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

Kun Liu, Hillol Kargupta, 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, 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. It explores Independent Component Analysis as a possible tool for breaching privacy in multiplicative perturbation-based model of privacy-protection. The paper also proposes a random projection-based technique to transform the data while preserving its certain statistical characteristics. The paper presents extensive theoretical analysis and experimental results. Experiments demonstrate that the proposed technique is effective and can be successfully applied for different types of privacy-preserving data mining applications.

Index Terms

Random projection, multiplicative data perturbation, privacy preserving data mining

I. 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 counter-terrorism 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 health-care data for detection of bio-terrorism 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 re-identification attacks have emerged which can link different public data sets to relocate the original subjects [1]. This calls for well-designed techniques that pays careful attention to hiding privacy-sensitive information while preserving the inherent statistical dependencies which are important for data mining applications.

This paper considers a randomized multiplicative data perturbation technique for distributed privacy preserving data mining. It is motivated by the work presented elsewhere [2] 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 data. The transformed data in the new representation is released to the data miner. The approach is fundamentally based on the Johnson-Lindenstrauss lemma [3] 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 are maintained within an arbitrarily small factor. Therefore, by projecting the data onto a *random* subspace, we will dramatically change its original form while preserving much of its underlying statistical characteristics.

This paper studies the random projection-based multiplicative perturbation technique in the context of computing inner product matrix from distributed data, which 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. This paper first introduces a random orthogonal transformation-based data perturbation approach which preserves the length and distance between the original data vectors. Then it brings in Independent Component Analysis (ICA) as a possible tool for breaching privacy. The random projection-based perturbation technique is presented next as an extension to improve the privacy level. Several real data mining applications, e.g. distributed inner product/Euclidean distance estimation, distributed clustering, linear classification etc., and experiments are provided to justify the effectiveness of this technique. In our opinion, the field of privacy preserving data mining is still in its early stage; the techniques for proving the correctness and quantifying privacy preserving capabilities are still in the development. Although our analytical and experimental results look promising, we make cautious claims. The primary objective of this paper is to explore the properties of random projection-based multiplicative noise as a tool for preserving privacy.

The remainder of this paper is organized as follows. Section II offers an overview of the related work in privacy preserving data mining. Section III discusses the random orthogonal transformation-based perturbation technique in the context of inner

product computation. Section IV explores the potential vulnerability of this method from the perspective of Independent Component Analysis (ICA). Section V presents the leading role of this paper – a random projection-based multiplicative data perturbation technique for privacy preserving data mining. Section VI gives a further detailed analysis about the privacy issues. Section VII illustrates several data mining tasks that this technique can be applied to, together with experimental results from the real world data sets. Finally, Section VIII concludes this paper and outlines the future research.

II. Related Work

Preserving data privacy is becoming an increasingly important issue in many data mining applications. Privacy of the data can depend on many different aspects, often dictated by the characteristics of the domain. Sometimes, individual or organizational entities may not be willing to divulge the individual values of records; sometimes, the "patterns" detected by a data mining system may be used in a counter-productive manner that violates the privacy of an individual or a group of individuals. Given a set of privacy constraints, the goal of a privacy preserving data mining system is to extract well-defined class of patterns from the data without unnecessarily violating the privacy constraints. There exists a growing body of literature on this topic. In the following of this section, we provide a brief description of the various techniques and methodologies.

A. Data Perturbation

Data perturbation approaches fall into two main categories, which we call the probability distribution approach and the value distortion approach. The probability distribution approach replaces the data with another sample from the same (estimated) distribution [4] or by the distribution itself [5]. On the other hand, the value distortion approach perturbs the values of data elements or attributes directly by some additive or multiplicative noise before it is released to the data miner. Some randomized methods [6] for categorical data may also be classified under this category. A detailed discussion of this approach can be found elsewhere [7]. In this paper, we mainly focus on the value distortion techniques.

Agrawal et al. [8] proposed a value distortion technique to protect the privacy by adding random noise from a Gaussian distribution to the actual data. They showed that this technique appears to mask the data while allowing extraction of certain patterns like the original data distribution and decision tree models with good accuracy. This approach has been further extended in [9] where an Expectation-Maximization-based (EM) algorithm is applied for a better reconstruction of the distribution. An information theoretic measure for quantifying the privacy is also discussed there. Evfimievski et al. [10], [11] and Rizvi [12] considered the same approach in the context of association rule mining and suggested techniques for limiting privacy breaches. More recently, Kargupta et al. [2] 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. Given the large body of existing signal-processing literature on filtering random additive noise, the utility of random additive noise for privacy-preserving data mining is not quite clear.

The possible drawback of additive noise makes one wonder about the possibility of using multiplicative noise [13], [14] for protecting the privacy of the data while maintaining some of the original analytic properties. Two basic forms of multiplicative noise have been well studied in the statistics community. The first method is based on generating random numbers that have a truncated Gaussian distribution with mean one and small variance, and multiplying each element of the original data by the noise. The second method is to take a logarithmic transformation of the data first (for positive data only), compute the covariance, and generate random noise following a multivariate Gaussian distribution with mean zero and variance equaling a constant times the covariance computed in the last step, then add this noise to each element of the transformed data, and finally take the antilog of the noise-added data. Multiplicative perturbation overcomes the scale problem, and it has been proved that the mean and variance/covariance of the original data elements can be estimated from the perturbed version. In practice, the first method is good if the data disseminator only wants to make minor changes to the original data; however the second method assures higher security than the first one and still maintains the data utility very well in the log-scale.

One of the main problems of the traditional additive perturbation and multiplicative perturbation is that they perturb each data element independently, and therefore the similarity between attributes or observations which are considered as vectors in the original data space is not well preserved. Many distance/similarity based data mining applications are thus hurt. In this paper, we propose an alternative approach to perturb data using multiplicative noise. Instead of applying noise to each element of the data, we make use of random projection matrices for constructing a perturbed representation of the data. This technique not only protects the confidentiality but also preserves certain statistical properties of the data, e.g., the inner product, the angles, the correlation, and the Euclidean distance. Moreover, this kind of perturbation also allows dimension reduction, which is well suited for distributed data mining problem where multiple parties want to collaboratively conduct computation on the union of their private data with as little communication as possible.

B. Data Swapping

The basic idea of data swapping, which was first proposed by Dalenius and Reiss [15], is to transform the database by switching a subset of attributes between selected pairs of records so that the lower order frequency counts or marginal

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 [16] for a thorough treatment.

C. Secure Multi-party Computation

The Secure Multi-party Computation (SMC) [17], [18] approach considers the problem of evaluating a function of two or more parties' secret inputs, such that each party finally gets the designated function output and nothing else is revealed, except what is implied by the party's own inputs and outputs. The circuit evaluation protocol, oblivious transfer and oblivious polynomial evaluation [19], homomorphic encryption [20], commutative encryption [21], Yao's millionaire problem (secure comparison) [22] and some other cryptographic techniques serve as the building blocks of SMC. Detailed discussions of Secure Multi-party Computation framework can be found elsewhere [23], [24]. It has been shown that any function that can be expressed by an arithmetic circuit over GF(2) is privately computable. However, the general approach does not scale well to large asynchronous distributed environments with voluminous data sets.

The work presented in [25] introduced a variety of new SMC applications and open problems for a spectrum of cooperative computation domains. These problems include privacy preserving information retrieval [26], privacy preserving geometric computation [27], privacy preserving statistical analysis [28], [29], and privacy preserving scientific computations [30], etc. The work in [31] discussed several SMC protocols to securely compute sum, set union, size of set intersection and inner product. These protocols are directly applied for privacy preserving association rule mining from vertically partitioned data [32] and horizontally partitioned data [33], clustering with distributed EM mixture modeling [34], and K-Means clustering over vertically partitioned data [35]. The SMC ideas have also been applied for privacy preserving distributed decision tree induction [36], [37], naive Bayes classification [38] for horizontally partitioned data, privacy preserving Bayesian network structure computation for vertically partitioned data [39], and many else.

D. Distributed Data Mining

The distributed data mining (DDM) [40]–[42] 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 [43] proposed a paradigm for clustering distributed privacy sensitive data in an unsupervised or a semi-supervised scenario. In this algorithm, each local data site builds a model and transmit only the parameters of the model to the central site where a global clustering model is constructed. Several other distributed algorithms, e.g., the meta-learning approach [44], the Fourier spectrum-based decision tree integration approach [45], and the collective principal component analysis-based clustering algorithm [46], are also potentially suitable for privacy preserving mining from multi-party data. A distributed privacy-preserving algorithm for Bayesian network parameter learning is reported elsewhere [47].

E. Rule Hiding

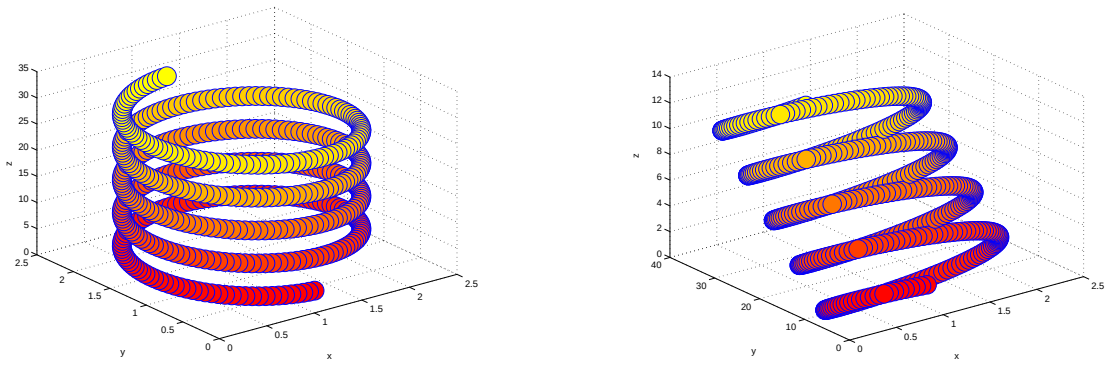
The main objective of rule hiding is to transform the database such that the sensitive rules are masked, and all the other underlying patterns can still be discovered.

The work in [48] 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 [49], [50] 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 [51] 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.

The work in [52] presented a new framework for dealing with inference problem which arises from parsimonious downgrading. Here "parsimonious downgrading" refers to the phenomenon of trimming out sensitive information from a data set when it is transferred from a secure environment (referred to as High) to a public domain (referred to as Low). This framework has been applied for decision tree classification analysis such that the receiver of the data will unable to build informative models for the data that is not downgraded. The following section considers random orthogonal transformations and explores the privacy-preserving (or lack of it) properties of such transformations.

III. Random Orthogonal Transformation

This section presents a 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.



(a) (b)

Fig. 1. (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.

An orthogonal transformation [53] is a linear transformation $R : \Re^m \rightarrow \Re^m$ 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 considering the following linear transformation of the two data sets.

$$U = XR, \quad \text{and} \quad V = YR; \quad \text{then we have,} \tag{1}$$

$$UU^T = XX^T, \quad \text{and} \quad VV^T = YY^T \quad \text{and} \tag{2}$$

$$UV^T = XRR^TY^T = XY^T$$

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 $\begin{pmatrix} X \\ Y \end{pmatrix}$ can still be perfectly computed there, where $\begin{pmatrix} X \\ Y \end{pmatrix}$ 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.

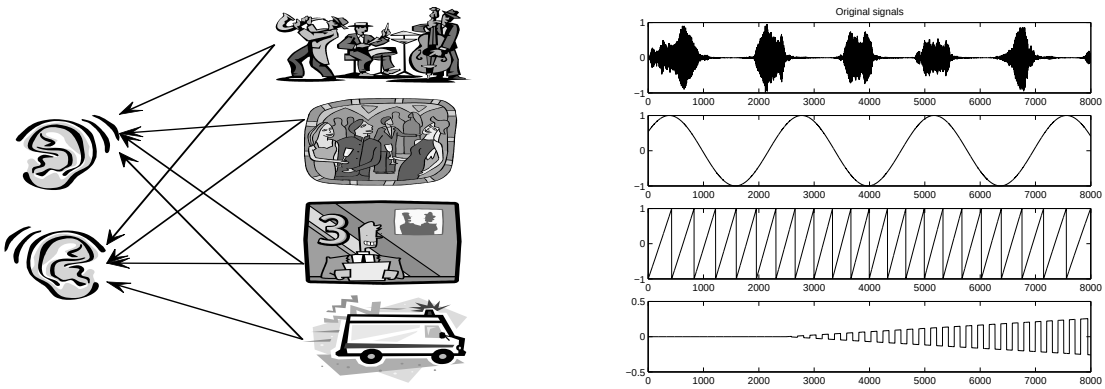
In the meantime, we can imagine both the privacy sensitive data and the transformation procedure are inside a black box, the perturbed data is the only output to the third party, the observer. For any orthogonal matrix $\hat{R}$, there always exists a matrix $\hat{X} = U\hat{R}^T$, such that $\hat{X}\hat{R}$ is still equal to U. This means that there are actually infi nite number of inputs and transformation procedures that can simulate the output, while the observer has no idea what has happened inside. So random orthogonal transformation seems to be a good way to protect the data privacy while preserving its utility. A similar approach using random rotation [1] to protect the data privacy has also been considered in [54].

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 identify the real values of the data through a proper rotation. Figure 1(a) and 1(b) illustrate how the random orthogonal transformation works in a 3D space. It can be seen that the data are not very well masked after transformation. 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 perturbation technique to improve the privacy level while preserving the data utilities.

IV. Independent Component Analysis

Independent Component Analysis (ICA) [55], [56] 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

[1]Rotation is not necessarily an orthogonal transformation, however, an orthogonal transformation corresponds to a rotation.



(a) (b)

Fig. 2. (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.

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 Figure 2(a)). Imagine you are in a cocktail party, even if different kinds of background sounds are mixed together, e.g., music, other people's chat, television news report, or even a siren from the passing-by ambulance, you still have no problem to identify 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.

A. ICA Model

The basic ICA model can be defined as follows:

$$u(t) = Rx(t) \tag{3}$$

where $x(t) = (x_1(t), x_2(t), \ldots, x_m(t))^T$ denotes the m-dimensional vector collecting the m independent unknown source signals $x_i(t), i = 1, 2, \ldots, m$ at time t. 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 [2], 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 statically independent audio signals, denoted as a 4×8000 matrix (shown in Figure 2(b)). A linear mixture of these signals (shown in Figure 3(a)) is generated by pre-multiplying a random 4×4 non-singular matrix to them. The goal of ICA is to recover the original signals using only the mixture. Figure 3(b) gives the estimated signals through ICA. It can be observed that the basic structure of the original signals are very well reconstructed. However, the order and the amplitude of the recovered signals are not necessarily the same as those of the original ones.

B. 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,

$$y(t) = Bu(t) \tag{4}$$

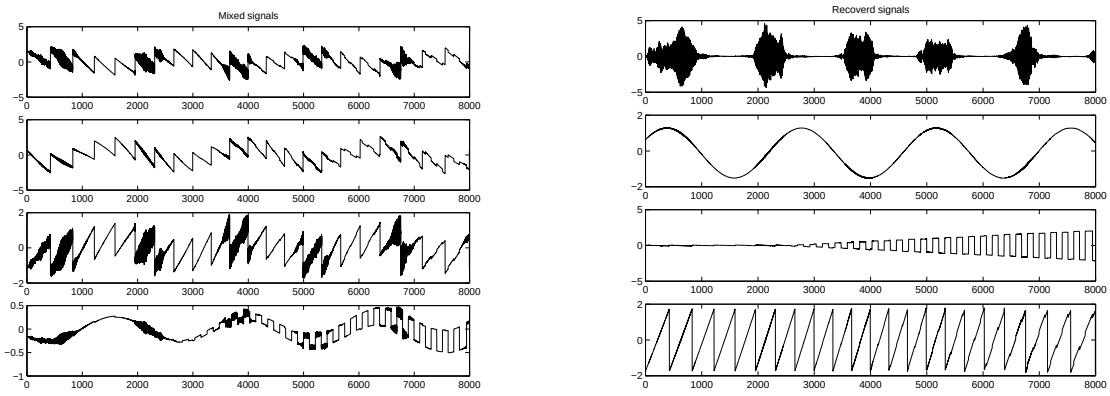
where B is an $l \times k$ dimensional separating matrix. Combining Eq. 3 and Eq. 4 together, we get

$$y(t) = BRx(t) = Zx(t), \tag{5}$$

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

[2]A matrix obtained by permuting the i-th and j-th rows of the identity matrix with $i < j$ is called permutation matrix.



(a) (b)

Fig. 3. (a) Linear mixture of the original source signals using a square random matrix. (b) Recovered signals using ICA. It can be observed that the basic structure of the original signals are very well reconstructed. However, the order and the amplitude of the recovered signals are not necessarily the same as those of the original ones.

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 kind of B if it is proved to exist. We will focus on the first step.

In general, by imposing the following fundamental restrictions [55], [57]–[59], 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:

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

  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 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 ICA can find a good approximation of the original signals from their perturbed version. Figure 3(a) and 3(b) 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 are more than one Gaussian signals, 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. Detailed analysis can be found elsewhere [57].

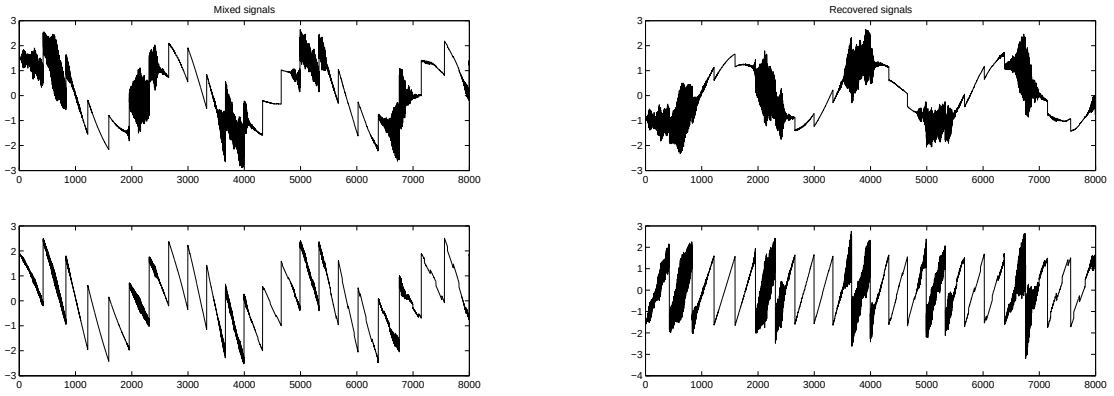
When $l \leq k < m$ (i.e., the number of sources is greater than the number of receivers), [3] it is generally not possible to design linear filters to simultaneously recover all these signals. This kind of separation problem is termed as overcomplete ICA or under-determined source separation. Cao et al. [57] analyzed the conditions for the existence of the separating matrix B.

We first introduce two definitions (Definition 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): [57] 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, j-th 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.

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

[3] 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.



(a) (b)

Fig. 4. (a) Linear mixture of the original four source signals (as shown in Figure 2(b)) with 50% 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.

Definition 4.2 (l-row Decomposable): [57] 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.,

$$y_i \quad = \quad \sum_{j \in S_i} z_{i,j} x_j, \text{ i = 1, 2, . . . , l}$$

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: [57] 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 [57]. ■

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. Then it is not possible to separate out any of the source signals. The following theorem characterized this property:

Theorem 4.4: Any $k \times m$ random matrix with $m \geq 2k - 1$ ($m \geq 2$) and independent random entries chosen from continuous distribution 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 non-empty sets, at least one set will have at least k members. Thus this set of columns contains a $k \times k$ sub-matrix. The determinant of the sub-matrix is a polynomial function of the matrix entries. Except for a set of measure zero, the probability that the polynomial is zero is 0. So with probability 1, this $k \times k$ sub-matrix will be nonsingular, i.e., the k column vectors of the sub-matrix span $\Re^k$ Euclidean space. Therefore there is always one vector in one group that belongs to the space spanned by the other group, which does not satisfy Theorem 4.3. ■

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. Figure 4(a) and 4(b) depict this property. It can be seen that after 50% 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 non-square 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 non-one (element) -to-one (element) map, which is totally different from the traditional data perturbation technique, and therefore it is even harder for the adversary to re-identify the sensitive data.

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 scale 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 receiver 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 independent random entries chosen from continuous distribution), there is no linear method to find a matrix B that can separate out any of the source signals.

C. 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 the it is unique up to a right multiplication by a diagonal and a permutation matrix [59, 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. Lewicki et al. [60] proposed a generalized method for learning overcomplete ICA in which the source signals were assumed to have a sparse distribution, e.g., Laplacian distribution. Several other similar solutions to the separation of independent components from their overcomplete mixtures have been proposed for super-Gaussian sources [61]–[63]. However, if any Gaussian signals were allowed, the mixing matrix is not identifiable [58], and the distribution of the source signals are not unique [59, Example 2 and 4]. Again, if the sources are correlated, they will cluster in the same group and only the real independent components hidden behind them are possible to be found. In the following section, we propose a random projection-based multiplicative perturbation technique. By letting the random matrix super non-square, 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.

V. Random Projection-based Multiplicative Perturbation

This section revisits the idea of multiplicative perturbation in the context of lessons we learnt by studying the properties of ICA. It particularly studies random projection matrices in the context of computing inner product and Euclidean distance without allowing direct access to the original data.

A. 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 [3] 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 $\Re^m$, there is a map $f : \Re^m \to \Re^k$ such that for all $x, y \in S$,

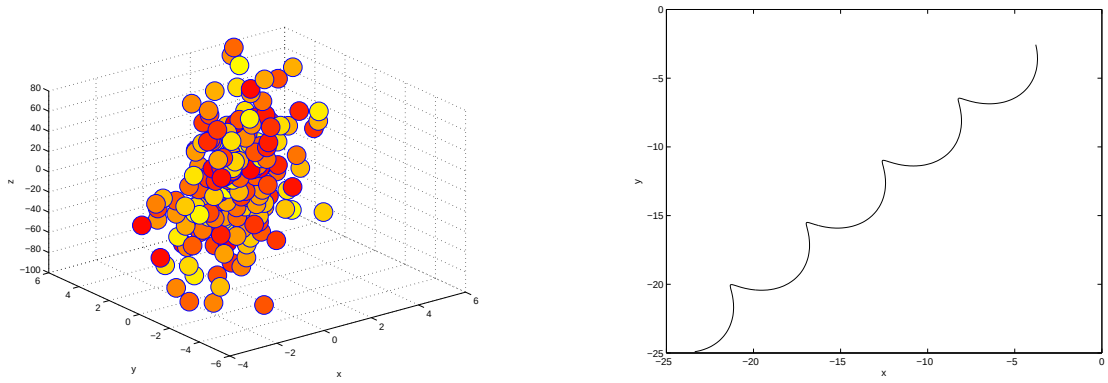
$$(1-\epsilon)||x-y||^2 \leq ||f(x)-f(y)||^2 \leq (1+\epsilon)||x-y||^2$$

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 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 how the random projection technique perturbs the data, we did both row-wise and column-wise projection of the sample data given in Figure 1(a). The results are shown in Figure 5(a) and 5(b). 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 VI. 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,

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



(a) (b)

Fig. 5. (a) The perturbed data after a row-wise random projection which reduces 50% 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 N(0,1) and the original data is given in Figure 1(a).

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

$$\begin{aligned} \epsilon_{i,j} &= \sum_{t=1}^{p} r_{t,i} r_{t,j} \\ E[\epsilon_{i,j}] &= E[\sum_{t=1}^{p} r_{t,i} r_{t,j}] \\ &= \sum_{t=1}^{p} E[r_{t,i} r_{t,j}] \end{aligned}$$

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,

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

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

Intuitively, this result echoes the observation made elsewhere [64] that in a high-dimensional space, vectors with random directions are almost orthogonal. A similar result was proved elsewhere [65].

Corollary 5.3: 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 each column vector (resp., row vector) of R is normalized to have a unit length, we have

$$E[R^T R] = I \text{ (resp., } E[RR^T] = I)$$

Lemma 5.2 can be used to prove the following results.

Lemma 5.4: **[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

$$U = \frac{1}{\sqrt{k}\sigma_r} RX, \quad \text{and} \quad V = \frac{1}{\sqrt{k}\sigma_r} RY; \quad \text{then} \quad E[U^T V] = X^T Y \tag{6}$$

Lemma 5.5: **[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

$$U = \frac{1}{\sqrt{k}\sigma_r} XR, \qquad \text{and} \qquad V = \frac{1}{\sqrt{k}\sigma_r} YR; \quad \text{then} \tag{7}$$
$$E[UV^T] = XY^T$$

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 if the data is properly normalized, the inner product is directly related to the cosine angle, the correlation, and even the Euclidean distance of the vectors. To be more specific,

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

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

the Euclidean distance of x and y is

$$dist(x,y) = \sqrt{\sum_i (x_i - y_i)^2} = \sqrt{\sum_i x_i^2 + \sum_i y_i^2 - 2\sum_i x_i y_i} = \sqrt{2 - 2x^T y}$$

.

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

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

Thus if the data owners project their data to a lower dimensional *random* space and then only release the perturbed version to a third party, the statistical properties of the original data are still maintained. On the one hand, given 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 VI, we shall give a further analysis about the privacy.

B. 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 independent and identically 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,j-th 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 $E[\epsilon_{i,j}] = 0$, $Var[\epsilon_{i,j}] = k\sigma_r^4$, $\forall i, j, i \neq j$. The following lemma gives the mean and variance of the projection error.

Lemma 5.6: Let x, y be two data vectors in $\Re^m$. 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 variance σ_r^2. Further let

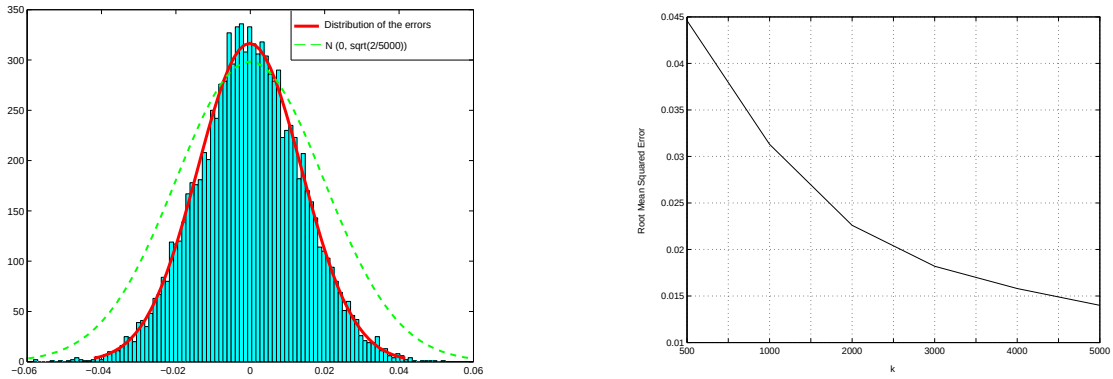
$$u = \frac{1}{\sqrt{k}\sigma_r} Rx, \qquad \text{and} \qquad v = \frac{1}{\sqrt{k}\sigma_r} Ry \text{ , then}$$
$$E[u^T v - x^T y] = 0$$
$$Var[u^T v - x^T y] = \frac{1}{k}(\sum_i x_i^2 \sum_i y_i^2 + (\sum_i x_i y_i)^2)$$

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 the Appendix. ■

Lemma 5.6 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 distortion also has a 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 10000 observations and 100 attributes. Rows of the data set correspond to observations, columns correspond to attributes. We normalize all the attributes to unity and compare the column-wise inner



(a) (b)

Fig. 6. (a) Distribution of the error of the estimated inner product matrix over two distributed data sets. Each data set contains 10000 records and 100 attributes. $k = 50\% \times 10000 = 5000$ (50% 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.

product of these two data sets before and after random projection. Figure 6(a) gives the results, and it depicts that even under 50% data projection rate (when $k = 5000$), the inner product still preserves very well after perturbation, and the error indeed follows Gaussian distribution with mean zero and variance less than $k/2$. Figure 6(b) 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 [66], 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. In Section VII, we will show this technique works pretty well for real world data sets even when they are not normalized.

By applying Lemma 5.6 to the vector $x - y$, we have

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

If x and y are normalized to unity,

$$Var[||u - v||^2 - ||x - y||^2] \leq \frac{8}{k}$$

where $||x - y||^2 = (x - y)^T(x - y)$ is the square of the Euclidean distance of x and y. Moreover, as a generalization of [65][Theorem 2], we also have the probability bound of the Euclidean distance as follows:

Lemma 5.7: Let x, y be two data vectors in $\Re^m$. 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 variance σ_r^2. Further let

$$u = \frac{1}{\sqrt{k}\sigma_r} Rx, \qquad \text{and} \qquad v = \frac{1}{\sqrt{k}\sigma_r} Ry \text{ , then}$$

$$Pr\{(1-\epsilon)||x-y||^2 \leq \quad ||u-v||^2 \quad \leq (1+\epsilon)||x-y||^2\} \geq 1 - 2e^{-(\epsilon^2 - \epsilon^3)\frac{k}{4}}$$

for any $0 < \epsilon < 1$.

Proof: Directly follows the proof of [65][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. In the next section, we shall give a detailed analysis about the privacy.

VI. Privacy Analysis

In the previous sections, we claimed that the perturbation method guarantees that a complete disclosure will not occur, which is based on the premise that the possible solutions are infinite when the number of equations is less than the number of unknowns. In this section, we shall give a further analysis on how much confidentiality the perturbation method can protect. First we shall prove that it is impossible to find the exact values of any elements of the original data even if the secret key, i.e. the random matrix is completely disclosed. We will also show how closely one can estimate the original data from its perturbed version in this case. Then we shall analyze whether one can recover the original data by a random guess of the

random matrix if he or she knows the probability density function of the random matrix. Finally we summarize our discussions and suggest some procedures that can further improve the privacy level.

A. The Specific Realization of the Random Matrix Is Disclosed

Consider the model $U = RX$, where $R \in \Re^{k\times m}$ with $k < m$, and $X \in \Re^{m\times n}$. The 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, the solution is never unique. In practice, the system can be analyzed by the QR factorization [67], [68] 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 $||x||_2$ [4]:

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

where $R^\dagger$ is nothing but the pseudo-inverse 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 $u = Rx$.

This 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 [29, Definition 4.1].

A coefficient matrix R is said to be l-secure if by removing any l columns from R, the remaining sub-matrix still has full row rank, which guarantees that any non-zero linear combination of the row vectors of R contains at least $l + 1$ non-zero elements. Otherwise, assume there are at most l non-zero 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 sub-matrix, we will get a zero vector, which means the row vectors of this sub-matrix are linearly dependent and the rank of this sub-matrix 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 non-zero 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 $l + 1$ non-zero 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 [29, 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 [5].

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, $\Upsilon^{'}$ is a $(l + 1) \times (m - (l + 1))$ matrix, and $(I : \Upsilon^{'})$ is a vertical concatenation of I and $\Upsilon^{'}$. Since R is l-secure, each row of $(I : \Upsilon^{'})$ contains at least $l + 1$ non-zero 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 non-zero entries in $\Upsilon^{'}$. Thus the whole system contains at least $2l + 1$ unknowns, with $l + 1$ unknowns being contributed by I, and at least l unknowns from $\Upsilon^{'}$. ■

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

[5] 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.

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 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 sub-matrix 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 the gear 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 V, 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. We have

$$\begin{aligned} \frac{1}{\sqrt{k}\hat{\sigma}_r} R^T u &= \frac{1}{\sqrt{k}\hat{\sigma}_r} R^T \frac{1}{\sqrt{k}\sigma_r} Rx \\ &= \frac{1}{k\hat{\sigma}_r\sigma_r} R^T Rx \end{aligned}$$

The estimation for the i-th 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,j-th entry of $R^T R$. Hence, the expectation of the estimated element $\hat{x}_i$ is

$$\begin{aligned} E[\hat{x}_i] &= E[\frac{1}{k\hat{\sigma}_r\sigma_r} \sum_t \epsilon_{i,t} x_t] \\ &= \frac{1}{k\hat{\sigma}_r\sigma_r} E[\sum_t \epsilon_{i,t} x_t] \\ &= \frac{1}{k\hat{\sigma}_r\sigma_r} k\sigma_r^2 x_i \\ &= \frac{\sigma_r}{\hat{\sigma}_r} x_i \end{aligned}$$

The variance of $\hat{x}_i$ can be computed as

$$\begin{aligned} Var[\hat{x}_i] &= E[\hat{x}_i^2] - (E[\hat{x}_i^2])^2 \\ &= \frac{1}{k^2\hat{\sigma}_r^2\sigma_r^2} E[(\sum_t \epsilon_{i,t} x_t)^2] - (\frac{\sigma_r}{\hat{\sigma}_r} x_i)^2 \\ &= \frac{1}{k^2\hat{\sigma}_r^2\sigma_r^2} E[\sum_t \epsilon_{i,t}^2 x_t^2 + \sum_{p \neq q} \epsilon_{i,p} x_p \epsilon_{i,q} x_q] - (\frac{\sigma_r}{\hat{\sigma}_r} x_i)^2 \\ &= \frac{1}{k^2\hat{\sigma}_r^2\sigma_r^2} \left((2k + k^2)\sigma_r^4 x_i^2 + k\sigma_r^4 \sum_{t, t \neq i} x_t^2 \right) - (\frac{\sigma_r}{\hat{\sigma}_r} x_i)^2 \end{aligned}$$

Here we use the fact that $E[\epsilon_{i,i}^2] = (2k + k^2)\sigma_r^4$, $E_{i \neq j}[\epsilon_{i,j}^2] = k\sigma_r^4$, and $E_{p \neq q}[\epsilon_{i,p}\epsilon_{i,q}] = 0$. When the estimated variance $\hat{\sigma}_r^2 \approx \sigma_r^2$, we have

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

In summary, when the random matrix is completely disclosed, one still cannot find the exact values of the original data, even not for partial elements. 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.

Note that the above discussions are based on the assumption that R is completely disclosed to the adversary by some chance. However, in our design, R is owned by the data owner and nobody else should know that. An adversary can only try to guess the the realization of random matrix R if its probability density function (PDF) is revealed. The following section investigates this behavior.

B. The Distribution of the Random Matrix Is Disclosed

Consider the random matrix $R \in \Re^{k\times m}$ with i.i.d. entries chosen from continuous distribution. The PDF of R, denoted by $f_{\mathbf{R}}(R)$ can be expressed as the joint PDF of all the random elements in the matrix:

$$\begin{aligned} f_{\mathbf{R}}(R) &= f_{\mathbf{r_{1,1}},\mathbf{r_{1,2}},\ldots\mathbf{r_{k,m}}}(r_{1,1}, r_{1,2}, \ldots r_{k,m}) \\ &= \prod_{i=1}^{km} f_{\mathbf{r}}(r) \end{aligned}$$

where $f_{\mathbf{r}}(r)$ is the PDF of a single random element in the matrix. Given $f_{\mathbf{R}}(R)$, there are infi nite number of realizations in the real domain. If the adversary knows the original dimensionality of the data, i.e., m, and the PDF of R, whether he or she can get a good estimation of the original data through a random guess of the random matrix?

Assume the adversary generated a random matrix $\hat{R}$ according to the PDF. Given $u = Rx$, he or she can estimate x by multiplying on the left 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,j-th 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.

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

The expectation of $\hat{x}_i$ is

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

Here we use the fact that $E[\hat{\epsilon}_{i,j}] = 0$.

The variance of $\hat{x}_i$ can be computed as

$$\begin{aligned} Var[\hat{x}_i] &= E[\hat{x}_i^2] - 0 \\ &= \frac{1}{k^2\sigma_r^4}E[(\sum_t \hat{\epsilon}_{i,t}x_t)^2] \\ &= \frac{1}{k^2\sigma_r^4}E[\sum_t \hat{\epsilon}_{i,t}^2 x_t^2 + \sum_{p\neq q}\hat{\epsilon}_{i,p}x_p\hat{\epsilon}_{i,q}x_q] \\ &= \frac{1}{k^2\sigma_r^4}k\sigma_r^4\sum_t x_t^2 \\ &= \frac{1}{k}\sum_t x_t^2 \end{aligned}$$

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

Thus the estimation of x_i is zero on average, and its variance is $\frac{1}{k}\sum_t x_t^2$. 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.

C. Summary

Random projection-based multiplicative perturbation technique provides a reasonable protection of the data privacy while preserving certain statistical properties of the original data. In practice, both the privacy sensitive data and the random matrix (the distribution and the specific realization) are secret and only the projected data is released for secondary usage, so it is not possible to find the exact values of the original data. Moreover, since the projection-based perturbation technique is not a one (element) to one (element) mapping, a third party even does not know the original dimensionality of the data by only viewing the released version, which further reduces the chance of re-identification.

If by any chance, the specific realization of the random matrix is revealed, the adversary still cannot find the exact value of any elements from the original data. An approximation can be made. The variance of the approximation error is $\frac{2}{k}x_i^2 + \frac{1}{k}\sum_{t,t\neq i} x_t^2$ if the adversary has a good estimation of the distribution of the random matrix. 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 this variance, which in turn, changes the privacy level. Recall in Section V, the projection error is also related to the magnitude of the data, so this leads to the tradeoff between privacy and accuracy of the data mining results. The higher the privacy, the lower the accuracy, and vice versa.

If the adversary only knows the PDF of the random matrix, he or she still cannot recover the original data by a random guess of the realization from the whole domain. What he or she can finally find will be approximately a null matrix with all entries being around 0.

Although ICA may pose a threat to this technique under some conditions, if the dimensionality of the random matrix satisfies some constrains, namely, $m \geq 2k - 1$ (which can be easily controlled by the data owners), there will be no linear ICA filter that can recover any of the original signals even if the data completely satisfies the fundamental restrictions of ICA. If the data set has more than one Gaussians and correlated components, which is not unrealistic for many real world data, ICA will simply not be able to separate them out. Actually, ICA can be considered as a generalization of the PCA and the Projection Pursuit. While PCA seeks uncorrelated principle components for a new representation of the original data, ICA seeks statistically independent components/underlying factors of the data. If the components of the original data themselves are not independent, ICA will finally get the "real" hidden factors of the original data, but not the data itself. Therefore, random projection-based data perturbation may be an interesting approach for privacy sensitive data mining that needs further exploration.

Some procedures can be applied to further improve the privacy level. For example, the data owner can change the variance of the random noise within a specific interval centered in 1 but not rule the variance out in the projection. That is, the projection model will be changed to $U = \frac{1}{\sqrt{k}}RX$ without the factor σ_r. This will obscure the data much more, with some loss of accuracy. The second possibility is to combine the projection model together with some other geometric transformations, e.g., translation, scaling and rotation. Since these geometric transformations will perfectly preserve the similarities between vectors, they will not influence the accuracy but only help random projection to improve the privacy level. The next section presents some real privacy preserving data mining applications that random projection-base perturbation technique can apply to, together with the experimental results.

VII. Applications

Random projection maps the original data onto a randomly chosen lower dimensional subspace while preserving the similarities pretty well. This technique has been successfully applied to a number of applications. The work in [66] has presented experimental results in using the random mapping for clustering the textual documents. The work in [69], [70] used random projections for nearest-neighbor search in a high dimensional Euclidean space, and also presented theoretical insights. In [71], the author has studied random projections in learning high-dimensional Gaussian mixture models. Random projection has also been investigated in the context of learning half spaces, learning intersections of half spaces, learning polynomial surfaces [65], etc. In our recent work [72], we studied a randomized algorithm for constructing decision trees from distributed heterogeneous data, where a random projection is used to help computing the information gain. The experimental results show that by using only 20% of the communication cost necessary to centralize the data we can achieve trees with accuracy at least 80% of the trees produced by the centralized version. In this section, we illustrate several examples together with the experimental results. In order to evaluate the performance of our technique in a general setting, we choose all the data sets from the UCI Machine Learning Repository or the UCI KDD Archive, and we will not perform any normalization on the data. The random matrices are chosen from Gaussian distribution with mean 0 and variance 4, and the random projection rate is always less than 50%.

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.
2) Alice and Bob generate an $k \times m$ random matrix R using this seed, respectively.

TABLE I

RELATIVE ERRORS IN COMPUTING THE INNER PRODUCT OF THE 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 II

RELATIVE ERRORS IN COMPUTING THE SQUARE OF THE EUCLIDEAN DISTANCE OF THE 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

3) Alice projects her data to $\Re^k$ using R and release the perturbed version $U = \frac{1}{\sqrt{k}\sigma_r} RX$ to a third party.
4) Bob projects his data to $\Re^k$ and release the perturbed version $V = \frac{1}{\sqrt{k}\sigma_r} RY$ to a third party.
5) The third party computes the inner product matrix using the perturbed data U and V and gets $\begin{pmatrix} U^TU & U^TV \\ V^TU & V^TV \end{pmatrix} \approx \begin{pmatrix} X^TX & X^TY \\ Y^TX & Y^TY \end{pmatrix}$.

Similarly for Euclidean distance.

Discussions: When the data is properly normalized, the inner product matrix is nothing but the cosine angle matrix or the correlation coeffi cient matrix of $(X : Y)$.

Experiments: We consider the Adult database from the UCI Machine Learning Repository for the experiment. This data set was original extracted from the 1994 census bureau database. Without loss of generality, we select the fi rst $10,000$ rows of the data with only two attributes (fnlwgt, education-num) and shows how the random projection preserves the inner product and (the square of) the Euclidean distance of them. Table I and II present the results over 20 runs. Here k is the dimensionality of the perturbed vector, 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% of its original size, the relative error of the estimated inner product and (the square of) Euclidean distance is only around 1.80%, which is pretty good. Figure 7 illustrates how the original data is perturbed.

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 $\begin{pmatrix} X \\ Y \end{pmatrix}$ without directly accessing the raw data.

Algorithm:

1) Alice and Bob cooperatively generate a secret random seed.
2) Alice and Bob generate an $n \times k$ random matrix R using this seed, respectively.
3) Alice and Bob project their data to $\Re^k$ using R and release the perturbed version $U = \frac{1}{\sqrt{k}\sigma_r} XR$, $V = \frac{1}{\sqrt{k}\sigma_r} YR$.

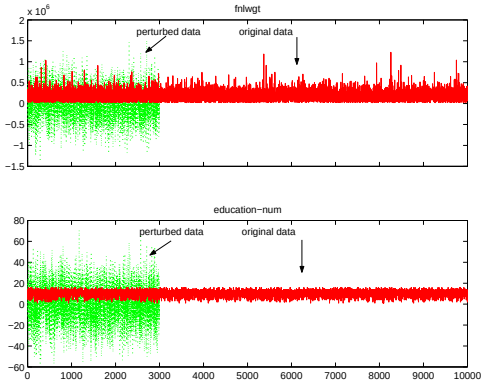
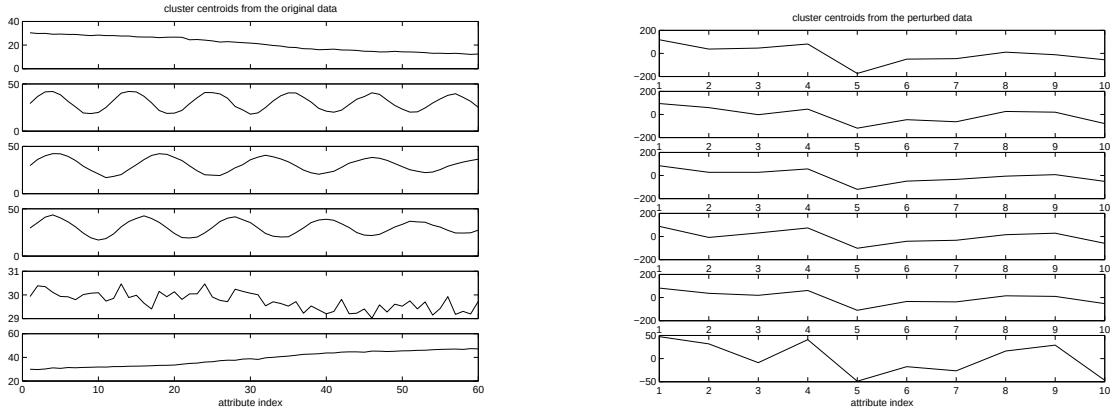


Fig. 7. Original data attributes and their perturbed counterparts. The random projection rate is 30%.

TABLE III

K-MEANS CLUSTERING FROM THE ORIGINAL DATA AND THE PERTURBED DATA.

#Attributes	Clustered Instances						Error 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%



(a) (b)

Fig. 8. (a) Cluster centroids found from the original data. (b) Cluster centroids found from the perturbed data with 17% projection rate.

4) The third party does K-Means clustering over the data set $\begin{pmatrix} U \\ V \end{pmatrix}$.

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 III shows the results. It can be seen that the clustering results are pretty good; even with 17% projection rate (the number of attributes is reduced from 60 to 10), the clustering error rate is still as low as 4.33%. Figure 8(a) and 8(b) give the cluster centroids from the original data and the perturbed data (17% projection rate), respectively. It can be observed that the perturbed data is really different from its original version in terms of magnitude, shape, and the number of attributes per example.

Linear Classification

Problem: Given a collection of sensitive data points x_i $(i = 1, 2, \ldots)$ in $\Re^n$, each labelled 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 $\Re^k$ using R such that $x_i^{'} = \frac{1}{\sqrt{k}\sigma_r} x_i R$, $\forall i$, and releases the perturbed data.
2) Run the perceptron algorithm in $\Re^k$:
   a) Let $w' = 0$. Do until all the examples are correctly classified
      i) Pick an arbitrary misclassified example $x_i^{'}$ and let $w' \leftarrow w' + classlabel(x_i^{'})x_i^{'}$.

Discussions: Note that in this algorithm, the class labels are not perturbed. Future example x is labelled positive if $w'(\frac{1}{\sqrt{k}\sigma_r} xR)^T > 0$ and negative otherwise. This is actually the same as checking whether $(w'\frac{1}{\sqrt{k}\sigma_r} R^T)x^T > 0$, namely, a linear separator in

TABLE IV

CLASSIFICATION ON THE PERTURBED IRIS PLANT DATA OVER 10-FOLD CROSS VALIDATION

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

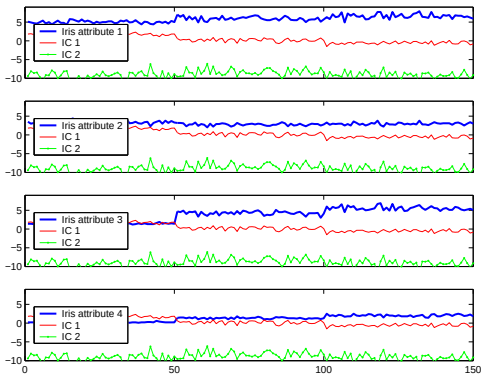


Fig. 9. Original data attributes together with the recovered independent components (ICs) found by ICA.

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 (Eq. 8) of the SVM task, the relationship of the original data points is completely quantified by inner product.

$$\begin{aligned} &\min_\alpha \tfrac{1}{2} \textstyle\sum_i \sum_j y_i y_j \alpha_i \alpha_j x_i x_j^T - \sum_i \alpha_i \\ &s.t. \quad \textstyle\sum_i y_i \alpha_i = 0 \\ &\alpha_i \geq 0 \quad i = 1, \ldots \end{aligned} \tag{8}$$

where $y_i = \pm 1$ is the class label of point x_i.

Experiments: We select the Iris Plant Database from the UCI Machine Learning Repository. This is a very simple data set with 150 instances and only 4 numeric attributes. We will show even for such small data set, our algorithm still works well. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant (Iris-setosa, Iris-versicolor, 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%, 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%. The classification results on the perturbed data over 10-fold cross validation are demonstrated in Table IV. It shows that the accuracy on the perturbed data over 10-fold cross validation is 86.67%, which is 91.55% as good as the results over the original data. To verify whether ICA can help recover the original data, we conduct ICA on the two perturbed attributes, only two independent components (ICs) have been found. We plot each attribute from the original data with the two ICs together to see how different they are. The results are shown in Figure 9. It can be seen that none of the ICs can approximate the true data very well.

The following section concludes this paper.

VIII. CONCLUSIONS AND FUTURE WORK

Privacy-preserving data mining from multi-party distributed data is playing an increasingly important role in many different application domains. Many of the recent efforts in this field are motivated by the idea of perturbing the data using randomized techniques. This paper explores the use of random projection matrices as a tool for privacy preserving data mining. The analysis shows that it is difficult to recover the exact values of any elements of the data even when the secret random matrix is disclosed. If the probability distribution of the random matrix is revealed by chance, an adversary still may not be able to identify the original data; most likely the adversary is going to get an approximately null matrix. Although ICA may pose a threat to this technique under some conditions, if we control the dimensionality of the random matrix to make it super non-square, then ICA will simply not work. 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 offer another interesting direction.

A primary direction for future work is motivated by the fact that there is still not a good benchmark to quantify the privacy and data utility. The models used in multi-antenna channel communication field [73] have strong connections with ours. There exists related work on computing the maximum mutual information between the original signals and the received signals in a multiplicative model. This represents the maximum reduction in uncertainty about the original signals given the received ones, and it might be used as an upper bound of the information that an adversary can learn through the perturbed data. We need more work for exploring this possibility in the future.

APPENDIX
PROOF OF LEMMA 5.6

Lemma 5.6: Let x, y be two data vectors in $\Re^m$. 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 variance σ_r^2. Further let

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

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: Using Lemma 5.2, the expectation of projection distortion is

$$\begin{aligned} E[u^T v - x^T y] &= E[\frac{1}{k\sigma_r^2} x^T R^T R y - x^T y] \\ &= \frac{1}{k\sigma_r^2} x^T E[R^T R] y - x^T y \\ &= \frac{1}{k\sigma_r^2} k\sigma_r^2 x^T y - x^T y \\ &= 0 \end{aligned}$$

To compute the variance of the distortion, let us first express the inner product between the projected vectors as

$$\begin{aligned} u^T v &= \frac{1}{\sqrt{k}\sigma_r} x^T R^T \frac{1}{\sqrt{k}\sigma_r} R y \\ &= \frac{1}{k\sigma_r^2} x^T R^T R y \\ &= \frac{1}{k\sigma_r^2} (\sum_i x_i \epsilon_{i,i} y_i + \sum_{i \neq j} x_i \epsilon_{i,j} y_j) \\ &= \frac{1}{k\sigma_r^2} \sum_i x_i \epsilon_{i,i} y_i + \frac{1}{k\sigma_r^2} \sum_{i \neq j} x_i \epsilon_{i,j} y_j \end{aligned}$$

Denote $\frac{1}{k\sigma_r^2} \sum_i x_i \epsilon_{i,i} y_i$ as Φ and $\frac{1}{k\sigma_r^2} \sum_{i \neq j} x_i \epsilon_{i,j} y_j$ as Ψ. Then $Var[u^T v] = Var[\Phi] + Var[\Psi] + Cov[\Phi, \Psi]$.

Now let us compute $Cov[\Phi, \Psi]$.

$$Cov[\Phi, \Psi] = E[\Phi\Psi] - E[\Phi]E[\Psi]$$

Since $E[\epsilon_{i,j}] = 0 \ \forall i, j, i \neq j$, so $E[\Psi] = 0$. Hence,

$$\begin{aligned} Cov[\Phi, \Psi] &= E[\Phi\Psi] - 0 \\ &= \frac{1}{k^2\sigma_r^4} E[\sum_i x_i \epsilon_{i,i} y_i \times \sum_{p \neq q} x_p \epsilon_{p,q} y_q] \end{aligned}$$

It is straightforward to verify that $E[\epsilon_{i,i}\epsilon_{p,q}] = 0$ when $p \neq q$. So $Cov[\Phi, \Psi] = 0$. ■

The variance of Φ is

$$\begin{aligned} Var[\Phi] &= Var[\frac{1}{k\sigma_r^2}\sum_i x_i\epsilon_{i,i}y_i] \\ &= \frac{1}{k^2\sigma_r^4}Var[\sum_i x_i\epsilon_{i,i}y_i] \\ &= \frac{1}{k^2\sigma_r^4}(E[(\sum_i x_i\epsilon_{i,i}y_i)^2] - (E[\sum_i x_i\epsilon_{i,i}y_i])^2) \\ &= \frac{1}{k^2\sigma_r^4}(E[\sum_i x_i^2\epsilon_{i,i}^2y_i^2 + \sum_{p\neq q} x_py_p\epsilon_{p,p}x_qy_q\epsilon_{q,q}] - (E[\sum_i x_i\epsilon_{i,i}y_i])^2) \end{aligned}$$

Since $E[\epsilon_{i,i}] = k\sigma_r^2$, $E[\epsilon_{i,i}^2] = (2k+k^2)\sigma_r^4$ and $E[\epsilon_{p,p}\epsilon_{q,q}] = k^2\sigma_r^4$, we have

$$\begin{aligned} Var[\Phi] &= \frac{1}{k^2\sigma_r^4}(2k+k^2)\sigma_r^4\sum_i x_i^2y_i^2 + \sum_{p\neq q} x_py_px_qy_q - (\sum_i x_iy_i)^2 \\ &= (\frac{2}{k}+1)\sum_i x_i^2y_i^2 + \sum_{p\neq q} x_py_px_qy_q - (\sum_i x_iy_i)^2 \end{aligned}$$

The variance of Ψ is

$$\begin{aligned} Var[\Psi] &= \frac{1}{k^2\sigma_r^4}Var[\sum_{i\neq j} x_i\epsilon_{i,j}y_j] \\ &= \frac{1}{k^2\sigma_r^4}(E[(\sum_{i\neq j} x_i\epsilon_{i,j}y_j)^2] - (E[\sum_{i\neq j} x_i\epsilon_{i,j}y_j])^2) \\ &= \frac{1}{k^2\sigma_r^4}(E[(\sum_{i\neq j} x_i\epsilon_{i,j}y_j)^2] - 0 \\ &= \frac{1}{k^2\sigma_r^4}\sum_{i\neq j}\sum_{p\neq q} x_iy_jx_py_qE[\epsilon_{i,j}\epsilon_{p,q}] \end{aligned}$$

Since $E[\epsilon_{i,j}\epsilon_{p,q}] = 0$ unless $i=p$ and $j=q$, or $i=q$ and $j=p$. Therefore,

$$\begin{aligned} Var[\Psi] &= \frac{1}{k^2\sigma_r^4}(\sum_{i\neq j} x_i^2y_j^2 + \sum_{i\neq j} x_iy_jx_jy_i)E_{i\neq j}[\epsilon_{i,j}^2] \\ &= \frac{1}{k^2\sigma_r^4}(\sum_i x_i^2\sum_{j\neq i} y_j^2 + \sum_i x_iy_i\sum_{j\neq i} x_jy_j)k\sigma_r^4 \\ &= \frac{1}{k}(\sum_i x_i^2\sum_i y_i^2 - \sum_i x_i^2y_i^2 + (\sum_i x_iy_i)^2 - \sum_i x_i^2y_i^2) \\ &= \frac{1}{k}(\sum_i x_i^2\sum_i y_i^2 + (\sum_i x_iy_i)^2 - 2\sum_i x_i^2y_i^2) \end{aligned}$$

Thus,

$$\begin{aligned} Var[u^Tv] &= Var[\Phi] + Var[\Psi] + 0 \\ &= (\frac{2}{k}+1)\sum_i x_i^2y_i^2 + \sum_{p\neq q} x_py_px_qy_q - (\sum_i x_iy_i)^2 + \frac{1}{k}(\sum_i x_i^2\sum_i y_i^2 + (\sum_i x_iy_i)^2 - 2\sum_i x_i^2y_i^2) \\ &= \frac{1}{k}(\sum_i x_i^2y_i^2 + (\sum_i x_iy_i)^2) + (\sum_i x_i^2y_i^2 + \sum_{p\neq q} x_py_px_qy_q - (\sum_i x_iy_i)^2) \\ &= \frac{1}{k}(\sum_i x_i^2y_i^2 + (\sum_i x_iy_i)^2) \end{aligned}$$

This gives the final result $Var[u^Tv - x^Tv] = \frac{1}{k}(\sum_i x_i^2y_i^2 + (\sum_i x_iy_i)^2)$.

ACKNOWLEDGMENT

This research is supported by the United States National Science Foundation Grant IIS-0329143. The second author would also like to acknowledge support from the United States National Science Foundation CAREER award IIS-0093353.

REFERENCES

[1] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, 2002. [Online]. Available: http://privacy.cs.cmu.edu/people/sweeney/kanonymity.html

[2] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proceedings of the IEEE International Conference on Data Mining*, Melbourne, FL, November 2003.

[3] W. B. Johnson and J. Lindenstrauss, "Extensions of lipshitz mapping into hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984.

[4] C. K. Liew, U. J. Choi, and C. J. Liew, "A data distortion by probability distribution," *ACM Transactions on Database Systems (TODS)*, vol. 10, no. 3, pp. 395–411, 1985. [Online]. Available: http://portal.acm.org/citation.cfm?id=4017

[5] E. Lefons, A. Silvestri, and F. Tangorra, "An analytic approach to statistical databases," in *Proceedings of the 9th International Conference on Very Large Data Bases*. Florence, Italy: Morgan Kaufmann Publishers Inc., November 1983, pp. 260–274. [Online]. Available: http://portal.acm.org/citation.cfm?id=673617

[6] S. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, pp. 63–69, 1965.

[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. [Online]. Available: http://portal.acm.org/citation.cfm?id=76895

[8] R. Agrawal and R. Srikant, "Privacy preserving data mining," in *Proceedings of the ACM SIGMOD Conference on Management of Data*, Dallas, TX, May 2000, pp. 439–450.

[9] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems*, Santa Barbara, CA, 2001, pp. 247–255. [Online]. Available: http://portal.acm.org/citation.cfm?id=375602

[10] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in *Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*, July 2002.

[11] A. Evfimevski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proceedings of the ACM SIGMOD/PODS Conference*, San Diego, CA, June 2003.

[12] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in *Proceedings of the 28th VLDB Conference*, Hong Kong, China, August 2002.

[13] J. J. Kim and W. E. Winkler, "Multiplicative noise for masking continuous data," Statistical Research Division, U.S. Bureau of the Census, Washington D.C., Tech. Rep. Statistics #2003-01, April 2003.

[14] K. Muralidhar, D. Batrah, and P. J. Kirs, "Accessibility, security, and accuracy in statistical databases: The case for the multiplicative fixed data perturbation approach," *Management Science*, vol. 41, no. 9, pp. 1549–1584, 1995.

[15] T. Dalenius and S. P. Reiss, "Data-swapping: A technique for disclosure control," *Journal of Statistical Planning and Inference*, vol. 6, pp. 73–85, 1982.

[16] S. E. Fienberg and J. McIntyre, "Data swapping: Variations on a theme by dalenius and reiss," National Institute of Statistical Sciences, Research Triangle Park, NC, Tech. Rep., 2003.

[17] A. C. Yao, "How to generate and exchange secrets," in *Proceedings 27th IEEE Symposium on Foundations of Computer Science*, 1986, pp. 162–167.

[18] O. Goldreich, S. Micali, and A. Wigderson, "How to play any mental game," in *Proceedings of the 19th annual ACM symposium on Theory of Computing*, 1987, pp. 218–229.

[19] M. Naor and B. Pinkas, "Oblivious transfer and polynomial evaluation," in *Proceedings of the 31th Annual Symposium on the Theory of Computing*, Atlanta, GA, May 1999, pp. 245–254.

[20] K. Sako and M. Hirt, "Efficient receipt-free voting based on homomorphic encryption," in *Proceedings of Advances in Cryptology (EUROCRYPT2000)*, Bruges, Belgium, May 2000, pp. 539–556.

[21] J. C. Benaloh and M. D. Mare, "One-way accumulators: A decentralized alternative to digital signatures," *Advances in Cryptology – EUROCRYPT'93. Workshop on the Theory and Application of Cryptographic Techniques. Lecture Notes in Computer Science.*, vol. 765, pp. 274–285, May 1993.

[22] A. C. Yao, "Protocols for secure computation," in *Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science*, 1982.

[23] O. Goldreich, *Secure Multi-Party Computation (Working Draft)*, Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, June 1998.

[24] B. Pinkas, "Cryptographic techniques for privacy preserving data mining," *SIGKDD Explorations*, vol. 4, no. 2, pp. 12–19, 2002. [Online]. Available: http://portal.acm.org/citation.cfm?id=772865

[25] W. Du and M. J. Atallah, "Secure multi-party computation problems and their applications: A review and open problems," in *Proceedings of the 2001 Workshop on New Security Paradigms*. Cloudcroft, NM: ACM Press, September 2001, pp. 13–22.

[26] ——, "Protocols for secure remote database access with approximate matching," in *7th ACM Conference on Computer and Communications Security(ACMCCS 2000). The first workshop on Security of Privacy in E-Commerce*, Athens, Greece, November 2000.

[27] M. J. Atallah and W. Du, "Secure multi-party computational geometry," in *WADS2001: Seventh International Workshop on Algorithms and Data Structures*, Providence, Rhode Island, August 2001, pp. 165–179.

[28] W. Du and M. J. Atallah, "Privacy-preserving cooperative statistical analysis," in *Proceedings of the 17th Annual Computer Security Applications Conference*, New Orleans, LA, December 2001.

[29] W. Du, Y. S. Han, and S. Chen, "Privacy-preserving multivariate statistical analysis: Linear regression and classification," in *Proceedings of 2004 SIAM International Conference on Data Mining (SDM04)*, Lake Buena Vista, FL, April 2004. [Online]. Available: http://www.cis.syr.edu/~wedu/Research/paper/sdm2004_privacy.pdf

[30] W. Du and M. J. Atallah, "Privacy preserving cooperative scientific computations," in *Proceedings of the 14th IEEE Computer Security Foundations Workshop*, Nova Scotia, Canada, June 2001, pp. 273–282.

[31] 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.

[32] J. S. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, July 2002.

[33] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," in *ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02)*, June 2002.

[34] X. Lin, C. Clifton, and Y. Zhu, "Privacy preserving clustering with distributed em mixture modeling," 2004, international Journal of Knowledge and Information Systems. To appear.

[35] J. Vaidya and C. Clifton, "Privacy-preserving k-means clustering over vertically partitioned data," in *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, D.C., August 2003.

[36] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *Advances in Cryptology (CRYPTO'00)*, ser. Lecture Notes in Computer Science, vol. 1880. Springer-Verlag, 2000, pp. 36–53.

[37] W. Du and Z. Zhan, "Building decision tree classifier on private data," in *Proceedings of the IEEE International Conference on Privacy, Security and Data Mining*. Maebashi City, Japan: Australian Computer Society, Inc., December 2002, pp. 1–8. [Online]. Available: http://portal.acm.org/citation.cfm?id=850784

[38] M. Kantarcoglu and J. Vaidya, "Privacy preserving naive bayes classifier for horizontally partitioned data," in *IEEE ICDM Workshop on Privacy Preserving Data Mining*, Melbourne, FL, November 2003, pp. 3–9.

[39] R. Wright and Z. Yang, "Privacy-preserving bayesian network structure computation on distributed heterogeneous data," in *Proceedings of the Tenth ACM SIGKDD Conference (SIGKDD'04)*, Seattle, WA, August 2004. [Online]. Available: http://www.cs.stevens.edu/~rwright/Publications/

[40] B.-H. Park and H. Kargupta, "Distributed data mining," in *The Handbook of Data Mining*, ser. Human Factors and Ergonomics, N. Ye, Ed. Lawrence Erlbaum Associates, Inc., 2003, pp. 341–358. [Online]. Available: http://www.cs.umbc.edu/~hillol/PUBS/review.pdf

[41] H. Kargupta, B.-H. Park, D. Hershberger, and E. Johnson, "Collective data mining: A new perspective toward distributed data analysis," in *Advances in Distributed and Parallel Knowledge Discovery*, H. Kargupta and P. Chan, Eds. AAAI/MIT Press, 2000, ch. 5, pp. 133–184. [Online]. Available: http://www.cs.umbc.edu/~hillol/PUBS/bc.pdf

[42] K. Liu, H. Kargupta, J. Ryan, and K. Bhaduri, "Distributed data mining bibliography," August 2004. [Online]. Available: http://www.cs.umbc.edu/~hillol/DDMBIB/

[43] S. Merugu and J. Ghosh, "Privacy-preserving distributed clustering using generative models," in *Proceedings of the Third IEEE International Conference on Data Mining ICDM'03*, Melbourne, FL, November 2003.

[44] S. Stolfo, A. L. Prodromidis, S. Tselepis, W. Lee, and D. W. Fan, "Jam: Java agents for meta-learning over distributed databases," in *Proceedings of the third International Conference on Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press, 1997, pp. 74–81.

[45] H. Kargupta and B.-H. Park, "A fourier spectrum-based approach to represent decision trees for mining data streams in mobile environments," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 2, pp. 216–229, 2004. [Online]. Available: http://www.cs.umbc.edu/~hillol/PUBS/mdmL.pdf

[46] H. Kargupta, W. Huang, S. Krishnamoorthy, and E. Johnson, "Distributed clustering using collective principal component analysis," *Knowledge and Information Systems*, vol. 3, no. 4, pp. 422–448, 2000. [Online]. Available: http://www.cs.umbc.edu/~hillol/PUBS/cpcad.pdf

[47] D. Ma, K. Sivakumar, and H. Kargupta, "Privacy sensitive bayesian network parameter learning," in *Proceedings of The Fourth IEEE International Conference on Data Mining (ICDM'04)*. Brighton, UK: IEEE Press, November 2004. [Online]. Available: http://www.cs.umbc.edu/~hillol/pubs.html

[48] M. J. Atallah, E. Bertino, A. K. Elmagarmid, M. Ibrahim, and V. S. Verykios, "Disclosure limitation of sensitive rules," in *Proceedings of the IEEE Knowledge and Data Engineering Workshop*, 1999, pp. 45–52.

[49] S. Oliveira and O. R. Zaiane, "Privacy preserving frequent itemset mining," in *Proceedings of the IEEE International Conference on Privacy, Security and Data Mining*. Maebashi City, Japan: Australian Computer Society, Inc., 2002, pp. 43–54. [Online]. Available: http://portal.acm.org/citation.cfm?id=850789

[50] V. S. Verykios, A. K. Elmagarmid, B. Elisa, Y. Saygin, and D. Elena, "Association rule hiding," in *IEEE Transactions on Knowledge and Data Engineering*, 2003.

[51] 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, December 2001. [Online]. Available: http://portal.acm.org/citation.cfm?id=604271

[52] L. Chang and I. S. Moskowitz, "Parsimonious downgrading and decision tree applied to the inference problem," in *Proceedings of the 1998 New Security Paradigms Workshop*, Charlottesville, VA, September 1998, pp. 82–89. [Online]. Available: http://citeseer.ist.psu.edu/376053.html

[53] E. W. Weisstein *et al.*, "Orthogonal transformation," MathWorld–A Wolfram Web Resource, 2004. [Online]. Available: http://mathworld.wolfram.com/OrthogonalTransformation.html

[54] S. R. M. Oliveira and O. R. Zaïane, "Privacy preserving clustering by data transformation," in *Proceedings of the 18th Brazilian Symposium on Databases*, Manaus, Amazonas, Brazil, October 2003, pp. 304–318. [Online]. Available: http://www.cs.ualberta.ca/~zaiane/postscript/sbbd03.pdf

[55] P. Common, "Independent component analysis: A new concept?" *IEEE Transactions on Signal Processing*, vol. 36, pp. 287–314, 1994.

[56] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4, pp. 411–430, June 2000.

[57] X.-R. Cao and R.-W. Liu, "A general approach to blind source separation," *IEEE Transactions on Signal Processing*, vol. 44, pp. 562–571, 1996.

[58] A. Taleb and C. Jutten, "On underdetermined source separation," in *Proceedings of 24th International Conference on Acoustics, and Signal Processing*, vol. 3, Phoenix, AZ, March 1999, pp. 1445–1448.

[59] J. Eriksson and V. Koivunen, "Identifiability and separability of linear ica models revisited," in *Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 2003.

[60] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[61] P. Bofill and M. Zibulevsky:, "Underdetermined blind source separation using sparse representation," *IEEE Transactions on Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[62] F. J. Theis and E. W. Lang:, "Geometric overcomplete ica," in *Proceedings of the 10th European Symposium on Artificial Neural Networks (ESANN'2002)*, Bruges, Belgium, April 24-26 2002, pp. 217–223.

[63] K. Waheed and F. M. Salam, "Algebraic overcomplete independent component analysis," in *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Source Separation*, Nara, Japan, April 1-4 2003.

[64] R. Hecht-Nielsen, "Context vectors: General purpose approximate meaning representations self-organized from raw data," *Computational Intelligence: Imitating Life*, pp. 43–56, 1994.

[65] R. I. Arriaga and S. Vempala, "An algorithmic theory of learning: Robust concepts and random projection," in *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*. New York, NY: IEEE Computer Society, October 1999, pp. 616–623. [Online]. Available: http://www-math.mit.edu/~vempala/papers/robust.ps

[66] S. Kaski, "Dimensionality reduction by random mapping: Fast similarity computation for clustering," in *Proceedings of International Joint Conference on Neural Networks (IJCNN'98)*, vol. 1, Piscataway, NJ, 1998, pp. 413–418. [Online]. Available: http://citeseer.ist.psu.edu/kaski98dimensionality.html

[67] J. W. Demmel and N. J. Higham, "Improved error bounds for underdetermined system solvers," Computer Science Department, University of Tennessee, Knoxville, TN, Tech. Rep. CS-90-113, August 1990. [Online]. Available: http://www.netlib.org/lapack/lawnspdf/lawn23.pdf

[68] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, ser. Prentice-Hall Series in Automatic Computation. Englewood Cliffs, New Jersey: Prentice-Hall, 1974, ch. 13.

[69] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proceedings of the 30th Symposium on Theory of Computing*. Dallas, TX: ACM Press, 1998, pp. 604–613. [Online]. Available: http://portal.acm.org/citation.cfm?id=276876

[70] J. M. Kleinberg, "Two algorithms for nearest-neighbor search in high dimensions," in *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing*. ACM Press, 1997, pp. 599–608. [Online]. Available: http://portal.acm.org/citation.cfm?id=258653

[71] S. Dasgupta, "Experiments with random projection," in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2000, pp. 143–151. [Online]. Available: http://portal.acm.org/citation.cfm?id=719759

[72] C. Giannella, K. Liu, T. Olsen, and H. Kargupta, "Communication efficient construction of decision trees over heterogeneously distributed data," in *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04)*, Brighton, UK, November 2004. [Online]. Available: http://www.cs.umbc.edu/~hillol/PUBS/ddt.pdf

[73] I. E. Telatar, "Capacity of multi-antenna gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, Nov/Dec 1999. [Online]. Available: http://mars.bell-labs.com/papers/proof/

Kun Liu Biography text here.

Hillol Kargupta Biography text here.

Jessica Ryan Biography text here.