Random Multiplication based Data Perturbation for Privacy Preserving Distributed Data Mining - 1

Prof. Ja-Ling Wu
Dept. CSIE & GINM
National Taiwan University
Data and User privacy calls for well designed techniques that pay careful attention to hiding privacy-sensitive information, while preserving the inherent statistical dependencies which are important for data analysis and data mining applications.

- Privacy Preserving Information Processing
  - PP-based Data Mining and Search
  - PP-based Signal Processing and Compression
  - PP-based Object Detection and Recognition
Suppose there are $N$ organizations $O_1; O_2; \ldots; O_N$; each organization $O_i$ has a **private transaction database** $DB_i$. A **third party data miner** wants to learn certain statistical properties of the **union** of these databases $\bigcup_{i=1}^{N} DB_i$.

These organizations are comfortable with this, but they are reluctant to disclose their raw data.

How could the data miner perform data analysis without compromising the privacy of the data?

> This is generally referred to as the **census problem**.
In this scenario, the data is usually arithmetically distorted and its new representation is released; anybody has arbitrary access to the published data. Fig. 1 illustrates a distributed two party-input case as well as a single-party-input case.

Fig. 1. (a) Distributed two-party-input computation model. (b) Single-party-input computation model.
This talk considers a randomized multiplicative data perturbation technique for this problem. It is motivated by the work presented in [1] which pointed out some of the problems of additive random perturbation. In other words, we explore the possibility of using multiplicative random matrices for constructing a new representation of the data. The transformed data is released to the data miner.

It is our target that the **inner product** and **Euclidean distance** are **preserved in the new data**. The approach is fundamentally based on the **Johnson-Lindenstrauss lemma** [2] which notes that **any set of s points in m-dimensional Euclidean space can be embedded into k-dimensional subspace, where k is logarithmic in s, such that the pair-wise distance of any two points is maintained within an arbitrarily small factor.**

Therefore, by projecting the data onto a random subspace, we can dramatically change its original form while preserving much of its underlying distance-related statistical characteristics.
Assuming that the private data is taken from the same continuous real domain and all the parties are semi-honest (which means there is no collusion between parties and all the parties follow the protocol properly).

Without loss of generality, let’s investigate the technique in a two-party-input scenario where Alice and Bob, each owning a private database, want a third party to analyze their data without seeing the raw information. The resultant technique can be easily modified and applied to other input cases.
2 Related Technology

2.1 Data Perturbation

Data perturbation approaches can be grouped into two main categories: the probability distribution approach and the value distortion approach. The probability distribution approach replaces the data with another sample from the same (or estimated) distribution, and the value distortion approach perturbs data elements or attributes directly by either additive noise, multiplicative noise, or some other randomization procedures. In this talk, we mainly focus on the value distortion approach.
The work in [3] proposed an additive data perturbation technique for building decision tree classifiers. Each data element is randomized by adding some random noise chosen independently from a known distribution such as Gaussian distribution. The data miner reconstructs the distribution of the original data from its perturbed version (using, e.g., an Expectation Maximization-based algorithm) and builds the classification models.

More recently, Kargupta et al. [1] questioned the use of random additive noise and pointed out that additive noise can be easily filtered out in many cases that may lead to compromising the privacy.

The possible drawback of additive noise makes one wonder about the possibility of using **multiplicative noise** for protecting the privacy of the data.

Two basic forms of multiplicative noise have been well studied in the **statistics community** [4]. One is to multiply each data element by a random number that has a truncated Gaussian distribution with mean one and small variance.

The other one is to take a logarithmic transformation of the data first, add predefined multivariate Gaussian noise, and take the antilog of the noise-added data.

In practice, the first method is good if the data disseminator only wants to make minor changes to the original data; the second method assures higher security than the first one but maintains the data utility in the log scale.
A potential problem of traditional additive and multiplicative perturbation is that each data element is perturbed independently, therefore the pair-wise similarity of records is not guaranteed to be maintained. In this talk, we consider an alternate approach that proves to preserve much of the underlying statistical aggregates of the data.
2.2 Data Swapping

- The basic idea of **data swapping** is to transform the database by switching a subset of attributes between selected pairs of records so that the lower order frequency counts or marginals are preserved and data confidentiality is not compromised.

- This technique could equally as well be classified under the **data perturbation** category. A variety of refinements and applications of data swapping have been addressed since its initial appearance. We refer to [5] for a thorough treatment.

2.3 k-Anonymity

- The k-Anonymity model [6] considers the problem that a data owner wants to share a collection of person-specific data without revealing the identity of an individual. To achieve this goal, data generalization and suppression techniques are used to protect the sensitive information.

- All attributes (termed as quasi-identifier) in the private database that could be used for linking with external information would be determined, and the data is released only if the information for each person contained in the release cannot be distinguished from at least k - 1 other people.

2.4 Secure Multiparty Computation

- The **Secure Multiparty Computation (SMC)** [7] technique considers the problem of evaluating a function of the secret inputs from two or more parties, such that no party learns anything but the designated output of the function.

- A large body of cryptographic protocols, including (garbled) circuit evaluation protocol, oblivious transfer, functional encryption, homomorphic encryption, and commutative encryption, serve as the building blocks of SMC.

- The work in [8] offered a broad view of SMC framework and its applications to data mining. The work in [9] detailed a rigorous introduction to SMC.

Partial Homomorphic Encryption ...> Fully Homomorphic Encryption
It was shown that any function that can be expressed by an arithmetic circuit is privately computable using a generic circuit evaluation (Garbled Circuits) protocol. However, the communication and computational complexity of doing so makes this general approach infeasible for large data sets.

A collection of SMC tools useful for large-scale privacy preserving data mining (e.g., secure sum, set union, and inner product) are discussed in [10]. An overview of the state-of-the-art privacy preserving data mining techniques is presented in [11].


This section presents a deterministic multiplicative perturbation method using random orthogonal matrices in the context of computing inner product matrix. Later, we shall analyze the deficiency of this method and then propose a more general case that makes use of random projection matrices for better protection of the data privacy.
An **orthogonal transformation** is a linear transformation $R : \mathbb{R}^n \rightarrow \mathbb{R}^n$, which preserves the length of vectors as well as the angles between them. Usually, orthogonal transformations correspond to and may be represented using orthogonal matrices.

Let $X$ and $Y$ be two data sets owned by Alice and Bob, respectively. $X$ is an $m_1 \times n$ matrix, and $Y$ is an $m_2 \times n$ matrix. Both of them observe the same attributes. Let $R$ be an $n \times n$ random orthogonal matrix.
Now, consider the following linear transformation of the two data sets:

\[ U = XR, \quad \text{and} \quad V = YR; \quad \text{then we have} \]
\[ UU^T = XXT, \quad VV^T = YYT, \quad UV^T = XRR^T Y^T = XY^T. \]

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 \( (\begin{array}{c} X \\ Y \end{array}) \) can still be perfectly computed there, where \( (\begin{array}{c} X \\ Y \end{array}) \) 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 $U^T V = X^T Y$, and all the pair-wise distances and similarities between the columns vectors from the data $(X : Y)$ are fully preserved in the perturbed data, where $(X : Y)$ denotes a vertical concatenation of $X$ and $Y$.

Therefore, a third party can analyze the correlation of the attributes from heterogeneously distributed (vertically partitioned) data without accessing the raw data.
Since only the transformed data is released, there are actually an **infinite number of inputs and transformation procedures** that can **simulate the same output**, while the observer has no idea what is the real form of the original data. Therefore, random orthogonal transformation seems to be a good way to protect data’s privacy while preserving its utility.
However, from the geometric point of view, an orthogonal transformation is either a pure rotation when the determinant of the orthogonal matrix is 1 or a roto-inversion (a rotation followed by a flip) when the determinant is -1, and, therefore, it is possible to re-identify the original data through a proper rotation.
Figs. 2a and 2b illustrate how the random orthogonal transformation works in a 3D space. It can be seen that the data is not very well masked after transformation. In this regard, the security of a similar approach using random rotation to protect the data privacy is also questionable.

Moreover, if all the original data vectors are statistically independent and they do not follow Gaussian distribution, it is possible to estimate their original forms quite accurately using Independent Component Analysis (ICA).
Fig. 2. (a) A sample data set. (b) The perturbed data after a random orthogonal transformation. The transformation corresponds to a rotation of the original data about the x-axis by a random angle.
Independent Component Analysis (ICA) [12] is a technique for discovering independent hidden factors that are underlying a set of linear or nonlinear mixtures of some unknown variables, where the mixing system is also unknown. These unknown variables are assumed non-Gaussian and statistically independent, and they are called the independent components (ICs) of the observed data. These independent components can be found by ICA.

A classical example of ICA is the cocktail party problem (as illustrated in Fig. 3a). Imagine you are in a cocktail party. Although different kinds of background sounds are mixed together, e.g., music, other people’s chat, television news report, or even a siren from a passing-by ambulance, you still have no problem identifying the discussion of your neighbors. It is not clear how human brains can separate the different sound sources.

However, ICA is able to do it if there are at least as many “ears” or receivers in the room as there are different simultaneous sound sources.
Fig. 3. (a) An illustration of the cocktail problem. In this case, what the ears hear are two linear combinations of four audio signals, i.e., four signal are compressed into two. (b) A sample of four independent source signals.
4.1 ICA Model

- The basic ICA model can be defined as follows:

\[ u(t) = R \, x(t) \]  \hspace{1cm} (1)

where \( x(t) = (x_1(t), x_2(t), \ldots, x_m(t))^T \) denotes a \( m \)-dimensional vector collecting the \( m \) independent source signals \( x_i(t), \text{ } i = 1, 2, \ldots, m \). Here, \( t \) indicates the time dependence.

Each signal \( x_i(t) \) can be viewed as an outcome of a continuous-value random process. \( R \) is a constant \( k \times m \) unknown mixing matrix, which can be viewed as a mixing system with \( k \) receivers. \( U(t) = (u_1(t), u_2(t), \ldots, u_k(t))^T \) is the observed mixture.
The aim of ICA is to design a filter that can recover the original signals from only the observed mixture.

Since $u(t) = Rx(t) = (R\Lambda P)(P^{-1}\Lambda^{-1}x(t))$ for any diagonal matrix $\Lambda$ and permutation matrix $P$, the recovered signals $x(t)$ can never have completely unique representation.

So, the uniqueness of the recovered signals found by ICA can only be guaranteed up to permutation and scaling ambiguities.
As an illustration, consider four statistically independent audio signals, denoted as a $4 \times 8,000$ matrix $X$ (shown in Fig. 3b). Note that, for the sake of simplicity, some of the signals we are showing here are deterministic; however, ICA generally works with continuous-value random process.

A linear mixture of these signals (shown in Fig. 4a) is generated by pre-multiplying a $4 \times 4$ nonsingular random matrix to $X$. The goal of ICA is to recover the original signals using only the mixture. Fig. 4b gives the estimated signals through ICA. It can be observed that the basic structure of the original signals are recovered very well; however, the order and the amplitude of the recovered signals are not necessarily the same as those of the original ones.
Fig. 4. (a) Linear mixture of the original source signals using a square random matrix. (b) Recovered signals using ICA.
4.2 Decomposability

In practice, a linear filter is designed to get the recovered signals
\[ y(t) = (y_1(t), y_2(t), \ldots, y_l(t))^T \]
from a \( k \)-dimensional input \( u(t) = (u_1(t), u_2(t), \ldots, u_k(t))^T \).
In other words,
\[ y(t) = B u(t), \quad (2) \]
where \( B \) is an \( l \times k \)-dimensional separating matrix. Combining (1) and (2) together, we get
\[ y(t) = B R x(t) = Z x(t) \quad (3) \]
where \( Z = B x R \) 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 $Z = B \cdot R = I$, where $I$ is an identity matrix. Thus, we can recover all the signals up to scaling and permutation ambiguities.

Actually, to solve the problem, there are two steps to be done. The first step is to determine the existence of $B$ such that $Z$ can decompose the mixture. The second step is to find such a kind of $B$ if it is proved to exist.
In general, by imposing the following fundamental restrictions [13], 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) $f_{x(t)}(x_1(t), x_2(t), \ldots, x_m(t))$ is factorizable in the following way:

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

where $f_{x_i(t)}(x_i(t))$ denotes the marginal probability density of $x_i(t)$.

Conditions for the existence of $B$:

- All the signals must be **non-Gaussian** with the possible exception of one signal.
- The **number of observed signals** $k$ must be at least as large as the independent source signals, i.e., $k > m$.
- Matrix $R$ must be of **full-column rank**.
These restrictions actually have exposed the **potential dangers** of random orthogonal transformation or random rotation techniques where the **mixing matrix** is square and of full-column rank.

If the original signals are also **statistically independent** and there are no Gaussians, it is most likely that **ICA** can find a good approximation of the original signals from their perturbed version. Figs. 4a and 4b illustrated this situation.
Note that, if some of the source signals are correlated, they may be lumped in the same group and can never be separated out.

If there is more than one Gaussian signal, the problem becomes more complicated. The output of the filter may be either individual non-Gaussian signals, individual Gaussian signals, or a mixture of Gaussian signals. A detailed analysis can be found elsewhere [14].

When \( l < k < m \) (i.e., the number of sources is greater than the number of receivers), it is generally not possible to design linear filters to simultaneously recover all these signals. This kind of separation problem is termed as over-complete ICA or underdetermined source separation.

Cao et al. [14] analyzed the conditions for the existence of the separating matrix \( B \).
Cao et al.[14] 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.

It follows that, if one can control the structure of the mixing matrix $R$ such that $R$ is not two-row decomposable, then there is no linear method that can find a matrix $B$ for separating the source signals into two or more disjoint groups. In that case, it is not possible to separate out any of the source signals. The following theorem characterized this property:
Theorem 4.4. Any $k \times m$ ($m \geq 2k - 1, m \geq 2$) random matrix with entries independent and identically chosen from some continuous distribution in the real domain is not two-row decomposable with probability 1.
**Proof.** For a $k \times m$ random matrix with $m \geq 2k - 1$ and any partition of its columns into two nonempty sets, at least one set will have at least $k$ members. Thus, this set of columns contains a $k \times k$ submatrix, denoted as $M$. If $M$ is nonsingular, then the $k$ column vectors of the submatrix span $\mathbb{R}^k$ Euclidean space. Thus, there is always at least one vector in one group belonging to the space spanned by the other group, which does not satisfy Theorem 4.3.
Matrix $R$ is \textit{l-row decomposable} if and only if its \textit{columns} can be grouped into \textit{l disjoint groups} such that the \textit{column vectors in each group are linearly independent of the vectors in all the other groups}.
Now, let us show $M$ is indeed nonsingular with probability 1. It has been proved in [15, Theorem 3.3] that the probability that $MM^T$ is positive definite is 1. Since a matrix is positive definite if and only if all the eigenvalues of this matrix are positive, and a matrix is nonsingular if and only if all its eigenvalues are nonzero, we have that $MM^T$ is nonsingular with probability 1. Further note that $\text{rank}(M^T) = \text{rank}(MM^T) = \text{rank}(M^TM)$, therefore $M$ is nonsingular with probability 1. This completes the proof.

The above non-singularity property of a random matrix has also been proved in [16, Theorem 3.2.1] when the random matrix is Gaussian.

Thus, by letting $m \gg k$, there is no linear filter that can separate the observed mixtures into two or more disjoint groups, so it is not possible to recover any of the source signals.

Figs. 5a and 5b depict this property.

Fig. 5. (a) Linear mixture of the original four source signals (as shown in Fig. 3b) with 50 percent random projection rate. \((m =4, k =2)\) (b) Recovered signals. It can be observed that none of the original signals are reconstructed, and at most \(k=2\) independent components can be found by ICA.
It can be seen that, after 50 percent row-wise random projection, the original four signals are compressed into two, and ICA cannot recover any of them. Moreover, projecting the original data using a 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 many (elements)-to-one (element) map**, which is totally different from the traditional one-to-one data perturbation technique, and, therefore, it is even harder for the adversary to re-identify the sensitive data.
Some Remarks of this Talk:

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

Recently, over-complete ICA ($k < m$) has drawn much attention. It has been found that, even when $k < m$, if all the sources are non-Gaussian and statistically independent, it is still possible to identify the mixing matrix such that it is unique up to a right multiplication by a diagonal and a permutation matrix [17, 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 over-complete ICA problem has only been treated in particular cases, e.g., the source signals are assumed to have sparse distribution [18].

In [19], a random projection-based multiplicative perturbation technique has been proposed. In which, by letting the random matrix super non-square, one can get a solution to PP-based data mining even facing of the over-complete ICA model.

It showed that randomly generated projection matrices are likely to be more appropriate for protecting the privacy, compressing the data, and still maintaining its utility.

[19] Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining, Kun Liu, Hillol Kargupta, and Jessica Ryan, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 1, JANUARY 2006