Random Value Distortion: Does It Really Preserve Privacy?

Hillol Kargupta and Souptik Datta
Computer Science and Electrical Engineering Department
University of Maryland Baltimore County
Baltimore, Maryland 21250
{hillol, souptik1}@cs.umbc.edu

Krishnamoorthy Sivakumar
School of Electrical Engineering and Computer Science
Washington State University
Pullman, Washington, USA
siva@eecs.wsu.edu

UMBC Technical Report Number TR-CS-03-25

Abstract

Privacy is becoming an increasingly important issue in many data mining applications. This has resulted in the development of several privacy-preserving data mining techniques. The random value distortion technique is one among them. Several data mining researchers have recently adopted this approach to mine data without losing the privacy. Random value distortion attempts to hide the sensitive data by randomly modifying the data values. This paper questions the utility of the random value distortion technique in data mining. The paper first notes that random matrices have "predictable" structures in the spectral domain and then it develops a random matrix-based spectral filtering technique to retrieve original data from the dataset distorted by adding random values. The proposed method works by comparing the spectrum generated from the observed data with that of random matrices. The paper presents the theoretical foundation and extensive experimental results to demonstrate that in many cases the random value distortion technique may not preserve any data privacy after all.

1 Introduction

Many data mining applications deal with privacy-sensitive data. Financial transactions, health-care records, and network communication traffic are some examples. Data mining in such privacy-sensitive domains is facing growing concerns. Therefore, we need to develop data mining techniques that are sensitive to the privacy issue. This has fostered the development of a class of data mining algorithms [1, 9] that try to extract the data patterns without directly accessing the original data and guarantees that the mining process does not get sufficient information to reconstruct the original data.

This paper considers an existing approach for privacy-preserving data mining by randomly perturbing the data while preserving the underlying probabilistic properties. It explores the random value perturbation-based approach [1], a well-known technique for masking the data using random noise. This approach tries to preserve data privacy by adding random noise, while making sure that the random noise still preserves the "signal" from the data so that the patterns can still be

accurately estimated. This paper questions the privacy-preserving capability of the random value perturbation-based approach. It shows that in many cases, the original data (sometimes called "signal" in this paper) can be accurately estimated from the perturbed data using a spectral filter that exploits some theoretical properties of random matrices. It presents the theoretical foundation and provides experimental results to support this claim.

Section 2 offers an overview of the related literature on privacy preserving data mining. Section 3 describes the random data perturbation method proposed in [1]. Section 4 presents a discussion on the eigenvalues of random matrices that builds the foundation of the technique proposed here to compromise the privacy protection introduced by the random value perturbation-based approach. Section 6 describes the proposed random matrix-based filtering technique to extract the original dataset. Section 7 applies the proposed technique and reports its performance for various data sets. Finally, Section 8 concludes this paper and outlines future research directions.

2 Related Work

There exists a growing body of literature on privacy-sensitive data mining. These algorithms can be divided into two different groups. One approach adopts a distributed framework; the other approach adds random noise to the data in such a way that the individual data values are distorted while still preserving the underlying distribution properties at a macroscopic level. The following part of this sections briefly discusses these two approaches.

The distributed approach supports computation of data mining models and extraction of "patterns" at a given node by exchanging only the minimal necessary information among the participating nodes without transmitting the raw data. The field of distributed data mining [10, 17] produced several distributed algorithms that are sensitive to privacy. For example the meta-learning based JAM system [24] was designed for mining multi-party distributed sensitive data such as financial fraud detection. The Fourier spectrum-based approach to represent and construct decision trees [11, 16], the Collective hierarchical clustering [7] are examples of additional distributed data mining algorithms that can be used with minor modifications for privacy-preserving mining from distributed data. In the recent past, several distributed techniques to mine multi-party data have been reported. A privacy preserving technique to construct decision trees [18] proposed elsewhere [13], multi-party secured computation framework [19], association rule mining from homogeneous [9] and heterogeneous [26] distributed data sets are some examples. There also exists a collection of useful privacy-sensitive data mining primitives such as secure sum computation [20], secure scalar product computation [26].

There is also a somewhat different approach and the algorithms belonging to this group works by first perturbing the data using randomized techniques. The perturbed data is then used to extracts the patterns and models. The randomized value distortion technique for learning decision trees [1] and association rule learning [3] are examples of this approach. Additional work on randomized masking of data can be found elsewhere [25].

This paper explores the second approach [1] that works by adding random noise to the data set in order to hide the individual data values of different attributes. It points out that in many cases the noise can be separated from the perturbed data by studying the spectral properties of the data and as a result its privacy can be seriously compromised. Before presenting the technique to do that, let us review the randomized value distortion [1] technique in details.

3 Random Value Perturbation Technique: A Brief Review

For the sake of completeness, we now briefly review the random data perturbation method suggested in [1]. We also discuss the procedure for reconstructing the original data distribution, as suggested in [1].

3.1 Perturbing the Data

The random value perturbation method attempts to preserve privacy of the data by modifying values of the sensitive attributes using a randomized process [1]. The authors explore two possible approaches — Value-Class Membership and Value Distortion — and emphasize the Value Distortion approach. In this approach, the owner of a dataset returns a value $u_i + v$, where u_i is the original data, and v is a random value drawn from a certain distribution. Most commonly used distributions are the uniform distribution over an interval $[-\alpha, \alpha]$ and Gaussian distribution with mean $\mu = 0$ and standard deviation σ. The n original data values $u_1, u_2, \ldots, u_n$ are viewed as realizations of n independent and identically distributed (i.i.d.) random variables U_i, $i = 1, 2, \ldots, n$, each with the same distribution as that of a random variable U. In order to perturb the data, n independent samples $v_1, v_2, \ldots, v_n$, are drawn from a distribution V. The owner of the data provides the perturbed values $u_1 + v_1, u_2 + v_2, \ldots, u_n + v_n$ and the cumulative distribution function $F_V(r)$ of V. The reconstruction problem is to estimate the distribution $F_U(x)$ of the original data, from the perturbed data.

3.2 Estimation of Distribution Function from the Perturbed Dataset

The authors [1] suggest the following method to estimate the distribution $F_U(u)$ of U, given n independent samples $w_i = u_i + v_i$, $i = 1, 2, \ldots, n$ and $F_V(v)$. Using Bayes' rule, the posterior distribution function $F'_U(u)$ of U, given that $U + V = w$, can be written as

$$F'_U(u) = \frac{\int_{-\infty}^{u} f_V(w - z) f_U(z) dz}{\int_{-\infty}^{\infty} f_V(w - z) f_U(z) dz},$$

which upon differentiation with respect to u yields the density function

$$f'_U(u) = \frac{f_V(w - u) f_U(u)}{\int_{-\infty}^{\infty} f_V(w - z) f_U(z) dz},$$

where $f_U(.)$, $f_V(.)$ denote the probability density function of U and V respectively. If we have n independent samples $u_i + v_i = w_i$, $i = 1, 2, \ldots, n$, the corresponding posterior distribution can be obtained by averaging:

$$f'_U(u) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_V(w_i - u) f_U(u)}{\int_{-\infty}^{\infty} f_V(w_i - z) f_U(z) dz}. \quad (1)$$

For sufficiently large number of samples n, we expect the above density function to be close to the real density function $f_U(u)$. In practice, since the true density $f_U(u)$ is unknown, we need to modify the right-hand side of equation 1. The authors suggest an iterative procedure where at each step $j = 1, 2, \ldots$, the posterior density $f_U^{j-1}(u)$ estimated at step $j - 1$ is used in the right-hand side of equation 1. The uniform density is used to initialize the iterations. The iterations are carried out until the difference between successive estimates becomes small. In order to speed up computations, the authors also discuss approximations to the above procedure using partitioning of the domain of data values.

4 Randomness and Patterns

The random perturbation technique "apparently" distorts the sensitive attribute values and still allows estimation of the underlying distribution information. However, does this apparent distortion fundamentally prohibit us from extracting the hidden information? In this section we explore this question.

Randomness may not necessarily imply uncertainty. Random events can often be analyzed and their properties can be explained using probabilistic frameworks. Statistics, randomized computation, and many other related fields are full of theorems, laws, and algorithms that rely on probabilistic characterization of random processes. Randomly generated structures like graphs also demonstrate interesting properties.

Random matrices [15] also exhibit many interesting properties that are often exploited in high energy physics [15], signal processing [22], and even data mining [12]. The random noise added to the data can be viewed as a random matrix and therefore its properties can be understood by studying the properties of random matrices. In this paper we shall develop a spectral filter designed based on random matrix theory for extracting the hidden data from the data perturbed by random noise. Our filtering approach is based on the observation that the distribution of eigenvalues of random matrices [15] exhibit some well known characteristics. The rest of this section discusses some of the important spectral properties of random matrices.

A random matrix is a matrix whose elements are random variables with given probability laws. The theory of random matrices deals with the statistical properties of the eigenvalues of such matrices. Eigenvalues of random matrices offer many interesting properties. For example, Wigner's semi-circle law [28], which says if V is an $n \times n$ matrix and has i.i.d. entries with zero mean and unit variance, the distribution of eigenvalues of $\frac{V+V'}{2\sqrt{2n}}$ has a probability density function given by

$$f(x) = \begin{cases} \frac{1}{\pi}(2n - x^2)^{1/2}, & |x| < \sqrt{2n} \\ 0, & \text{otherwise.} \end{cases}$$

In this paper, we are mainly concerned about distribution of eigenvalues of the sample covariance matrix obtained from a random matrix. Let V be a random $m \times n$ matrix whose entries are V_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$, are i.i.d. random variables with zero mean and variance σ^2. The covariance matrix of X is given by $Y = \frac{1}{m}V'V$. Clearly, Y is an $n \times n$ matrix. Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues of Y. Let

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} U(x - \lambda_i),$$

be the empirical cumulative distribution function (c.d.f.) of the eigenvalues λ_i, $(1 \leq i \leq n)$, where

$$U(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

is the unit step function. In order to consider the asymptotic properties of the c.d.f. $F_n(x)$, we will consider the dimensions $m = m(N)$ and $n = n(N)$ of matrix X to be functions of a variable N. We will consider asymptotics such that in the limit as $N \to \infty$, we have $m(N) \to \infty$, $n(N) \to \infty$, and $\frac{m(N)}{n(N)} \to Q$, where $Q \geq 1$. Under these assumptions, it can be shown that [8] the empirical c.d.f. $F_n(x)$ converges in probability to a continuous distribution function $F_Q(x)$ for every x, whose

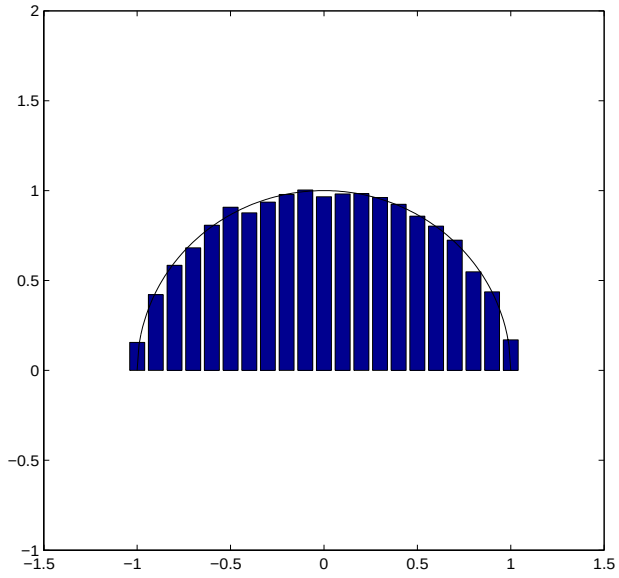

Figure 1: Wigner's semi-circle law: Distribution of the eigenvalues of $\frac{V+V'}{2\sqrt{2n}}$ where V is a random matrix takes the shape of a semi-circle.

probability density function (p.d.f.) is given by

$$f_Q(x) = \begin{cases} \frac{Q\sqrt{(x-\lambda_{\min})(\lambda_{\max}-x)}}{2\pi\sigma^2 x} & \lambda_{\min} < x < \lambda_{\max} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are as follows:

$$\begin{aligned} \lambda_{\min} &= \sigma^2(1 - 1/\sqrt{Q})^2. \\ \lambda_{\max} &= \sigma^2(1 + 1/\sqrt{Q})^2. \end{aligned} \tag{3}$$

Further refinements of this result and other discussions can be found in [22, 5, 14, 2, 4, 29, 21].

5 Separating the Data from the Noise

Consider an $m \times n$ data matrix U and a noise matrix V with same dimensions. The random value perturbation technique generates a modified data matrix $U_p = U + V$. Our objective is to extract U from U_p. Although the noise matrix V may introduce seemingly significant difference between U and U_p, it may not be successful in hiding the data.

Random noise has well defined probabilistic properties that may be used to identify the noise component of the perturbed data matrix U_p in an appropriate representation. The rest of this paper argues that the spectral representation of the data allows us to do exactly that.

Consider the covariance matrix of U_p:

$$\begin{aligned} U_p^T U_p &= (U+V)^T(U+V) \\ &= U^T U + V^T U + U^T V + V^T V. \end{aligned} \tag{4}$$

Now note that when the signal random vector (rows of U) and noise random vector (rows of V) are uncorrelated, we have $E[U^T V] = E[V^T U] = 0$. The uncorrelated assumption is valid in practice

since the noise that V that is added to the data U is generated by a statistically independent process. Recall that the random value perturbation technique discussed in the previous section introduces uncorrelated noise to hide the signal or the data. If the number of observations is sufficiently large, we have that $U^TV \sim 0$. Equation 4 can now be simplified as follows:

$$U_p^T U_p = U^T U + V^T V \tag{5}$$

Since the correlation matrices U^TU, $U_p^TU_p$, and V^TV are symmetric and positive semi-definite, let

$$\begin{aligned} U^TU &= Q_u \Lambda_u Q_u^T, \\ U_p^TU_p &= Q_p \Lambda_p Q_p^T, \text{ and} \\ V^TV &= Q_v \Lambda_v Q_v^T, \end{aligned} \tag{6}$$

where Q_u, Q_p, Q_v are orthogonal matrices whose column vectors are eigenvectors of U^TU, $U_p^TU_p$, V^TV, respectively, and Λ_u, Λ_p, Λ_v are diagonal matrices with the corresponding eigenvalues on their diagonals.

The following result from matrix perturbation theory [27] gives a relationship between Λ_u, Λ_v, and Λ_v.

Theorem 1 *[27] Suppose $\lambda_{1,(a)} \geq \lambda_{2,(a)} \geq \ldots \lambda_{n,(a)} \geq 0$, $a \in \{u, p, v\}$ are the eigenvalues of U^TU, $U_p^TU_p$, and V^TV, respectively. Then, for $i = 1, \ldots, n$,*

$$\lambda_{i,(p)} \in [\lambda_{i,(u)} + \lambda_{n,(v)}, \lambda_{i,(u)} + \lambda_{1,(v)}].$$

This theorem provides us a bound on the change in the eigenvalues of the data correlation matrix U^TU in terms of the minimum and maximum eigenvalues of the noise correlation matrix V^TV. Now let us take a step further and explore the properties of the eigenvalues of the perturbed data matrix U_p for large values of m.

Lemma 1 *Let data matrix U and noise matrix V be of size $m \times n$ and $U_p = U + V$. Let Q_u, Q_p, Q_v be orthogonal matrices and Λ_u, Λ_p, Λ_v be diagonal matrices as defined in 6. If $m/n \to \infty$ then $\Lambda_p = \Delta \Lambda_u \Delta^T + \Lambda_v$ where $\Delta = Q_p^T Q_u$.*

Proof:
Using Equations 5 and 6 we can write,

$$\begin{aligned} Q_p \Lambda_p Q_p^T &= Q_u \Lambda_u Q_u^T + Q_v \Lambda_v Q_v^T \\ \Rightarrow \Lambda_p &= Q_p^T Q_u \Lambda_u Q_u^T Q_p + Q_p^T Q_v \Lambda_v Q_v^T Q_p \\ &= \Delta \Lambda_u \Delta^T + Q_p^T Q_v \Lambda_v Q_v^T Q_p \end{aligned} \tag{7}$$

Let the minimum and maximum eigenvalues of V be $\lambda_{\min,(v)}$ and $\lambda_{\max,(v)}$ respectively. It follows from equation 2 that $m/n \to \infty$ all the eigenvalues in Λ_v become identical since $\lim_{m/n=Q\to\infty} \lambda_{\max,(v)} = \lim_{m/n=Q\to\infty} \lambda_{\min,(v)} = \sigma^2$ (say). This implies that, as $m/n \to \infty$, $\Lambda_v \to \sigma^2 I$, where I is the $n \times n$ identity matrix. Therefore, if the number of observations m is large enough (note that, in practice, number of features n is fixed), $V^TV = Q_v \Lambda_v Q_v^T = \sigma^2 Q_v Q_V^T = \sigma^2 I$. Therefore Equation 7 becomes

$$\begin{aligned} \Lambda_p &= \Delta \Lambda_u \Delta^T + Q_p^T Q_p \Lambda_v Q_p^T Q_p \\ \Lambda_p &= \Delta \Lambda_u \Delta^T + \Lambda_v. \end{aligned} \tag{8}$$

■

If the norm of the perturbation matrix V is small, the eigenvectors Q_p of $U_p^T U_p$ would be close to the eigenvectors $Q_u^T Q_u$ of $U^T U$. Indeed, matrix perturbation theory provides precise bounds on the angle between eigenvectors (and invariant subspaces) of a matrix U and that of its perturbation $U_p = U + V$, in terms of the norms of the perturbation matrix V. For example, let (x_u, λ_u) be an eigenvector-eigenvalue pair for matrix $U^T U$ and $\epsilon = \|V^T V\|_2 = \sigma_{\max}(V^T V)$ be the two-norm of the perturbation, where $\sigma_{\max}(V^T V)$ is the largest singular value of $V^T V$. Then there exists an eigenvalue-eigenvector pair (x_p, λ_p) of $U_p^T U_p$ satisfying [27, 23]

$$\tan(\angle(x_u, x_p) < 2\frac{\epsilon}{\delta - \epsilon},$$

where δ is the distance between λ_u and the closest eigenvalue of $U^T U$, provided $\epsilon < \delta$. This shows that the eigenvalues of $U^T U$ and $U_P^T U_P$ are in general close, for small perturbations. Moreover,

$$|\lambda_u - x^* U_p x_u| < 2\frac{\epsilon^2}{\delta - \epsilon},$$

where x^* is the conjugate-transpose of x. Consequently, the product $\Delta = Q_p^T Q_u$, which is the matrix of inner products between the eigenvectors of $U^T U$ and $U_p^T U_p$ would be close to an identity matrix; i.e., $\Delta = Q_p^T Q_u \approx I$. Thus equation 8 becomes

$$\Lambda_p \approx \Lambda_u + \Lambda_v. \tag{9}$$

Suppose the signal covariance matrix has only a few dominant eigenvalues, say $\lambda_{1,(u)} \geq \cdots \geq \lambda_{k,(u)}$, with $\lambda_{i,(u)} \leq \epsilon$ for some small value ϵ and $i = k+1, \ldots, n$. This condition is true for many real-world signals. Suppose $\lambda_{k,(u)} > \lambda_{1,(v)}$, the largest eigenvalue of the noise covariance matrix. It is then clear that we can separate the signal and noise eigenvalues Λ_u, Λ_v from the eigenvalues Λ_p of the observed data by a simple thresholding at $\lambda_{1,(v)}$.

Note that equation 9 is only an approximation. However, in practice, one can design a filter based on this approximation to filter out the perturbation from the data. Experimental results presented in the following sections indicate that this provides a good recovery of the data.

6 Random Matrix-Based Data Filtering

This section describes the proposed filter for extracting the original data from the noisy perturbed data. Suppose actual data U is perturbed by a randomly generated noise matrix V in order to produce $U_p = U + V$. Let $u_{p,i} = \mathbf{u}_i + \mathbf{v}_i$, $i = 1, 2, \ldots, m$, be m (perturbed) data points, each being a vector of n features.

The proposed filtering technique first calculates the covariance matrix of the perturbed data U_p. Using the distribution of eigenvalues of the covariance matrix, and the theory of random matrices, the covariance matrix of U_p is decomposed into a noise part and an actual data part. The eigenstates corresponding to actual data are then used to reconstruct the actual data.

In the following section, we discuss the proposed filtering procedure. We first explore the case where the distribution $F_V(v)$ of the random noise V (including the variance) is known, as required by the random value perturbation scheme [1]. Next we discuss how the noise variance can be estimated from the eigenvalue distribution of the perturbed data. The reader should note that the random value perturbation scheme provides information about the noise distribution. So estimation of the noise variance is not necessary. We explored that case in order to develop a broader understanding about the performance of the proposed filtering technique.

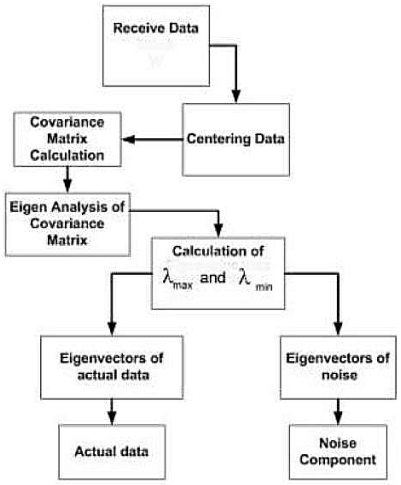

Figure 2: Flowchart of the spectral filtering technique.

6.1 Known Noise Variance

When the noise distribution $F_V(v)$ of V is completely known (as required by the random value perturbation technique [1]), the noise variance σ^2 is first calculated from the given distribution. Equation 2 is then used to calculate λ_{max} and λ_{min} which provide the theoretical bounds of the eigenvalues corresponding to noise matrix V. From the perturbed data, we compute the eigenvalues of its covariance matrix Y, say $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Then we identify the noisy eigenstates $\lambda_i \leq \lambda_{i+1} \leq \cdots \leq \lambda_j$ such that $\lambda_i \geq \lambda_{min}$ and $\lambda_j \leq \lambda_{max}$. The remaining eigenstates are the eigenstates corresponding to actual data. Let, $\Lambda_v = \text{diag}\ (\lambda_i, \lambda_{i+1}, \ldots, \lambda_j)$ be the diagonal matrix with all noise-related eigenvalues, and A_v be the matrix whose columns are the corresponding eigenvectors. Similarly, let Λ_u be the eigenvalue matrix for the actual data part and A_u be the corresponding eigenvector matrix which is an $n \times k$ matrix ($k \leq n$). Based on these matrices, we decompose the covariance matrix Y into two parts, Y_s and Y_r with $Y = Y_s + Y_r$, where $Y_r = A_v \Lambda_r A_v^T$, is the covariance matrix corresponding to random noise part, and $Y_s = A_u \Lambda_u A_u^T$, is the covariance matrix corresponding to actual data part. An estimate $\hat{U}$ of the actual data U is obtained by projecting the data U_p on to the subspace spanned by the columns of A_u. In other words, $\hat{U} = U_p A_u A_u^T$.

6.2 Unknown Noise Variance

When the noise variance σ^2 is unknown, we first estimate it using the perturbed data. The estimated noise variance is then used to filter the perturbed data. In order to estimate the noise variance σ^2 we first compute the eigenvalues of the covariance matrix Y of the perturbed data W. A histogram of the eigenvalue distribution is plotted and compared to that of the theoretical noise eigenvalue density function $f_Q(x)$ given in equation 2. Note that the density function $f_Q(x)$ depends on the variance σ^2. Typically, the theoretical density function $f_Q(x)$ is a good fit to the left portion of the histogram of the computed eigenvalues, corresponding to small eigenvalues. The larger eigenvalues that do not fit this theoretical density function correspond to the actual information part of the perturbed data. An iterative procedure is employed to obtain the value of σ that results in the best fit of $f_Q(x)$ to the observed histogram.

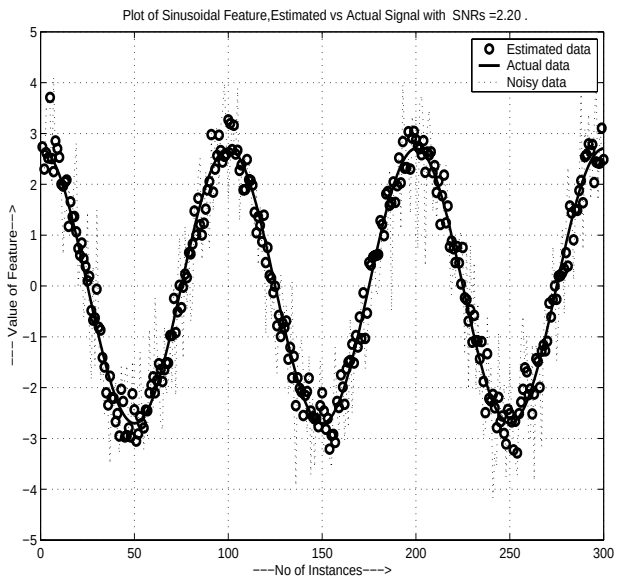

Figure 3: Estimation of original sinusoidal data with known random noise variance.

7 Experimental Results

Our proposed method is used on datasets of different sizes which have some trend in their values. The actual dataset is distorted by adding Gaussian noise (Normally distributed random numbers with zero mean and specific variance), and our proposed technique is applied to recover the actual data from the perturbed data with the knowledge of noise distribution (noise variance in particular). Experimental results show this method estimates the pattern and gives close estimation of individual values of actual data. Figure 3 shows one such estimation of data when the actual data has sinusoidal trend.

The distribution of eigenvalues shows (Figure 4) the method accurately distinguishes between noisy eigen values and eigenvalues corresponding to actual data. Note that the estimated eigenvalues of actual data is very close to eigenvalues of actual data and almost overlap with them above λ_{max}. The eigenvalues of actual data below the λ_{min} are of very small values and are negligible. Thus , even though there are no estimations corresponding to them, the estimation of actual data is fairly accurate.

The theory of random matrix based filtering can be extended for dataset having single feature, i.e when the dataset is in the form of a vector containing data points.The whole vector is divided into a fixed no of vectors having equal length, and all of these vectors are added to form a matrix. The matrix is then distorted by adding random noise in the same way as before and is disclosed. The same method is applied to this matrix to estimate the original data. After the data matrix is estimated, its columns are concatenated to form the single vector. The performance of the estimation remains nearly same as compared to multi-featured data in the form of a matrix.

We tried to replicate the experiment reported in [1] using our method to recover the triangular distribution. We used a vector data of 10000 values having a triangular distribution. The individual values of actual data are within 0 and 1. We split the data vector into 50 columns, each having 200 values, and added a Gaussian random variable with mean 0 and standard deviation $\sigma = 0.25$ to each data value. We then applied our algorithm to recover the actual data from the distorted data with the known noise $\sigma = 0.25$. Figure 5 shows estimation of the distribution. The distorted distribution looks nowhere close to the actual triangular distribution, but the estimated distribution looks very

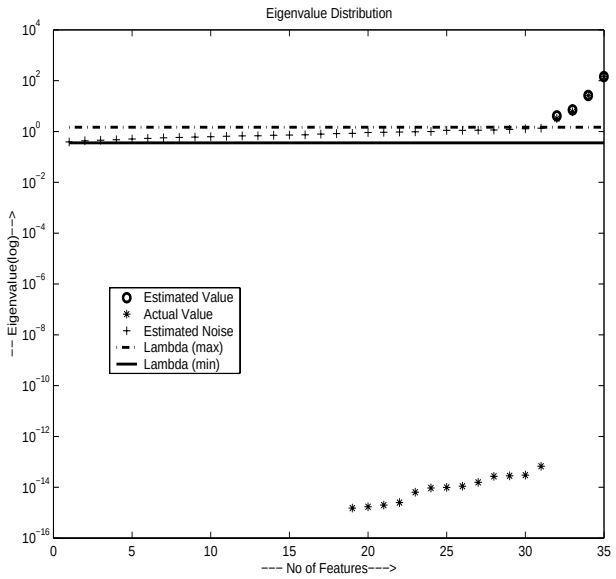

Figure 4: Distribution of eigenvalues of actual data, and estimated eigenvalues of random noise and actual data.

similar to the original distribution. This shows that our method recovers the original distribution closely. Figure 6 shows a chunk of data, their values after distortion by random numbers, and their estimated values. Note, the estimated values are pretty close to actual values, whereas the distorted values are far apart. Figure 7 shows the error in estimation of of the actual data for the whole dataset (10000 points). The estimation error remains within -0.3 to 0.3 in this experiment. So, our method can recover the distribution of data as well as their individual values closely. That is where it questions the privacy preserving ability of randomized value perturbation technique reported in [1].

Quality of recovery depends upon relative noise content of the data. If the relative noise compared to actual dataset increases very much, the recovery method performs poorly. We define the term 'Signal-to-Noise Ratio' (SNR) to quantify the relative amount of noise added to actual data to perturb it.

$$\text{SNR} = \frac{\text{Value of Actual Data}}{\text{Value of Noise Added to the Data}}$$

As the noise added to the actual value increases, the SNR decreases. Our experiments show that this method predicts the actual data reasonably well up to a SNR value of 1.0 (i.e. 100% noise). The results shown in Figure 3 is the case of mean SNR value nearly 2, i.e. noise content is 50%. Figure 6 shows a datablock where mean SNR is 1.9. As the SNR goes below 1, the estimation becomes too erroneous. Figure 8 shows the difference in estimation accuracy as the SNR increases from 1. The dataset used has sinusoidal trend in its values.The upper figure shows the estimation corresponding to 23% noise(mean SNR = 4.3), and the lower figure shows estimation corresponding to 100% noise (mean SNR = 1.0). Figure 9 shows the variation of estimation error with change in SNR value. As the SNR value decreases, mean error in estimation shows an increasing trend, though the variation is not linear and has some erratic behavior.

In case of unknown noise distribution, the method estimates the noise variance first. From the eigenvalues of covariance matrix of actual data, a histogram of the eigenvalue distribution is

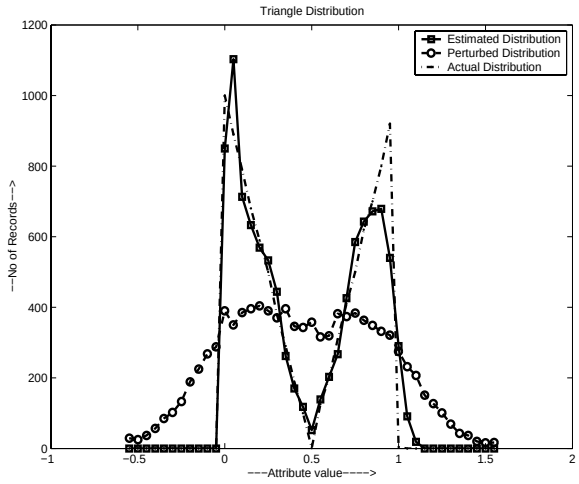

Figure 5: Original data has 'Triangular' distribution.Though perturbed data distribution does not look like a triangular distribution, our method estimates the distribution very closely.

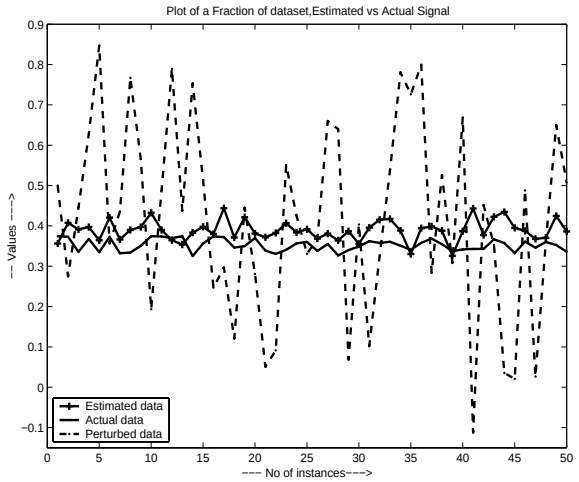

Figure 6: Plot of individual values of a fraction of dataset used in Figure 5.Our method gives close estimation of individual values.

obtained, and this is compared with best possible theoretical density function given by Equation 2. The variance corresponding to the best fit gives the estimation of the noise variance.

To get the best estimation of variance, the algorithm estimates noise variance from the best fit curve several times. In each trial , the variance estimation algorithm starts with a very small variance value near zero, create the theoretically generated distribution and measures the mean square error between it and histogram of eigenvalues of actual data. It then increases variance by a small value, again computes the mean square error and compares it with the previous error to get the minimum error and corresponding variance. The algorithm does the said operation up-to a threshold value of variance, and stores of the variance corresponding to minimum mean square error between theoretically generated density function curve and histogram of eigenvalues of actual data.That value of variance is treated as the estimated value of noise variance for that particular trial.In our experiment, we used 100 such trials for each variance estimation. After the set of estimates are calculated from all trials, the distribution of estimated variances is checked for

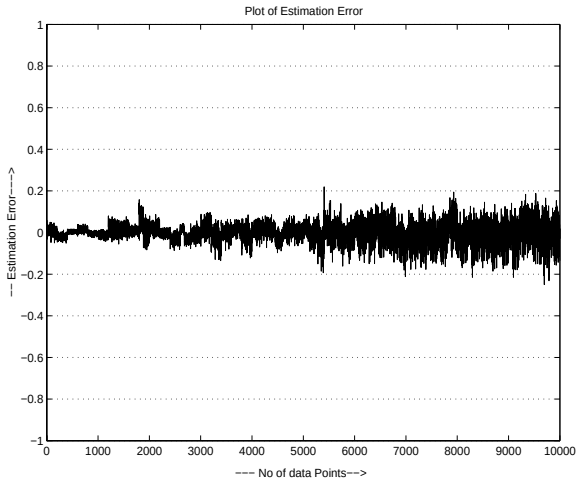

Figure 7: Plot of estimation error for experiment reported in Figure 5. The error is limited within -0.25 and +0.25.

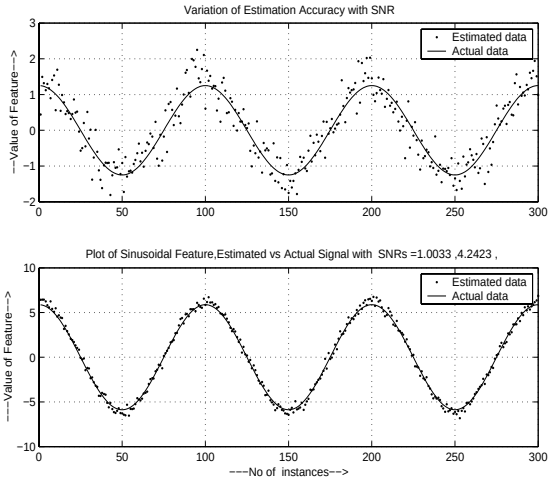

Figure 8: A higher noise content (low SNR)leads to less accurate estimation.SNR in upper figure is 1, while that for lower figure is 4.3.

outliers in them. The mean μ_1 and standard deviation σ_1 of the estimates are calculated , and values lying outside the span $\mu_1 \pm 2\sigma_1$ are discarded. During each trial, if the algorithm does not get best fit within a predefined threshold value of variance, it stores that threshold value of variance as the estimation. These values are also treated as outliers at the end and are discarded.

After discarding the outlier estimations, an average of the rest of the estimates are taken to get the actual estimate of noise variance. We have noticed that discarding the outliers and taking average of the remaining number of estimate improves the estimation accuracy to a large extent. Once the noise variance is estimated, the same technique is applied as before to estimate the original data. Figure 10 shows the estimation of actual data having 300 values and a sawtooth trend with SNR value of 4.25 when distribution of noise is not known. The average over 100 estimates of noise variance after discarding the outliers gave an estimated variance of 0.83452 where the actual noise variance is 0.85. Although not all the estimates are always so close, on an average, the difference between the estimated variance and true variance remains within 10% of the actual variance in all

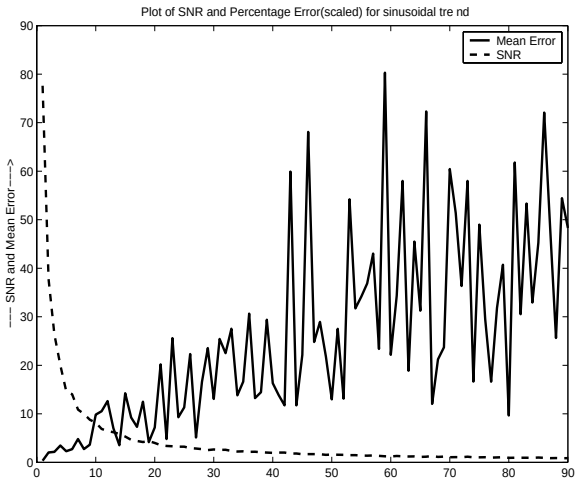

Figure 9: Variation of mean estimation error with change in SNR values. As SNR decreases, mean estimation error shows an increasing trend.

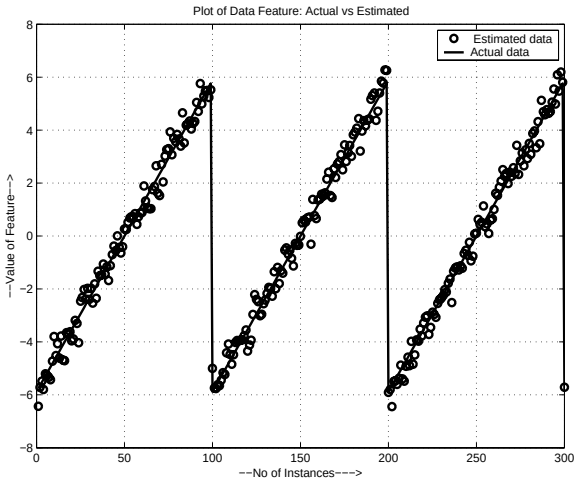

Figure 10: Estimation of actual data when the noise distribution is not known.

our experiments. Figure 11 shows the distribution of eigenvalues of actual data, estimated noise and estimated data for experimental result reported in Figure 10.

8 Conclusion and Future Work

Preserving privacy in data mining activities is a very important issue in many applications. This paper illustrates a noise filtering technique by which true data values can be estimated from the perturbed values (by random noise). This raises questions against the claim of preserving privacy by perturbing data with random numbers and disclosing the perturbed dataset as well as the probability distribution of the random number generator. The proposed approach works by comparing the empirically observed eigenvalue distribution of the given data with that of the known distribution of random matrices. The theoretically known values of upper and lower limits of the spectrum (eigenvalues) are used to identify the boundary between the eigen-states due to noise and that of the actual data.

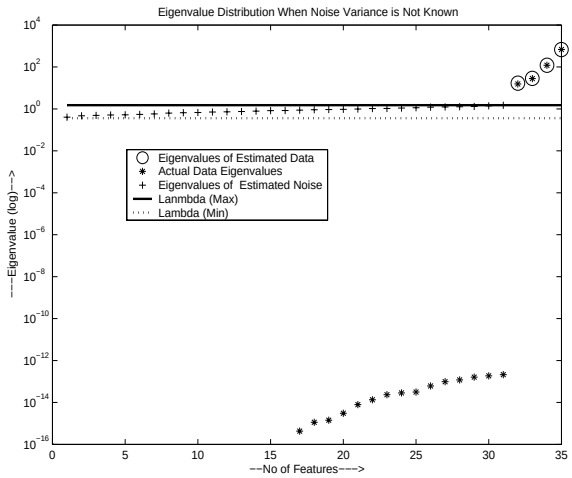

Figure 11: Distribution of eigenvalues of actual data, estimated data and estimated noise when noise variance is not known.

This random matrix based approach to separating the information bearing and noisy eigenstates has potential computational advantages. Indeed, since the upper bound $\lambda_{\max}$ of the noisy eigenvalues is known a priori, one can easily use a suitable numerical technique (e.g., power method [6]) to compute just the few largest eigenvalues. Once these eigenvalues and corresponding eigenvectors are computed, one can obtain the actual-data-part of the covariance matrix, which can be subtracted off from the total covariance to isolate the noise-part of the covariance. The proposed approach is simple, and retrieves actual data with reasonable precision. For the datasets considered in this paper, our experimental results support this claim. So, the method of perturbing data with random number to hide their original value is not a very reliable method to preserve privacy.

This work points out a potential problem in the existing literature. However, it leaves open the problem of coming up with methods which can actually preserve privacy without destroying statistical properties of the original dataset. We believe that this can be done by first narrowing down the specific pattern that we want to preserve through randomized perturbation. We hope that this work will encourage data mining researchers to design privacy-preserving techniques that pay careful attention to the properties of random noise and their effect on preserving privacy.

Acknowledgments

The authors acknowledge supports from the United States National Science Foundation CAREER award IIS-0093353, NASA (NRA) NAS2-37143, and TEDCO, Maryland Technology Development Center.

References

[1] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceeding of the ACM SIGMOD Conference on Management of Data*, pages 439–450, Dallas, Texas, May 2000. ACM Press.

[2] Z. D. Bai, J. W. Silverstein, and Y. Q. Yin. A note on the largest eigenvalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis*, 26(2):166–168, August 1988.

[3] S. Evfimievski. Randomization techniques for privacy preserving association rule mining. In *SIGKDD Explorations*, volume 4(2), Dec 2002.

[4] S. Geman. A limit theorem for the norm of random matrices. *The Annals of Probability*, 8(2):252–261, April 1980.

[5] U. Grenander and J. W. Silverstein. Spectral analysis of networks with random topologies. *SIAM Journal on Applied Mathematics*, 32(2):499–519, 1977.

[6] J. E. Jackson. *A User's Guide to Principal Components*. John Wiley, 1991.

[7] E. Johnson and H. Kargupta. Collective, hierarchical clustering from distributed, heterogeneous data. In *Lecture Notes in Computer Science, volume 1759*, pages 221–244, 1999.

[8] D. Jonsson. Some limit theorems for the eigenvalues of a sample covariance matrix. *Journal of Multivariate Analysis*, 12:1–38, 1982.

[9] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. In *SIGMOD Workshop on DMKD*, Madison, WI, June 2002.

[10] H. Kargupta, B. Park, D. Hershberger, and E. Johnson. Collective data mining: a new perspective towards distributed data mining. In *Advances in Distributed and Parallel Knowledge Discovery, Eds: Kargupta, Hillol and Chan, Philip*. AAAI/MIT Press, 2000.

[11] H. Kargupta, H. Park, S. Pittie, L. Liu, D. Kushraj, and K. Sarkar. MobiMine: Monitoring the stock market from a PDA. *ACM SIGKDD Explorations*, 3:37–47, 2001.

[12] H. Kargupta, K. Sivakumar, and S. Ghosh. Dependency detection in mobimine and random matrices. In *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 250–262. Springer, 2002.

[13] Y. Lindell and B. Pinkas. Privacy preserving data mining. In *Advances in Cryptology CRYPTO 2000*, pages 36–54, August 2000.

[14] V. A. Marcenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR — Sbornik*, 1(4):457–483, 1967.

[15] M. L. Mehta. *Random Matrices*. Academic Press, London, 2 edition, 1991.

[16] B. Park, A. R., and H. Kargupta. A fourier analysis-based approach to learn classifier from distributed heterogeneous data. In *Proceedings of the First SIAM Internation Conference on Data Mining*, Chicago, US, 2001.

[17] B. H. Park and H. Kargupta. Distributed data mining: Algorithms, systems, and applications. In *In Data Mining Handbook*, To be published, 2002.

[18] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[19] J. R. Quinlan. Induction of decision trees. In *Machine Learning*, pages 81 – 106, 1986.

[20] B. Schneier. Applied cryptography. John Wiley and Sons, 1995.

[21] J. W. Silverstein. On the weak limit of the largest eigenvalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis*, 30(2):307–311, August 1989.

[22] J. W. Silverstein and P. L. Combettes. Signal detection via spectral theory of large dimensional random matrices. *IEEE Transactions on Signal Processing*, 40(8):2100–2105, 1992.

[23] G. W. Stewart. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Review*, 15(4):727–764, October 1973.

[24] S. Stolfo et al. Jam: Java agents for meta-learning over distributed databases. In *Proceedings Third International Conference on Knowledge Discovery and Data Mining*, pages 74–81, Menlo Park, CA, 1997. AAAI Press.

[25] J. F. Traub, Y. Yemini, and H. Woz'niakowski. The statistical security of a statistical database. *ACM Transactions on Database Systems (TODS)*, 9(4):672–679, 1984.

[26] J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In *The Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, CA, July 2002.

[27] H. Weyl. Inequalities between the two kinds of eigenvalues of a linear transformation. In *Proceedings of the National Academy of Sciences*, volume 35, pages 408–411, 1949.

[28] E. P. Wigner. On the statistical distribution of the widths and spacings of nuclear resonance levels. *Proceedings of the Cambridge Philosophical Society*, 47:790–798, 1952.

[29] Y. Q. Yin, Z. D. Bai, and P. R. Krishnaiah. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields*, 78(4):509–521, August 1988.