

# **Matrix Decomposition-Based Data Distortion Techniques for Privacy Preservation in Data Mining**

**Jun Zhang\*, Jie Wang**

Department of Computer Science  
University of Kentucky  
773 Anderson Tower  
Lexington, KY 40506-0046  
USA

voice: +1 859-257-3892

fax: +1 859-323-1971

email: [jzhang@cs.uky.edu](mailto:jzhang@cs.uky.edu), [jwanga@csr.uky.edu](mailto:jwanga@csr.uky.edu)

**Shuting Xu**

Department of Computer Information Systems  
Virginia State University  
Petersburg, VA 23806  
USA

voice: +1 804-524-6719

email: [sxu@vsu.edu](mailto:sxu@vsu.edu)

(\* Corresponding author)

**Technical Report TR 472-07, Department of Computer Science,  
University of Kentucky, Lexington, KY, 2007**

# Matrix Decomposition-Based Data Distortion Techniques for Privacy Preservation in Data Mining

Jun Zhang and Jie Wang, University of Kentucky, USA  
Shuting Xu, Virginia State University, USA

## INTRODUCTION

Data mining technologies have now been used in commercial, industrial, and governmental businesses, for various purposes, ranging from increasing profitability to enhancing national security. The widespread applications of data mining technologies have raised concerns about trade secrecy of corporations and privacy of innocent people contained in the datasets collected and used for the data mining purpose. It is necessary that data mining technologies designed for knowledge discovery across corporations and for security purpose towards general population have sufficient privacy awareness to protect the corporate trade secrecy and individual private information. Unfortunately, most standard data mining algorithms are not very efficient in terms of privacy protection, as they were originally developed mainly for commercial applications, in which different organizations collect and own their private databases, and mine their private databases for specific commercial purposes.

In the cases of inter-corporation and security data mining applications, data mining algorithms may be applied to datasets containing sensitive or private information. Data warehouses and government agencies may potentially have access to many databases collected from different sources and may extract any information from these databases. This potentially unlimited access to data and information raises the fear of possible abuse and promotes the call for privacy protection and due process of law.

Privacy-preserving data mining techniques have been developed to address these concerns. The general goal of the privacy-preserving data mining techniques is defined as to hide sensitive individual data values from the outside world or from unauthorized persons, and simultaneously preserve the underlying data patterns and semantics so that a valid and efficient decision model based on the distorted data can be constructed. In the best scenarios, this new decision model should be equivalent to or even better than the model using the original data from the viewpoint of decision accuracy. There are currently at least two broad classes of approaches to achieving this goal. The first class of approaches attempts to distort the original data values so that the data miners (analysts) have no means (or greatly reduced ability) to derive the original values of the data. The second is to modify the data mining algorithms so that they allow data mining operations on distributed datasets without knowing the exact values of the data or without direct accessing the original datasets. This article only discusses the first class of approaches. Interested readers may consult (Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., and Zhu, M., 2003) and the references therein for discussions on distributed data mining approaches.

## BACKGROUND

The input to a data mining algorithm in many cases can be represented by a vector-space model, where a collection of records or objects is encoded as an  $n \times m$  object-attribute matrix (Frankes, & Baeza-Yates, 1992). For example, the set of vocabulary (words or terms) in a dictionary can be the items forming the rows of the matrix, and the occurrence frequencies of all terms in a document are listed in a column of the matrix. A collection of documents thus forms a

term-document matrix commonly used in information retrieval. In the context of privacy-preserving data mining, each column of the data matrix can contain the attributes of a person, such as the person's name, income, social security number, address, telephone number, medical records, etc. Datasets of interest often lead to a very high dimensional matrix representation (Achlioptas, 2004). It is observable that many real-world datasets have nonnegative values for attributes. In fact, many of the existing data distortion methods inevitably fall into the context of matrix computation. For instance, having the longest history in privacy protection area and by adding random noise to the data, additive noise method can be viewed as a random matrix and therefore its properties can be understood by studying the properties of random matrices (Kargupta, Sivakumar, & Ghosh, 2002; Mahta, 1991).

Matrix decomposition in numerical linear algebra typically serves the purpose of finding a computationally convenient means to obtain a solution to a linear system. In the context of data mining, the main purpose of matrix decomposition is to obtain some form of simplified low-rank approximation to the original dataset for understanding the structure of the data, particularly the relationship within the objects and within the attributes and how the objects relate to the attributes (Hubert, Meulman, & Heiser, 2000). The study of matrix decomposition techniques in data mining, particularly in text mining, is not new, but the application of these techniques as data distortion methods in privacy-preserving data mining is a recent interest (Xu, Zhang, Han, & Wang, 2005). A unique characteristic of the matrix decomposition techniques, a compact representation with reduced-rank while preserving dominant data patterns, stimulates researchers' interest in utilizing them to achieve a win-win task both on high degree privacy-preserving and high level data mining accuracy.

## MAIN FOCUS

Data distortion is one of the most important parts in many privacy-preserving data mining tasks. The desired distortion methods must preserve data privacy, and at the same time, must keep the utility of the data after the distortion (Verykios, Bertino, Fovino, Provenza, Saygin, & Theodoridis, 2004). The classical data distortion methods are based on the random value perturbation (Agrawal, & Srikant, 2000). The more recent ones are based on the data matrix-decomposition strategies (Wang, Zhong, & Zhang, 2006; Wang, Zhang, Zhong, & Xu, 2007; Xu, Zhang, Han, & Wang, 2006).

### Uniformly Distributed Noise

The original data matrix  $A$  is added with a uniformly distributed noise matrix  $E_u$ . Here  $E_u$  is of the same dimension as that of  $A$ , and its elements are random numbers generated from a continuous uniform distribution on the interval from  $C_1$  to  $C_2$ . The distorted data matrix  $A_u$  is denoted as:  $A_u = A + E_u$ .

### Normally Distributed Noise

Similar to the previous method, here the original data matrix  $A$  is added with a normally distributed noise matrix  $E_n$ , which has the same dimension as that of  $A$ . The elements of  $E_n$  are random numbers generated from the normal distribution with a parameter mean  $\mu$  and a standard deviation  $\rho$ . The distorted data matrix  $A_n$  is denoted as:  $A_n = A + E_n$ .

## Singular Value Decomposition

Singular Value Decomposition (SVD) is a popular matrix factorization method in data mining and information retrieval. It has been used to reduce the dimensionality of, (and remove the noise in the noisy), datasets in practice (Berry, Drmac, & Jessup, 1999). The use of SVD technique in data distortion is proposed in (Xu, Zhang, Han, & Wang, 2005). In (Wang, Zhang, Zhong, & Xu, 2007), the SVD technique is used to distort portions of the datasets.

The SVD of the data matrix  $A$  is written as

$$A = U\Sigma V^T$$

where  $U$  is an  $n \times n$  orthonormal matrix,  $\Sigma = \text{diag}[\sigma_1, \sigma_2, \dots, \sigma_s]$  ( $s = \min\{m, n\}$ ) is an  $n \times m$  diagonal matrix whose nonnegative diagonal entries (the singular values) are in a descending order, and  $V^T$  is an  $m \times m$  orthonormal matrix. The number of nonzero diagonal entries of  $\Sigma$  is equal to the rank of the matrix  $A$ .

Due to the arrangement of the singular values in the matrix  $\Sigma$  (in a descending order), the SVD transformation has the property that the maximum variation among the objects is captured in the first dimension, as  $\sigma_1 \geq \sigma_i$  for  $i \geq 2$ . Similarly, much of the remaining variations is captured in the second dimension, and so on. Thus, a transformed matrix with a much lower dimension can be constructed to represent the structure of the original matrix faithfully. Define

$$A_k = U_k \Sigma_k V_k^T$$

where  $U_k$  contains the first  $k$  columns of  $U$ ,  $\Sigma_k$  contains the first  $k$  nonzero singular values, and  $V_k^T$  contains the first  $k$  rows of  $V^T$ . The rank of the matrix  $A_k$  is  $k$ . With  $k$  being usually small, the dimensionality of the dataset has been reduced dramatically from  $\min\{m, n\}$  to  $k$  (assuming all attributes are linearly independent). It has been proved that  $A_k$  is the best  $k$  dimensional approximation of  $A$  in the sense of the Frobenius norm.

In data mining applications, the use of  $A_k$  to represent  $A$  has another important implication. The removed part  $E_k = A - A_k$  can be considered as the noise in the original dataset (Xu, Zhang, Han, & Wang, 2006). Thus, in many situations, mining on the reduced dataset  $A_k$  may yield better results than mining on the original dataset  $A$ . When used for privacy-preserving purpose, the distorted dataset  $A_k$  can provide protection for data privacy, at the same time, it keeps the utility of the original data as it can faithfully represent the original data structure.

## Sparsified Singular Value Decomposition

After reducing the rank of the SVD matrices, we can further distort the data matrices by removing their small size entries. This can be done with a threshold strategy. Given a threshold value  $\varepsilon$ , we set any data entry in the matrices  $U_k$  and  $V_k^T$  to be zero if its absolute value is smaller than  $\varepsilon$ . We refer to this operation as the dropping operation (Gao, & Zhang, 2003). For example, we set  $u_{ij} = 0$  in  $U_k$  if  $|u_{ij}| < \varepsilon$ . Similar operation is applied to the entries in  $V_k^T$ . Let  $\overline{U}_k$  denote  $U_k$  with dropped entries and  $\overline{V}_k^T$  denote  $V_k^T$  with dropped entries, we can represent the distorted dataset as

$$\overline{A}_k = \overline{U}_k \Sigma_k \overline{V}_k^T$$

The sparsified SVD method is equivalent to further distorting the dataset  $A_k$ . Denoting  $E_\epsilon = A_k - \overline{A_k}$ , we have

$$A = \overline{A_k} + E_k + E_\epsilon$$

The dataset provided to the data mining analysts is  $\overline{A_k}$ , which is twice distorted in the sparsified SVD strategy. Without the knowledge of  $E_k$  and  $E_\epsilon$ , it will be difficult for the data mining analysts to recover the exact values of  $A$ , based on the disclosed values of  $A_k$ .

### Nonnegative Matrix Factorization

Given an  $n \times m$  nonnegative matrix dataset  $A$  with  $A_{ij} \geq 0$  and a prespecified positive integer  $k \leq \min\{n, m\}$ , the nonnegative matrix factorization (NMF) finds two nonnegative matrices  $W \in R^{n \times k}$  with  $W_{ij} \geq 0$  and  $H \in R^{k \times m}$  with  $H_{ij} \geq 0$ , such that  $A \approx WH$  and the objective function

$$f(W, H) = \frac{1}{2} \|A - WH\|_F^2$$

is minimized. Here  $\|\cdot\|_F$  is the Frobenius norm. The matrices  $W$  and  $H$  may have many other desirable properties in data mining applications. Several algorithms to compute nonnegative matrix factorizations for some applications of practical interests are proposed in (Lee, & Seung, 1999; Pascual-Montano, Carazo, Kochi, Lehmann, & Pascual-Marqui, 2006). Some of these algorithms are modified in (Wang, Zhong, & Zhang, 2006) to compute nonnegative matrix factorizations for enabling privacy-preserving in datasets for data mining applications. Similar to the sparsified SVD techniques, sparsification techniques can be used to drop small size entries from the computed matrix factors to further distort the data values (Wang, Zhong, & Zhang, 2006).

In text mining, NMF has an advantage over SVD in the sense that if the data values are nonnegative in the original dataset, NMF maintains their nonnegativity, but SVD does not. The nonnegativity constraints can lead to a parts-based representation because they allow only additive, not subtractive, combinations of the original basis vectors (Lee, & Seung, 1999). Thus, dataset values from NMF have some meaningful interpretations in the original sense. On the contrary, data values from SVD are no longer guaranteed to be nonnegative. There has been no obvious meaning for the negative values in the SVD matrices. In the context of privacy-preserving, on the other hand, the negative values in the dataset may actually be an advantage, as they further obscure the properties of the original datasets.

### Utility of the Distorted Data

Experimental results obtained in (Wang, Zhang, Zhong, & Xu, 2007; Wang, Zhong, & Zhang, 2006; Xu, Zhang, Han, & Wang, 2006; Xu, Zhang, Han, & Wang, 2005), using both synthetic and real-world datasets with a classification algorithm, show that both SVD and NMF techniques provide much higher degree of data distortion than the standard data distortion techniques based on adding uniformly distributed noise or normally distributed noise. In terms of the accuracy of the data mining algorithm, techniques based on adding uniformly distributed noise or normally distributed noise sometimes degrade the accuracy of the classification results, compared with applying the algorithm on the original, undistorted datasets. On the other hand, both SVD and NMF techniques can generate distorted datasets that are able to yield better

classification results, compared with applying the algorithm directly on the original, undistorted datasets. This is amazing, as we intuitively expect that data mining algorithms applied on the distorted datasets may produce less accurate results, than applied on the original datasets.

It is not clear why the distorted data from SVD and NMF are better for the data classification algorithm used to obtain the experimental results. The hypothesis is that both SVD and NMF may have some functionalities to remove the noise from the original datasets by removing small size matrix entries. Thus, the distorted datasets from SVD and NMF look like “cleaned” datasets. The distorted datasets from the techniques based on adding either uniformly distributed noise or normally distributed noise do not have this property. They actually generate “noisy” datasets in order to distort data values.

## **FUTURE TRENDS**

Using matrix decomposition-based techniques in data distortion for privacy-preserving data mining is a relatively new trend. This class of data privacy-preserving approaches has many desirable advantages over the more standard privacy-preserving data mining approaches. There are a lot of unanswered questions in this new research direction. For example, a classical problem in SVD-based dimensionality reduction techniques is to determine the optimal rank of the reduced dataset matrix. Although in the data distortion applications, the rank of the reduced matrix does not seem to sensitively affect the degree of the data distortion or the level of the accuracy of the data mining results (Wang, Zhang, Zhong, & Xu, 2007), it is still of both practical and theoretical interests to be able to choose a good rank size for the reduced data matrix.

Unlike the data distortion techniques based on adding either uniformly distributed noise or normally distributed noise, SVD and NMF does not maintain some statistical properties of the original datasets, such as the mean of the data attributes. Such statistical properties may or may not be important in certain data mining applications. It would be desirable to design some matrix decomposition-based data distortion techniques that maintain these statistical properties.

The SVD and NMF data distortion techniques have been used with the support vector machine based classification algorithms (Xu, Zhang, Han, & Wang, 2006). It is not clear if they are equally applicable to other data mining algorithms. It is certainly of interest for the research community to experiment these data distortion techniques with other data mining algorithms.

There is also a need to develop certain techniques to quantify the level of data privacy preserved in the data distortion process. Although some measures for data distortion and data utility are defined in (Xu, Zhang, Han, & Wang, 2006), they are not directly related to the concept of privacy-preserving in datasets.

## **CONCLUSION**

We have presented two classes of matrix decomposition-based techniques for data distortion to achieve privacy-preserving in data mining applications. These techniques are based on matrix factorization techniques commonly practiced in matrix computation and numerical linear algebra. Although their application in text mining is not new, their application in data distortion with privacy-preserving data mining is a recent attempt. Previous experimental results have demonstrated that these data distortion techniques are highly effective for high accuracy privacy protection, in the sense that they can provide high degree of data distortion and maintain high level data utility with respect to the data mining algorithms.

The computational methods for SVD and NMF are well developed in the matrix computation community. Very efficient software packages are available either in standard matrix computation packages such as MATLAB or from several websites maintained by individual researchers. The availability of these software packages greatly accelerates the application of these and other matrix decomposition and factorization techniques in data mining and other application areas.

## REFERENCES

- Achlioptas, D. (2004). Random matrices in data analysis. *Proceedings of the 15<sup>th</sup> European Conference on Machine Learning*, pp. 1-8, Pisa, Italy.
- Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 439-450, Dallas, TX.
- Berry, M. W., Drmac, Z., & Jessup, E. R. (1999). Matrix, vector space, and information retrieval. *SIAM Review*, 41, 335-362.
- Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., and Zhu, M. (2003). Tools for privacy preserving distributed data mining. *ACM SIGKDD Explorations*, 4(2), 1-7.
- Frankes, W., & Baeza-Yates, R. (1992). *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Englewood Cliffs, NJ.
- Gao, J., & Zhang, J. (2003). Sparsification strategies in latent semantic indexing. *Proceedings of the 2003 Text Mining Workshop*, pp. 93-103, San Francisco, CA.
- Hubert, L., Meulman, J., & Heiser, W. (2000). Two purposes for matrix factorization: a historical appraisal. *SIAM Review*, 42(4), 68-82.
- Kargupta, H., Sivakumar, K., & Ghosh, S. (2002). Dependency detection in mobility and random matrices. *Proceedings of the 6<sup>th</sup> European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 250-262, Helsinki, Finland.
- Lee, D. D., & Seung, H. S. (1999). Learning in parts of objects by non-negative matrix factorization. *Nature*, 401, 788-791.
- Mahta, M. L. (1991). *Random Matrices*. 2<sup>nd</sup> edition. Academic, London.
- Pascual-Montano, A., Carazo, J. M., Kochi, K., Lehmann, D., & Pascual-Marqui, P. D. (2006). Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28, 403-415.
- Verykios, V.S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record*, 3(1), 50-57.
- Wang, J., Zhong, W. J., & Zhang, J. (2006). NNMF-based factorization techniques for high-accuracy privacy protection on non-negative-valued datasets. *Proceedings of the IEEE Conference on Data Mining 2006, International Workshop on Privacy Aspects of Data Mining (PADM 2006)*, pp. 513-517, Hong Kong, China.
- Wang, J., Zhang, J., Zhong, W. J., & Xu, S. (2007). A novel data distortion approach via selective SSVD for privacy protection. *International Journal of Information and Computer Security*, to appear.
- Xu, S., Zhang, J., Han, D., & Wang, J. (2006). Singular value decomposition based data distortion strategy for privacy protection. *Knowledge and Information Systems*, 10(3), 383-397.

Xu, S., Zhang, J., Han, D., & Wang, J. (2005). Data distortion for privacy protection in a terrorist analysis system. *Proceedings of the 2005 IEEE International Conference on Intelligence and Security Informatics*, pp. 459-464, Atlanta, GA.

## **KEY TERMS AND THEIR DEFINITIONS**

**Privacy-Preserving Data Mining:** Extracting valid knowledge and information from the datasets without learning the underlying data values or data patterns, or without revealing the values or patterns of the private data.

**Matrix Decomposition:** A factorization of a matrix into some canonical form, usually in the form of a product of two or more matrices.

**Singular Value Decomposition:** The factorization of a rectangular matrix into the product of three matrices. The first and the third matrices are orthonormal. The second matrix is diagonal and contains the singular values of the original matrix.

**Nonnegative Matrix Factorization:** A class of algorithms that factor a (usually nonnegative) matrix into the product of two matrices, both have nonnegative entries. This type of factorization of matrices is not unique by incorporating different constraints.

**Data Distortion:** A systematic perturbation of data values in a database in order to mask the original data values, but allow certain properties of the database to be preserved.

**Data Utility:** A dataset's ability to maintain its performance with data mining algorithms after the data distortion process.