

Deriving Private Information from Randomized Data

Zhengli Huang, Wenliang Du and Biao Chen
Department of Electrical Engineering and Computer Science
Syracuse University, Syracuse, NY 13244
Tel: 315-443-9180 Fax: 315-443-1122
{zhuang,wedu,bichen}@ecs.syr.edu

ABSTRACT

Randomization has emerged as a useful technique for data disguising in privacy-preserving data mining. Its privacy properties have been studied in a number of papers. Kargupta et al. challenged the randomization schemes, and they pointed out that randomization might not be able to preserve privacy. However, it is still unclear what factors cause such a security breach, how they affect the privacy preserving property of the randomization, and what kinds of data have higher risk of disclosing their private contents even though they are randomized.

We believe that the key factor is the correlations among attributes. We propose two data reconstruction methods that are based on data correlations. One method uses the Principal Component Analysis (PCA) technique, and the other method uses the Bayes Estimate (BE) technique. We have conducted theoretical and experimental analysis on the relationship between data correlations and the amount of private information that can be disclosed based on our proposed data reconstructions schemes. Our studies have shown that when the correlations are high, the original data can be reconstructed more accurately, i.e., more private information can be disclosed.

To improve privacy, we propose a modified randomization scheme, in which we let the correlation of random noises “similar” to the original data. Our results have shown that the reconstruction accuracy of both PCA-based and BE-based schemes become worse as the similarity increases.

Keywords

Privacy-Preserving Data Mining, Randomization, PCA, and Bayes Estimate.

1. INTRODUCTION

With the advance of the information age, data collection and data analysis have exploded both in size and complexity. The attempt to extract important patterns and trends from the vast data sets has led to a challenging field called

Data Mining. When a complete data set is available, various statistical, machine learning and modeling techniques can be applied to analyze the data. In many contexts, data are distributed across different sites. Traditionally, the data warehousing approach has been used to mine distributed databases. It requires that data from all the participating sites are collected at a centralized warehouse. However, many data owners may be reluctant to share their data with others due to privacy and confidentiality concerns. This is a serious impediment to perform mutually beneficial data mining tasks.

Privacy-Preserving Data Mining (PPDM) has emerged to address this issue. The research of PPDM is aimed at bridging the gap between collaborative data mining and data confidentiality. It involves many areas such as statistics, computer sciences, and social sciences. It is of fundamental importance to homeland security, modern science, and to our society in general.

Agrawal and Srikant first proposed using randomization to solve PPDM problems [2]. In their randomization scheme, a random number is added to the value of a sensitive attribute. For example, if x_i is the value of a sensitive attribute, $x_i + r$, rather than x_i , will appear in the database, where r is a random value drawn from some distribution. It is shown that given the distribution of random noises, recovering the distribution of the original data is possible. The randomization techniques have been used for a variety of privacy preserving data mining work [1, 21, 9, 7].

Kargupta et al. challenged the randomization schemes, and they pointed out that randomization might not be secure [16]. They proposed a random matrix-based Spectral Filtering (SF) technique to recover the original data from the perturbed data. Their results have shown that the recovered data can be reasonably close to the original data. The results indicate that for certain types of data, randomization might not preserve privacy as much as we have believed.

Motivated by Kargupta et al's work, we want to answer a series of important questions that are still unanswered: *what are the key factors that decide the accuracy of the data reconstruction? what are the conditions that make data less privacy preserving using randomization? can we improve randomization to achieve better privacy?* Being able to answer these questions is important to understand how secure the randomization schemes are: first it tells us what types of data should not use the randomization to disguise; second, this understanding gives us a clear direction on how to improve the randomization to achieve better privacy preservation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGMOD 2005 June 14-16, 2005, Baltimore, Maryland, USA
Copyright 2005 ACM 1-59593-060-4/05/06 \$5.00.

We hypothesize that the relationships among data attributes might be the key factor that decides how much privacy can be preserved. Our hypothesis is motivated by the following intuitive extreme case: Assume that there are m numbers that have exactly the same values z . If each of them is disguised by an independent uniformly-random number with mean zero, we can estimate the value z by calculating the mean of these m perturbed numbers. As we know, the mean converges to z when m becomes large. Although the above example is unrealistic, it indicates that when data are highly correlated (thus redundant), we are able to derive, from the disguised data, more accurate information about the original data. In other words, there exists a strong relationship between the correlation and the randomization’s privacy-preserving property.

The goal of this paper is to find out such a relationship, then based on which to understand how well the randomization works in terms of privacy preserving. We have developed two methods that exploit the correlations among data to reconstruct the original data from a randomized data set.

Our first scheme is based on Principal Component Analysis (PCA) method, which provides a framework for us to control the degree of redundancy, we choose to use a scheme that is directly based on PCA theory. Kargupta’s scheme is also based on PCA, but part of it is based on Matrix perturbation theory, which makes it difficult to achieve a clear understanding of the correlation-vs-privacy relationship. The choice of directly basing on PCA is not motivated by the performance (actually, both schemes have similar performance under some conditions), it is rather motivated by its simplicity and being able to give an intuitive theoretical explanation on why it works. We call both schemes the PCA-based schemes.

We have also developed a method that is more general than the PCA-base schemes. In this scheme, we formulate the data reconstruction problem as an estimation problem, i.e., given the disguised data Y , we find X , such that the posterior probability $P(X | Y)$ is maximized. This is exactly the Bayes estimate [20]. We use the Bayes estimate techniques to solve X , and then use X as the final reconstructed data. Our results show that this method can obtain more accurate results than the PCA-based schemes.

Based on our conclusion that correlation can reveal private information, we propose a modified randomization scheme, in which we let correlations of random noise “similar” to the original data. We have shown that the results of data reconstructions based on both PCA-based and BE-based schemes become worse when the correlation of noise becomes more and more “similar” to the original data.

The rest of the paper is organized as follows. We discuss the related work in Section 2. In Section 3, we summarize the factors that can affect the privacy of randomization. In Section 4, we show a univariate data reconstruction scheme that does not exploit data correlations. The result of this scheme is used as the baseline data for our comparison. In Section 5 and 6, we present our PCA-based data reconstruction scheme and BE-based data reconstruction scheme, respectively. The experiment results are presented in Section 7. Section 8 describes an improved randomization scheme and its experiment results. Finally we summarize our work in Section 9.

2. RELATED WORK

There are two general approaches to privacy preserving data mining: the randomization approach and the Secure Multi-party Computation (SMC) approach. In the randomization approach, random noises are added to the original data, and only the disguised data are shared [2, 1, 21, 9, 7]. There are two different randomization methods: the Random Perturbation scheme and the Randomized Response scheme.

Agrawal and Srikant proposed a scheme for privacy-preserving data mining using random perturbation [2]. This work has been extended by Agrawal and Aggarwal [1]. Under the scheme, Evfimievski et al. proposed an approach to conduct Privacy-Preserving Association Rule Mining [9].

The randomized response is mainly used to deal with categorical data. Rizvi and Haritsa presented a scheme called *MASK* to mine associations with secrecy constraints [21]; Du and Zhan proposed an approach to conduct Privacy-Preserving Decision Tree Building [7]. All these approaches are based on the Randomized Response technique proposed by Warner [26].

Privacy is analyzed in most of the above studies, in addition, two studies have focused on the privacy analysis. The first one is due to Evfimievski et al. [8], and the other is due to Kargupta et al. [16]. In their paper, Evfimievski et al. presented a formula of privacy breaches and a methodology to limit the breaches in the field of association rule mining.

Kargupta et al. pointed out an important issue: arbitrary randomization is not safe [16]. Inspired by their work, we study why and how correlations affect privacy. In addition to correlations, we identify other potential factors that can influence privacy.

Another approach to achieve Privacy-Preserving Data Mining is to use Secure Multi-party Computation (SMC) techniques. Briefly, an SMC problem deals with computing certain function on multiple inputs, in a distributed network where each participant holds one of the inputs; SMC ensures that no more information is revealed to a participant in the computation than what can be inferred from the participant’s input and the final output [11].

Several SMC-based privacy-preserving data mining schemes have been proposed [17, 19, 23, 24, 5]. Lindell and Pinkas used SMC to build decision trees over the horizontally partitioned data [17]. Vaidya and Clifton proposed the solutions to the clustering problem [24] and the association rule mining problem [23] for vertically partitioned data. Several SMC tools and fundamental techniques are also proposed in the literature [19, 5]. Some more schemes were presented in recent conferences as follows. Wright et al. [27] and Meng et al. [18] used SMC to solve privacy-preserving Bayesian network problems. Gilburd et al. proposed a new privacy model, k -privacy, for real-world large-scale distributed systems [10]. Sanil et al. described a privacy-preserving algorithm of computing regression coefficients [22]. Du et al. have developed building blocks to solve secure two-party Multivariate Linear Regression and Classification problems [6]. Wang et al. used an iterative bottom-up generalization to generate data, which remains useful to classification but difficult to disclose private sources [25].

3. DERIVING PRIVATE INFORMATION

Kargupta et al. used a data reconstruction approach to derive private information from a disguised data set [16]. Namely, a new data set X^* is reconstructed from the disguised data using certain algorithms, and the difference between X^* and the actual original data set X indicates how much private information can be disclosed. The further apart X^* is from X , the higher level of the privacy preservation is achieved. Therefore, the difference between X^* and X can be used as the measure to quantify how much privacy is preserved. Our work also uses data reconstruction approaches, but we propose two different data reconstruction algorithms.

A variety of information can lead to the disclosure of private information in a disguised data set. We summarize several of them in the following:

- **Attribute Dependency:** Attributes in many data sets are not independent, and some attributes might have a strong correlation among themselves. It is important to understand how such relationship can cause private information disclosure.
- **Sample Dependency:** For certain types of data sets, such as the time series data, there exists serial dependency among the samples. Even after perturbing the data with random noise, this dependency can still be recovered. For instance, various techniques are available from the signal processing literature to de-noise the contaminated signals. One interesting research problem is: for different types of data, what kind of dependency relationships will help the adversaries reconstruct the original data?
- **Partial Value Disclosure:** In practice, it is possible that the values of some attributes can be disclosed (via other channels). For example, assume we have a medical database that is disguised by randomization schemes. Knowing that the patient Alice has diabetes and heart problems, we might be able to estimate the other information about her. How to quantify privacy under these circumstances?
- **Data Mining Results:** In the SMC approach, all the participating parties can see the final results. These results contain aggregate information about the data, which can lead to possible privacy breaches. For example, in the association rule mining, assume that there is a rule saying that A implies B with 90% of support. Even if one party knows only A and the association rule results, he or she will be able to infer B with high confidence. How do various data mining results, including classification models, association rules, and clustering affect individual privacy? Kantarcioglu et al. has initiated studies on this issue [15].

The scope of the problems described above are broader than what we have covered in this paper. In this paper, we focus on the first problem, i.e., how to use data correlation information to derive private information?

4. UNIVARIATE DATA RECONSTRUCTION

In this section, we describe two data reconstruction methods derived from the existing work on randomization. The

first approach is only based on the distribution of noise. It does not consider the distribution of the original data X . The second approach bases its guess on the posterior distribution $P(X|Y)$, which can be estimated from the disguised data. Because both reconstruction methods treat each attribute independently without considering the dependency relationship among attributes, we treat X and Y as if they are one attribute.

We assume that the adversaries have the disguised data $Y = X + R$, where X is the original data, and R is the noise with a zero mean. Let X have n records or objects, which are considered as realizations of n independent identically distributed (i.i.d.) random variables or i.i.d. random vectors (when there are multiple attributes). Let R have the same size of data values as X . They are the realizations of n independent random variables or random vectors, drawn from a certain distribution.

4.1 Using Noise Distribution

This is a naive guessing method: for each disguised data item y , the adversaries always use y as its guess of the original, i.e., the adversaries always guess the value of the random noise to be zero. We call this method the *Noise Distribution-based Reconstruction (NDR)*.

Let $y_i = x_i + r_i$ for $i = 1, \dots, n$, where x_i , y_i , r_i are samples of X , Y , and R , respectively. The mean square error (m.s.e.) of the NDR scheme can be derived in the following:

$$m.s.e. = \frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2 = \frac{1}{n} \sum_{i=1}^n r_i^2 = \frac{1}{n} \sum_{i=1}^n (r_i - 0)^2$$

From the above equation, the m.s.e. of NDR is exactly the variance of the random numbers. When the random numbers have a large variance, the reconstruction accuracy of NDR is low.

4.2 Using Univariate Distribution

NDR scheme is not good for reconstructing the original data. It does not consider the distribution of X and Y , which can be helpful for data reconstruction.

In this subsection, we show how to reconstruct the original data for each attribute based on some distributions. Since we treat each attribute of the data set independently, we call this method the *Univariate Distribution-based Reconstruction (UDR)*.

Let f_X , f_Y , f_R represent the distribution of X , Y , and R , respectively. We first derive the posterior distribution $P(X|Y)$, which gives us the probability for different values of X after having observed the value of Y . Since our goal is to reconstruct the original data, we need to pick a value that can minimize the overall mean square error. Our next theorem indicates that picking the expected value of the distribution achieves the minimum mean square error:

THEOREM 4.1. *Given a distribution $f(x)$, let \bar{x} be the expected value of x . Let z be a constant. The mean square error $e = \int_{-\infty}^{\infty} (x - z)^2 f(x) dx$ is minimized when $z = \bar{x}$.*

PROOF. If we want to find what value of z makes the $e = \int_{-\infty}^{\infty} (x - z)^2 f(x) dx$ minimum, we can differentiate the equation twice on z . Then we find a value which makes the first derivative equal to zero and the second derivative larger than zero. This value indeed minimizes the value of e . The

first derivative is:

$$\frac{\partial \int_{-\infty}^{\infty} (x-z)^2 f(x) dx}{\partial z} = 0$$

then

$$\begin{aligned} \int_{-\infty}^{\infty} 2 * (x-z) f(x) dx &= 0. \\ 2 * \int_{-\infty}^{\infty} x f(x) dx - 2z * \int_{-\infty}^{\infty} f(x) dx &= 0 \\ \int_{-\infty}^{\infty} x f(x) dx - z &= 0 \\ z &= \bar{x} \end{aligned} \quad (1)$$

The second derivative is:

$$\frac{\partial^2 \int_{-\infty}^{\infty} (x-z)^2 f(x) dx}{\partial^2 z} = 2 > 0 \quad (2)$$

Therefore to minimize mean square errors, z must be the expected value of x . \square

Next we will show how to compute the posterior distribution $P(X|Y)$ and its expected value. To compute $P(X|Y)$, we need to know the distributions f_X , f_Y , and f_R . R 's distribution f_R is public. Y 's distribution f_Y can be estimated from the samples, i.e., the disguised data set. X 's distribution f_X is unknown; however, it has been shown by the studies in the privacy preserving data mining area that f_X can be estimated from the disguised data [2]. Therefore, in our next analysis, we assume all three distributions f_X , f_Y , and f_R are known. We have the following:

$$P(x|Y = y) = \frac{f(y|x)f_X(x)}{f_Y(y)} = \frac{f_R(y-x)f_X(x)}{f_Y(y)}. \quad (3)$$

Therefore the expected value of X given the disguised value $Y = y$ is the following:

$$\begin{aligned} E(x|Y = y) &= \int_{-\infty}^{\infty} x \frac{f_X(x)f_R(y-x)}{f_Y(y)} dx \\ &= \frac{\int_{-\infty}^{\infty} x f_X(x)f_R(y-x) dx}{f_Y(y)}. \end{aligned} \quad (4)$$

We thus use $E(x|Y = y)$ as our guess to reconstruct the original data.

It should be noted that UDR only considers the distribution of one dimension; it does not use any correlation between different attributes. If the attributes are highly correlated, the use of the correlations will greatly help the adversaries' estimations. In the following sections, we will study how to take advantage of the correlations among the attributes.

5. PCA-BASED DATA RECONSTRUCTION

In this section, we will present a different estimation technique which is based on PCA (principal component analysis). We called this technique PCA-Based Data Reconstruction (PCA-DR). To help readers understand PCA-DR, we briefly describe how PCA works.

5.1 Principal Components Analysis

Principal Component Analysis (PCA)[14] is a way to reduce the dimensionality of a data set with interrelated variables, but still contain as much variance of the data set as possible. If a data set has m variables, each of which has n implementations, PCA can transform the data set to a new data set with $p \leq m$ variables, which are uncorrelated and are ordered by the variances they contain. We usually say there is a strong trend along a direction if the variance in the direction is large.

Let D be a data set of n records of m variables (also called attributes). It can be viewed as a transposed vector of m variables. Let us start to search for the first principal component (PC), which presents the largest variance of the original data set in the direction of a certain vector. We look for a linear function De_1 of the variables of D which has maximum variance, where e_1 is a vector of m constants. To get the second PC, we look for a linear function De_2 , uncorrelated with De_1 , and having maximum variance. Accordingly, a linear function De_p , uncorrelated with De_1, \dots, De_{p-1} , is found which has maximum variance. The result vectors De_1, De_2, \dots, De_p are called principal components (PCs). Since the value of p is always smaller than or equal to m , PCA is used for compression. The variances in the directions of the vectors decrease from De_1 to De_p .

If some variables have significant correlations among them, the first few generated PCs will count most of the variances in D . Accordingly, the subsequently-generated PCs will count a smaller portion of the variances of D .

In order to find PCs, the covariance matrix C is computed. This is the matrix whose (i, j) -th entry is the covariance between the i th and j th variables of D (when $i = j$, it is the variance of the i th attribute of D). Then, e_k is an eigenvector of C corresponding to its k th largest eigenvalue λ_k . The k th PC is De_k , the variance of which is equal to λ_k . We briefly introduce the procedures of decreasing the number of the data dimensions and of restoring data below.

5.1.1 Decreasing the Dimensionality

Let the original data set be D as before. The mean of each attribute of the data set is 0 due to the requirement of PCA. A non-0-mean data set can always be adjusted to a 0-mean data set by subtracting the mean of each attribute from it. Then all operations can be executed on the adjusted data set. When the operations are done and restoring the data set is wanted, the mean will be added back. For the simplicity of presentation, we ignore the adjustment steps and consider all data sets we use here are 0-mean data sets.

From D , its covariance matrix C can be computed. Using C , eigenvectors $[e_1, e_2, \dots, e_m]$ can then be obtained, so is their corresponding eigenvalues: $\lambda_1, \dots, \lambda_m$, where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$.

Assume that the data have large variances along the directions of the first p eigenvectors, and small variances along the directions of the other $m - p$ eigenvectors. Let $E = [e_1, e_2, \dots, e_p]$ be a matrix of size $m * p$. The following equation describes the transformation that can reduce dimensions.

$$D_n = DE, \quad (5)$$

where D_n is a new data matrix of size $n * p$. Thus, the number of the dimensions of the data set decreases from m

to p .

5.1.2 Restoring the Original Data

Next we restore the original data from the new data. If $p = m$, E is orthogonal, meaning that its transpose is its inverse. We have:

$$D = D_n(E)^{-1} = D_n E^T$$

If E is only composed of p ($< m$) eigenvector, the above equation is only an approximation and becomes

$$\hat{D} = D_n E^T \quad (6)$$

where \hat{D} is the estimated original data set.

When we use reduced E (e.g. the first p eigenvectors), the restored data will not reflect the variances along the directions of the other $m - p$ eigenvectors. However, the restored data still contains most information of the data set because the variances in the directions of the principal eigenvectors are maintained. We call an eigenvector ‘‘principal’’ if it is used to calculate the principal components. We call an eigenvalue ‘‘principal’’ when it is corresponding to a principal eigenvector.

5.2 PCA-Based Data Reconstruction

Existing methods have not exploited the correlations when quantifying privacy. We believe that the correlations can help the adversaries make more accurate guess. For example, let A_i be an attribute in the data set. If several attributes are highly correlated with A_i , then we have redundant information about A_i . We should be able to estimate A_i with better accuracy based on this information.

When a data set has strong correlations among its data, the data set has large variances in the directions of some vectors but small variances in the other directions. The addition of noise does not change the trends too much, because if the random numbers added to the original data are independent, their variances will be evenly distributed among all the directions.

Principal Component Analysis is a technique that identifies those trends. Let us consider the information loss when we only select p principals out of the total m . There will be information loss. All the variance along the other $m - p$ directions will be lost. However, when the data are highly correlated, the variances along the first p directions are much larger than the variances along the rest $m - p$ directions. Thus removing those $m - p$ directions during the transformation does not cause much information loss.

The information loss for the random numbers is different. In randomization scheme, random numbers are independent for each attribute. Their correlations are zero. Therefore, their variance will be evenly distributed to those m directions. If we remove $m - p$ directions in the PCA transformation, we are able to remove $\frac{m-p}{m}$ portion of the random numbers’ variance. The more variance of random numbers we remove, the better.

Therefore, if the data is highly correlated, then more dimensions can be reduced without causing too much information loss for the original data; at the same time, the information loss for the noise increases. Intuitively speaking, during the PCA transformation using the first $p < m$ principals, we can filter out a portion of the random numbers.

Based on the above observation, we present a PCA-based data reconstruction scheme. Since PCA introduces infor-

mation loss, it is important to understand how much of the original information is lost and how much of the noise is lost. Using the PCA-based scheme proposed by Kargupta et al. [16], the information loss and noise loss cannot be clearly quantified, because of matrix perturbation theory also used in that scheme complicates the analysis. Therefore, to be able to understand how correlation affect the privacy of the randomization scheme, we choose to use PCA directly to reconstruct the data. Namely we reconstruct the original eigenvalues and eigenvectors, and then based on the original eigenvalues, we select the principal components. Since the original eigenvalues reflect the degree of correlations among data attributes, the number of principal components and the sum of their variances indicate how much of the original information is kept via PCA.

5.2.1 Estimating Covariance Matrix

To conduct PCA, we need to know the covariance matrix for the original data. How do we get the covariance matrix for the original data? The following theorem provides the answer.

THEOREM 5.1. *Let X_i and X_j represent two variables in the data set. $Cov(X_i + R_i, X_j + R_j)$ represents the (i, j) -th entry of the covariance matrix from the disguised data set, where R_i and R_j are random variables with zero means, and they are independent from X_i or X_j . $Cov(X_i, X_j)$ represents the (i, j) -th entry of the covariance matrix from the original data set. Let σ be the standard deviation of R_i . We have the following relationship:*

$$\begin{aligned} & Cov(X_i + R_i, X_j + R_j) \\ = & \begin{cases} Cov(X_i, X_j) + \sigma^2, & \text{for } i = j \\ Cov(X_i, X_j), & \text{for } i \neq j \end{cases} \end{aligned}$$

PROOF. Based on the definition of the covariance, we have the following equations:

$$\begin{aligned} & Cov(X_i + R_i, X_j + R_j) \\ = & E((X_i + R_i)(X_j + R_j)) - E(X_i + R_i)E(X_j + R_j) \\ = & E(X_i X_j) + E(R_i)E(X_j) + E(X_i)E(R_j) \\ & + E(R_i R_j) - E(X_i)E(X_j) \\ = & E(X_i X_j) + 0 + 0 + E(R_i R_j) - E(X_i)E(X_j) \\ = & Cov(X_i, X_j) + E(R_i R_j). \end{aligned}$$

When $i \neq j$, R_i and R_j are independent, so $E(R_i R_j) = E(R_i)E(R_j) = 0$; when $i = j$, $E(R_i R_i) = E((R_i - 0)^2) = \sigma^2$, where σ is the standard deviation of the random variables R_i . Combining this with the above equation, we have proved the theorem. \square

When the sample size is large enough, the relationship presented in the above theorem carries over to the sample covariance matrices (PCA is applied to sample covariance matrices). The above theorem indicates that we can derive the covariance matrix of the original data based on the disguised data. All we need to do is to subtract σ^2 from the diagonal elements of the covariance matrix that is derived from the disguised data. Although the derived matrix is only an approximation, when the number of samples becomes larger, the approximation becomes more accurate.

5.2.2 Applying PCA

After getting the approximated sample covariance matrix for the real data, we can use the following PCA-based method to reconstruct the original data (recall that Y represents the disguised data. We use C to represent the sample covariance matrix derived from Y):

1. Conduct PCA to get $C = Q\Lambda Q^T$, where Λ is a diagonal matrix consisting of eigenvalues, and Q a matrix formed by eigenvectors.
2. Let p be the number of principal components to be selected. Set \hat{Q} to be the first p columns of Q .¹
3. Reconstruct the data using $\hat{X} = Y\hat{Q}\hat{Q}^T$.

Let m be the number of attributes in the data set. If $p = m$ (i.e., we do not reduce the dimension), the above reconstruction procedure gets back to Y , and nothing is filtered out. When $p < m$, \hat{X} will lose information. If we lose the same amount of information on both X and R , then such a transformation is not helpful. However, as we have discussed earlier, when the data are highly correlated, we do not lose much information on X ; more importantly, we lose information about R , and the amount of information loss with regard to R should, intuitively speaking, be proportional to the ratio of $\frac{p}{m}$. In the next sub-section, we will formally quantify the information loss on R .

5.3 Analysis

For the sake of simplicity, we only analyze PCA-DR using covariance matrix from the original data. That is, the covariance matrix is directly obtained from the original data, rather than being estimated from the disguised data. There are only minor differences between the covariance matrices from original data and the estimated one. From the last step of PCA procedure described earlier, we have

$$\hat{X} = Y\hat{Q}\hat{Q}^T,$$

so we get

$$\begin{aligned}\hat{X} &= (X + R)\hat{Q}\hat{Q}^T \\ &= X\hat{Q}\hat{Q}^T + R\hat{Q}\hat{Q}^T.\end{aligned}\quad (7)$$

The error between \hat{X} and the original data X comes from two sources: one is the error caused by $X\hat{Q}\hat{Q}^T$, the other is the error caused by $R\hat{Q}\hat{Q}^T$. The former is decided by the correlations among data and the number of principals included in \hat{Q} . The latter can be quantified by the following theorem:

THEOREM 5.2. *Let m be the number of the attributes of the original data set, p be the number of principal components being used in PCA-DR. The variance of the random noise is σ^2 , and its mean is 0. Let δ^2 be the mean square error of PCA-DR caused by $R\hat{Q}\hat{Q}^T$. We have the following relationship:*

¹There are a number of ways to select principal components. We can fix the number of selected principal components; we can also fix the portion of the original information that we want to keep; we can also choose the dominant eigenvalues by finding the largest gap between the dominant eigenvalues and the non-dominant ones. The last method is used in our experiments.

$$\delta^2 = \sigma^2 \frac{p}{m}. \quad (8)$$

PROOF. See Appendix A. \square

Due to Theorem (5.2), the mean square error that is caused by R in PCA-DR scheme is proportional to the variance of the random numbers and the ratio of $\frac{p}{m}$. This confirms our intuitive explanation. We will also use experiments to verify these relationships.

6. BAYES-ESTIMATE-BASED DATA RECONSTRUCTION

The PCA-based reconstruction works well when the data are highly correlated. However, when the correlations of data are not high enough, the non-principal components will not be many; according to Theorem (5.2), we will not be able to filter out significant amount of the noise. Our results in Section 7 will show that when the correlations become low, the accuracy achieved by the PCA-based scheme is even worse than the univariate data reconstruction scheme.

In this section, we describe a more accurate data reconstruction method. We want to fully take advantage of the relationship among data. Unlike the PCA-based schemes, which use this relationship to filter out noise, We formulate the data reconstruction problem as an estimation problem, i.e., based on the disguised data that we have observed and on the data relationship that we know, we use a value that can most likely produce such an observation as our reconstructed data. In other words, given the disguised data Y , we search for X , such that the posterior distribution $P(X | Y)$ is maximized. We then use X as the data reconstructed from the disguised data. This is the idea of the Bayes estimate [20]. We call our scheme the *Bayes-Estimate-based Data Reconstruction (BE-DR)*.

To simplify derivation, we assume that the original data have multivariate normal distribution. This assumption is reasonable since this distribution is a good approximate distribution in many situations [13]. Because of the appealing properties of multivariate normal distribution, the calculation of Bayes estimate is simplified to computation of matrices and vectors even though the form of the distribution is more complicated. As we will explain later, this assumption can be relaxed.

6.1 Data Reconstruction

Suppose that random noise used for each attribute has normal distribution and it is independent from those used for other attributes. Suppose that an adversary only has the disguised data set and the distribution of the random noise. Let the original data have m attributes and n records. They can be considered as observations of a random vector of length m . Let the random vector of the original data be \vec{X} . Similarly let the random vector of the noise be \vec{R} , and the random vector of the disguised data \vec{Y} . Let $f_{\vec{X}}$, $f_{\vec{Y}}$, $f_{\vec{R}}$ represent the distributions of \vec{X} , \vec{Y} , \vec{R} , respectively. The distribution of \vec{X} and \vec{R} are described in the following:

$$\begin{aligned}f_{\vec{X}}(\vec{x}) &= \frac{1}{(2\pi)^{m/2} |\Sigma_x|^{1/2}} e^{-\frac{1}{2}(\vec{x} - \mu_x)^T \Sigma_x^{-1} (\vec{x} - \mu_x)} \\ f_{\vec{R}}(\vec{r}) &= \frac{1}{(2\pi\sigma)^{m/2}} e^{-\frac{1}{2}(\vec{r} - \mu_r)^T (\vec{r} - \mu_r) / \sigma^2},\end{aligned}$$

where Σ_x is the covariance matrix of the original data, σ^2 is the variance of the random noise, $\vec{\mu}_x$, $\vec{\mu}_r$ are the mean vectors of the original data and the noise data.

We use the posterior distribution $P(\vec{X}|\vec{Y})$ to represent the probabilities for different values of \vec{X} given an observation of \vec{Y} . By using Bayesian rule, we get the following formulae:

$$P_{\vec{X},\vec{Y}}(\vec{x}|\vec{y}) = \frac{f_{\vec{X}}(\vec{x})f_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x})}{f_{\vec{Y}}(\vec{y})}, \quad (9)$$

where $f_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x})$ represents the probability of getting \vec{y} from \vec{x} , which is exactly the probability of the random number $(\vec{y} - \vec{x})$. Therefore, $f_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x}) = f_{\vec{R}}(\vec{y} - \vec{x})$.

We want to find a value of \vec{X} , such that $P_{\vec{X},\vec{Y}}(\vec{x}|\vec{Y} = \vec{y})$ is maximized. Noticing that the denominator $f_{\vec{Y}}(\vec{y})$ does not change when \vec{X} changes, we only need to consider the numerator. Thus we only need to maximize:

$$\begin{aligned} f_{\vec{X}}(\vec{x})f_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x}) &= f_{\vec{X}}(\vec{x})f_{\vec{R}}(\vec{y} - \vec{x}) \\ &= \frac{1}{(2\pi)^{m/2}|\Sigma_x|^{1/2}} e^{-\frac{1}{2}(\vec{x} - \vec{\mu}_x)^T \Sigma_x^{-1} (\vec{x} - \vec{\mu}_x)} \cdot \\ &\quad \frac{1}{(2\pi\sigma)^{m/2}} e^{-\frac{1}{2}(\vec{y} - \vec{x} - \vec{\mu}_r)^T (\vec{y} - \vec{x} - \vec{\mu}_r) / \sigma^2} \end{aligned} \quad (10)$$

Since the logarithm is a monotone one-to-one function, we could maximize the following function instead:

$$-\frac{1}{2}(\vec{x} - \vec{\mu}_x)^T \Sigma_x^{-1} (\vec{x} - \vec{\mu}_x) - \frac{1}{2}(\vec{y} - \vec{x})^T (\vec{y} - \vec{x}) / \sigma^2.$$

Note that in the above equation the constant terms are ignored because it does not affect computing the maximum estimator of \vec{x} ; the mean vector $\vec{\mu}_r$ is ignored too since it is assumed to be zero vector in randomization schemes.

We let the first derivative of the above equation with respect to \vec{x} be 0. We get:

$$\Sigma_x^{-1}(\vec{x} - \vec{\mu}_x) + (\vec{x} - \vec{y}) / \sigma^2 = 0.$$

After simplifying and rearranging the above equation, we have

$$\hat{\vec{x}} = (\Sigma_x^{-1} + 1/\sigma^2 \cdot I)^{-1} (\Sigma_x^{-1} \vec{\mu}_x + \vec{y} / \sigma^2), \quad (11)$$

where I is the identity matrix of the same size as Σ_x .

Equipped with Equation (11), we now describe our Bayes-Estimate-based data reconstruction scheme:

1. Derive Σ_x from Theorem (5.1) using disguised data Y .
2. Derive $\vec{\mu}_x$ by computing the mean vector of the disguised data. We know that $\vec{\mu}_x \approx \vec{\mu}_y$ because random noises have zero means.
3. For each \vec{y} , derive $\hat{\vec{x}}$ using Equation (11).
4. Use $\hat{\vec{x}}$ as the reconstructed value.

The BE-based scheme, the PCA-based scheme, and the univariate data reconstruction scheme have the following relationship: when the correlations among data are low, e.g., data are independent, the results of BE-DR should converge to the univariate data reconstruction. This is because when data are independent, data from one attribute cannot help the reconstruction of another attribute. Thus, the BE-DR

scheme is equivalent to the univariate data reconstruction. On the other hand, when the correlations among data become high, the results of BE-DR should be similar to those of PCA-DR, because they both fully exploit the correlation-ship among data.

Regarding our assumption on the multivariate normal distribution: although we have only shown the results for data sets that satisfy multi-normal distribution, the approach can be extended to data sets that satisfy other distribution. However, for other distributions, we might not be able to derive an equation with a simple analytic form for its first derivative. In such situations, the Bayes estimate must be sought using numerical methods, such as Gradient descent methods [12, 3]. We will study them in our future work.

7. EXPERIMENT

7.1 Methodology

We have designed a series of experiments to evaluate the PCA-DR scheme and the BE-DR scheme. Our goal is to find out how the correlations among the attributes affect the accuracy of these methods. Data correlations can be affected by a number of parameters, including the ratio of the number of principal components to the number of attributes and the variance of data on the principal and non-principal components. We have designed experiments to study how these parameters affect our schemes. We also compare our results with SF algorithm [16].

We decided to use synthetic data for our experiments, because it is difficult to find real data sets that bear properties pre-determined for each experiment. Our approach is to generate a covariance matrix first, then generate the synthetic data set based on the covariance matrix, and finally conduct the PCA-DR or BE-DR procedure. However, generating a covariance matrix with pre-determined properties is not a trivial task either. To better control the properties of the matrix, we generated the covariance matrix in a reverse way: we generated the eigenvalues and eigenvectors first, and then we computed the covariance matrix from the eigenvalues and eigenvectors. We can control the properties of covariance matrix by changing eigenvalues. Our procedure is described in the following:

1. We specify Λ as a diagonal matrix with the corresponding eigenvalues on its diagonal. The size of Λ is m by m .
2. By using Gram-Schmidt orthonormalization process [4], we generate an orthogonal matrix Q of size m by m . We consider each column of Q as an eigenvector.
3. We let the covariance matrix $C = Q \times \Lambda \times Q^T$.
4. We generate a data set based on the covariance matrix. In our experiments, we use `mvrnd` from Matlab to generate data from C . The function `mvrnd` generates a data set of multivariate normal distribution based on the provided covariance matrix and the mean vector. This resultant data set will be used as the original data set X .
5. We randomly generate a noise data set using normal distributions. This noise data set is then added to the original data set to form the disguised data set Y .

