Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining

Kun Liu, Hillol Kargupta, *Senior Member*, *IEEE*, and Jessica Ryan

Abstract—This paper explores the possibility of using multiplicative random projection matrices for privacy preserving distributed data mining. It specifically considers the problem of computing statistical aggregates like the inner product matrix, correlation coefficient matrix, and Euclidean distance matrix from distributed privacy sensitive data possibly owned by multiple parties. This class of problems is directly related to many other data-mining problems such as clustering, principal component analysis, and classification. This paper makes primary contributions on two different grounds. First, it explores Independent Component Analysis as a possible tool for breaching privacy in deterministic multiplicative perturbation-based models such as random orthogonal transformation and random rotation. Then, it proposes an approximate random projection-based technique to improve the level of privacy protection while still preserving certain statistical characteristics of the data. The paper presents extensive theoretical analysis and experimental results. Experiments demonstrate that the proposed technique is effective and can be successfully used for different types of privacy-preserving data mining applications.

Index Terms—Random projection, multiplicative data perturbation, privacy preserving data mining.

1 Introduction

Privacy is becoming an increasingly important issue in many data-mining applications that deal with health care, security, financial, behavioral, and other types of sensitive data. It is particularly becoming important in counterterrorism and homeland defense-related applications. These applications may require creating profiles, constructing social network models, and detecting terrorist communications among others from privacy sensitive data. For example, mining healthcare data for detection of bioterrorism may require analyzing clinical records and pharmacy transactions data of certain off-the-shelf drugs. However, combining such diverse data sets belonging to different parties may violate the privacy laws. Although health organizations are allowed to release data as long as the identifiers (e.g., name, SSN, address, etc.,) are removed, it is not considered safe enough since reidentification attacks may be constructed for linking different public data sets to identify the original subjects [1]. This calls for well-designed techniques that pay careful attention to hiding privacy-sensitive information, while preserving the inherent statistical dependencies which are important for data-mining applications.

The problem we are interested in and discuss in this paper can be defined as follows: Suppose there are N organizations $O_1, O_2, \ldots, O_N$; each organization O_i has a private transaction database DB_i. A third party data miner wants to learn certain statistical properties of the union of these databases $\bigcup_{i=1}^{N} DB_i$. These organizations are comfortable with this, but they are reluctant to disclose their raw data. How could the data miner perform data analysis without compromising the privacy of the data? This is generally referred to as the *census problem* [2]. In this scenario, the data is usually distorted and its new representation is released; anybody has arbitrary access to the published data. Fig. 1 illustrates a distributed two-party-input case as well as a single-party-input case.

This paper considers a randomized multiplicative data perturbation technique for this problem. It is motivated by the work presented elsewhere [3] that pointed out some of the problems of additive random perturbation. Specifically, this paper explores the possibility of using multiplicative random projection matrices for constructing a new representation of the data. The transformed data is released to the data miner. It can be proved that the inner product and Euclidean distance are preserved in the new data. The approach is fundamentally based on the Johnson-Lindenstrauss lemma [4] which notes that any set of s points in m-dimensional Euclidean space can be embedded into k-dimensional subspace, where k is logarithmic in s, such that the pair-wise distance of any two points is maintained within an arbitrarily small factor. Therefore, by projecting the data onto a *random* subspace, we can dramatically change its original form while preserving much of its underlying distance-related statistical characteristics.

In this paper, we assume that the private data is from the same continuous real domain and all the parties are semihonest (which means there is no collusion between parties and all the parties follow the protocol properly). Without loss of generality, we demonstrate our technique in a two-party-input scenario where Alice and Bob, each owning a private database, want a third party to analyze

• *The authors are with the Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, Baltimore, MD 21250. E-mail: {kunliu1, hillol, jryan4}@cs.umbc.edu.*

Manuscript received 5 Mar. 2005; revised 31 May 2005; accepted 14 June 2005; published online 18 Nov. 2005.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0067-0304.

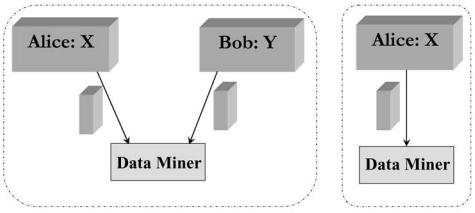


Fig. 1. (a) Distributed two-party-input computation model. (b) Single-party-input computation model.

their data without seeing the raw information. Our technique can be easily modified and applied to other input cases.

The remainder of this paper is organized as follows: Section 2 offers an overview of the related work in privacy preserving data mining. Section 3 discusses the random orthogonal transformation-based perturbation technique in the context of distributed inner product computation. This is computationally equivalent to many problems such as computing Euclidean distance, correlation, angles, or even covariance between a set of vectors. These statistical aggregates play a critical role in many data-mining techniques such as clustering, principal component analysis, and classification. Section 4 explores the potential vulnerability of this method from the perspective of Independent Component Analysis (ICA). Section 5 presents a random projection-based multiplicative data perturbation technique as an extension to enhance the privacy level. Section 6 gives a further detailed analysis about the privacy issues. Section 7 compares our technique with other existing secure matrix product protocols. Several real data mining applications, e.g., distributed inner product/Euclidean distance estimation, distributed clustering, linear classification, etc., and experiments are provided in Section 8 to justify the effectiveness of this technique. Finally, Section 9 concludes this paper and outlines the future research.

2 Related Work

This section presents a brief overview of the literature on privacy preserving data mining.

2.1 Data Perturbation

Data perturbation approaches can be grouped into two main categories: the probability distribution approach and the value distortion approach. The probability distribution approach replaces the data with another sample from the same (or estimated) distribution [5] or by the distribution itself [6], and the value distortion approach perturbs data elements or attributes directly by either additive noise, multiplicative noise, or some other randomization procedures [7]. In this paper, we mainly focus on the value distortion approach.

The work in [8] proposed an additive data perturbation technique for building decision tree classifiers. Each data element is randomized by adding some random noise chosen independently from a known distribution such as Gaussian distribution. The data miner reconstructs the distribution of the original data from its perturbed version (using, e.g., an Expectation Maximization-based algorithm) and builds the classification models. More recently, Kargupta et al. [3] questioned the use of random additive noise and pointed out that additive noise can be easily filtered out in many cases that may lead to compromising the privacy.

The possible drawback of additive noise makes one wonder about the possibility of using multiplicative noise for protecting the privacy of the data. Two basic forms of multiplicative noise have been well studied in the statistics community [9]. One is to multiply each data element by a random number that has a truncated Gaussian distribution with mean one and small variance. The other one is to take a logarithmic transformation of the data first, add predefined multivariate Gaussian noise, and take the antilog of the noise-added data. In practice, the first method is good if the data disseminator only wants to make minor changes to the original data; the second method assures higher security than the first one but maintains the data utility in the log-scale. A potential problem of traditional additive and multiplicative perturbation is that each data element is perturbed independently, therefore the pair-wise similarity of records is not guaranteed to be maintained. In this paper, we propose an alternate approach that proves to preserve much of the underlying statistical aggregates of the data.

Additive and multiplicative perturbation usually deal with numeric data only. Perturbation for categorical data was initially considered in [10], where a randomized response method was developed for the purpose of data collection through interviews. The work in [11] considered categorical data perturbation in the context of association rule mining. This work was extended in [12], where a framework for quantifying privacy breaches was introduced. The framework uses the concept of γ-amplification and applies it without any assumption about the underlying distribution from which the original data is drawn. The work in [13] considered this framework again and showed how to optimally set the perturbation parameters for reconstruction while maintaining γ-amplification.

2.2 Data Swapping

The basic idea of data swapping, which was first proposed by Dalenius and Reiss [14], is to transform the database by switching a subset of attributes between selected pairs of records so that the lower order frequency counts or marginals are preserved and data confidentiality is not compromised. This technique could equally as well be classified under the data perturbation category. A variety of refinements and applications of data swapping have been addressed since its initial appearance. We refer readers to [15] for a thorough treatment.

2.3 *k*-Anonymity

The k-Anonymity model [1] considers the problem that a data owner wants to share a collection of person-specific data without revealing the identity of an individual. To achieve this goal, data generalization and suppression techniques are used to protect the sensitive information. All attributes (termed as quasi-identifier) in the private database that could be used for linking with external

information would be determined, and the data is released only if the information for each person contained in the release cannot be distinguished from at least $k-1$ other people.

2.4 Secure Multiparty Computation

The Secure Multiparty Computation (SMC) [16] technique considers the problem of evaluating a function of the secret inputs from two or more parties, such that no party learns anything but the designated output of the function. A large body of cryptographic protocols, including circuit evaluation protocol, oblivious transfer, homomorphic encryption, and commutative encryption, serve as the building blocks of SMC. The work in [17] offered a broad view of SMC framework and its applications to data mining. The work in [18] detailed a rigorous introduction to SMC. It was shown that any function that can be expressed by an arithmetic circuit is privately computable using a generic circuit evaluation protocol. However, the communication and computational complexity of doing so makes this general approach infeasible for large data sets. A collection of SMC tools useful for large-scale privacy preserving data mining (e.g., secure sum, set union, and inner product) are discussed in [19]. An overview of the state-of-the-art privacy preserving data mining techniques is presented in [20].

2.5 Distributed Data Mining

The distributed data mining (DDM) [21], [22] approach supports computation of data mining models and extraction of "patterns" at a given node by exchanging only the minimal necessary information among the participating nodes. The work in [23] proposed a paradigm for clustering distributed privacy sensitive data in an unsupervised or a semisupervised scenario. In this algorithm, each local data site builds a model and transmits only the parameters of the model to the central site where a global clustering model is constructed. A distributed privacy-preserving algorithm for Bayesian network parameter learning is reported elsewhere [24].

2.6 Rule Hiding

The main objective of rule hiding is to transform the database such that the sensitive rules are masked, and all the other underlying patterns can still be discovered. The work in [25] gave a formal proof that the optimal sanitization is an NP-hard problem for the hiding of sensitive large item sets in the context of association rule mining. For this reason, some heuristic approaches have been applied to address the complexity issues. For example, the perturbation-based association rule hiding technique [26] is implemented by changing a selected set of 1-values to 0-values or vice versa so that the frequent item sets that generate the rule are hidden or the support of sensitive rules is lowered to a user-specified threshold. The blocking-based association rule hiding approach [27] replaces certain attributes of the data with a question mark. In this regard, the minimum support and confidence will be altered into a minimum interval. As long as the support and/or the confidence of a sensitive rule lies below the middle in these two ranges, the confidentiality of data is expected to be protected.

3 Random Orthogonal Transformation

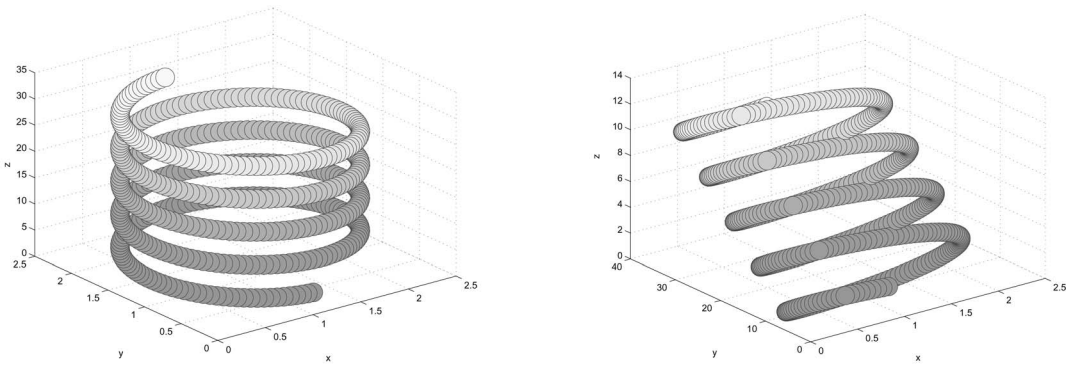
This section presents a deterministic multiplicative perturbation method using random orthogonal matrices in the context of computing inner product matrix. Later, we shall analyze the deficiency of this method and then propose a more general case that makes use of random projection matrices for better protection of the data privacy.

An orthogonal transformation [28] is a linear transformation $R : \mathbb{R}^n \rightarrow \mathbb{R}^n$, which preserves the length of vectors as well as the angles between them. Usually, orthogonal transformations correspond to and may be represented using orthogonal matrices. Let X and Y be two data sets owned by Alice and Bob, respectively. X is an $m_1 \times n$ matrix, and Y is an $m_2 \times n$ matrix. Both of them observe the same attributes. Let R be an $n \times n$ random orthogonal matrix. Now, consider the following linear transformation of the two data sets:

$$U = XR, \quad \text{and} \quad V = YR; \quad \text{then we have}$$
$$UU^T = XX^T, VV^T = YY^T, UV^T = XRR^TY^T = XY^T.$$

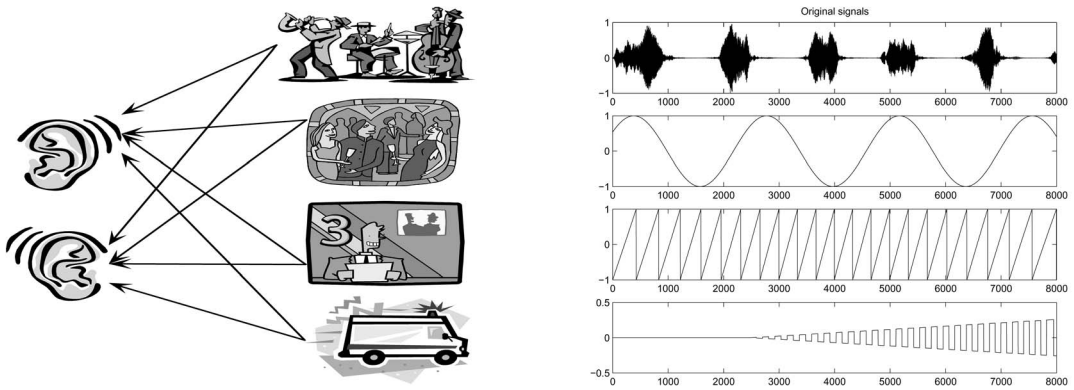
So, if both Alice and Bob transform their data using a secret orthogonal matrix, and only release the perturbed version to a third party, all the pair-wise angles/distances between the row vectors from data $\binom{X}{Y}$ can still be perfectly computed there, where $\binom{X}{Y}$ is a horizontal concatenation of X and Y. Therefore, it is easy to implement a distance-based privacy preserving data-mining application in a third party for homogeneously distributed (horizontally partitioned) data. Similarly, if we transform the data in a way such that $U = RX$, $V = RY$, we will have $U^TV = X^TY$, and all the pair-wise distances and similarities between the columns vectors from the data $(X : Y)$ are fully preserved in the perturbed data, where $(X : Y)$ denotes a vertical concatenation of X and Y. Therefore, a third party can analyze the correlation of the attributes from heterogeneously distributed (vertically partitioned) data without accessing the raw data.

Since only the transformed data is released, there are actually an infinite number of inputs and transformation procedures that can simulate the output, while the observer has no idea what is the real form of the original data. Therefore, random orthogonal transformation seems to be a good way to protect data's privacy while preserving its utility. However, from the geometric point of view, an orthogonal transformation is either a pure rotation when the determinant of the orthogonal matrix is 1 or a rotoinversion (a rotation followed by a flip) when the determinant is -1, and, therefore, it is possible to reidentify the original data through a proper rotation. Figs. 2a and 2b illustrate how the random orthogonal transformation works in a 3D space. It can be seen that the data is not very well masked after transformation. In this regard, the security of a similar approach using random rotation [29] to protect the data privacy is also questionable. Moreover, if all the original data vectors are statistically independent and they do not follow Gaussian distribution, it is possible to estimate their original forms quite accurately using Independent Component Analysis (ICA). In the following sections, we shall briefly discuss the properties of ICA and then propose a random projection-based multiplicative



(a) (b)

Fig. 2. (a) A sample data set. (b) The perturbed data after a random orthogonal transformation. The transformation corresponds to a rotation of the original data about the x-axis by a random angle.



(a) (b)

Fig. 3. (a) An illustration of the cocktail problem. In this case, what the ears hear are two linear combinations of four audio signals, i.e., four signals are compressed into two. (b) A sample of four independent source signals.

perturbation technique to improve the privacy level while preserving the data utilities.

4 Independent Component Analysis

Independent Component Analysis (ICA) [30] is a technique for discovering independent hidden factors that are underlying a set of linear or nonlinear mixtures of some unknown variables, where the mixing system is also unknown. These unknown variables are assumed non-Gaussian and statistically independent, and they are called the independent components (ICs) of the observed data. These independent components can be found by ICA. A classical example of ICA is the cocktail party problem (as illustrated in Fig. 3a). Imagine you are in a cocktail party. Although different kinds of background sounds are mixed together, e.g., music, other people's chat, television news report, or even a siren from a passing-by ambulance, you still have no problem identifying the discussion of your neighbors. It is not clear how human brains can separate the different sound sources. However, ICA is able to do it if there are at least as many "ears" or receivers in the room as there are different simultaneous sound sources.

4.1 ICA Model

The basic ICA model can be defined as follows:

$$u(t) = Rx(t), \tag{1}$$

where $x(t) = (x_1(t), x_2(t), \ldots, x_m(t))^T$ denotes a m-dimensional vector collecting the m independent source signals $x_i(t), i = 1, 2, \ldots, m$. Here, t indicates the time dependence. Each signal $x_i(t)$ can be viewed as an outcome of a continuous-value random process. R is a constant $k \times m$ unknown mixing matrix, which can be viewed as a mixing system with k receivers. $u(t) = (u_1(t), u_2(t), \ldots, u_k(t))^T$ is the observed mixture. The aim of ICA is to design a filter that can recover the original signals from only the observed mixture. Since $u(t) = Rx(t) = (R\Lambda P)(P^{-1}\Lambda^{-1}x(t))$ for any diagonal matrix Λ and permutation matrix P, the recovered signals $x(t)$ can never have completely unique representation. So, the uniqueness of the recovered signals found by ICA can only be guaranteed up to permutation and scaling ambiguities.

As an illustration, consider four statistically independent audio signals, denoted as a $4 \times 8,000$ matrix X (shown in Fig. 3b). Note that, for the sake of simplicity, some of the signals we are showing here are deterministic; however, ICA generally works with continuous-value random process. A linear mixture of these signals (shown in Fig. 4a) is generated by premultiplying a 4×4 nonsingular random matrix to X. The goal of ICA is to recover the original signals using only the mixture. Fig. 4b gives the estimated signals

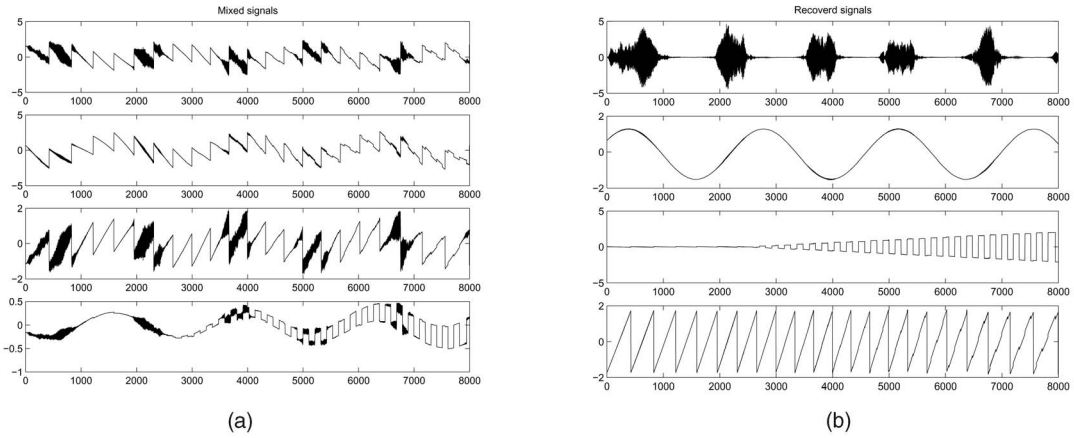


Fig. 4. (a) Linear mixture of the original source signals using a square random matrix. (b) Recovered signals using ICA.

through ICA. It can be observed that the basic structure of the original signals are recovered very well; however, the order and the amplitude of the recovered signals are not necessarily the same as those of the original ones.

4.2 Decomposability

In practice, a linear filter is designed to get the recovered signals $y(t) = (y_1(t), y_2(t), \ldots, y_l(t))^T$ from a k-dimensional input $u(t) = (u_1(t), u_2(t), \ldots, u_k(t))^T$. In other words,

$$y(t) = Bu(t), \tag{2}$$

where B is an $l \times k$-dimensional separating matrix. Combining (1) and (2) together, we get

$$y(t) = BRx(t) = Zx(t), \tag{3}$$

where $Z = BR$ is an $l \times m$ matrix. Each element of $y(t)$ is thus a linear combination of $x_i(t)$ with weights given by $z_{i,j}$.

Ideally, when $k \geq m$ (i.e., the number of receivers is greater than or equal to the number of source signals), if the mixing matrix R has full column rank, there always exists an $l \times k$ separating matrix B such that $Z = BR = I$, where I is an identity matrix. Thus, we can recover all the signals up to scaling and permutation ambiguities. Actually, to solve the problem, there are two steps to be done. The first step is to determine the existence of B such that Z can decompose the mixture. The second step is to find such a kind of B if it is proved to exist. We will focus on the first step.

In general, by imposing the following fundamental restrictions [31], all the source signals can be separated out up to scaling and permutation ambiguities:

- The source signals are statistically independent, i.e., their joint probability density function (PDF) $f_{\mathbf{x(t)}}(x_1(t), x_2(t), \ldots, x_m(t))$ is factorizable in the following way:

$$f_{\mathbf{x(t)}}(x_1(t), x_2(t), \ldots, x_m(t)) = \prod_{i=1}^{m} f_{\mathbf{x_i(t)}}(x_i(t)),$$

  where $f_{\mathbf{x_i(t)}}(x_i(t))$ denotes the marginal probability density of $x_i(t)$.
- All the signals must be non-Gaussian with the possible exception of one signal.
- The number of observed signals k must be at least as large as the independent source signals, i.e., $k \geq m$.
- Matrix R must be of full-column rank.

These restrictions actually have exposed the potential dangers of random orthogonal transformation or random rotation techniques where the mixing matrix is square and of full-column rank. If the original signals are also statistically independent and there are no Gaussians, it is most likely that ICA can find a good approximation of the original signals from their perturbed version. Figs. 4a and 4b illustrated this situation.

Note that, if some of the source signals are correlated, they may be lumped in the same group and can never be separated out. If there is more than one Gaussian signal, the problem becomes more complicated. The output of the filter may be either individual non-Gaussian signals, individual Gaussian signals, or a mixture of Gaussian signals. A detailed analysis can be found elsewhere [32].

When $l \leq k < m$ (i.e., the number of sources is greater than the number of receivers),[1] it is generally not possible to design linear filters to simultaneously recover all these signals. This kind of separation problem is termed as overcomplete ICA or underdetermined source separation. Cao et al. [32] analyzed the conditions for the existence of the separating matrix B.

We first introduce two definitions (Definitions 4.1 and 4.2) and one theorem (Theorem 4.3) from the original materials without any proof. They serve as important building blocks in our solutions.

Definition 4.1 (Partition Matrix) [32]. *A set of m integers $S = \{1, 2, \ldots, m\}$ can be partitioned into l ($l \leq m$) disjoint subsets S_i, $i = 1, 2, \ldots, l$. An $l \times m$ matrix Z is called a partition matrix if its i, jth entry $z_{i,j} = 1$ when $j \in S_i$, and $z_{i,j} = 0$ otherwise. Z is called a generalized partition matrix if it is a product of an $l \times m$ partition matrix and an $m \times m$ nonsingular diagonal matrix.*

1. This implies that the number of recovered signals will be less than or equal to the number of the original signals. This is reasonable since we cannot get more signals than the original ones.

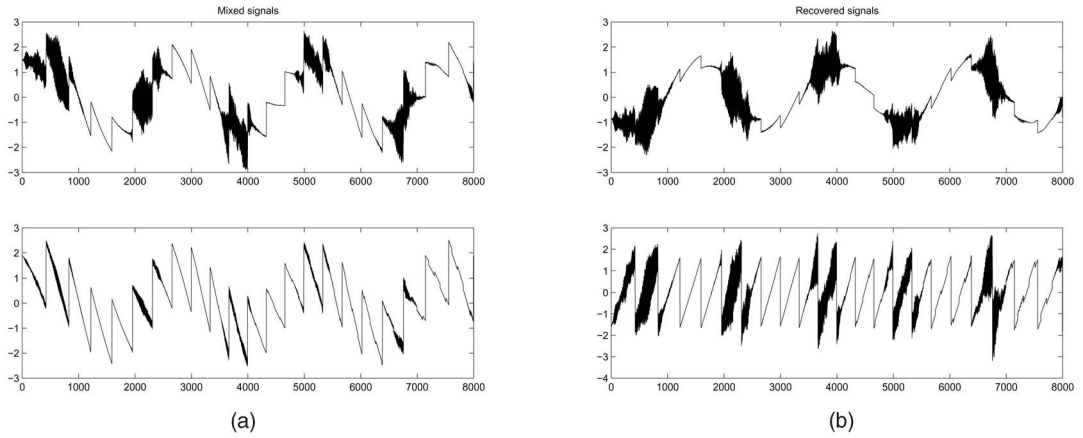


Fig. 5. (a) Linear mixture of the original four source signals (as shown in Fig. 3b) with 50 percent random projection rate. ($m = 4, k = 2$). (b) Recovered signals. It can be observed that none of the original signals are reconstructed, and at most $k = 2$ independent components can be found by ICA.

When none of the subsets S_i are empty, Z is simply a matrix in which each column has only one nonzero entry, and each row has at least one nonzero entry.

Definition 4.2 (*l*-row Decomposable) [32]. *$A\ k \times m$ matrix R is called l-row decomposable if there exists an $l \times k$ matrix B such that $Z = B \times R$ is an $l \times m$ generalized partition matrix.*

Therefore, if R is l-row decomposable, there exists a matrix B that enables Z to separate the source signals into l disjoint subgroups; each output $y_i(t), i = 1, 2, \ldots, l$ is a linear combination of the source signals in one subgroup, i.e.,

$$y_i = \sum_{j \in S_i} z_{i,j} x_j,\ \mathrm{i} = 1, 2, \ldots, 1.$$

If for some i, $S_i = \{p\}$, then $y_i = z_{i,p} x_p$, i.e., by using Z, we can separate out one signal x_p up to scaling ambiguities. If the number of the disjoint subgroups is m ($l = m$), then every subset S_i, $i = 1, \ldots, l$, contains only one element, we will have a complete separation. Also, note that, if R is l-row decomposable, it must be $(l-1)$-row decomposable since we can add two outputs $y_i(t)$ and $y_j(t)$ together to get $l - 1$ subgroups.

Theorem 4.3 [32]. *Matrix R is l-row decomposable if and only if its columns can be grouped into l disjoint groups such that the column vectors in each group are linearly independent of the vectors in all the other groups.*

Proof. Please see the proof of Theorem 1 in [32]. □

Cao et al. proved that, with $k < m$, the source signals can at most be separated into k disjoint groups from the observed mixture, and at most $k - 1$ signals (independent components) can be separated out.

Our claim is that, if we can control the structure of the mixing matrix R such that R is not *two*-row decomposable, then there is no linear method that can find a matrix B for separating the source signals into two or more disjoint groups. In that case, it is not possible to separate out any of the source signals. The following theorem characterized this property:

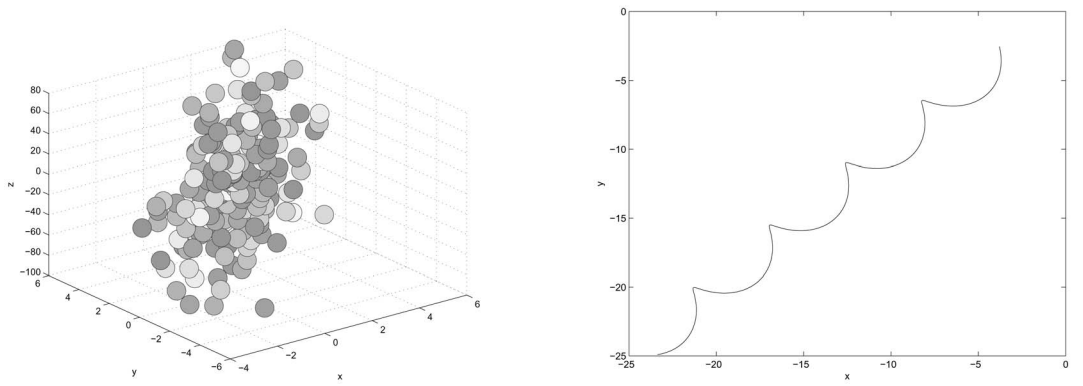
Theorem 4.4. *Any $k \times m$ ($m \geq 2k - 1, m \geq 2$) random matrix with entries independent and identically chosen from some continuous distribution in the real domain is not two-row decomposable with probability 1.*

Proof. For a $k \times m$ random matrix with $m \geq 2k - 1$ and any partition of its columns into two nonempty sets, at least one set will have at least k members. Thus, this set of columns contains a $k \times k$ submatrix, denoted as M. If M is nonsingular, then the k column vectors of the submatrix span $\mathbb{R}^k$ Euclidean space. Thus, there is always at least one vector in one group belonging to the space spanned by the other group, which does not satisfy Theorem 4.3.

Now, let us show M is indeed nonsingular with probability 1. It has been proved in [33, Theorem 3.3] that the probability that MM^T is positive definite is 1.[2] Since a matrix is positive definite if and only if all the eigenvalues of this matrix are positive, and a matrix is nonsingular if and only if all its eigenvalues are nonzero [34, Theorem 1.2.2], we have that MM^T is nonsingular with probability 1. Further note that $rank(M) = rank(MM^T) = rank(M^T M)$ [35], therefore M is nonsingular with probability 1. This completes the proof. □

The above nonsingularity property of a random matrix has also been proved in [34, Theorem 3.2.1] when the random matrix is Gaussian. Thus, by letting $m >> k$, there is no linear filter that can separate the observed mixtures into two or more disjoint groups, so it is not possible to recover any of the source signals. Figs. 5a and 5b depict this property. It can be seen that, after 50 percent row-wise random projection, the original four signals are compressed into two, and ICA cannot recover any of them. Moreover, projecting the original data using a nonsquare random matrix has two more advantages. One is to compress the data, which is very suited for distributed computation applications; the other one is to realize a many (elements)-to-one (element) map, which is totally different from the traditional one-to-one data perturbation technique, and, therefore, it is even harder for the adversary to reidentify the sensitive data.

2. We can get this result by replacing the matrix A in [33, Theorem 3.3] with an identity matrix.



(a) (b)

Fig. 6. (a) The perturbed data after a row-wise random projection which reduces 50 percent of the data points. (b) The perturbed data after a column-wise random projection which maps the data from 3D space onto 2D space. The random matrix is chosen from $\mathrm{N}(0,1)$ and the original data is given in Fig. 2a.

The discussion in this section summarizes as:

- If the components of the original data themselves are not statistically independent, that is, the original data $X = MC$, where M is another mixing matrix and C is the real independent components, after perturbed by a random matrix R, we will get a new mixing model $U = RX = (RM)C$. Even if ICA works perfectly for this model, what we finally get is the underlying independent components C (up to scaling and permutation ambiguities), but not X. If there are more than one Gaussian signals, the output of the filter may be either individual non-Gaussian signals, individual Gaussian signals, or a mixture of Gaussian signals, which are totally indeterministic.
- When $k \geq m$ (i.e., the number of receivers is greater than or equal to the number of source signals), and all the source signals are statistically independent, they can be separated out from the mixture up to scaling and permutation ambiguities if and only if the mixing matrix R is of full-column rank and at most one source signal is Gaussian.
- When $l \leq k < m$ (i.e., the number of receivers is less than the number of sources), the source signals can at most be separated into k disjoint groups from the mixtures, and at most $k-1$ signals can be separated out. Especially, when the mixing matrix R is not *two*-row decomposable ($m \geq 2k-1, m \geq 2$, and with i.i.d. entries chosen from continuous distribution), there is no linear method that can find a matrix B to separate out any of the source signals.

4.3 Recent Work on Overcomplete ICA

Recently, overcomplete ICA ($k < m$) has drawn much attention. It has been found that, even when $k < m$, if all the sources are non-Gaussian and statistically independent, it is still possible to identify the mixing matrix such that it is unique up to a right multiplication by a diagonal and a permutation matrix [36, Theorem 3.1]. If it is also possible to determine the distribution of $x(t)$, we could reconstruct the source signals in a probabilistic sense. However, despite its high interest, the overcomplete ICA problem has only been treated in particular cases, e.g., the source signals are assumed to have sparse distribution [37]. In the following section, we propose a random projection-based multiplicative perturbation technique. By letting the random matrix super nonsquare, we get an overcomplete ICA model. It shows that randomly generated projection matrices are likely to be more appropriate for protecting the privacy, compressing the data, and still maintaining its utility.

5 Random Projection-Based Multiplicative Perturbation

This section studies random projection-based multiplicative perturbation in the context of computing inner product and Euclidean distance without allowing direct access to the original data.

5.1 Basic Mechanism

Random projection refers to the technique of projecting a set of data points from a high-dimensional space to a randomly chosen lower-dimensional subspace. The key idea of random projection arises from the Johnson-Lindenstrauss Lemma [4] as follows:

Lemma 5.1 (Johnson-Lindenstrauss Lemma). *For any $0 < \epsilon < 1$ and any integer s, let k be a positive integer such that $k \geq \frac{4 \ln s}{\epsilon^2/2 - \epsilon^3/3}$. Then, for any set S of $s = |S|$ data points in $\mathbb{R}^m$, there is a map $f : \mathbb{R}^m \rightarrow \mathbb{R}^k$ such that, for all $x, y \in S$,*

$$(1-\epsilon)||x-y||^2 \leq ||f(x)-f(y)||^2 \leq (1+\epsilon)||x-y||^2,$$

where $||.||$ denotes the vector 2-norm.

This lemma shows that any set of s points in m-dimensional Euclidean space can be embedded into an $O(\frac{\log s}{\epsilon^2})$-dimensional space such that the pair-wise distance of any two points are maintained within an arbitrarily small factor. This beautiful property implies that it is possible to change the data's original form by reducing its dimensionality but still maintains its statistical characteristics. In this section, we shall demonstrate how random matrices can be used for this kind of map. To give the reader a general idea of how the random projection technique perturbs the data, we did both row-wise and column-wise projection of the sample data given in Fig. 2a. The results are shown in Figs. 6a and 6b. It can be seen that the original structure of

the data has been dramatically obscured. A further analysis about the privacy is given in Section 6. In the following part of this section, we discuss some interesting properties of the random matrix and random projection, which are good for maintaining the data utility.

Lemma 5.2. *Let R be a $p \times q$ random matrix such that each entry $r_{i,j}$ of R is independent and identically chosen from some unknown distribution with mean zero and variance σ_r^2, then*

$$E[R^T R] = p\sigma_r^2 I \quad \text{and} \quad E[RR^T] = q\sigma_r^2 I.$$

Proof. Let $r_{i,j}$ and $\epsilon_{i,j}$ be the i, ith entries of matrix R and $R^T R$, respectively,

$$\epsilon_{i,j} = \sum_{t=1}^{p} r_{t,i} r_{t,j}$$

$$E[\epsilon_{i,j}] = E\left[\sum_{t=1}^{p} r_{t,i} r_{t,j}\right] = \sum_{t=1}^{p} E[r_{t,i} r_{t,j}].$$

Since the entries of random matrix are independent and identically distributed (i.i.d.),

$$E[\epsilon_{i,j}] = \begin{cases} \sum_{t=1}^{p} E[r_{t,i}]E[r_{t,j}] & \text{if } i \neq j; \\ \sum_{t=1}^{p} E[r_{t,i}^2] & \text{if } i = j. \end{cases}$$

Now, note that $E[r_{i,j}] = 0$ and $E[r_{i,j}^2] = \sigma_r^2$, therefore,

$$E[\epsilon_{i,j}] = \begin{cases} 0 & \text{if } i \neq j; \\ p\sigma_r^2 & \text{if } i = j. \end{cases} \Longrightarrow E[R^T R] = p\sigma_r^2 I.$$

Similarly, we have $E[RR^T] = q\sigma_r^2 I$. □

Intuitively, this result echoes the observation made elsewhere [38], that in a high-dimensional space, vectors with random directions are almost orthogonal. A similar result was proved elsewhere [39]. Lemma 5.2 can be used to prove the following results.

Lemma 5.3 (ROW-WISE PROJECTION). *Let X and Y be two data sets owned by Alice and Bob, respectively. X is an $m \times n_1$ matrix, and Y is an $m \times n_2$ matrix. Let R be a $k \times m (k < m)$ random matrix such that each entry $r_{i,j}$ of R is independent and identically chosen from some unknown distribution with mean zero and variance σ_r^2. Further, let*

$$U = \frac{1}{\sqrt{k}\sigma_r} RX, \quad \text{and} \quad V = \frac{1}{\sqrt{k}\sigma_r} RY; \quad \text{then} \quad E[U^T V] = X^T Y. \tag{4}$$

Lemma 5.4 (COLUMN-WISE PROJECTION). *Let X and Y be two data sets owned by Alice and Bob, respectively. X is an $m_1 \times n$ matrix and Y is an $m_2 \times n$ matrix. Let R be an $n \times k (k < n)$ random matrix such that each entry $r_{i,j}$ of R is independent and identically chosen from some unknown distribution with mean zero and variance σ_r^2. Further, let*

$$U = \frac{1}{\sqrt{k}\sigma_r} XR, \quad \text{and} \quad V = \frac{1}{\sqrt{k}\sigma_r} YR; \quad \text{then} \quad E[UV^T] = XY^T. \tag{5}$$

The above results show that the row-wise projection preserves the column-wise inner product and the column-wise projection preserves the row-wise inner product. The beauty of this property is that inner product is directly related to many other distance-related metrics. To be more specific:

- The Euclidean distance of x and y is

$$||x - y|| = \sqrt{(x-y)^T (x-y)}.$$

- If the data vectors have been normalized to unity, then the cosine angle of x and y is

$$\cos\theta = \frac{x^T y}{||x|| \cdot ||y||} = x^T y.$$

- If the data vectors have been normalized to unity with zero mean, the sample correlation coefficient of x and y is

$$\rho_{x,y} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{m}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{m}\right)\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{m}\right)}} = x^T y.$$

Thus, if the data owner reduces the number of attributes of the data by projection, the statistical dependencies among the observations will be maintained; if the data owner compresses the observations, the relationship between the attributes will be preserved. On the one hand, given only the perturbed data U or V, one cannot determine the values of the original data X or Y, which is based on the premise that the possible solutions are infinite when the number of equations is less than the number of unknowns. On the other hand, we can directly apply common data-mining algorithms on the perturbed data without accessing the original sensitive information.

In the next section, we will discuss some nice bounds about the inner product and Euclidean distance preserved by the random projection, and, in Section 6, we shall give a further analysis about the privacy.

5.2 Error Analysis

In practice, due to the cost of communication and security concerns, we always use one specific realization of the random matrix R. Therefore, we need to know more about the distribution of $R^T R$ (similarly, for RR^T) in order to quantify the utility of the random projection-based perturbation technique.

Assume entries of the $k \times m$ random matrix R are i.i.d. and chosen from Gaussian distribution with mean zero and variance σ_r^2, we can study the statistical properties of the estimation of the inner product.

Let $\epsilon_{i,j}$ be the i, jth entry of matrix $R^T R$. It can be proved that $\epsilon_{i,j}$ is approximately Gaussian, $E[\epsilon_{i,i}] = k\sigma_r^2$, $Var[\epsilon_{i,i}] = 2k\sigma_r^4$, $\forall i$ and $E[\epsilon_{i,j}] = 0$, $Var[\epsilon_{i,j}] = k\sigma_r^4$, $\forall i, j, i \neq j$ (please see Appendix I for the proof which can be found on the Computer Society Digital Library at http://www.computer.org/tkde/archives.htm). The following lemma gives the mean and variance of the projection error.

Lemma 5.5. *Let x, y be two data vectors in $\mathbb{R}^m$. Let R be a $k \times m$ random matrix. Each entry of R is independent and*

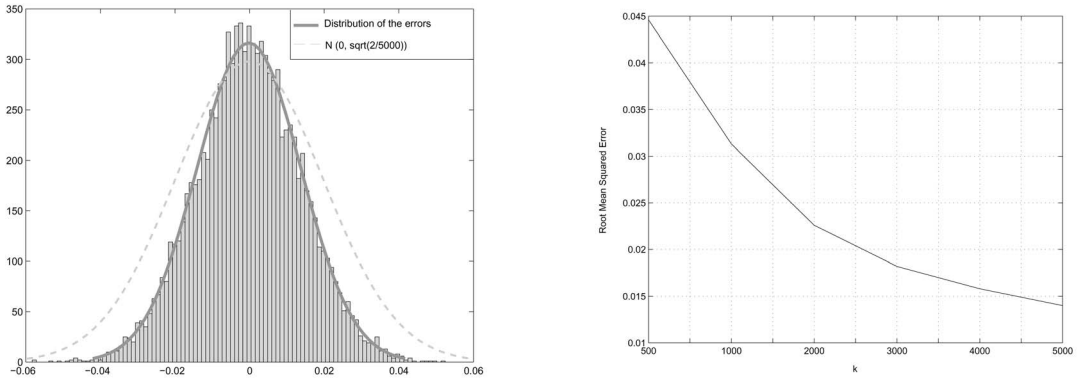


Fig. 7. (a) Distribution of the error of the estimated inner product matrix over two distributed data sets. Each data set contains 10,000 records and 100 attributes. $k = 50\% \times 10,000 = 5,000$ (50 percent row-wise projection). The random matrix is chosen from $N(0, 2)$. Note that the variance of the error is even smaller than the variance of distribution $N(0, \sqrt{2/k})$. (b) Root Mean Squared Error (RMSE) of the estimated inner product matrix with respect to the dimensionality of the reduced subspace.

identically chosen from Gaussian distribution with mean zero and variance σ_r^2. Further, let

$$u = \frac{1}{\sqrt{k}\sigma_r} Rx, \quad \text{and} \quad v = \frac{1}{\sqrt{k}\sigma_r} Ry; \text{ then}$$
$$E[u^T v - x^T y] = 0,$$
$$Var[u^T v - x^T y] = \frac{1}{k}\left(\sum_i x_i^2 \sum_i y_i^2 + (\sum_i x_i y_i)^2\right).$$

In particular, if both x and y are normalized to unity, $\sum_i x_i^2 \sum_i y_i^2 = 1$ and $(\sum_i x_i y_i)^2 \leq 1$. We have the upper bound of the variance as follows:

$$Var[u^T v - x^T y] \leq \frac{2}{k}.$$

Proof. Please see Appendix II which can be found on the Computer Society Digital Library at http://www.computer.org/tkde/archives.htm. □

Lemma 5.5 shows that the error ($u^T v - x^T y$) of the inner product matrix produced by random projection-based perturbation technique is zero, on average, and the variance is at most the inverse of the dimensionality of the reduced space multiplied by 2 if the original data vectors are normalized to unity. Actually, since $\epsilon_{i,j}$ is approximately Gaussian, the error also has an approximate Gaussian distribution, namely, $N(0, \sqrt{2/k})$. To validate the above claim, we choose two randomly generated data sets from a uniform distribution in $[0, 1]$, each with 10,000 observations and 100 attributes. We normalize all the attributes to unity and compare the column-wise inner product of these two data sets before and after row-wise random projection. Fig. 7a gives the results and it depicts that, even under 50 percent data projection rate (when $k = 5,000$), the inner product still preserves very well after perturbation, and the error indeed approximates Gaussian distribution with mean zero and variance less than $2/k$. Fig. 7b shows the Root Mean Squared Error (RMSE) of the estimated inner product matrix with respect to the dimensionality of the reduced subspace. It can be seen that, as k increases, the error goes down exponentially, which means that the higher the dimensionality of the data, the better this technique works. This lemma also echoes the results found in [40], where entries of R are independent and identically chosen from some unknown distribution with mean zero and each column vector of R is normalized to have a unit length.

By applying Lemma 5.5 to the vector $x - y$, we have

$$E[||u - v||^2 - ||x - y||^2] = 0.$$

If x and y are normalized to unity,

$$Var[||u - v||^2 - ||x - y||^2] \leq \frac{32}{k},$$

where $||x - y||^2 = (x - y)^T (x - y)$ is the square of the Euclidean distance of x and y. Note that this bound defines the maximum variance of the distortion. As a generalization of [39, Theorem 2], we also have the probability bound of the Euclidean distance as follows:

Lemma 5.6. *Let x, y be two data vectors in $\mathbb{R}^m$. Let R be a $k \times m$-dimensional random matrix. Each entry of the random matrix is independent and identically chosen from Gaussian distribution with mean zero and variance σ_r^2. Further, let*

$$u = \frac{1}{\sqrt{k}\sigma_r} Rx, \quad \text{and} \quad v = \frac{1}{\sqrt{k}\sigma_r} Ry; \text{ then}$$
$$Pr\{(1-\epsilon)||x-y||^2 \leq ||u-v||^2 \leq (1+\epsilon)||x-y||^2\}$$
$$\geq 1 - 2e^{-(\epsilon^2 - \epsilon^3)\frac{k}{4}}$$

for any $0 < \epsilon < 1$.

Proof. Directly follows the proof of [39, Theorem 2] with the exception that random matrix is chosen independently according to $N(0, \sigma_r)$. □

This result also shows that as the reduced dimensionality k increases, the distortion drops exponentially, which echoes the above observations that the higher the dimensionality of

the data, the better the random projection works. Many applications of random projection can be found in the literature, e.g., image and text clustering [40] and distributed decision tree construction [41]. In the next section, we shall give a detailed analysis about the privacy.

6 Privacy Analysis

Generally speaking, the random projection-based multiplicative perturbation technique guarantees that both the *dimensionality* and the *exact value of each element* of the original data are kept confidential. These properties are based on the assumptions that both data and random noise are from the continuous real domain and all the participating parties are semihonest.

In this section, we shall give a more rigorous analysis on how much privacy our perturbation technique can preserve when the adversary has different kinds of prior knowledge of the data and when the basic assumptions of this technique are not satisfied.

6.1 The Specific Realization of the Random Matrix is Disclosed

Consider the model $U = RX$, where $R \in \mathbb{R}^{k\times m}$ with $k < m$, and $X \in \mathbb{R}^{m\times n}$. This model can be viewed as a set of underdetermined systems of linear equations (more unknowns than equations), each with the form $u = Rx$, where x is an $m \times 1$ column vector from X and u is the corresponding column vector from U. For each linear system, assume both R and u are known, so the solution is never unique. In practice, the system can be analyzed by the QR factorization [42] of R^T such that

$$R^T = Q\begin{pmatrix}\overline{\mathcal{R}}\\ 0\end{pmatrix},$$

where Q is an $m \times m$ orthogonal matrix and $\overline{\mathcal{R}}$ is a $k \times k$ upper triangular matrix. If R has full row rank, i.e., $rank(R) = k$, there is a unique solution x_{min_norm} that minimizes $||x||_2$:[3]

$$\begin{aligned} x_{min\_norm} &= Q\begin{pmatrix}\overline{\mathcal{R}}^{T^{-1}}u\\ 0\end{pmatrix} = Q\begin{pmatrix}\overline{\mathcal{R}}\\ 0\end{pmatrix}(\overline{\mathcal{R}}^T\overline{\mathcal{R}})^{-1}u\\ &= R^T(RR^T)^{-1}u = R^{\dagger}u, \end{aligned}$$

where $R^{\dagger}$ is nothing but the pseudoinverse of R. This solution x_{min_norm} serves as a starting point to the underdetermined system $u = Rx$. The complete solution set can be characterized by adding an arbitrary vector from the null space of R, which can be constructed by the rational basis for the null space of R, denoted by N. It can be confirmed that $RN = 0$ and that any vector x, where

$$x = x_{min\_norm} + Nv$$

for an arbitrary vector v satisfies $u = Rx$.

These results prove that, even if the random matrix R is known to the adversary, it is impossible to find the exact values of ***all the elements*** *in vector x of each underdetermined system of linear equations. The best we can do is to find the minimum norm solution.* However, one may ask whether it is possible to completely identify *some* elements in the vector x. Obviously, if we can find as many linearly independent equations as some unknown elements, we can partially solve the system. In the following, we will discuss this possibility by using the "l-secure" definition introduced in [43, Definition 4.1].

A coefficient matrix R is said to be l-secure if, by removing any l columns from R, the remaining submatrix still has full row rank, which guarantees that any nonzero linear combination of the row vectors of R contains at least $l + 1$ nonzero elements. Otherwise, assume there are at most l nonzero elements. Then, if we remove these l corresponding columns from R and apply the same linear combination on all the row vectors of this remaining submatrix, we will get a zero vector, which means the row vectors of this submatrix are linearly dependent and the rank of this submatrix is not of full row rank, which contradicts the l-secure definition. So, if a coefficient matrix is l-secure, each unknown variable in a linear equation is disguised by at least l other unknown variables no matter what kind of nonzero linear combination produces this equation. Now, the question is whether we can find $l + 1$ linearly independent equations that just involve these $l + 1$ unknowns? The answer is *No*. It can be proved that any $l + 1$ nonzero linear combinations of the equations contains at least $2l + 1$ unknown variables if these $l + 1$ vectors are linearly independent. The following theorem formalizes this property (which can be viewed as a generalization of [43, Theorem 4.3]).

Theorem 6.1. *Let Υ be an $(l + 1) \times m$ matrix, where each row of Υ is a nonzero linear combination of row vectors in R. If R is l-secure, the linear equations system $u = \Upsilon x$ involves at least $2l + 1$ unknown variables if these $l + 1$ vectors are linearly independent.*[4]

Proof. Since row vectors of Υ are all linearly independent, $u = \Upsilon x$ can be transformed into $u = (I : \Upsilon')x$ through a proper Gaussian elimination, where I is the $(l + 1) \times (l + 1)$ identity matrix, Υ' is a $(l + 1) \times (m - (l + 1))$ matrix, and $(I : \Upsilon')$ is a vertical concatenation of I and Υ'. Since R is l-secure, each row of $(I : \Upsilon')$ contains at least $l + 1$ nonzero entries, which corresponds to $l + 1$ unknowns. Because in each row of $(I : \Upsilon')$, there is a single 1 from I, there are at least l nonzero entries in Υ'. Thus, the whole system contains at least $2l + 1$ unknowns, with $l + 1$ unknowns being contributed by I, and at least l unknowns from Υ'. □

In summary, if a coefficient matrix is l-secure, any linear combinations of the equations contains at least $l + 1$ variables and it is not possible to find $l + 1$ linearly independent equations that just involve the same $l + 1$ variables, thus the solutions to any partial unknown variables are infinite.

Now, consider the $k \times m$ random projection matrix and the restrictions of ICA we discussed in the previous sections. When $m = 2k - 1$, after removing any $k - 1$ columns from mixing matrix R, according to the proof of Theorem 4.4, the remaining square matrix has full row rank with

3. This problem is referred to as finding a minimum norm solution to an underdetermined system of linear equations.

4. If these $l + 1$ vectors are not linearly independent, the $l + 1$ equations contain $\Gamma + l$ unknown variables. Here, Γ denotes the rank of the matrix formed by these $l + 1$ vectors.

probability 1. That means the system is $(k-1)$-secure with probability 1 when the mixing matrix R is known to the adversary, i.e., theoretically, each unknown variable is disguised by at least $k-1$ variables, and we cannot find $k-1$ linearly independent equations that just involve these variables, so the solutions are infinite. When $m > 2k-1$, the security level is even higher because we can remove more columns while keeping the submatrix full row rank (however, the accuracy of the random projection will probably be compromised if k is too small).

This result shows that, even if the random matrix R is known to the adversary, if R is $(k-1)$-secure, each unknown variable is masked by at least $k-1$ other unknown variables no matter how the equations are linear combined. So, it is impossible to find the exact value of ***any element*** *in the original data.*

Since the exact values of the original data cannot be identified, let us change gears and see how well can we estimate them if both the perturbed data and the specific random matrix are known (however, we assume the adversary does not know the true variance of the random entries, and, in practice, an estimated one may be used instead.).

Recall the projection model described in Section 5. If entries of the $k \times m$ random matrix R are independent and identically chosen from Gaussian distribution with mean zero and variance σ_r^2, given $u = \frac{1}{\sqrt{k}\sigma_r} Rx$, we can estimate x by multiplying on the left by $\frac{1}{\sqrt{k}\hat{\sigma}_r} R^T$, where $\hat{\sigma}_r$ is the estimated variance of the random entries. Note that, in practice, since the specific realization of R is disclosed, an adversary can compute $\hat{\sigma}_r$ by computing the sample variance of $r_{i,j}$. Therefore, in the following equations, we view $\hat{\sigma}_r$ as a constant. We have

$$\frac{1}{\sqrt{k}\hat{\sigma}_r} R^T u = \frac{1}{k\hat{\sigma}_r \sigma_r} R^T R x.$$

The estimation for the ith data element of vector x, denoted by $\hat{x}_i$, can be expressed as

$$\hat{x}_i = \frac{1}{k\hat{\sigma}_r \sigma_r} \sum_t \epsilon_{i,t} x_t,$$

where $\epsilon_{i,j}$ is the i,jth entry of $R^T R$. With simple mathematical derivation, we have the expectation and variance of the estimation as follows:

$$E[\hat{x}_i] = \frac{\sigma_r}{\hat{\sigma}_r} x_i,$$

$$Var[\hat{x}_i] = \frac{1}{k^2 \hat{\sigma}_r^2 \sigma_r^2} \left((2k + k^2)\sigma_r^4 x_i^2 + k\sigma_r^4 \sum_{t,t \neq i} x_t^2 \right) - \left(\frac{\sigma_r}{\hat{\sigma}_r} x_i \right)^2.$$

When the estimated variance $\hat{\sigma}_r^2 \approx \sigma_r^2$, we have

$$E[x_i - \hat{x}_i] \approx 0,$$

$$Var[x_i - \hat{x}_i] \approx \frac{2}{k} x_i^2 + \frac{1}{k} \sum_{t,t \neq i} x_t^2.$$

In summary, when the random matrix is completely disclosed, one cannot find the exact value of any element of the original data. However, by exploring the properties of the random matrix R, we can find an approximation of the original data. The distortion is zero on average, and its variance is approximately $\frac{2}{k} x_i^2 + \frac{1}{k} \sum_{t,t \neq i} x_t^2$. We view this variance as a privacy measure in the worst case. By controlling the magnitude of the vector x (which can be done by simply multiplying a scalar to each element of the vector), we can adjust the variance of the distortion of the estimation, which, in turn, changes the privacy level.

6.2 The Dimensionality and the Distribution of the Random Matrix Are Disclosed

This section studies whether an adversary can get a good estimation of the original data through a random guess of the random matrix if he or she knows the probability density function (PDF) of R and its dimensionality m.

Assume the adversary generated a random matrix $\hat{R}$ according to the PDF. Given $u = Rx$, the adversary can estimate x by multiplying on the left of u by $\frac{1}{\sqrt{k}\sigma_r} \hat{R}^T$

$$\frac{1}{\sqrt{k}\hat{\sigma}_r} \hat{R}^T u = \frac{1}{\sqrt{k}\sigma_r} \hat{R}^T \frac{1}{\sqrt{k}\sigma_r} Rx.$$

Let $\hat{\epsilon}_{i,j}$ denote the i,jth entry of $\hat{R}^T R$ such that $\hat{\epsilon}_{i,j} = \sum_t \hat{r}_{t,i} r_{t,j}$ $\forall i,j$. Let $\hat{x}_i$ denote the estimation of x_i, we have

$$\hat{x}_i = \frac{1}{k\sigma_r^2} \sum_t \hat{\epsilon}_{i,t} x_t.$$

The expectation and variance of $\hat{x}_i$ are

$$E[\hat{x}_i] = E\left[\frac{1}{k\sigma_r^2} \sum_t \hat{\epsilon}_{i,t} x_t \right] = 0,$$

$$Var[\hat{x}_i] = E\left[\frac{1}{k^2 \sigma_r^4} \left(\sum_t \hat{\epsilon}_{i,t} x_t \right)^2 \right] = \frac{1}{k} \sum_t x_t^2.$$

Here, we use the fact that $E[\hat{\epsilon}_{i,j}] = 0, E_{p \neq q}[\hat{\epsilon}_{i,p} \hat{\epsilon}_{i,q}] = 0$ and $E[\hat{\epsilon}_{i,t}^2] = k\sigma_r^4$.

This fact indicates that the adversary cannot identify the original data by a random guess of the random matrix, all she or he can get is approximately a null matrix with all entries being around 0.

6.3 The Data Inputs are Restricted to Boolean

In the discussion of Section 6.1, we do not assume any prior knowledge of the original data with the exception that it is from the continuous real domain. However, when the data inputs are restricted to Boolean, our protocol will be at a high disclosure risk. For example, suppose the adversary knows the random matrix is $(0.1, 0.3, 0.5)$ and the perturbation equation is $0.1d_1 + 0.3d_2 + 0.5d_3 = 0.9$, where (d_1, d_2, d_3) is the original data. Then, even though there is just one equation, the adversary will know that $d_1 = d_2 = d_3 = 1$. Actually, if the system of linear equations has a unique solution (either for all the unknowns or for partial unknowns), the adversary could try all possible combinations of 1 and 0 for all the data elements to obtain the correct solution. Similar results will occur if the data is discrete and the adversary knows exactly all the possible candidates. However, we need to note that, in practice, both the dimensionality of the data and the random matrix are kept secret, so the adversary does not know the equation "$0.1d_1 + 0.3d_2 + 0.5d_3 = 0.9$," but only a single number 0.9. Therefore, the random projection-based perturbation offers a reasonable protection for boolean and other discrete data.

6.4 The Distribution of the Data is Revealed

Recall in Section 4.3, we stated that, if all the sources are non-Gaussian and statistically independent, it is possible for overcomplete ICA to identify the mixing matrix up to scaling and permutation ambiguities. If the adversary also happens to know the distribution of the original data sources under this situation, overcomplete ICA could possibly reconstruct the sources in a probabilistic sense. However, in the literature, overcomplete ICA has only been treated in particular cases, and an exact recovery is still impossible. Actually, in practice, the data sets usually have more than one Gaussians and correlated components, ICA can only find the "real" hidden independent factors behind the original data, but not the data itself.

6.5 The Trouble with Malicious Parties

The perturbation technique we proposed assumes a semi-honest model, which means all the parties follow the protocol properly and there is no collusion. However, it is possible that the data miner and one of the data owners are malicious and they want to cooperatively extract the sensitive information from the other party. For example, to probe Bob's private data, Alice may reveal the secret random matrix to the data miner or the data miner may send Bob's perturbed data back to Alice. These behaviors are actually the same as disclosing the specific realization of the random matrix, which is well studied in Section 6.1.

The next section compares our perturbation technique with other existing secure inner product protocols.

7 Comparison with Other Secure Matrix Product Protocols

This paper studies the random projection-based multiplicative perturbation technique in the context of computing inner product matrix from distributed privacy sensitive data. Recently, there has been a growing body of research on secure inner product computation [43], [44], [45], [46], which looks similar with ours. However, our work distinguishes with other existing protocols in the following aspects.

First of all, the problem we are dealing with is different. Most of the existing techniques are handling a Secure Two-Party Computation model, where two parties, Alice and Bob, each having a private database, want to cooperatively conduct data-mining operations on the union of their data. However, the problem we are interested in is how a data owner can release a version of its private data with guarantees that the original sensitive information cannot be reidentified while the analytic properties of the data are preserved?

Second, the methodology for privacy protection we are investigating is different. In the SMC-based model, the inner product of two parties, Alice and Bob, is usually divided into two secret pieces, with one piece going to Alice and the other going to Bob. The computation of each inner product requires the cooperation of the two parties. However, our work explores the data perturbation technique. The private data is masked by multiplicative noise only once and, then, released to the data miner. The data owner will not participate in future data-mining activities at all.

Third, our technique requires lower communication cost when computing the inner product. By mapping the data to a lower-dimensional random space, we compress the data quite a lot, which is well suited for distributed computation problem. However, most of the existing SMC-based inner product protocol are synchronous and requires several rounds of communications between two parties for each inner product computation; therefore, they do not scale very well to large data set. Table 1 compares the communication cost of several existing secure inner product protocols with ours.

TABLE 1
Comparison of Several Secure Inner Product Protocols

Protocol	Communication Cost
Our Work	$2k$ $(k \leq 0.5n)$
[43, Protocol 2]	n
[46]	$2n$
[45]	$2n$
[44, Protocol 3]	$4mn$ (n^m is large enough)

Finally, it should be noted that most of the existing SMC-based inner product computations do not deal with the situation where one party is malicious and lies about its input. For example, if Alice replaces her input vector with $(1, 0, \ldots, 0)$, the result of the inner product tells Alice the exact value of the first element of the other party's data. However, in our model, the inner product is known to the data miner. Giving spurious input to the protocol could not let one party derive the other party's private information if the data miner does not collude with the adversary. In the worst case, Alice may reveal the secret random matrix to the data miner or the data miner may send Bob's perturbed data back to Alice. These behaviors are actually the same as disclosing the specific realization of the random matrix. In that case, the adversary still cannot compute the exact values of the original data, but only an approximation.

8 Applications

In this section, we illustrate several applications of the random projection-based perturbation technique together with the experimental results. All the data sets are chosen from the UCI Machine Learning Repository and KDD Archive without any normalization. The random matrices are generated from Gaussian distribution with mean 0 and variance 4.

8.1 Inner Product/Euclidean Distance Estimation from Heterogeneously Distributed Data

Problem. *Let X be an $m \times n_1$ data matrix owned by Alice and Y be an $m \times n_2$ matrix owned by Bob. Compute the column-wise inner product and Euclidean distance matrices of the data $(X : Y)$ without directly accessing it.*

Algorithm:

1. Alice and Bob cooperatively generate a secret random seed and use this seed to generate a $k \times m$ random matrix R.
2. Alice and Bob project their data onto $\mathbb{R}^k$ using R and release the perturbed version $U = \frac{1}{\sqrt{k}\sigma_r} RX$ and $V = \frac{1}{\sqrt{k}\sigma_r} RY$ to a third party.
3. The third party computes the inner product matrix using the perturbed data U and V and gets

TABLE 2
Relative Errors in Computing the Inner Product of Two Attributes

k	Mean(%)	Var(%)	Min(%)	Max(%)
100 (1%)	9.91	0.41	0.07	23.47
500 (5%)	5.84	0.25	0.12	18.41
1000 (10%)	2.94	0.05	0.03	7.53
2000 (20%)	2.69	0.04	0.01	7.00
3000 (30%)	1.81	0.03	0.27	6.32

TABLE 3
Relative Errors in Computing the Square of the Euclidean Distance of Two Attributes

k	Mean(%)	Var(%)	Min(%)	Max(%)
100 (1%)	10.44	0.67	1.51	32.58
500 (5%)	4.97	0.29	0.23	18.32
1000 (10%)	2.70	0.05	0.11	7.21
2000 (20%)	2.59	0.03	0.31	6.90
3000 (30%)	1.80	0.01	0.61	3.91

$$\begin{pmatrix} U^TU & U^TV \\ V^TU & V^TV \end{pmatrix} \approx \begin{pmatrix} X^TX & X^TY \\ Y^TX & Y^TY \end{pmatrix}.$$

Discussions: Similarly, the third party can compute the Euclidean distance on the perturbed data. When the data is properly normalized, the inner product matrix is nothing but the cosine angle matrix or the correlation coefficient matrix of $(X : Y)$.

Experiments: We consider the Adult database from the UCI Machine Learning Repository for the experiment. This data set was originally extracted from the 1994 census bureau database. Without loss of generality, we select the first 10,000 rows of the data with only two attributes (fnlwgt, education-num) and show how random projection preserves the inner product and (the square of) the Euclidean distance of them. Tables 2 and 3 present the results over 20 runs. Here, k is the dimensionality of the perturbed vector, and k is also represented as the percentage of the dimensionality of the original vector. It can be seen that, when the vector is reduced to 30 percent of its original size, the relative error of the estimated inner product and (the square of) Euclidean distance is only around 1.80 percent, which is pretty good. Fig. 8 illustrates how the original data is perturbed.

8.2 K-Means Clustering from Homogeneously Distributed Data

Problem. *Let X be an $m_1 \times n$ data matrix owned by Alice and Y be an $m_2 \times n$ matrix owned by Bob. Cluster the union of these two data sets $\binom{X}{Y}$ without directly accessing the raw data.*

Algorithm:

1. Alice and Bob cooperatively generate a secret random seed and use this seed to generate an $n \times k$ random matrix R.
2. Alice and Bob project their data onto $\mathbb{R}^k$ using R and release the perturbed version $U = \frac{1}{\sqrt{k}\sigma_r} XR$ and $V = \frac{1}{\sqrt{k}\sigma_r} YR$.
3. The third party does K-Means clustering over the data set $\binom{U}{V}$.

Discussions: The above algorithm is based on the fact that column-wise projection preserves the distance of row vectors. Actually, random projection maps the data to a lower-dimensional random space while maintaining much of its variance just like PCA. However, random projection only requires $O(mnk)(k << n)$ computations to project an $m \times n$ data matrix into $k \times n$ dimensions, while the computation complexity of estimating the PCA is $O(n^2m) + O(n^3)$. This algorithm can be generalized for other distance-based data-mining applications such as nested-loop outlier detection, k-nearest-neighbor search, etc. Moreover, by doing a column-wise projection and then concatenating the perturbed data vertically, we can also apply clustering algorithm on heterogeneously distributed data.

Experiments: For this task, we choose the Synthetic Control Chart Time Series data set from the UCI KDD Archive. This data set contains 600 examples of control charts, each with 60 attributes. There are six different classes of control charts: normal, cyclic, increasing trend, decreasing trend, upward shift, and downward shift. We horizontally partition the data into two subsets, perform random projections, and then conduct K-Means clustering on the union of the projected data. Table 4 shows the results. It can be seen that the clustering results are pretty good; even with a 17 percent projection rate (the number of attributes is reduced from 60 to 10), the clustering error rate is still as low as 4.33 percent.

8.3 Linear Classification

Problem. *Given a collection of sensitive data points $x_i (i = 1, 2, \ldots)$ in $\mathbb{R}^n$, each labeled as positive or negative, find a weight vector w such that $wx_i^T > 0$ for all positive points x_i and $wx_i^T < 0$ for all negative points x. Here, we assume $x_i (i = 1, 2, \ldots)$ is a row vector.*

Algorithm:

1. The data owner generates an $n \times k$ random matrix R and projects the data to $\mathbb{R}^k$ using R such that $x_i' = \frac{1}{\sqrt{k}\sigma_r} x_i R$, $\forall i$, and releases the perturbed data.

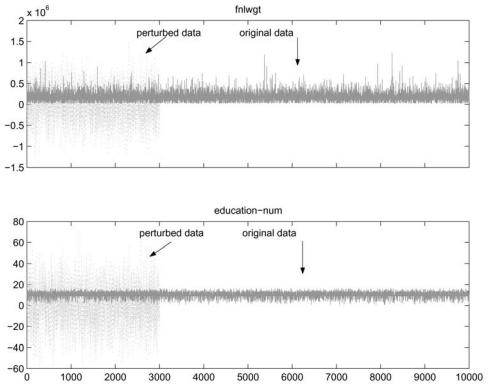


Fig. 8. Original data attributes and their perturbed counterparts. The random projection rate is 30 percent.

TABLE 4
K-Means Clustering from the Original and the Perturbed Data

#Attributes	Clustered Instances						Err Rate
	1	2	3	4	5	6	
60 (Original data)	187	25	41	34	117	196	0.00%
30 (50% Projection)	188	25	40	34	117	196	0.17%
20 (33% Projection)	182	29	36	32	128	193	2.50%
10 (17% Projection)	182	19	65	36	108	190	4.33%

2. Run the perceptron algorithm in $\mathbb{R}^k$:
   a. Let $w' = 0$. Do until all the examples are correctly classified
      - Pick an arbitrary misclassified example x'_i and let
      $$w' \leftarrow w' + \eta \cdot \texttt{classlabel}(x'_i) \cdot x'_i.$$
      Here, η is the learning rate.

Discussions: Note that, in this algorithm, the class labels are not perturbed. Future example x is labeled positive if $w'\left(\frac{1}{\sqrt{k}\sigma_r} xR\right)^T > 0$ and negative otherwise. This is actually the same as checking whether $\left(w'\frac{1}{\sqrt{k}\sigma_r} R^T\right)x^T > 0$, namely, a linear separator in the original n-dimensional space. This also implies that w' is nothing but the projection of w such that $w' = \frac{1}{\sqrt{k}\sigma_r} wR$ and, therefore,

$$w' x_i'^T = \frac{1}{\sqrt{k}\sigma_r} wR \frac{1}{\sqrt{k}\sigma_r} R^T x_i^T \approx w x_i^T.$$

This algorithm can be easily generalized for Support Vector Machine (SVM) because, in the Lagrangian dual problem of the SVM task, the relationship of the original data points is completely quantified by inner product.

Experiments: We select the Iris Plant Database for the experiment. This is a very simple data set with 150 instances and only four numeric attributes. We will show that, even for such a small data set, our algorithm still works well. The data set contains three classes of 50 instances each, where each class refers to a type of iris plant (Iris-setosa, Iris-versicolor, and Iris-virginica). We manually merge Iris-setosa and Iris-versicolor together so that we can do a binary classification on this data. The projection rate is 50 percent; hence, the data has only two attributes left after perturbation. We perform a voted perceptron learning on both the original data and the perturbed data. The accuracy on the original data over 10-fold cross validation is 94.67 percent. The classification results on the perturbed data over 10-fold cross validation are demonstrated in Table 5. It shows that the average accuracy on the perturbed data is 86.67 percent, which is 91.55 percent as good as the results over the original data.

The following section concludes this paper.

TABLE 5
Classification on the Perturbed Iris Plant Data over 10-Fold Cross Validation

Accuracy(%)	1	2	3	4	5
	66.67	80.00	100.00	80.00	93.33
	6	7	8	9	10
	86.67	80.00	93.33	93.33	93.33
Mean(%)	86.67				
Std(%)	9.43				

9 Conclusions and Future Work

This paper explores the use of random projection matrices as a tool for privacy preserving data mining. It proves that, after perturbation, the distance-related statistical properties of the original data are still well maintained without divulging the dimensionality and the exact data values. The experimental results demonstrate that this technique can be successfully applied to different kinds of data mining tasks, including inner product/Euclidean distance estimation, correlation matrix computation, clustering, outlier detection, linear classification, etc. The random projection-based technique may be even more powerful when used with some other geometric transformation techniques like scaling, translation, and rotation. Combining this with SMC-based techniques offers another interesting direction.

Acknowledgments

This research is supported by the US National Science Foundation Grant IIS-0329143. Hillol Kargupta would also like to acknowledge support from the US National Science Foundation CAREER award IIS-0093353.

References

[1] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *Int'l J. Uncertainty, Fuzziness, and Knowledge-Based Systems,* vol. 10, no. 5, pp. 557-570, 2002.
[2] S. Chawla, C. Dwork, and F. McSherry, "Toward Privacy in Public Databases," *Proc. Second Theory of Cryptography Conf. (TCC'05),* Feb. 2005.
[3] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," *Proc. IEEE Int'l Conf. Data Mining,* Nov. 2003.
[4] W.B. Johnson and J. Lindenstrauss, "Extensions of Lipshitz Mapping into Hilbert Space," *Contemporary Math.,* vol. 26, pp. 189-206, 1984.
[5] C.K. Liew, U.J. Choi, and C.J. Liew, "A Data Distortion by Probability Distribution," *ACM Trans. Database Systems (TODS),* vol. 10, no. 3, pp. 395-411, 1985.
[6] E. Lefons, A. Silvestri, and F. Tangorra, "An Analytic Approach to Statistical Databases," *Proc. Ninth Int'l Conf. Very Large Data Bases,* pp. 260-274, Nov. 1983.
[7] N.R. Adam and J.C. Worthmann, "Security-Control Methods for Statistical Databases: A Comparative Study," *ACM Computing Surveys (CSUR),* vol. 21, no. 4, pp. 515-556, 1989.
[8] R. Agrawal and R. Srikant, "Privacy Preserving Data Mining," *Proc. ACM SIGMOD Conf. Management of Data,* pp. 439-450, May 2000.
[9] J.J. Kim and W.E. Winkler, "Multiplicative Noise for Masking Continuous Data," Technical Report Statistics #2003-01, Statistical Research Division, US Bureau of the Census, Washington D.C., Apr. 2003.
[10] S. Warner, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *J. Am. Statistical Assoc.,* vol. 60, pp. 63-69, 1965.
[11] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," *Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD'02),* July 2002.

[12] A. Evfimevski, J. Gehrke, and R. Srikant, "Limiting Privacy Breaches in Privacy Preserving Data Mining," *Proc. ACM SIGMOD/PODS Conf.,* June 2003.

[13] S. Agrawal and J.R. Haritsa, "A Framework for High-Accuracy Privacy-Preserving Mining," *Proc. 21st Int'l Conf. Data Eng. (ICDE'05),* pp. 193-204, Apr. 2005.

[14] T. Dalenius and S.P. Reiss, "Data-Swapping: A Technique for Disclosure Control," *J. Statistical Planning and Inference,* vol. 6, pp. 73-85, 1982.

[15] S.E. Fienberg and J. McIntyre, "Data Swapping: Variations on a Theme by Dalenius and Reiss," technical report, Nat'l Inst. of Statistical Sciences, Research Triangle Park, NC, 2003.

[16] A.C. Yao, "How to Generate and Exchange Secrets," *Proc. 27th IEEE Symp. Foundations of Computer Science,* pp. 162-167, 1986.

[17] B. Pinkas, "Cryptographic Techniques for Privacy Preserving Data Mining," *SIGKDD Explorations,* vol. 4, no. 2, pp. 12-19, 2002.

[18] O. Goldreich, *The Foundations of Cryptography,* vol. 2, chapter 7. Cambridge Univ. Press, 2004.

[19] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Zhu, "Tools for Privacy Preserving Distributed Data Mining," *ACM SIGKDD Explorations,* vol. 4, no. 2, 2003.

[20] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-Art in Privacy Preserving Data Mining," *ACM SIGMOD Record,* vol. 3, no. 1, pp. 50-57, Mar. 2004.

[21] B.-H. Park and H. Kargupta, "Distributed Data Mining," *The Handbook of Data Mining, ser. Human Factors and Ergonomics,* pp. 341-358, N. Ye, ed., Lawrence Erlbaum Associates, Inc., 2003.

[22] K. Liu, H. Kargupta, J. Ryan, and K. Bhaduri, "Distributed Data Mining Bibliography," http://www.csee.umbc.edu/~hillol/DDMBIB/, 2004.

[23] S. Merugu and J. Ghosh, "Privacy-Preserving Distributed Clustering Using Generative Models," *Proc. Third IEEE Int'l Conf. Data Mining (ICDM'03),* Nov. 2003.

[24] D. Meng, K. Sivakumar, and H. Kargupta, "Privacy Sensitive Bayesian Network Parameter Learning," *Proc. Fourth IEEE Int'l Conf. Data Mining (ICDM'04),* Nov. 2004.

[25] M.J. Atallah, E. Bertino, A.K. Elmagarmid, M. Ibrahim, and V.S. Verykios, "Disclosure Limitation of Sensitive Rules," *Proc. IEEE Knowledge and Data Eng. Workshop,* pp. 45-52, 1999.

[26] V.S. Verykios, A.K. Elmagarmid, B. Elisa, Y. Saygin, and D. Elena, "Association Rule Hiding," *IEEE Trans. Knowledge and Data Eng.,* 2003.

[27] Y. Saygin, V.S. Verykios, and C. Clifton, "Using Unknowns to Prevent Discovery of Association Rules," *SIGMOD Record,* vol. 30, no. 4, pp. 45-54, Dec. 2001.

[28] E.W. Weisstein et al., "Orthogonal Transformation," MathWorld-A Wolfram Web Resource, 2004.

[29] S.R.M. Oliveira and O.R. Zaïane, "Privacy Preserving Clustering by Data Transformation," *Proc. 18th Brazilian Symp. Databases,* pp. 304-318, Oct. 2003.

[30] P. Common, "Independent Component Analysis: A New Concept?" *IEEE Trans. Signal Processing,* vol. 36, pp. 287-314, 1994.

[31] A. Hyvärinen and E. Oja, "Independent Component Analysis: Algorithms and Applications," *Neural Networks,* vol. 13, no. 4, pp. 411-430, June 2000.

[32] X.-R. Cao and R.-W. Liu, "A General Approach to Blind Source Separation," *IEEE Trans. Signal Processing,* vol. 44, pp. 562-571, 1996.

[33] M.L. Eaton and M.D. Perlman, "The Non-Singularity of Generalized Sample Covariance Matrices," *The Annals of Statistics,* vol. 1, no. 4, pp. 710-717, 1973.

[34] A.K. Gupta and D.K. Nagar, *Matrix Variate Distributions,* H. Brezis, R.G. Douglas, and A. Jeffrey, eds. Chapan & Hall/CRC, 1999.

[35] W. Hardle and L. Simar, *Applied Multivariate Statistical Analysis.* chapter 2.1, pp. 57-63, Springer, 2003.

[36] J. Eriksson and V. Koivunen, "Identifiability and Separability of Linear ICA Models Revisited," *Proc. Fourth Int'l Symp. Independent Component Analysis and Blind Signal Separation (ICA2003),* Apr. 2003.

[37] M.S. Lewicki and T.J. Sejnowski, "Learning Overcomplete Representations," *Neural Computation,* vol. 12, no. 2, pp. 337-365, 2000.

[38] R. Hecht-Nielsen, "Context Vectors: General Purpose Approximate Meaning Representations Self-Organized from Raw Data," *Computational Intelligence: Imitating Life,* pp. 43-56, 1994.

[39] R.I. Arriaga and S. Vempala, "An Algorithmic Theory of Learning: Robust Concepts and Random Projection," *Proc. 40th Ann. Symp. Foundations of Computer Science,* pp. 616-623, Oct. 1999.

[40] S. Kaski, "Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering," *Proc. Int'l Joint Conf. Neural Networks (IJCNN'98),* vol. 1, pp. 413-418, 1998.

[41] C. Giannella, K. Liu, T. Olsen, and H. Kargupta, "Communication Efficient Construction of Decision Trees over Heterogeneously Distributed Data," *Proc. Fourth IEEE Int'l Conf. Data Mining (ICDM'04),* Nov. 2004.

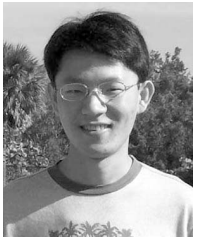
[42] J.W. Demmel and N.J. Higham, "Improved Error Bounds for Underdetermined System Solvers," Technical Report CS-90-113, Computer Science Dept., Univ. of Tennessee, Knoxville, TN, Aug. 1990.

[43] W. Du, Y.S. Han, and S. Chen, "Privacy-Preserving Multivariate Statistical Analysis: Linear Regression and Classification," *Proc. 2004 SIAM Int'l Conf. Data Mining (SDM04),* Apr. 2004.

[44] M.J. Atallah and W. Du, "Secure Multi-Party Computational Geometry," *Proc. WADS2001: Seventh Int'l Workshop on Algorithms and Data Structures,* pp. 165-179, Aug. 2001.

[45] W. Du and Z. Zhan, "Building Decision Tree Classifier on Private Data," *Proc. IEEE Int'l Conf. Privacy, Security, and Data Mining,* pp. 1-8, Dec. 2002.

[46] J.S. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," *Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* July 2002.



Kun Liu received the BS degree from the Department of Computer Science and Technology at Nankai University in 2001. He is currently a PhD candidate in the Department of Computer Science and Electrical Engineering at the University of Maryland, Baltimore County. His research interests include privacy preserving data mining, distributed data mining, and machine learning.

Hillol Kargupta received the PhD degree in computer science from the University of Illinois at Urbana-Champaign in 1996. He is an associate professor in the Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County. He is also a cofounder of Agnik LLC, a ubiquitous data intelligence company. His research interests include mobile and distributed data mining and computation in biological processes of gene expression. Dr. Kargupta won a US National Science Foundation CAREER award in 2001 for his research on ubiquitous and distributed data mining. He, along with his coauthors, received the best paper award at the 2003 IEEE International Conference on Data Mining for a paper on privacy-preserving data mining. He won the 2000 TRW Foundation Award and the 1997 Los Alamos Award for Outstanding Technical Achievement. His research has been funded by the US National Science Foundation, US Air Force, Department of Homeland Security, NASA, and various other organizations. He has published more than 80 peer-reviewed articles in journals, conferences, and books. He has coedited two books: *Advances in Distributed and Parallel Knowledge Discovery*, AAAI/MIT Press, and *Data Mining: Next Generation Challenges and Future Directions*, AAAI/MIT Press. He is an associate editor of the *IEEE Transactions on Knowledge and Data Engineering* and the *IEEE Transactions on Systems, Man, and Cybernetics, Part B*. He regularly serves on the organizing and program committees of many data mining conferences. More information about him can be found at http://www.cs.umbc.edu/~hillol. He is a senior member of the IEEE.

Jessica Ryan received the undergraduate degree in computer engineering from the University fo Maryland, Baltimore County. She has published several papers on data mining and related research areas.