

# Wavelet-Based Data Distortion for Privacy-Preserving Collaborative Analysis \*

Lian Liu, Jie Wang, Zhenmin Lin, and Jun Zhang<sup>†</sup>

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

June 8, 2007

## Abstract

With the rapid development of modern data collection and data warehouse technologies, data mining is becoming more and more a standard practice. Accompanying this trend, preserving privacy in certain data becomes a challenge to data mining applications in many fields, especially in medical, financial and homeland security fields. We present a class of novel privacy-preserving data distortion methods in the collaborative analysis situations based on wavelet transformation, which provides an effective and efficient balance between data utilities and privacy protection beyond its fast run time. We also provide a new privacy breach algorithm in the collaborative analysis which could threaten the data privacy, even with the distorted data values, in the single basis wavelet transformation case. Thus, we further propose a multi-basis wavelet data distortion strategy for better privacy preserving in these situations. Through experiments on real-life datasets, we conclude that the multi-basis wavelet data distortion method is a very promising privacy-preserving technique.

---

\*Technical Report No. 482-07, Department of Computer Science, University of Kentucky, Lexington, KY, 2007. The research work of the authors was supported in part by the National Science Foundation under grant CCF-0527967, in part by the National Institutes of Health under grant 1-R01-HL086644-01, in part by the Kentucky Science and Engineering Foundation under grant KSEF-148-502-06-186, and in part by the Alzheimer's Association under grant NIGR-06-25460.

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

## 1 Introduction

With the widespread availability of digital data in the information age, data collection as well as data mining is becoming more and more a standard practice whose goal is to efficiently and correctly discover patterns, association rules, or relationships hidden in a large number of different formats and multiparty data, and then combine the historical patterns and the modern understanding to predict future trends. We consider the result of the former function as information, and the latter one as knowledge. Because of the potential benefits of the information and knowledge from the data mining practice, a great volume of data mining applications and publications has continuously been emerging in many different fields.

Although with such a broad and attractive prospect, data mining techniques undoubtedly face a challenge to their legal survivals. That is how to protect the privacy of certain crucial data such as medical records, private financial messages, and homeland security information. Due to the private business profit and social security consideration, it is reported in [7] that few people and organizations are willing to disclose their original and private data to the public without some mechanisms for privacy protection. For example, to comply with the Health Insurance Portability and Accountability Act (HIPAA) [1], individual persons and organizations do not have to reveal their medical data for the public use without the privacy protection guarantee in any case. Therefore, one of the most urgent tasks facing the data mining community is

to develop new techniques that can achieve a good balance between accurate data utilities and sufficient privacy preservation.

Especially in commercial data analysis fields, in order to maximize business profit return and to provide better customer services, different business organizations may reach a multiparty agreement that each party is willing to share its own commercial data with others. In such cases, we need develop multiparty data mining models based on accurate collaborative data analysis. At the same time, we have to take concrete steps to ensure that certain private information in each owner's data is not disclosed to the other parties.

We consider two real-life scenarios where different companies can share their data. For simplicity, we assume that there are two companies called Company A and Company B that decide to share their data beyond the boundary of their individual entities.

1. The database of Company A has exactly the same customer set as that of Company B, but the two dataset attribute sets are different.
2. The attribute set of the database of Company A is identical to that of Company B, but the two companies target at different customer sets.

In both scenarios, the collaborative analysis is considered as an essential approach to gaining more comprehensive knowledge from the combined databases. Thus, the former is referred to as vertically collaborative analysis [25] and the latter as horizontally collaborative analysis [16]. The scenario with different customer sets and different attribute sets will be considered in our future study.

In this paper, we present a class of novel privacy-preserving collaborative analysis methods based on wavelet transformation. Wavelets are a set of functions which are localized, scaled and well-organized in order to satisfy certain requirements. Wavelet transformation is widely used in signal processing [6, 8] and noise suppression [22].

The remaining parts of this paper are arranged as follows. We provide a brief review of related work in Section 2. The background knowledge and the detailed procedures about the wavelet-based data privacy-preserving algorithm are discussed in Section 3. The experimental

results and analyses will be presented in Section 4. Finally a brief conclusion is given in Section 5.

## 2 Related Work

In the past decade, there have been a large number of privacy-preserving data mining literatures. We can divide these literatures into 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 to the original datasets. In this paper, we will not deal with privacy-preserving data mining methods in this category. Interested readers may consult [5] and the references therein for discussions on distributed data mining approaches. Methods in the other category modify the values of the datasets to protect privacy of the data values. There are several classes of data distortion or perturbation methods in this category. For example, one class is focused on data anonymization [17, 23, 26, 27, 33]. The other class pays more attention to perturbing the whole dataset or the confidential parts of the dataset using certain distribution of random noises [4, 9, 11, 14, 19].

Briefly, on one hand, the data anonymization strategy removes certain parts of the dataset such as unique and confidential identifiers, e.g., social security numbers or driver's license numbers or credit card numbers. Sweeney [24] demonstrated that this strategy may not be safe to guarantee identification because the intruders can discover certain secret information by exploiting relationships between other attributes.

On the other hand, the data randomization perturbation preserves data utilities such as patterns and association rules by using the additive random noise. However, Kargupta *et al.* [15] showed that it is highly possible to differentiate the original true values from the additive randomization noise perturbed datasets.

Recently, matrix decomposition and factorization techniques have been used to distort numerical valued datasets in the applications of privacy-preserving data mining. In particular, singular value decomposition (SVD) [31, 32] and nonnegative matrix factorization (NMF) [28] have been shown to be very effective in providing high level data privacy preservation and maintaining high degree

data utilities.

In addition to the above documents, signal transformation methods related to Fourier or wavelet transformation have been used as strategies for data perturbation [3, 18, 30]. Both transformation based privacy preserving distortion methods seem to have a very good property on privacy protection and data utility preservation.

The run time complexity of the wavelet-based transformation is  $O(n)$  which is better than the  $O(n \log n)$  run time of the Fourier transformation, where  $n$  is the maximum level number of wavelet or Fourier decompositions, to be defined later. Thus, data analysts may prefer the wavelet-based methods which have a very attractive merit, fast run time, in dealing with very large datasets. However, in [3], the wavelet perturbed dataset in the transformed space has different dimensions from that in the original space. This might create a problem when a third party data miner or the collaborative analyst has data parts from different sources to match each other. There is certainly an advantage to consider the transformed dataset that keeps the same dimension as the original dataset in the collaborative data analysis situation. Therefore, we propose a different set of data distortion, suppression and reconstruction (transformation back) strategies based on wavelets to keep the dimensions of the original and distorted datasets.

### 3 Algorithms

In this section, we present the detailed procedures of the privacy-preserving data distortion method based on wavelet transformation for collaborative analysis which can achieve a desirable balance between accurate data utilities and good privacy protection.

#### 3.1 Assumptions

The matrix representation (vector-space format) is one of the most popular ways to encode the object-attribute relationships in many real-life datasets. In this format, a 2-dimensional (2D) matrix is used to store the dataset in which each row of the matrix stands for an individual object and each column represents a particular attribute of these objects. Apparently, in this matrix, the privacy is a set of all confidential attributes represented by columns

and all secret objects represented by rows. In such a matrix, we assume that every element is fixed, discrete, and numerical. Any missing element is not allowed.

With the matrix representation, we assume that the original database is only accessible by the data owners and authorized users. The corresponding dataset outsourced to other analysts and the public is the distorted dataset that has no harm to either data mining accuracy or data privacy safety.

#### 3.2 Wavelet Decomposition

In mathematical terms, a discrete wavelet transformation (DWT) is a wavelet transformation for which the input discrete samples are divided into approximation coefficients and detail coefficients which correspond to the low frequency and high frequency decompositions of the original samples, respectively. Such wavelet decomposition process is applied recursively with high and low passing filters on the approximation coefficients of the previous level and then down-sampled. A brief transformation process is illustrated in Figure 1:

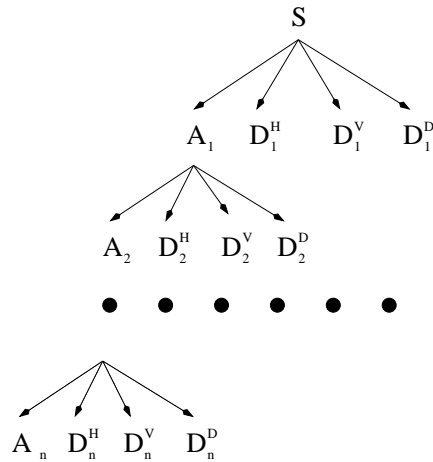


Figure 1: The DWT decomposition schema.  $S$  is the original 2D matrix.  $A_i$  contains the  $i$ -th level approximation coefficients,  $D_i^H$ ,  $D_i^V$  and  $D_i^D$  are the horizontal, vertical and diagonal detail coefficients, respectively.

According to the above graph and introduction in [29], the 2D DWT decomposition firstly uses the high pass filters and low pass filters to process the original entire  $a * b$

matrix. Then the  $b$  columns with length  $a$  are passed to the first filter which operates on the columns horizontally. These filter outputs are then downsampled by a half. In other words, a half of the column elements are thrown away. The remaining half column outputs are further decomposed using the second filter which considers the input data as transposed  $a/2$  rows of length  $b$ . Similarly, the second filter throws away a half of the remaining column elements. After these processes, one approximation coefficient and three detail coefficient submatrices are produced at this level. For the next level DWT decomposition process, it just recursively processes the approximation coefficients of the previous level as the new input matrix.

Through the above description, it is clear that the discrete wavelet transformation only depends on the maximum decomposition level and the filters (wavelet basis). For a given wavelet basis, the maximum number of decomposition levels,  $n$ , of DWT mainly depends on the dimensions of the input signals. Although the standard 2D DWT decomposition needs the input matrix represented in  $2^a * 2^b$  dimensions, where  $a$  and  $b$  are two integers, we can still deal with matrices of any dimension size as follows [29].

For any  $a * b$  dimension matrix, the DWT decomposition can process and downsample all columns through the standard DWT filters, but the rows may not be sufficiently decomposed (for simplicity, we assume that  $a > b$ ). However, in the data distortion techniques, it does not matter because we can still suppress the entire detail coefficients and then reconstruct them and the approximation coefficients, to be introduced in the next section, to successfully distort the whole original data if  $n$  is large enough.

Thus, the maximum number of decomposition levels,  $n$ , of a data matrix of any dimension  $a * b$  is defined as:  $n = \lceil \log_2 \min(a, b) \rceil$ .

### 3.3 Coefficient Suppression and Wavelet Reconstruction

Although the original matrix could be replaced by the approximation coefficient matrix as the analysis target dataset, the dimension of the approximation coefficient matrix is downsized. The strategy proposed by Bapna and Gangopadhyay [3] will further remove some columns of the transformed data deemed as “less important”. So

there may be a problem to use the transformed data in the multiparty collaborative analysis situations, which require the dimensions of the individual datasets to match each other to facilitate analysis with respect to the corresponding object set or the corresponding attribute set. One way of maintaining dimensions of the dataset matrices is to transform the individual dataset matrices back to the original spaces and to reconstruct the original matrix formats. For our privacy-preserving purpose, we need the data to be distorted when the data entries are transformed back to the original space.

Therefore, we suppress the detail coefficients to reduce the high frequency “noise” which is hidden among the original data entries. Our proposed suppression procedure is:

$$d_i = \begin{cases} 0 & \text{if } |d_i| < \delta, \\ d_i + \delta & \text{if } d_i < 0 \text{ and } |d_i| > \delta, \\ d_i - \delta & \text{if } d_i > 0 \text{ and } |d_i| > \delta, \end{cases}$$

where  $d_i$  is the detail coefficient element of the original matrix and  $\delta$  is a predefined positive threshold value.

With this coefficient suppression process, we use the inverse discrete wavelet transformation (IDWT) on the approximation coefficients and modified detail coefficients to transform the data matrix back to the original space to obtain a new data matrix,  $S^*$ , which has the same dimension as the original data matrix  $S$ , but with different attribute values. The new data matrix not only preserves the data utilities such as classes and patterns, but also prevents intruders from guessing the original attribute values from the distorted matrix.

### 3.4 Breach Algorithm and Multi-Basis Wavelet Transformation

The single basis wavelet transformation distortion algorithm may efficiently prevent the public from guessing the true data values. In this section, however, we present a breach exploitation algorithm under a special circumstance of collaborative analysis.

For example, two parties of Company A and Company B reach an agreement to mutually share their business data. Before sharing the data, each company distorts its own original data using the process described in Sections 3.2 and 3.3. Then Company A and Company B exchange their distorted datasets. After the exchange, each

company has its original true data and the partner's distorted data. In the single basis wavelet distortion scenario, although it is very difficult for each company to guess the exactly true values of the other's distorted data, exploiting the true range of the partner's distorted data may be possible by using our breach exploitation algorithm.

The main reason why the privacy is kept and data is distorted in the wavelet transformation is that we suppress (modify) the detail coefficients of the original data. In other words, if we could reinstate the values of the modified detail coefficients, we can obtain the same exact data range as the original one. So, the core of our breach algorithm is that we step by step increase the detail coefficients of the distorted data, until we reach one that gives us the best analysis result, regulated by the analysis result of the company's own original data. The detailed procedure is presented in Algorithm 1.

---

**Algorithm 1** Breach Algorithm in the Single Basis Wavelet Transformation Scenario.

---

- 1: Assume matrix  $A$  is the original data of Company A, and matrix  $B^*$  is the distorted data of Company B
  - 2: Do the collaborative analysis on the combination of  $A$  and  $B^*$ , the analysis result corresponding to the matrix  $A$  is used as the guideline
  - 3: Wavelet decompose matrix  $B^*$  to get approximation coefficients,  $coApp$ , and detail coefficients,  $coDet$
  - 4: Assign  $t_{end}$  a large enough value and  $t_{step}$  the increasing value at each step.
  - 5:  $t = 0$
  - 6: **while**  $t < t_{end}$  **do**
  - 7:   Increase each  $coDet$  value by  $t$  if the value is positive, or decrease each  $coDet$  value by  $t$  if the value is negative. Then obtain the new detail coefficients,  $coDet^*$ .
  - 8:   Transform the intermediate matrix back to a new matrix  $B_t$  by using IDWT on the approximation coefficients and the new detail coefficients.
  - 9:   Do collaborative analysis on the combination of  $A$  and  $B_t$ , and then compare the new analysis result corresponding to the matrix  $A$  with the guideline.
  - 10:   Increase  $t$  by  $t_{step}$ .
  - 11: **end while**
- 

In Algorithm 1, the essential idea is to assume that Company A can correctly guess the wavelet basis used

by Company B to distort its data, and use a trials-and-errors procedure to figure out the range of the threshold value used to suppress true values of the detail coefficients. Since the number of existing wavelet basis is limited to not very many [29], this procedure is likely to succeed.

In Section 4, our experimental results will show that the above breach algorithm can efficiently estimate the approximate values of the distorted data which fall into a very close range to the true values of the original data.

The nature of the success of the breach algorithm is largely due to the fact that a single basis wavelet and a single suppression threshold value are used on the entire data matrix. There are several strategies that will make the exploitation of the breach algorithm much more difficult, if not impossible.

One way of increasing the difficulty of guessing the suppression threshold value is to use different threshold values to suppress detail coefficients of different columns or different rows. It will be difficult for Company A to adjust individual threshold values to get the best analysis result to figure out a small range for each of the threshold values.

A minor drawback of this counter-breach strategy is that the values of the thresholds are usually small. So the possible difference between these threshold values may not be very large, which provides a certain opportunity (although much more difficult than in the single threshold value case) for the breach algorithm.

We can further increase the difficulty for the breach algorithm by using multi-basis wavelets, coupled with multiple threshold values. To be specific, we can first randomly partition the data matrix into many submatrices, and use a different basis wavelet to decompose a different submatrix. We then transform each perturbed data submatrix back to the original subspace to combine them into the distorted data matrix to be outsourced.

Generally speaking, we can partition the data matrix in any way, into submatrices of blocks. To make things easier to understand and without loss of too much generality, we assume that the whole original matrix  $S$  is partitioned into  $k$  submatrices along rows or columns as the follows:

$$S = \begin{bmatrix} S_1 \\ \vdots \\ S_k \end{bmatrix},$$

or

$$S = [ S_1 \dots S_k ] .$$

Since the possibility of guessing the correct matrix partition, the possibility of guessing the correct choice of a particular basis wavelet for a particular submatrix, and the possibility of guessing a particular threshold value for a particular row or column of a particular submatrix are very remote, the use of multi-basis wavelet and multiple threshold values for data distortion can be very difficult to breach.

## 4 Experimental Results

### 4.1 Data Privacy Measures

We choose the five data distortion privacy measure metrics, VD, RP, RK, CP and CK, first defined in [31], and then in [32], to evaluate the proposed data distortion methods. We also develop a new measure metric, RangePer, to measure the preserved attribute values and the privacy range. The objective of these measure metrics is to evaluate the possibility of estimating the true values and range of the original data from the distorted data [2, 10, 21].

In brief, VD value is the ratio of the Frobenius norm of the difference between the original matrix  $S$  and the distorted matrix  $S^*$  to the Frobenius norm of  $S$ . RP value presents the ratio of the average change of ranks for all attributes to the number of total elements of the matrix. RM denotes the percentage of elements which keep their ranks of values in each column after the distortion. CP stands for change of ranks of the average values of the attributes. CK is defined to evaluate the percentage of the attributes that keep their ranks of average values after the distortion. The detailed definitions of these measure metrics are presented in [31].

For an  $a*b$  matrix, the metric RangePer is defined as:

$$\text{RangePer} = \frac{\sum_{i=1}^a \sum_{j=1}^b f(a_{i,j}, a_{i,j}^*)}{a * b},$$

where  $f(a_{i,j}, a_{i,j}^*)$  is computed as:

$$f(a_{i,j}, a_{i,j}^*) = \begin{cases} 1 & \text{if } \frac{|a_{i,j} - a_{i,j}^*|}{|a_{i,j}|} < \epsilon \\ 0 & \text{if } \frac{|a_{i,j} - a_{i,j}^*|}{|a_{i,j}|} > \epsilon \end{cases},$$

and  $a_{i,j}$  is the true value of the original data matrix,  $a_{i,j}^*$  is the corresponding value of the distorted data matrix, and  $\epsilon$  is a predefined value. The metric RangePer measures the percentage of the data elements whose relative differences between the distorted values and the true values are smaller than  $\epsilon$ .

According to their definitions, we know that a larger VD, RP, CP and RangePer value, and a smaller RK and CK value refer to a better privacy-preserving level.

### 4.2 Distortion Experiments

In the experiment section, we choose two real-life databases obtained from the University of California, Irvine (UCI), Machine Learning Repository [20]. They are the Wisconsin breast cancer original dataset (WBC) donated by Olvi Mangasarian, and the Wisconsin breast cancer diagnostic database (WDBC) donated by Nick Street. The summary of the two original databases is in Table 1.

In addition to the summary, the attributes of the two databases only have numerical values and no missing value. (In the original WBC database, there are a few missing values in the sixth column. We replace these missing values by 1 if the object belongs to the malignant class and 2 if the object is in the benign class, according to the standard classification provided by the UCI Repository.) Tables 2 and 3 demonstrate our privacy-preserving distortion experimental results. In the experiments, we choose the SVD-based data distortion method for comparison [31, 32]. We use the simplest SVD data distortion method, i.e., no sparsification strategy is implemented. For each database, we perform three wavelet transformations: the single basis wavelet transformation (S), the vertically partitioned multi-basis wavelet transformation (VP), and the horizontally partitioned multi-basis wavelet transformation (HP).

In the SVD data distortion experiment, we choose the reduced rank  $k$  value to be 5 in WBC and 15 in WDBC.

For simplicity, in the vertical and horizontal partitions, we only partition the original database into two submatrices and each submatrix is approximately a half of the original one in size.

In the single basis wavelet transformation (S) of both Tables 2 and 3, we choose the Haar basis wavelet for decomposition and reconstruction, and use 0.5 as our thresh-

Table 1: The summary of the WBC and WDBC databases.

Database	Number of Instances	Number of Features	Number of Classes
WBC	699	9	2
WDBC	569	30	2

old value ( $\delta = 0.5$ ). In the vertically (VP) and horizontally (HP) partitioned multi-basis wavelet transformations of both Tables 2 and 3, we select Haar basis wavelet for the first half partition and Daub-4 basis wavelet for the second half for decomposition and reconstruction, and also take 0.5 as our threshold value ( $\delta = 0.5$ ) in these cases.

With respect to the run time, we set two time measures, Total Time and Max Partition Time. The Total Time is the measure of the summed time of all transformations both in the single basis wavelet and the multi-basis wavelet processes. The Max Partition Time is only applied to the multi-basis wavelet transformation (in other words, there is no meaning for the SVD-based method and the single basis wavelet transformation to measure the Max Partition Time because they do not partition the data matrix), and it is the maximum transformation time of all partitions. In our particular cases, the Max Partition Time is the larger one of the run times for the two submatrix distortions.

The results of our experiments, especially the run time, 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 operation system. For the results reported in Tables 2 and 3, the support vector machine (SVM light) with a five-fold cross validation [12, 13] is employed as the standard classification tool which is used to measure the data utility accuracy in our experiments. According to Tables 2 and 3, we can draw the following conclusions:

1. The data accuracy level of the wavelet-based distortion methods is as good as that of the SVD and the original data.
2. The run time of the wavelet-based distortion methods is faster than that of the SVD-based method even in the multi-basis wavelet transformation. When the size of the dataset is larger, this advantage is more significant.

3. In the vertically and horizontally partitioned transformations, although their Total Times are larger than those of the single basis wavelet transformation, we could do parallel wavelet decomposition and suppression and reconstruction of each partition independently in distributed situations. Therefore, the Max Partition Time may be considered as a more reasonable and meaningful performance indicator than the Total Time in these cases.
4. Most of the privacy preservation metrics show that the wavelet-based distortion methods can keep a better privacy level than the standard SVD-based method. (We did not compare with the sparsified SVD data distortion method.)
5. In the three wavelet-based distortion methods (S, VP and HP), their analysis accuracy and privacy-preserving and run time performances (in terms of Max Partition Time) are similar.

### 4.3 Breach Experiments

In this part, we will show that the multi-basis wavelet distortion method is safer than the single basis wavelet distortion method in terms of the range value exploitation. We choose WBC to demonstrate the breach algorithm in a 2-party collaborative analysis situation.

The original WBC data is a  $699 \times 9$  matrix. We partition this matrix into 2 parts, a  $350 \times 9$  submatrix only owned by Company A and a  $349 \times 9$  submatrix exclusively accessible to Company B. Then, Company B decides to distort its  $349 \times 9$  submatrix only using the single Haar basis wavelet and 0.5 as the entire submatrix threshold and then gives it to Company A. Thus, Company A has the original  $350 \times 9$  submatrix and the Company B's distorted  $349 \times 9$  submatrix. Company A uses the breach algorithm proposed in Algorithm 1 to exploit Company B's original submatrix. The results of the breach algorithm are shown in Figure 2.

Table 2: Performance comparison of SVD and wavelet transformation on WBC.

Database	VD	RP	RK	CP	CK	Run Time (Seconds)		Accuracy
						Total Time	Max Partition Time	
Original								96.0%
SVD	0.2080	239.4	0.006358	1.556	0.4444	0.07882		95.9%
Wavelet(S)	0.2557	238.6	0.004769	1.333	0.5556	0.03081		96.0%
Wavelet(VP)	0.3526	247.1	0.005564	1.556	0.333	0.06362	0.03639	95.6%
Wavelet(HP)	0.3140	239.1	0.005087	2.000	0.333	0.05153	0.03089	96.1%

Table 3: Performance comparison between SVD and wavelet transformation on WDBC.

Database	VD	RP	RK	CP	CK	Run Time (Seconds)		Accuracy
						Total Time	Max Partition Time	
Original								85.4%
SVD	0.000035	121.3	0.3454	0	1.0000	0.13880		85.4%
Wavelet(S)	0.000843	165.3	0.1083	4.800	0.4000	0.05166		85.4%
Wavelet(VP)	0.001011	168.6	0.1041	4.733	0.4667	0.09274	0.05223	85.4%
Wavelet(HP)	0.000962	165.5	0.1141	3.267	0.4667	0.08177	0.05350	85.4%

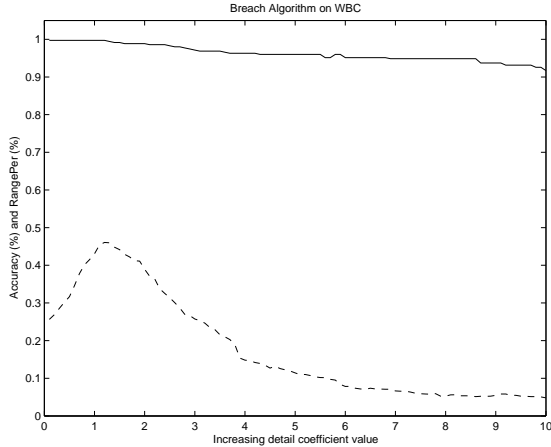


Figure 2: Breach exploitation results with the single basis wavelet transformation.

In Figure 2, we step by step increase the detail coefficients of Company B's distorted matrix by the value from 0 to 10. The step increasing value used in these experiments is 0.1. The solid line in the figure represents the accuracy percentage of analysis result corresponding to the original submatrix  $A$ . The dashed line denotes the per-

centage of the elements in Company B's distorted matrix which are fallen in the  $1 \pm 0.15$  ( $\epsilon = 0.15$ ) range of the true value of the original data submatrix  $B$ .

From this graph, we can see that the classification accuracy is mostly decreasing from the beginning to the end. The RangePer value is first increasing and then decreasing smoothly. The peak value of RangePer appears at near 1.2, rather than at the beginning stage. After reaching the peak value, the RangePer value is decreasing smoothly to a much lower level. The exploitation goal is to guess more and more elements which fall in the very close range of the true value. In other words, the peak value is the best obtainable exploitation which has the maximum number of close elements compared to the true values.

Therefore, Company A could reconstruct a database of Company B which has more elements in the close range of the true values than the distorted matrix when the accuracy of the classification result is at a comparatively high level, namely at a high level in the accuracy line.

However, such a breach exploitation process will not be effective in the multi-basis wavelet distortion method. We still use the WBC database as the experimental example, and partition it as in the previous case. This time, Company B further partitions the  $349 \times 9$  submatrix into 2 parts, one is  $174 \times 9$  and another is  $175 \times 9$ . Company B distorts

the  $174 \times 9$  submatrix using the Haar basis wavelet and the 0.2 threshold value, and transforms the  $175 \times 9$  submatrix using the Doub-4 basis wavelet and the 0.6 threshold value. Then Company B combines them and gives the distorted submatrix to Company A. Thus, Company A has the original  $350 \times 9$  submatrix and Company B's distorted  $349 \times 9$  submatrix. Company A still uses the breach algorithm proposed in Algorithm 1 to exploit Company B's original matrix. The results of the breach algorithm are shown in Figure 3.

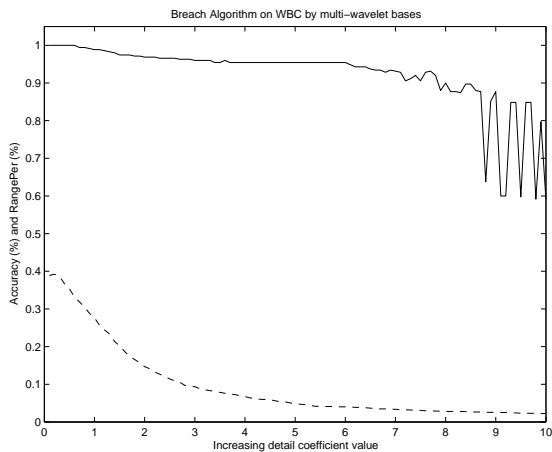


Figure 3: Breach exploitation results with the multi-basis wavelet transformation.

In Figure 3, unfortunately, it seems that no success is achieved in the exploitation process. On one hand, the accuracy line is decreasing in most cases and becomes oscillatory to the lower level at the end. On the other hand, the RangePer value is monotonically decreasing from the beginning to the end. The peak value appears at the beginning. In other words, it does not seem to be possible for Company A to exploit a closer range matrix by increasing the detail coefficients than the distorted matrix.

According to the comparison and analysis, choosing the multi-basis wavelet distortion not only can keep the data utilities and the attribute value privacy, but also may prevent the intruders from exploiting the data value range in the collaborative analysis situations.

## 5 Concluding Remarks

In this paper, we propose a class of new privacy preserving data distortion methods based on wavelet transformation. Through experiments, we demonstrate that the wavelet-based data distortion methods, especially the multi-basis wavelet transformation, can effectively and efficiently render a balance between data utilities and data privacy beyond its remarkable fast run time in comparison with the SVD-based distortion method which has already been demonstrated as a promising privacy preserving data distortion method [31].

In addition, we provide a new privacy breach algorithm in the collaborative analysis situations which could threaten the data privacy, even with the distorted values, without knowing the relationship between the distorted matrix and original matrix. Based on this observation, we propose a multi-basis wavelet transformation for enhanced data distortion and compare the single basis wavelet distortion method with the multi-basis wavelet distortion method in terms of the data value range.

Our major contributions in this paper can be summarized as follows:

1. We use discrete wavelet transformation (DWT) to simultaneously decompose the original data into approximation coefficients and detail coefficients, and then suppress the high frequency detail coefficients to achieve data distortion.
2. More importantly, we use inverse discrete wavelet transformation (IDWT) on approximation coefficients and suppressed detail coefficients to transform back the dataset to keep the same dimension as the original dataset. Clearly, the purpose of keeping dimension is to facilitate collaborative analysis in two cases under our consideration. In the first scenario, the dimension of the customer sets must be the same for all parties. In the second scenario, the dimension of the attribute sets has to be the same for all parties.
3. We find that the classification analysis results of the distorted data using only single basis wavelet and the partitioned distorted data using multi-basis wavelet are as good as that of the original one.

4. According to our experimental results, it may not be very safe for people to share their data and protect the data privacy by using only a single basis wavelet transformation. Otherwise, decoding the range of one party's original data is feasible by using other party's original data.
5. With respect to the complexity of the run time, the wavelet transformation is of  $O(n)$  which is significantly smaller than most of the other data distortion algorithms such as SVD-transformation [31] and FFT-transformation [30].

Further research work along this line can be carried out to select more appropriate threshold values for different databases. The better choice of the threshold values should be related to the size of the data elements, i.e., a relative threshold value should be computed for each column, according to a certain norm of the column. There is also interest in studying which wavelet basis achieves the best performance in terms of data distortion and data utilities. Comparing the wavelet transformation based data distortion methods with the more aggressive SVD-based data distortion methods, such as the sparsified SVD-based methods [31, 32], should be of interest.

## References

- [1] *Standard for privacy of individually identifiable health information* [<http://www.hhs.gov/ocr/hipaa/finalmaster.html>]. Federal Register, 66(40), 2001.
- [2] D. Agrawal and C. C. Aggarwal. *On the design and quantification of privacy preserving data mining algorithms*. In Proceedings of the 20th ACM SIGACT-SIGMODSIGART Symposium on Principles of Database Systems, pp. 247-255, Santa Barbara, CA, 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] 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, 2005.
- [5] 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.
- [6] R. Coifman, Y. Meyer and V. Wickerhauser. *Wavelet analysis and signal processing*. In: Wavelets and Their Applications, Edited by M. B. Ruskai, Jones and Bartlett Publishers, Sudbury, MA, 1991.
- [7] L. Cranor. *Special issue on internet privacy*. Communications of the ACM, 42(2):28-38, 1999.
- [8] D. Donoho. *Nonlinear wavelet methods for recovery of signals, densities and spectra from indirect and noisy data*. In Proceedings of Symposia in Applied Mathematics, American Mathematical Society, 47:173-205, 1993.
- [9] A. Evfimievski. *Randomization in privacy preserving data mining*. ACM SIGKDD Explorations Newsletter, 4(2):43-48, 2002.
- [10] A. Evfimievski, J. Gehrke and R. Srikant. *Limiting privacy breaches in privacy preserving data mining*. In Proceedings of PODS 2003, pp. 211-222, San Diego, CA, 2003.
- [11] 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.
- [12] T. Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*. Kluwer Academic Publisher, Norwell, MA, 2002.
- [13] T. Joachims. *Making large-scale SVM learning practical*. In Advances in Kernel Methods - Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), MIT Press, Cambridge, MA, 1999.

- [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, 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] S. Meregu and J. Ghosh. *Privacy-preserving distributed clustering using generative models*. In Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), pp. 211-218, Melbourne, FL, 2003.
- [17] A. Meyerson and R. Williams. *General k-anonymization is hard*. Carnegie Mellon University, School of Computer Science Tech Report, 03-113, 2003.
- [18] S. Mukherjee, Z. Chen and A. Gangopadhyay. *A privacy preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms*. The VLDB Journal, 15(4):293-315, 2006.
- [19] K. Muralidhar and R. Sarathy. *Security of random data perturbation methods*. ACM Transactions on Database Systems, 24(4):487-493, 1999.
- [20] D. J. Newman, S. Hettich, C. L. Blake and C. J. Merz. *UCI repository of machine learning databases* [<http://www.ics.uci.edu/mllearn/MLRepository.html>]. University of California, Department of Information and Computer Science, Irvine, CA, 1998.
- [21] H. Polat and W. Du. *SVD-based collaborative filtering with privacy*. In the 20th ACM Symposium on Applied Computing, Track on E-commerce Technologies, pp. 791-795, Santa Fe, NM, 2005.
- [22] N. Saito. *Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion*. In: Wavelets in Geophysics, Foufoula-Georgiou and Kumar (eds.), pp. 224-235, Academic Press, Burlington, MA, 1994.
- [23] L. Sweeney. *Achieving k-anonymity privacy protection using generalization and suppression*. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):571-588, 2002.
- [24] L. Sweeney. *Guaranteeing anonymity when sharing medical data, the DataFly system*. Journal of the American Medical Informatics Association, Suppl. S, pp. 51-55, 1997.
- [25] J. Vaidya and C. Clifton. *Privacy-preserving K-means clustering over vertically partitioned data*. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 206-215, Washington, DC, 2003.
- [26] K. Wang, B. C. M. Fung and G. Dong. *Integrating private databases for data analysis*. In Proceedings of the 2005 IEEE International Conference on Intelligence and Security Informatics (ISI 2005), pp. 171-182, Atlanta, GA, 2005.
- [27] K. Wang, P. S. Yu, and S. Chakraborty. *Bottom-up generalization: a data mining solution to privacy protection*. In Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 2004), pp. 249-256, 2004.
- [28] J. Wang, W. J. Zhong and J. Zhang. *NNMF-based factorization techniques for high-accuracy privacy protection on non-negative-valued datasets*. In Proceedings of the 2006 IEEE Conference on Data Mining, International Workshop on Privacy Aspects of Data Mining (PADM 2006), pp. 513-517, Hong Kong, China, 2006.
- [29] M. Weeks and M. A. Bayoumi. *Three-dimensional discrete wavelet transform architectures*. IEEE Transactions on Signal Processing, 50(8):2050-2063, 2002.
- [30] 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.

- [31] 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.
- [32] 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.
- [33] S. Zhong, Z. Yang and R. N. Wright. *Privacy-enhancing k-anonymization of customer data*. In Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 139-147, 2005.