

Privacy Vulnerabilities with Background Information in Data Perturbation*

Lian Liu, Jie Wang, and Jun Zhang[†]

Laboratory for High Performance Scientific Computing and Computer Simulation,
Department of Computer Science, University of Kentucky,
Lexington, KY 40506–0046, USA

June 23, 2008

Abstract

The issue of data privacy is considered a significant hindrance to the development and industrial applications of database publishing and data mining algorithms. Among many privacy-preserving methodologies, data perturbation is a popular technique for achieving the balance between data utilities and information privacy and security. It is known that the attacker’s background or reference information about the original data can play a significant role in breaching data privacy. In this paper, we study the situation in which data privacy may be compromised with the leakage of a few original data records. In detail, we consider one situation in which the data owner publishes a perturbed database and the attacker knows exactly one or a few records of the original data. We find out that the remaining original data may be breached by a combination of the attacker’s reference information and the perturbed data. We consider a potential privacy vulnerability with reference information in privacy-preserving database publishing and data mining based on the eigenspace of the perturbed data under some constraints. We then show that a general data perturbation model is vulnerable from this type of reference privacy breach.

1 Introduction

Database publishing and data mining techniques enable the discovery of valuable data patterns and knowledge in collected and shared data and increase business profitability

*Technical Report CMIDA-HiPSCCS 005-08, Department of Computer Science, University of Kentucky, KY, 2008. The research work of J. Zhang was supported in part by NSF under grants CCF-0527967 and CCF 0727600, in part by NIH under grant 1R01HL086644-01, in part by Alzheimer’s Association under grant NIGR-06-25460, and in part by KSEF under grant KSEF-148-502-06-186.

[†]Corresponding author. E-mail: jzhang@cs.uky.edu. URL: <http://www.cs.uky.edu/~jzhang>.

and enhance national security. The precondition of useful data analysis is the collection of large amounts of data, which has been made possible by the recent availability of relatively inexpensive means of electronic data collections. On the other hand, we also face the challenge of controlling the level of knowledge disclosure and securing certain confidential patterns within the data, without (noticeably) affecting the utilities of the data for intended purposes of data analysis. The difficulty of data security increases considerably if we aim to achieve the goal of maintaining confidential data privacy and data utilities at the same time, in privacy-preserving database publishing and data mining.

Data privacy and security can be compromised from many different ways, both inside and outside the data collection organizations. Even within the data collection organizations, different people are assigned different levels of trustworthiness, usually through the privileges of the computer accounts they use. To protect data privacy and security from being compromised intendedly or unintendedly, it is preferable that data is preprocessed appropriately before it is distributed for analysis or made to the public. One of the most useful data preprocessing techniques is data perturbation, which attempts to hide the true values of the original data in an effort to preserve data privacy.

In this paper, we theoretically analyze data privacy in a reference background situation and develop strategies to extract original information from the perturbed data. A reference record is a piece (or pieces) of original data record(s) exactly known by an attacker. Such additional information can be used by the attacker to compromise other records in the original data, with the availability of the public perturbed data.

Let us start with a fictitious situation. An organization collects many records from thousands of hundreds of persons including Bob, and compiles such records into a well-defined dataset as the original matrix A and distorts the original dataset to a perturbed dataset as a matrix \tilde{A} and finally publishes this perturbed dataset \tilde{A} to the public. As for Bob, he knows the exact original values of his record in A and the corresponding perturbed values of his record in the perturbed dataset \tilde{A} . We consider the theoretical possibility that Bob may use his original data values and the perturbed dataset to breach the privacy of other records in the original dataset.

The remaining parts of this paper are arranged as follows. A brief account of the related works and some popular data perturbation techniques is presented in Section 2. Our theoretical analysis based on a general data perturbation model is contained in Section 3 and experimental results are discussed in Section 4, respectively. Finally a brief conclusion is given in Section 5.

2 Related Works

In privacy preserving data mining and database security, many researchers attempt to develop techniques to maintain data utilities without disclosing the original data and to produce data analysis results that are as close to those based on the original data as possible. Among those techniques, there are two main categories. Methods in the first category modify data mining algorithms so that they allow data mining operations on distributed datasets without knowing the exact values of the data or without directly accessing the original datasets. In this paper, we will not deal with privacy-preserving data mining methods in this category. Interested readers may consult [8] for more information. Methods in the other category perturb the values of the datasets to protect privacy of the data attributes. These methods pay more attention to perturbing the whole dataset or the confidential parts of the dataset using distributions of certain noises [6, 7, 9, 13, 14, 17].

Many techniques in the second category are easy to implement and practically useful. For instance, Tendick [19] perturbed each attribute in the dataset independently of the other attributes by the addition of a multivariate normal distribution e with the mean 0 in the form of $\tilde{A} = A + e$.

Bapna et al. [3] and Xu et al. [20] used wavelet and Fourier transformations to decompose the original matrix A and then used the transformed matrix as a perturbed matrix \tilde{A} , respectively. In essence, in both Fourier and wavelet decompositions, the original data matrix is multiplied by an orthonormal matrix to generate the perturbed matrix.

Chen et al. [6, 7] used a complicated rotation technique to perturb the original dataset as: $\tilde{A} = RA + \Psi + \Delta$, where R is an orthogonal matrix, Ψ is a random translation matrix, and Δ is a Gaussian noise matrix $N(0, \beta^2)$. Each vector of the matrix $N(0, \beta^2)$ can be defined by two parameters, the mean 0 and the variance (standard deviation squared) β^2 .

In recent years, however, it is noticed that the perturbed or distorted datasets from certain data perturbation techniques may not be safe if an attacker has some background information about the original datasets [11, 12, 14, 16, 18]. Furthermore, Kargupta et al. [15] showed that it is highly possible to differentiate the original true values from the additive randomization noise perturbed datasets. Guo and Wu [12, 11] calculated a useful upper bound and lower bound about the difference between the original dataset and the estimated dataset which is computed from the perturbed dataset by spectral filtering techniques.

Their works have mentioned the use of background information probably possessed by the attacker, but they just obtained their analytical results exclusively from the perturbed public dataset. In fact, it is very unlikely that an attacker has no idea about the perturbed dataset other than the public version. The common sense, statistical measure, reference, and even a small amount of leakage may dramatically help the attacker weaken the privacy

of the dataset. In the next section, we will focus our attention on privacy of the perturbed dataset with the background of the reference records.

3 Reference Information Analysis

We begin by giving some useful mathematical preparations and then generalize our notations in this paper.

3.1 Singular Value Decomposition

First, we introduce a very useful tool in our analysis, Singular Value Decomposition (SVD), which is a popular matrix factorization method in matrix computation and is widely used in data mining and information retrieval. It has been used to reduce the dimensionality of databases in practice and remove the noise in noisy databases [4]. The use of SVD techniques in data perturbations for privacy-preserving data mining is proposed in [21, 22].

The SVD of the original $n * m$ data matrix A is written as

$$A = USV^T. \tag{1}$$

Here U is an $n*n$ orthonormal matrix, $S = \text{diag}[\sigma_1, \dots, \sigma_s]$ ($s = \min(n, m)$, without the loss of generality, we assume $n \geq m$) is an $n*m$ diagonal matrix whose nonnegative diagonal entries are in a non-increasing order. We call the diagonal entries $\sigma_1, \dots, \sigma_s$ the singular values. And V^T is also an orthonormal matrix with dimension $m * m$. The number of nonzero diagonal entries of S is equal to the rank of the matrix A .

Define

$$A_k = U_k S_k V_k^T, \text{ for a positive integer } k \leq \min(n, m),$$

where U_k only contains the first k columns of U , S_k contains the first k nonzero singular values, and V_k^T contains the first k rows of V^T . Obviously, the rank of the matrix A_k is k , and A_k is often called the truncated SVD. A_k has a well-known property that it is the best k -dimensional (rank- k) approximation of A in terms of the Frobenius norm.

In information retrieval, $E_k = A - A_k$ can be considered as the noise of the original data matrix. In privacy-preserving data mining, A_k can be used as a perturbed version of A [21, 22]. So, A_k represents a good approximation which keeps similar patterns of A , while provides protection for data privacy [21, 22].

3.2 Preliminaries and Notations

To generalize the perturbation techniques to include those discussed previously as well as many other methods which can be obtained from a general model to perform the perturbation process on the original datasets, we define a theoretical general data perturbation

model as the follows:

$$\tilde{A} = RA + E, \quad (2)$$

where R is an orthogonal matrix and E is a Gaussian noise matrix with the mean 0 and an arbitrary variance.

For simplicity, in the following discussion, we use Matlab notations to represent matrix columns, rows, and entries, respectively.

A	the original matrix
\tilde{A}	the perturbed matrix
$A^{i,j}$	the entry (i,j) of A
$A^{i,:}$	the i -th row of A , simplified as A^i
$A^{:,j}$	the j -th column of A
$A^{i_1:i_2,j_1:j_2}$	the submatrix of A from the i_1 -th row to the i_2 -th row and from the j_1 -th column to the j_2 -th column
a^i	$A^i * V_k$
\tilde{A}_k^i	the i -th row of \tilde{A}_k as in Theorem 5

In the following contents, $\|\cdot\|$ is referred to as the 2-norm (Euclidean norm) unless otherwise explicitly stated.

3.3 Stability of Perturbation Angle

The major work of this paper shows that the attacker has a high possibility to figure out the other perturbed matrix records based on one reference record which is an original matrix record exactly known by this attacker. From the mathematical viewpoint, the perturbation model in Equation (2) preserves not only the angles between the entire rows or columns of the original dataset and those of the perturbed dataset, but also the angles between the sub-rows or sub-columns before and after the perturbation.

Lemma 1. [10] *If H is an orthogonal matrix of appropriate dimension, then for any matrix A ,*

$$\|AH\| = \|HA\| = \|A\|.$$

The following lemma is easy to prove.

Lemma 2. *If the singular value decomposition of the matrix A is*

$$A = USV^T,$$

then the following equations hold:

1. $AV=US$,
2. $\|A^i\|=\|U^i S\|$,
3. $\|A^i V_k\|=\|U^i S_k\|$, here S_k is an $n * m$ diagonal matrix which only contains the first k singular values of S .

Theorem 3. Let A and $\tilde{A} = A + E$ be $n * m$ real matrices, and

$$A = USV^T, \quad \tilde{A} = \tilde{U}\tilde{S}\tilde{V}^T$$

are the SVDs of A and \tilde{A} , respectively.

$$A_k = U_k S_k V_k^T, \quad \tilde{A}_k = \tilde{U}_k \tilde{S}_k \tilde{V}_k^T$$

are the rank- k best approximations to A and \tilde{A} , respectively.

Assume that $\|E\| \leq (\tilde{\sigma}_k - \sigma_{k+1})$ and $\|E\| \leq \sigma_k$, σ_k and $\tilde{\sigma}_k$ are the k -th singular values of A and \tilde{A} , respectively. We define

$$a^i = A^i V_k, \quad \tilde{a}^i = \tilde{A}^i \tilde{V}_k, \quad \text{and} \quad e^i = E^i \tilde{V}_k.$$

Then

$$\|a^i\| \approx \|A^i\| \quad \text{and} \quad \|e^i\| \leq \|a^i\|.$$

Proof.

$$\begin{aligned} \|a^i\| &= \|A^i * V_k\| \\ &= \|U^i * S_k\| \quad (\text{Lemma 2.3}) \\ &\approx \|U^i * S\| \quad (\text{Because } \sigma_{k+1} \text{ is small relative to } \sum_{i=1}^k \sigma_i) \\ &= \|A^i\|. \quad (\text{Lemma 2.2}) \end{aligned}$$

We define SVD of E as $E = U_E S_E V_E^T$, and $S_E = \text{diag}(\sigma_E^1, \sigma_E^2, \dots, \sigma_E^m)$. Since $\|E\| = \sigma_E^1 \leq \sigma_k$, (σ_k is the k -th singular value of A), we have

$$\begin{aligned} \|E^i\|^2 &= \|U_E^i S_E\|^2 \quad (\text{Lemma 2.2}) \\ &= \sum_{j=1}^m (U_E^{ij})^2 * (\sigma_E^j)^2 \\ &\leq \sum_{j=1}^m (U_E^{ij})^2 * (\sigma_E^1)^2 \\ &= (\sigma_E^1)^2, \quad (\because U_E^i \text{ is an unitary vector}) \\ &\leq (\sigma_k)^2, \quad (\because \|E\| = \sigma_E^1 \leq \sigma_k) \\ &= \sum_{j=1}^m (U_E^{ij})^2 * (\sigma_k)^2 \end{aligned}$$

$$\begin{aligned}
&\leq \sum_{j=1}^m (U^{ij})^2 * (\sigma_j)^2, (\because \sigma_k \leq \sigma_1, \dots, \sigma_{k-1}) \\
&= \|U^i S V^T\|^2 \\
&= \|A^i\|^2.
\end{aligned} \tag{3}$$

Then from Inequality (3), it is easy to obtain $\|e^i\| \leq \|a^i\|$. \square

Corollary 4. *Let A and $\tilde{A} = A + E$ be $n * m$ real matrices. Assume that $\|E\| \leq (\tilde{\sigma}_k - \sigma_{k+1})$ and $\|E\| \leq \sigma_k$. We define*

$$a^{i,j_1:j_2} = A^{i,j_1:j_2} V_k^{j_1:j_2}, \quad \tilde{a}^{i,j_1:j_2} = \tilde{A}^{i,j_1:j_2} \tilde{V}_k^{j_1:j_2}, \quad \text{and} \quad e^{i,j_1:j_2} = E^{i,j_1:j_2} \tilde{V}_k^{j_1:j_2}.$$

Then

$$\|a^{i,j_1:j_2}\| \approx \|A^{i,j_1:j_2}\| \quad \text{and} \quad \|e^{i,j_1:j_2}\| \leq \|a^{i,j_1:j_2}\|.$$

Corollary 4 is a sub-row version of Theorem 3. The mathematical proof is very similar to the proof of Theorem 3. For example, we just need to replace a^i , \tilde{a}^i , and e^i with $a^{i,j}$, $\tilde{a}^{i,j}$ and $e^{i,j}$ (j is in $j_1:j_2$), respectively.

Theorem 5. [2] *Let A and $\tilde{A} = A + E$ be $n * m$ real matrices and R is an orthogonal matrix. Assume that $\|E\| \leq (\tilde{\sigma}_k - \sigma_{k+1})$ and $\|E\| \leq \sigma_k$. a^i , \tilde{a}^i and e^i are defined as in Theorem 3. Then*

$$\|a^i - R\tilde{a}^i\| \leq \|a^i\| \quad \text{and} \quad \|A_k^i - \tilde{A}_k^i\| \leq \|A_k^i\|.$$

Theorem 5 is a simplified version of Theorem 2 in [2] which needs to verify the conditions $\|a^i\| \approx \|A^i\|$ and $\|e^i\| \leq \|a^i\|$. If the conditions are met, Theorem 2 in [2] is true. However, in practice, A^i and E^i are unknown, it is not practical for this verification. We simplify these conditions to verify if $\|E\| \leq \sigma_k$ and $\|E\| \leq (\tilde{\sigma}_k - \sigma_{k+1})$ instead of $\|a^i\| \approx \|A^i\|$ and $\|e^i\| \leq \|a^i\|$. If $\|E\| \leq \sigma_k$ and $\|E\| \leq (\tilde{\sigma}_k - \sigma_{k+1})$, Theorem 2 in [2] as well as the simplified version, Theorem 5, are always true. Later, we will discuss how to verify $\|E\| \leq \sigma_k$ and $\|E\| \leq (\tilde{\sigma}_k - \sigma_{k+1})$ in practice. The proof details of Theorem 5 can be found in [2].

Theorem 5 establishes a link between the original data A_k^i and the perturbed data \tilde{A}_k^i and bounds the difference between them.

Based on Theorem 5, the following corollary is straightforward.

Corollary 6. *Let A and $\tilde{A} = A + E$ be $n * m$ real matrices. Assume that $\|E\| \leq (\tilde{\sigma}_k - \sigma_{k+1})$ and $\|E\| \leq \sigma_k$. $a^{i,j_1:j_2}$, $\tilde{a}^{i,j_1:j_2}$, and $e^{i,j_1:j_2}$ are defined as in Corollary 4. Then*

$$\|a^{i,j_1:j_2} - R\tilde{a}^{i,j_1:j_2}\| \leq \|a^{i,j_1:j_2}\| \quad \text{and} \quad \|A_k^{i,j_1:j_2} - \tilde{A}_k^{i,j_1:j_2}\| \leq \|A_k^{i,j_1:j_2}\|.$$

Based on Theorem 5 and Corollary 6, we can establish a connection between an original data pair and a perturbed data pair, as in the next corollary.

Corollary 7. *If a^p , a^q , $a^{p,j_1:j_2}$, and $a^{q,j_1:j_2}$ satisfy the conditions in Theorem 3 and Corollary 4, then*

1. $[2] \|\angle(A_k^p, A_k^q) - \angle(\tilde{A}_k^p, \tilde{A}_k^q)\| \leq \epsilon$,
2. $\|\angle(A_k^{p,j_1:j_2}, A_k^{q,j_1:j_2}) - \angle(\tilde{A}_k^{p,j_1:j_2}, \tilde{A}_k^{q,j_1:j_2})\| \leq \epsilon$.

Here, ϵ is a small positive number, and $\angle(A_k^p, A_k^q)$ denotes the angle between A_k^p and A_k^q .

Remark 8.

1. Due to the disclosure of the perturbed dataset, all \tilde{A} related information, such as \tilde{A}_k^p , \tilde{A}_k^q , $\tilde{A}_k^{p,j_1:j_2}$ and $\tilde{A}_k^{q,j_1:j_2}$ in Corollary 7, are known. In a reference information case, one or more original records are assumed to be known as the reference information. We assume that the attacker, Bob, knows the exact original value of A^p . Therefore, in Corollary 7, A_k^p , $A_k^{p,j_1:j_2}$, \tilde{A}_k^p , \tilde{A}_k^q , $\tilde{A}_k^{p,j_1:j_2}$, and $\tilde{A}_k^{q,j_1:j_2}$ are all known. Only A_k^q and $A_k^{q,j_1:j_2}$ are unknown which are the attacker's breach target.

2. When the perturbed dataset satisfies the conditions $\|E\| \leq (\tilde{\sigma}_k - \sigma_{k+1})$ and $\|E\| \leq \sigma_k$, it is highly possible for Bob to work out the unknown original record A_k^q through Corollary 7, if A_k^q is either entire row highly similar or sub-row highly similar to the A_k^p record.

3. In practice, the attacker, Bob, only needs to calculate the angle between \tilde{A}_k^q and any \tilde{A}_k^p or the sub rows. If the angle is very close to $\pi/2$ (the cosine value is very close to 1), then the corresponding entire rows or sub-rows of the original records A^p and A^q are very similar. In such a case, the attacker can see that A^q is entire row or sub-row close to A^p , and can directly figure out A^q based on the known A^p .

So, practically, if we can figure out $\|E\|$, $\tilde{\sigma}_k$, and σ_k , we may establish the connection between the original data pairs and the perturbed data pairs in Corollary 7. The only remaining problem is how to verify whether the perturbed dataset satisfies the conditions $\|E\| \leq (\tilde{\sigma}_k - \sigma_{k+1})$ and $\|E\| \leq \sigma_k$. If the verification is positive, the practical strategy in Remark 8 can be used to breach the data privacy in a reference case. We can use the following Proposition 9 and Theorem 10 to approximate $\|E\|$, $\tilde{\sigma}_k$ and σ_k .

Proposition 9. *[10] For any matrix A , if R is an orthogonal matrix, then $\tilde{A} = RA$ does not change the angle between any rows or any columns of A .*

The proof is very simple and can be found in any book on matrix algorithm. From this proposition, we know that the multiplication of the original matrix and an orthogonal matrix does not change the angle distribution of the original matrix.

Theorem 10. [14] Let A and $\tilde{A} = A + E$ be $n * m$ real matrices. If $n/m \rightarrow \infty$, A and E are uncorrelated, and the norm of the matrix E is small relative to the norm of A , then

$$\tilde{S} \approx S + S_E.$$

Here S and S_E are defined in Theorem 3 and in the corresponding proof.

In [5], it is stated that the norm of a random matrix whose entries are independent random variables with the mean zero is almost close to $\sqrt{m+n}$. Therefore, we can use $\sqrt{m+n}$ as an approximation to the norm of E if E is a Gaussian noise matrix with the mean 0. If the approximated norm of E , i.e., $\sqrt{m+n}$, is small relative to $\tilde{\sigma}_{k+1}$ of the perturbed matrix for a certain k , (all $\tilde{\sigma}_i, 1 \leq i \leq k$, are known), we can consider that $\tilde{\sigma}_1, \dots, \tilde{\sigma}_{k+1}$ are very close to $\sigma_1, \dots, \sigma_{k+1}$, per Theorem 10. Therefore, we can use $\tilde{\sigma}_k, \tilde{\sigma}_{k+1}$ and $\sqrt{n+m}$ to approximately verify the satisfaction of conditions $\|E\| \leq (\tilde{\sigma}_k - \sigma_{k+1})$ and $\|E\| \leq \sigma_k$.

4 Experimental Results

In the experiment section, we choose two real databases obtained from the University of California, Irvine (UCI), Machine Learning Repository [1].

The first one is Bupa Liver-disorders Research Database donated by Richard S. Forsyth. It has 5 numerical-valued attributes in 345 instances (male patients) which are all blood test results and are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. In addition to the first 5 numerical values, there are 2 attributes: drinks and selectors. The former represents the number of half-pint equivalents of alcoholic beverages drunk per day, while the latter denotes the field used to split data into two sets. So in our experiment, the Bupa dataset is a 345*7 numerical matrix whose first 5 columns are numerical values and the last 2 columns are categorical numbers (drinks and selectors).

The second dataset is Wine Recognition Database Donated by Stefan Aeberhard whose purpose is to use chemical analysis to determine the origin of wines. The dimension of this matrix is 178*14, representing 13 constituents found in each of the three types of wines and a wine category.

The purpose of our experiments is to use only the public perturbed dataset to check the satisfaction of the conditions $\|E\| \leq (\tilde{\sigma}_k - \sigma_{k+1})$ and $\|E\| \leq \sigma_k$, then further examine the preservation property of the angles between records before and after the general perturbation model in Equation (2).

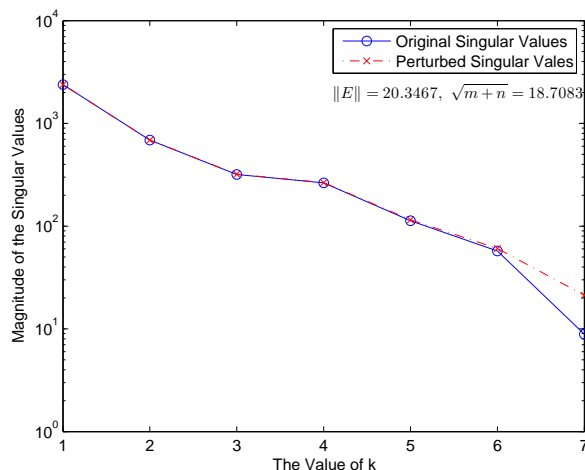
The following results of our experiments are averaged values of five repeated experiments, obtained from a Dell desktop workstation with a P4-2.8GHz CPU, 40G harddisk, and 256MB memory in Matlab 6.5.0.180913a with a Linux operating system.

4.1 Approximation of $\|E\|$, σ_k and $\tilde{\sigma}_k$

According to Lemma 1 and Theorem 10, the characteristics of singular value (eigenvalue) distribution of the data perturbation model in Equation (2) are as follows:

1. An orthogonal matrix R will not change the original singular values (eigenvalues).
2. A Gaussian noise matrix E will perturb the original singular values (eigenvalues) at most $\sqrt{m+n}$ which is an approximation of $\|E\|$.
3. The singular values (eigenvalues) of the perturbed matrix are approximately equal to the sum of the singular values (eigenvalues) of the original matrix and those of the Gaussian noise matrix.

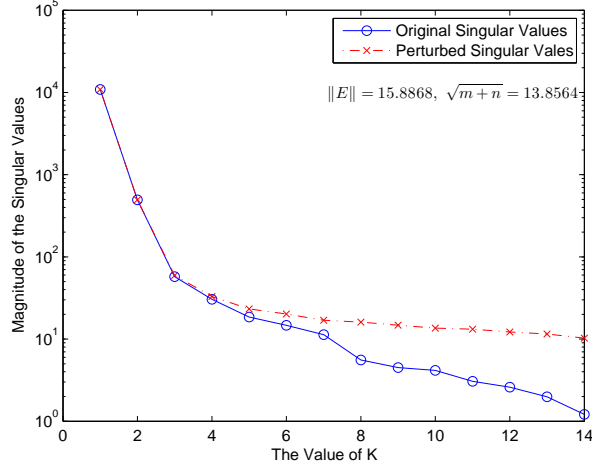
Figure 1: Distribution of singular values of the Bupa dataset before and after the perturbation.



In the experiments about the singular value distribution before and after the perturbation, we use the general perturbation model in Equation (2) to show the correctness of our mathematical analysis. Theoretically, the matrix R can be any orthogonal matrix with the appropriate dimension to multiply with A . In our experiments, we use the U matrix of the SVD of A in Equation (1) to be R and a random matrix from the standard Gaussian noise matrix $N(0, 1)$ ($\beta^2=1$) as E . Our experimental results are shown in Figure 1 for the Bupa dataset and in Figure 2 for the Wine dataset.

Since there can be many different choices for R and E , the results in Figures 1 and 2 are not unique. However, we believe that the basic trend of the singular value distributions using other choices of the R and E matrices should be similar to those shown in the two figures.

Figure 2: Distribution of singular values of the Wine dataset before and after the perturbation.



Based on the above two figures, it is clear that the singular values (eigenvalues) of the perturbed datasets are very close to those of the original datasets. In Figure 1, the first 6 perturbed singular values are almost the same as those of the original ones, (the two lines are overlapped at the beginning, and they form a fork starting at the 6-th point). In Figure 2, the first 4 perturbed singular values are almost identical to the original ones, (they are overlapped from the first point to the 4-th point). The difference between the last perturbed singular value and the corresponding original one is still very small, (please note that the y -axes of these figures are in a logarithmic scale). Therefore, we can use the perturbed singular values $\tilde{\sigma}_i$ to approximate the first few original singular values ($\tilde{\sigma}_i \approx \sigma_i$).

From the two figure legends, we can see that there is no big difference between $\|E\|$ and $\sqrt{m+n}$, given the comparatively large singular values. For example, for the Bupa dataset, $\|E\| = 20.3476$ and $\sqrt{m+n} = 18.7083$, their difference is much smaller than the singular values, $\sigma_1 = 2385.5$ and $\sigma_1^* = 2388.2$, ($\sigma_4 = 263.6$ and $\sigma_4^* = 264.6$). Hence, we can use $\sqrt{m+n}$ to approximate the norm of the Gaussian noise matrix ($\sqrt{m+n} \approx \|E\|$), and then use the following formula to determine the exact value of k .

$$\begin{aligned}
 k &= \min\{i \mid \|E\| \leq \sigma_i \text{ and } \|E\| \leq \tilde{\sigma}_i - \sigma_{i+1}\} \\
 &\approx \min\{i \mid \sqrt{m+n} \leq \tilde{\sigma}_i \text{ and } \sqrt{m+n} \leq \tilde{\sigma}_i - \tilde{\sigma}_{i+1}\}.
 \end{aligned}$$

4.2 Angle Preservation

After determining the value of k , we can calculate the angle between any perturbed data pair $(\tilde{A}_k^p, \tilde{A}_k^q)$, $\angle(\tilde{A}_k^p, \tilde{A}_k^q)$, or the sub-row counterpart, and know that $\angle(\tilde{A}_k^p, \tilde{A}_k^q)$ is very

similar to $\angle(A_k^p, A_k^q)$ by Corollary 7.

In practice, we should choose a small positive value for ϵ . In the following experiments, we choose three different values of ϵ , $\frac{\pi}{90}$, $\frac{\pi}{180}$ and $\frac{\pi}{360}$, and list our results in Table 1.

For any given pair (p, q) , we call this pair accurately computable if $|\angle(\tilde{A}_k^p, \tilde{A}_k^q) - \angle(A_k^p, A_k^q)| \leq \epsilon$.

Table 1: Percentage of angle preservation between A_k and \tilde{A}_k .

Bupa ($k = 6$)		Wine ($k = 4$)	
ϵ	Accuracy	ϵ	Accuracy
$\frac{\pi}{90}$	91.86%	$\frac{\pi}{90}$	90.70%
$\frac{\pi}{180}$	91.27%	$\frac{\pi}{180}$	87.64%
$\frac{\pi}{360}$	90.96%	$\frac{\pi}{360}$	85.74%

In Table 1, the percentage numbers in the accuracy columns denote the ratio of the accurately computable pairs to all pairs. It can be seen that the accuracy ratio is still very large even when the angle difference, ϵ , is very small (e.g., the accuracy ratio is around 91% in Bupa and around 87% in Wine, when $\epsilon = \pi/180$). In other words, the angle between the perturbed data pairs and the corresponding original data pairs are accurately preserved. In practice, we know all \tilde{A}_k^i , ($i = 1, \dots, n$), and a given reference original data A_k^p . If $\angle(\tilde{A}_k^p, \tilde{A}_k^q)$ is close to $\pi/2$, it is highly possible that $\angle(A_k^p, A_k^q)$ is $\pi/2$. Then A_k^q is probably the same as A_k^p , which is known. So A_k^q is leaked, and the privacy of the dataset is compromised.

Therefore, according to our experimental results, we can draw the following conclusions:

1. The magnitude of $\sqrt{m+n}$ is a very useful quantity to approximate the norm of a Gaussian noise matrix with the mean 0 when the ratio of the number of rows to that of columns is large enough.
2. The distribution of singular values of the perturbed dataset is highly similar to those of the original matrix when the Gaussian noise matrix is not related to the original dataset, $n/m \rightarrow \infty$, and $\sqrt{m+n}$ is small relatively to the norm of the perturbed dataset.
3. It is very easy and practical to determine the value of k simply by the combination of $\sqrt{m+n}$ and the distribution of singular values of the perturbed dataset.
4. The angle preservation of the general perturbation model is very good for many of the data records. This is not a desirable property for databases in privacy-preserving data publishing and data mining if some original records are leaked in a reference

information case. In other words, developers and researchers should pay more attention to take this property and situation into consideration in the future development of database publishing systems and privacy-preserving data mining algorithms.

5 Conclusion and Future Research

In consideration of the privacy issue in database publishing and data mining, researchers and users are concerned with the possibility that the potential attacker has background information to breach the privacy. We studied a reference background situation in which the attacker knows at least one exact record in the original dataset.

Data owners hope that dataset privacy can be perfectly kept even in the above situation. Our theoretical analysis and experimental results, however, show that the angle between different records in the dataset is accurately preserved before and after the perturbation in the general data perturbation model. Moreover, the angle is not only preserved in the original space, but also in the sub-row or sub-column spaces. Obviously, this is extremely undesirable for the privacy protection of databases. For example, if the attacker discovers that one perturbed record (or some attributes of this record) and the perturbed reference record (or some corresponding attributes of this reference record) are similar, through our theoretical analysis, it is highly possible that the attacker will figure out that the two records (or some attributes of them) in the original dataset are similar or even identical.

Further research work can extend our analysis to establish specific and concrete formulas to approximate a useful range of the value of ϵ which is difficult to decide in practice. Different datasets may tolerate different range of the ϵ values in terms of the meaning and magnitude of the particular data records. More importantly, we hope to develop a privacy-preserving database publishing and data mining method to break the property of angle preservation while keeping a good balance of data utilities and data privacy.

References

- [1] A. Asuncion, and D. J. Newman. *UCI machine learning repository* [<http://www.ics.uci.edu/~mllearn/MLRepository.html>]. Department of Information and Computer Science, University of California, Irvine, CA, 2007.
- [2] Y. Azar, A. Fiat, A. Karlin, F. Mcsherry, and J. Saia. *Spectral analysis of data*. In Proceedings of the 33rd Symposium on Theory of Computing, ACM, pp. 619-626, New York, NY, 2001.

- [3] S. Bapna, and A. Gangopadhyay. *A wavelet-based approach to preserve privacy for classification mining*. Decision Sciences Journal, 37(4):623-642, 2006.
- [4] M. W. Berry, Z. Drmac, and E. R. Jessup. *Matrix, vector space, and information retrieval*. SIAM Review, 41: 335-362, 1999.
- [5] R. B. Boppana. *Eigenvalues and graph bisection: an average-case analysis*. In Proceedings of the 28th Annual FOCS, pp. 280-285, Los Angeles, CA, 1987.
- [6] K. Chen, and L. Liu. *Privacy preserving data classification with rotation perturbation*. In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), pp. 589-592, Houston, Texas, 2005.
- [7] K. Chen, G. Sun, and L. Liu. *Towards attack-resilient geometric data perturbation*. In Proceedings of the 2007 SIAM International Conference on Data Mining (SDM 2007), pp. 78-89, Minneapolis, MN, 2007.
- [8] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Zhu. *Tools for privacy preserving distributed data mining*. ACM SIGKDD Explorations, 4(2):1-7, 2003.
- [9] A. Evfimievski. *Randomization in privacy preserving data mining*. ACM SIGKDD Explorations Newsletter, 4(2):43-48, 2002.
- [10] G. H. Golub, C. F. Van Loan. *Matrix computations, 3rd Edition*. John Hopkins University, Columbia, MD, 1996.
- [11] S. Guo, and X. Wu. *On the use of spectral filtering for privacy preserving data mining*. In Proceedings of the 21st ACM Symposium on Applied Computing, pp. 622-626, Dijon, France, 2006.
- [12] S. Guo, X. Wu, and Y. Li. *On the lower bound of reconstruction error for spectral filtering based privacy preserving data mining*. Knowledge Discovery in Databases: PKDD 2006, 4213: 520-527, 2006.
- [13] Z. Huang, W. Du, and B. Chen. *Deriving private information from randomized data*. In Proceedings of the 2005 ACM SIGMOD Conference, pp. 37-48, Baltimore, MD, 2005.
- [14] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. *On the privacy preserving properties of random data perturbation techniques*. In Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), pp. 99-106, Melbourne, Florida, 2003.

- [15] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. *Random-data perturbation techniques and privacy-preserving data mining*. Knowledge and Information Systems, 7(4):387-414, 2005.
- [16] T. Jiang. *How many entries of a typical orthogonal matrix can be approximated by independent normals?* Annals of Probability, 34(4): 1497-1529, 2006.
- [17] K. Muralidhar, and R. Sarathy. *Security of random data perturbation methods*. ACM Transactions on Database Systems, 24(4):487-493, 1999.
- [18] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern. *Worst-case background knowledge for privacy-preserving data publishing*. In Proceedings of the 23rd International Conference on Data Engineering (ICDE), pp. 126-135, Istanbul, Turkey, 2007.
- [19] P. Tendick. *Optimal noise addition for preserving confidentiality in multivariate data*. J. Statist. Planning and Inference, 27(2): 341-353, 1991.
- [20] S. Xu and S. Lai. *Fast Fourier transform based data perturbation method for privacy protection*. In Proceedings of the 2007 IEEE International Conference on Intelligence and Security Informatics, pp. 221-224, New Brunswick, NJ, 2007.
- [21] S. Xu, J. Zhang, D. Han and J. Wang. *Data distortion for privacy protection in a terrorist analysis system*. In Proceedings of the 2005 IEEE International Conference on Intelligence and Security Informatics, pp. 459-464, Atlanta, GA, 2005.
- [22] S. Xu, J. Zhang, D. Han and J. Wang. *Singular value decomposition based data distortion strategy for privacy protection*. Knowledge and Information Systems, 10(3): 383-397, 2006.