given by

\[
\Pr(S_1 = 1|\bar{S}_0 = 1) = \frac{p_1 \pi_1}{p_1 \pi_1 + q_1 \pi_2}, \quad \Pr(S_1 = 1|\bar{S}_0 = 0) = \frac{(1-p_1) \pi_1}{(1-p_1) \pi_1 + (1-q_1) \pi_2}.
\]

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

\[
\Pr(S_1 = 1|Q_1) = \pi_1, \quad \Pr(\bar{S}_1 = S_1 = 1|Q_1)
\]

\[
= \pi_1 + \sum_{i \neq 1} \Pr(\bar{S}_1 = i|\bar{S}_0 = 1) \Pr(S_1 = 1|\bar{S}_0 = i)
\]

\[
= \pi_1 \left[ \frac{p_1^2}{p_1 \pi_1 + q_1 \pi_2} \right] + \frac{(1-p_1)^2}{(1-p_1) \pi_1 + (1-q_1) \pi_2}
\]

(16)

Taking the derivative with respect to \( p_1 \), we have (16) is minimized when \( p_1 = \frac{1}{d_i} \), and the minimal value is \( \pi_i^2 = \frac{\pi_{i0} \pi_{i1}}{\pi_{i0} \pi_{i1} + \pi_{i2} \pi_{i1}} \). Following similar strategies, we can prove the general case when we randomize both \( Q_1 \) and \( S_1 \).

Appendix 2. Proof of Result 2

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

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

\[
\text{Cov}(\eta) = \bar{X} \text{Cov}(\bar{\pi} - \pi) \bar{X}^T = \bar{X} \Sigma \bar{X} = \Lambda
\]

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

\[
E[\bar{\pi} - \pi^T] = E[(\bar{\pi} - \pi)(\bar{\pi} - \pi)^T] = E[\bar{X} \eta \eta^T \bar{X}^T]
\]

\[
= E(\eta \eta^T) = E(\Sigma \eta \eta^T) = \Sigma E(\eta \eta^T)
\]

\[
= \Sigma \text{Var}(\eta) + E(\eta \eta^T)
\]

\[
= \Sigma \lambda_i = \text{trace}(\Lambda) = \text{trace}(\Sigma).
\]

The last equality is due to the fact that

\[
\text{trace}(\Sigma) = \text{trace}(\bar{X} \Sigma \bar{X}^T) = \text{trace}(\bar{X} \bar{X}^T \Sigma) = \text{trace}(\Lambda)
\]

We proved the lemma.

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

\[
P_i^{-1} = \frac{1}{p_i - q_i} (I - q_i \bar{1} \bar{1}^T), \quad \left[ P_i^{-1} \right]_{F} = \frac{(d_i - 1)^3}{(d_i p_i - 1)^3} + 1.
\]

where \( 1 \) is the column vector, whose cells all equal 1.

**Proof.** We can re-write \( P_i \) as follows:

\[
P_i = (p_i - q_i) \bar{1} \bar{1}^T + q_i \bar{1} \bar{1}^T = (p_i - q_i) (I + \frac{q_i}{p_i} \bar{1} \bar{1}^T)
\]

With \( q_i = \frac{1-p_i}{d_i-1} \), \( \bar{1} \bar{1} = d_i \), and the binomial inverse theorem [45], we can immediately get \( P_i^{-1} \), and \( \left[ P_i^{-1} \right]_{F} \) can be directly calculated from \( P_i^{-1} \).

**Lemma 5.** Let \( (P_i^*)^T \), denote the i-th column (or row) of \( P_i^T \).

Then, for any \( i = 1, 2, \ldots, D \), we have \( \left[ (P_i^*)^T \right]_{F} = \frac{1}{D} \left[ P_i^T \right]_{F} \).
PROOF. With Lemma 4, we observe that every row (or column) of $P_\pi$ has the same components, except for a change of order. Since $P_i^{-1} = P_\pi^{(-1)} \otimes \cdots \otimes P_m^{(-1)}$, we can also conclude that all rows (or columns) of $P_\pi$ have the same components, except for a change of order. Therefore, for all $i$'s, $\| P_i \|_F = \frac{1}{D} \| P \|_F$.

Next, we prove the main result. With Lemma 3, when the distance is the squared Euclidean distance, to minimize $E[d(\hat{\pi}, \pi)]$ is equivalent to minimize $\text{trace}(\Sigma)$. With (2), we have

$$\text{trace}(\Sigma) = \text{trace} \left\{ \frac{1}{N} \sum_{i=1}^{D} (P_\pi)^T (P_\pi)^{-1} (P_\pi)^T - \pi \pi^T \right\}$$

$\propto \text{trace} \left\{ P_\pi^{T} (P_\pi)^{(-1)} (P_\pi)^{-1} \right\}$

$\propto \text{trace} \left\{ (P_\pi)^{T} (P_\pi)^{-1} (P_\pi)^{-1} \right\}$

$\propto \sum_{i=1}^{D} (P_\pi)_i^T (P_\pi)_i$ (with Lemma 5)

$\propto \| P \|_F = \prod_{i=1}^{D} \| P_i \|_F$.

Notice that the constraint function is an increasing function of $P_\pi$, 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.