

Data Distortion for Privacy Protection in a Terrorist Analysis System ^{*}

Shuting Xu ^{**}, Jun Zhang ^{***}, Dianwei Han [†], and Jie Wang [‡]

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

Abstract. Privacy-preserving is a major concern in the application of data mining techniques to datasets containing personal, sensitive, or confidential information. Data distortion is a critical component to preserve privacy in security-related data mining applications, such as in data mining-based terrorist analysis systems. We propose a sparsified Singular Value Decomposition (SVD) method for data distortion. We also put forth a few metrics to measure the difference between the distorted dataset and the original dataset and the degree of the privacy protection. Our experimental results using synthetic and real world datasets show that the sparsified SVD method works well in preserving privacy as well as maintaining utility of the datasets.

1 Introduction

With the widespread availability of modern computing technology, the advance of fast data collection techniques, and the affordability of vast volume of data storage devices, data of various kinds are collected at an unprecedented speed and scale. The need for understanding and making use of the collected data sparks renewed interest in studying and developing data mining techniques, i.e., the use of computer-aided statistical techniques to “comb” through large amount of data for automatic and semi-automatic exploration and pattern discovery. Today, data is one of the most important corporate assets of companies, governments, and research institutions [10] and is used for various private and public interest.

The use of data mining technologies in counterterrorism and homeland security has been flourishing since the U.S. Government encouraged the use of such technologies [22]. However, government access to and use of personal information in commercial databases raises concerns about the protection of privacy and due

^{*} Technical Report No. 432-05, Department of Computer Science, University of Kentucky, Lexington, KY, 2005. This research work was supported by the Kentucky New Economy Safety and Security (NESSI) Consortium.

^{**} E-mail: sxu2@csr.uky.edu.

^{***} The corresponding author. E-mail: jzhang@cs.uky.edu, URL: <http://www.cs.uky.edu/~jzhang>.

[†] E-mail: dianweih@csr.uky.edu.

[‡] E-mail: jwanga@csr.uky.edu.

process [9]. Recent privacy criticism from libertarians on DARPA's ¹ Terrorism Information Awareness Program led to the defunding of DARPA's Information Awareness Office. Thus, it is necessary that data mining technologies designed for counterterrorism and security purpose have sufficient privacy awareness to protect the privacy of innocent people. Unfortunately, most existing data mining technologies are not very efficient in terms of privacy protections, as they were originally developed mainly for commercial applications, in which different organizations collect and own their databases, and mine their databases for specific commercial purposes. In the cases of security and counterterrorism, data mining may mean a totally different thing. Government may potentially have access to any databases and may extract any information from these databases. This potentially unlimited access to data and information raises the fear of possible abuse.

Data can be collected at a centralized location or collected at different locations, but integrated at a centralized location (data warehousing). Alternatively, data can be collected and stored at distributed locations. Different data storage patterns may have different privacy concerns. If the data storage is centralized, the major privacy concern is to shield the exact values of the attributes from the data analysts. Thus, data distortion is a technique that is usually considered in such a situation [1, 16]. On the other hand, in a distributed database situation, the major privacy concern is to maintain independence of the distributed data ownership and to prevent the exchange of exact values of the attributes between different parties of the distributed database ownership. This concern is related to the issue of data mining in a distributed environment [3, 13]. This paper deals with the first situation, i.e., we study data distortion techniques for a centralized database.

We propose a class of methods for privacy protection in data processing that may be used in some terrorist analysis systems and other data mining applications. We assume that the vector-space model [11] is used to build the population datasets for analysis. A dataset can be represented by a matrix A . Each row of the matrix represents an object, and each column of the matrix represents an attribute. In modeling populations with individual persons, the dataset matrix is usually sparse, as many of the attributes are not taken by most of the objects simultaneously. The objects can be individual persons of the general population. The attributes can be a person's name, address, age, home address, credit card numbers, etc. Thus, information contained in such datasets is highly confidential. The confidentiality of the personal information should not be compromised in the process of data mining applications.

In order to preserve data privacy, we assume that no one except the data owner or authorized users have the right to access the original data. The analysts will only see the distorted dataset matrix \bar{A} , not the original dataset A . The distorted dataset matrix does not have an obvious meaning for the individual attributes. The matrix \bar{A} cannot be used to reconstruct the original matrix A ,

¹ DARPA stands for Defense Advanced Research Projects Agency, affiliated with the Department of Defense of the United States [23].

without knowing the error part $E = A - \bar{A}$. In this way, the analysts, who will run the data mining algorithms on the distorted dataset matrix \bar{A} , will not be able to know the original attributes or the distribution of the attributes, unless appropriate permission is granted by higher level officials to do so. Thus, data mining techniques applied on the distorted datasets will maintain the inherent property of privacy protection.

In this paper, we will discuss several data distortion methods that can be used in protecting privacy in some terrorist analysis systems. We propose a sparsified Singular Value Decomposition (SVD) method for data distortion. There are some publications about using SVD-related methods in counterterrorism data mining techniques, such as in detecting local correlation [18], social network analysis [19], novel information discovery [20] and information extraction [21], etc. However, to the best of our knowledge, there has been no work on using SVD-related methods in data distortion. We also propose some metrics to measure the difference between the distorted dataset and the original dataset and the degree of privacy protection. Our experimental results using both synthetic and real world datasets will show that the sparsified SVD method is very efficient in keeping both data privacy and data utility.

The structure of the paper is as follows: In Section 2, we introduce a simplified model of terrorist analysis system with privacy protection, some data distortion methods and the proposed sparsified SVD method. We also put forth some privacy measure metrics in Section 3. We briefly introduce the data utility measure in Section 4. The computational experiments are carried out and the results are discussed in Section 5. We sum up this paper in Section 6.

2 Analysis System and Data Distortion

2.1 A Simplified Model Terrorist Analysis System

A simplified model terrorist analysis system can be consisted of two parts, the data manipulation part and the data analysis part. As illustrated in Figure 1, only the data owner or authorized users can manipulate the original data. After the data distortion process, the original dataset is transformed into a completely different data matrix and is provided to the analysts. All actions in the data analysis part are operated on the distorted data matrix. For example, the analysts can apply data mining techniques such as classification, relationship analysis, or clustering, on the distorted data. As the data analysts have no access to the original database without the authorization of the data owner, the privacy contained in the original data is protected.

2.2 Data Distortion

Data distortion is one of the most important parts in the proposed model terrorist analysis system. The desired distortion methods must preserve the privacy, and at the same time, must keep the utility of the data after the distortion [26].

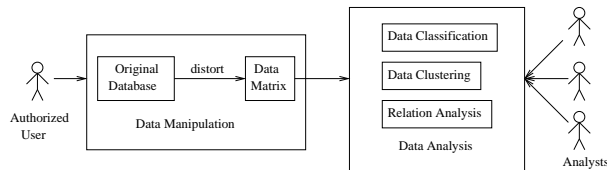


Fig. 1. Illustration of a simplified model terrorist analysis system.

Some data distortion methods based on random value have been proposed and applied in [2, 7, 16]. We will review two of the commonly used random value data distortion methods, as well as propose a class of SVD-based methods for data distortion in this section.

Uniformly distributed noise In this method, the original data matrix A is added by a uniformly distributed noise matrix N_u [2]. N_u is of the same size as A , and its elements are random numbers chosen from the continuous uniform distribution on the interval from C_1 to C_2 . The distorted matrix \bar{A}_u is: $\bar{A}_u = A + N_u$.

Normally distributed noise Similarly as the previous method, here the original data matrix A is added by a normally distributed noise matrix N_n , which has the same size as A [2]. The elements of N_n are random numbers chosen from the normal distribution with parameters mean μ and standard deviation σ . The distorted matrix \bar{A}_n is: $\bar{A}_n = A + N_n$.

SVD Singular Value Decomposition (SVD) [14] is a popular method in data mining and information retrieval [8]. It is usually used to reduce the dimensionality of the original dataset A . Here we use it as a data distortion method.

Let A be a sparse matrix of dimension $n \times m$ representing the original dataset. The rows of the matrix correspond to data objects and the columns to attributes. The singular value decomposition of the matrix A is [14]

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

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 are in a descending order, and V^T is an $m \times m$ orthonormal matrix. The number of nonzero diagonals of Σ is equal to the rank of the matrix A .

Due to the arrangement of the singular values in the matrix Σ (in a descending order), the SVD transformation has the property that the maximal 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 original matrix faithfully. Define

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

where U_k contains the first k columns of U , Σ_k contains the first k nonzero diagonals of Σ , 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 Frobenius norm.

In data mining applications, the use of A_k to represent A has another important function. The removed part $E_k = A - A_k$ can be considered as the noise in the original dataset A [4]. Thus, in many cases, mining on the reduced dataset A_k may yield better results than mining on the original dataset A . When used for privacy preserving, the distorted data A_k can provide protection for data privacy, at the same time it may keep the utility of the original data as it can faithfully represent the original data.

Sparsified SVD We propose a data distortion method that is better than SVD in preserving privacy: a sparsified SVD.

After reducing the rank of the SVD matrices, we set some small size entries, which are smaller than a certain threshold ϵ , in U_k and V_k^T , to zero. We refer to this operation as the dropping operation [12]. For example, given a threshold value ϵ , we drop u_{ij} in U_k if $|u_{ij}| < \epsilon$. Similarly, an element v_{ij} in V_k^T is also dropped if $|v_{ij}| < \epsilon$. Let \bar{U}_k denote U_k with dropped elements and \bar{V}_k^T denote V_k^T with dropped elements, we can represent the distorted data matrix \bar{A}_k , with

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

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

$$A = \bar{A}_k + E_k + E_\epsilon.$$

The data provided to the analysts is \bar{A}_k which is twice distorted in the sparsified SVD method.

The SVD sparsification concept was proposed by Gao and Zhang in [12] for reducing the storage cost and enhancing the performance of SVD in text retrieval applications. Several sparsification strategies were proposed and experimented in [12]. The one that we used in this paper is the simplest one.

3 Privacy Measures

Some privacy metrics have been proposed in literature [1, 2]. However, the metric used in [2] has been proved to be incomplete [1], and the one used in [1] needs to know the density function of each attribute *a priori*, which may be difficult to obtain for the real world datasets. We propose some privacy measures which only depend on the original matrix A and its distorted counterpart, \bar{A} .

3.1 Value Difference

After a data matrix is distorted, the value of its elements changes. The value difference (VD) of the datasets is represented by the relative value difference in the Frobenius norm. Thus VD is the ratio of the Frobenius norm of the difference of A and $|\bar{A}|$ to the Frobenius norm of A :

$$VD = \|A - |\bar{A}|\|_F / \|A\|_F.$$

For example, for the following dataset A_e , its distorted data matrix \bar{A}_e is obtained by applying the Sparsified SVD with $k = 2$ and $\epsilon = 0.001$. Then the VD value computed for this distortion is 0.3136.

$$A_e = \begin{bmatrix} 1 & 2.5 & 5 & 0.3 \\ 2 & 3.9 & 2 & 1.1 \\ 4 & 1.8 & 8 & 0.5 \\ 1 & 3.3 & 6 & 1.2 \end{bmatrix}, \quad \bar{A}_e = \begin{bmatrix} 1.7 & 0.8 & -5.3 & -0.1 \\ 0.2 & 2.8 & -3.8 & -0.8 \\ 3.6 & -0.7 & -8.3 & 0.5 \\ 1.9 & 1.4 & -6.5 & -0.2 \end{bmatrix}.$$

3.2 Position Difference

After a data distortion, the relative order of the value of the data elements changes, too. We use several metrics to measure the position difference of the data elements.

We use RP to denote the average change of order for all the attributes. After the elements of an attribute are distorted, the order of each element changes. Assume dataset A has n data objects and m attributes. Ord_j^i denotes the ascending order of the j th element in attribute i , and \overline{Ord}_j^i denotes the ascending order of the distorted element A_{ji} . Then RP is defined as:

$$RP = \left(\sum_{i=1}^m \sum_{j=1}^n |Ord_j^i - \overline{Ord}_j^i| \right) / (m * n).$$

If two elements have the same value, we define the element with the lower row index to have the higher order. In dataset A_e , the order vector for the first attribute can be represented as $Ord^1 = [1 \ 3 \ 4 \ 2]^T$. After the distortion, $\overline{Ord}^1 = [2 \ 1 \ 4 \ 3]^T$. The total change of order for this attribute is 3. We can calculate the total change of order for the other attributes and get $RP = 1.2$.

RK represents the percentage of elements that keep their orders of value in each column after the distortion. It is computed as:

$$RK = \left(\sum_{i=1}^m \sum_{j=1}^n Rk_j^i \right) / (m * n),$$

where Rk_j^i represents whether or not an element keeps its position in the order of values:

$$Rk_j^i = \begin{cases} 1, & \text{if } Ord_j^i = \overline{Ord}_j^i, \\ 0, & \text{otherwise.} \end{cases}$$

For example, the order vector of the second attribute in A_e is $[2\ 4\ 1\ 3]^T$, and after the distortion, it is still $[2\ 4\ 1\ 3]^T$. Thus all the elements keep their original order. RK for this example is 0.31.

One may infer the content of an attribute from its relative value difference compared with the other attributes. Thus it is desirable that the order of the average value of each attribute varies after the data distortion. Here we use the metric CP to define the change of order of the average value of the attributes:

$$CP = \left(\sum_{i=1}^m |OrdAV_i - \overline{OrdAV}_i| \right) / m,$$

where $OrdAV_i$ is the ascending order of the average value of attribute i , while \overline{OrdAV}_i denotes its ascending order after the distortion. For instance, the order vector of all attributes in matrix A_e is: $[2\ 3\ 4\ 1]^T$. The order vector for the distorted matrix \overline{A}_e is: $[4\ 3\ 1\ 2]^T$. Then the total change of order is 6, so CP is equal to 1.2.

Similarly as RK , we define CK to measure the percentage of the attributes that keep their orders of average value after the distortion. So it is calculated as:

$$CK = \left(\sum_{i=1}^m Ck^i \right) / m,$$

where Ck^i is computed as:

$$Ck^i = \begin{cases} 1, & \text{if } OrdAV_i = \overline{OrdAV}_i, \\ 0, & \text{otherwise.} \end{cases}$$

In the previous example, $CK = 0.25$.

The higher the value of RP and CP , and the lower the value of RK and CK , the more privacy is preserved.

4 Utility Measure

The data utility measures assess whether a dataset keeps the performance of data mining techniques after the data distortion, e.g., whether the distorted data can maintain the accuracy of the data mining techniques such as classification, clustering, etc. In this paper, we choose the accuracy in Support Vector Machine (SVM) classification as the data utility measure.

SVM is based on structural risk minimization theory [25]. It has been successfully applied to many applications like face identification, text categorization, bioinformatics, etc. [5, 6, 17].

In SVM classification, the goal is to find a hyperplane that separates the examples with maximum margin. Given l examples $(x_1, y_1), \dots, (x_l, y_l)$, with $x_i \in R^n$ and $y_i \in \{-1, 1\}$ for all i , SVM classification can be stated as a quadratic programming problem:

$$\text{minimize } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i$$

$$\text{subject to } \begin{cases} y_i(\langle w, x_i \rangle + b) \leq 1 - \xi_i \\ \xi_i \geq 0 \\ C > 0 \end{cases}$$

where C is a user-selected regularization parameter, and ξ_i is a slack variable accounting for errors. After solving the quadratic programming problem, we can get the following decision function:

$$f(x) = \sum_{i=1}^l \alpha_i y_i \langle x_i, x \rangle + b. \quad (1)$$

where $0 \leq \alpha_i \leq C$.

For the nonlinear case, we apply a mapping $\Phi : X \rightarrow F$ to map the input space into some feature space F . Here we use a kernel function, $K(x, x_i) = \langle \Phi(x), \Phi(x_i) \rangle$, which is a symmetric function and satisfies the Mercer's condition. We substitute $K(x, x_i)$ for the dot product, which maps the input space into some reproduced kernel feature space. Then Equation (1) can be rewritten as:

$$f(x) = \sum_{i=1}^l \alpha_i y_i K(x, x_i) + b. \quad (2)$$

5 Experiments and Results

We conduct some experiments to test the performance of the data distortion methods: SVD, sparsified SVD (SSVD), adding uniformly distributed noise (UD) and adding normally distributed (ND) noise.

5.1 Synthetic dataset

First, we compare the performance of the four data distortion methods using a synthetic dataset. The dataset is a 2000 by 100 matrix (Org), whose entries are randomly generated numbers within the interval $[1, 10]$ obeying a uniform distribution. We classify the dataset into two classes using a randomly chosen rule: If

$$|\sin(\text{Org}(i, 1)) - \text{Org}(i, 88)| * |\cos(\text{Org}(i, 45))| * \text{Org}(i, 78) > 15,$$

then record i is assigned to class 1, otherwise, it is assigned to class -1 . We use SVM classification [15] to construct the classifier and a 5-fold cross validation to obtain the classification results. The uniformly distributed noise is generated from the interval $[0, 0.8]$. The normally distributed noise is generated with $\mu = 0$ and $\sigma = 0.4$. For SVD and SSVD, the rank k is chosen to be 80, and in SSVD, the dropping threshold value ϵ is 5×10^{-3} .

Table 1 shows the value of the privacy measures and the utility measure of applying the data distortion methods. For easy comparison, we set the VD value of UD, ND, and SVD to be approximately equal. That is, we compare these

Table 1. Comparison of distortion methods for the synthetic dataset.

Data	VD	RP	RK	CP	CK	Accuracy
Org	-	-	-	-	-	77.6%
UD	0.1643	0	1	0	1	76.5%
ND	0.1719	1662.8	0.0026	0	1	75.8%
SVD	0.1721	1666.7	0.0008	11.22	0.57	88.2%
SSVD	0.7647	1667.3	0.0002	36.42	0	86.8%

methods under the condition that they loss approximately the same amount of value after the data distortion. In this experiment, UD keeps the relative order of each element and each attribute. Thus we think it provides the least protection for privacy. ND is better than UD in changing the order of the elements, but it keeps the order of the attributes. Both SVD and SSVD are better in keeping the privacy for the elements and the attributes. SSVD is even better than SVD. It has similar value with SVD in RP value, but its RK value is much lower, which means fewer elements keep their order after the distortion. Its CP value is more than three times higher than that of SVD. The CK value for SSVD is 0, which means all the attributes change their order in average value after the distortion. For SVD, only about a half of the attributes change their order.

Next, let us look at how these distortion methods keep the data utility. The Accuracy column in Table 1 shows the percentage of the correctly classified data records. The accuracy of classifying the original dataset is 77.6%. Using UD and ND lowers this accuracy a little bit. While using SVD and SSVD, the accuracy is actually improved. The accuracy of using SVD is raised to 88.2%, while using SSVD it is 86.8%.

5.2 Real world dataset

For a real world dataset, we download some information about 100 terrorists from a terrorist analysis web site [24]. We selected 42 attributes ($m = 42$), such as their nationality, different sibling relationships, pilot training, locations of temporary residency, wedding attendance, meeting attendance, etc. The original matrix is of dimension 100×42 . To test the real world dataset, the uniformly distributed noise is chosen from the interval $[0, 0.09]$. The normally distributed noise is generated with $\mu = 0$ and $\sigma = 0.05$. The rank k for SVD and SSVD is chosen to be 25. The dropping threshold value ϵ in SSVD is 10^{-3} .

In Table 2, we classify the terrorists into two groups, those are related with Bin Laden and those are not. The accuracy of classifying the original dataset is 67%. The UD and ND methods keep this accuracy, while using SVD the accuracy rises to 70%. The accuracy obtained by using SSVD is 69%, also improves a little bit. Thus SVD and SSVD are a little better in data utility.

In order to be fair in comparing the privacy metrics, we also make the VD value of UD, ND and SVD to be almost the same. For privacy protection, UD does not perform well. It does not change any order of the elements in attributes,

Table 2. Comparison of distortion methods for classification 1 (Bin Laden association).

Data	VD	RP	RK	CP	CK	Accuracy
Org	-	-	-	-	-	67%
UD	0.0566	0	1	11.4	0.15	67%
ND	0.0537	31.9	0.0298	12.2	0.27	66%
SVD	0.0525	31.2	0.0251	12.2	0.12	70%
SSVD	1.0422	37.5	0.0066	13.1	0.05	69%

and has the lowest CP and a high CK values. ND is better than UD but is significantly worse than SVD in CK measure. The CK value obtained by using ND is 0.27, while SVD reduces it more than a half to 0.12. Among the four distortion methods, SSVD is the best to preserve privacy in this experiment. It has the highest RP and RK values, which means it is the best in keeping the privacy of the order of the individual elements. It also has the best CP and CK values, which means it also exceeds other methods in changing the order of the attributes.

Table 3. Comparison of distortion methods for classification 2 (Al Qaeda association).

Data	VD	RP	RK	CP	CK	Accuracy
Org	-	-	-	-	-	67%
UD	0.0575	31.9	0.0166	9.5	0.07	66%
ND	0.0566	34.1	0.0390	12.0	0.07	64%
SVD	0.0525	31.2	0.0251	12.2	0.12	70%
SSVD	1.3829	35.0	0.0090	11.5	0.02	65%

Table 3 shows the results of performing another classification on the real world dataset. This time the terrorists are grouped according to whether or not they have relationship with the terrorist organization Al Qaeda. The previous target attribute about whether a person has relationship with Bin Laden is inserted into the data matrix and the attribute about whether a person has relationship with Al Qaeda is taken out as the target attribute. Thus the original data matrix for privacy analysis is a little different from the one used in the previous classification task. All the distorted matrices are generated from the new original matrix.

Again, we keep the VD value to be very similar for UD, ND and SVD. SVD is the best for the classification result, it improves the accuracy (70%) over using the original dataset (67%). The other three methods decrease the accuracy slightly. For the privacy protection, SSVD works best in keeping the privacy of each elements. It has the highest RP and lowest RK values. Its CP value is slightly lower than those of ND and SVD, but its CK value is more than three times lower than that of ND, and six times lower than that of SVD. SVD is

not outstanding in preserving privacy in this experiment. ND exceeds it in RP and CK values. In this example, SSVD is still the best in keeping the privacy of data.

6 Concluding Remarks

We proposed to use the sparsified SVD method for data distortion in a simplified model terrorist analysis system. The experimental results show that SSVD is the best in preserving data privacy. It is also efficient in keeping data utility. SVD works well, too. Both are better than the standard data distortion methods which add noise straightforwardly. We believe that the SVD-based methods can be used in data mining techniques for data distortion purpose in order to protect privacy and other sensitive information contained and visible in the original datasets. Further research can be done to test the choice of different (sparsified) SVD parameters, such as k and ϵ , on the effect of the data distortion. Other SVD sparsification strategies [12] can also be tested in the data distortion applications. It should be pointed out that the application of the proposed SVD-based data distortion methods is not limited to the terrorist analysis systems. Many data analysis processes in which there is a need for data distortion may benefit from the proposed SVD-based data distortion methods.

References

1. D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.*, Santa Barbara, California, USA, May 2001.
2. R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, Texas, May, 2000.
3. R. Agrawal, A. Evfimievski, and R. Srikant. Information sharing across private databases. in *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pp. 86-97, San Diego, CA, 2003.
4. M. W. Berry, Z. Drmac, and E. R. Jessup. Matrix, vector space, and information retrieval. *SIAM Rev.*, 41:335–362, 1999.
5. C. Burges. *A Tutorial on Support Vector Machine for Pattern Recognition*. Kluwer Academic Publishers, 1998.
6. C. Campbell. Kernel methods: A survey of current techniques. *Neurocomputing*, 48 (2002) 63-84.
7. S. Datta, H. Kargupta, and K. Sivakumar. Homeland defense, privacy-sensitive data mining, and random value distortion. In *Proceedings of the SIAM Workshop on Data Mining for Counter Terrorism and Security (SDM'03)*, San Francisco, CA, May 2003.
8. S. Deewester, S. Dumais, *et al.* Indexing by latent semantic analysis, *J. Amer. Soc. Infor. Sci.*, 41:391–407, 1990.

9. J. X. Dempsey and P. Rosenzweig. Technologies that can protect privacy as information is shared to combat terrorism. Legal Memorandum #11, The Heritage Foundation, May 26, 2004. Available at www.heritage.org/Research/HomelandDefense/lm11.cfm.
10. V. Estvill-Castro, L. Brankovic, and D. L. Dowe. Privacy in data mining. Australian Computer Society, NSW Branch, Australia. Available at www.acs.org.au/nsw/articles/1999082.html.
11. W. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Englewood Cliffs, NJ, 1992.
12. J. Gao and J. Zhang. Sparsification strategies in latent semantic indexing, in *Proceedings of the 2003 Text Mining Workshop*, M. W. Berry and W. M. Pottenger, (ed.), pp. 93–103, San Francisco, CA, May 3, 2003.
13. B. Gilburd, A. Schuster, and R. Wolff. K-TTP: a new privacy model for large-scale distributed environments. in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, Seattle, WA, USA, 2004.
14. G. H. Golub and C. F. van Loan. *Matrix Computations*, John Hopkins Univ., 3rd Ed., 1996.
15. T. Joachims. Making large-scale SVM learning practical. *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges and A. Smola (ed.), MIT-Press, 1999.
16. C. K. Liew, U. J. Choi, and C. J. Liew. A data distortion by probability distribution. *ACM Transactions on Database Systems*, Vol. 10, No. 3, September 1985, Pages 395-411.
17. Y. Li, S. Gong, H. Liddell. Support vector regression and classification based multi-view face detection and recognition. In *Proc. of the IEEE International Conference on Automatic Face and Gesture Recognition (FGR'00)*, Grenoble, France, 2000.
18. D. B. Skillicorn. Clusters within clusters: SVD and counterterrorism, in *Proceedings of 2003 Workshop on Data Mining for Counter Terrorism and Security*, 12 pages, San Francisco, CA, May 3, 2003.
19. D. B. Skillicorn. Social network analysis via matrix decompositions: applications to al Qaeda, Technical Report, School of Computing, Queen's University, Aug. 2004.
20. D. B. Skillicorn and N. Vats. Novel information discovery for intelligence and counterterrorism, Technical Report 2004-488, School of Computing, Queen's University, Sep. 2004.
21. A. Sun, M. Naing, *et al.* Using support vector machines for terrorism information extraction, *ISI*, 2003: 1-12.
22. K. A. Taipale. Data mining and domestic security: connecting the dots to make sense of data. *Colum. Sci. & Tech. Law Rev.*, 5 (2003) 1–83.
23. T. Tether. Statement before the Subcommittee on Technology, Information Policy, Intergovernmental Relations and the Census, Committee on Government Reform, U.S. House of Representatives, May 6, 2003. Available at www.fas.gov/irp/congress/2003_hr/050603tether.html.
24. www.trackingthethreat.com
25. V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.
26. V. S. Verykios, E. Bertino, I. N. Fovino, *et al.* State-of-the-art in privacy preserving data mining. *SIGMOD Record*, 33 (2004) 50–57.