EXPLOIT THE SCALE OF BIG DATA FOR DATA PRIVACY: AN EFFICIENT SCHEME BASED ON DISTANCE-PRESERVING ARTIFICIAL NOISE AND SECRET MATRIX TRANSFORM

Xiaohua Li and Zifan Zhang

Department of Electrical and Computer Engineering
State University of New York at Binghamton
Binghamton, NY 13902
Email: {xli,zzhang20}@binghamton.edu

ABSTRACT

In this paper we show that the extensive results in blind/non-blind channel identification developed within the community of signal processing in communications can play an important role in guaranteeing big data privacy. It is widely believed that the sheer scale of big data makes most conventional data privacy techniques ineffective for big data. In contrast to this pessimistic common belief, we propose a scheme that exploits the sheer scale to guarantee privacy. This scheme uses jointly artificial noise and secret matrix transform to scramble the source data. Desirable data utility can be supported because the noise and the transform preserve some important geometric properties of the source data. With a comprehensive privacy analysis, we use the blind/non-blind channel identification theories to show that the secret transform matrix and the source data can not be estimated from the scrambled data. The artificial noise and the sheer scale of big data are critical for this purpose. Simulations of collaborative filtering are conducted to demonstrate the proposed mascheme.

Index Terms— big data, privacy, signal processing, channel identification, blind source separation

1. INTRODUCTION

Big data refers to the seemingly unlimited data generated from various sources such as social media, web surfing or market transactions. We rely on big data analytics to discover useful information from the data [1]. Big data has some unique characteristics in terms of volume, variety and velocity [2], where volume denotes the sheer data scale or the huge data size, variety refers to the heterogeneous data structures, and velocity describes the time sensitive and time varying nature of the data.

The full potential of big data cannot be realized without massive data sharing. Nevertheless, sharing of big data faces the challenge of data privacy. Many data records contain personally identifiable information that can be linked to the data owner. Experience has indicated that it is not enough to just remove the identity of the data owner from the source data. With advanced data analysis and some outside knowledge, it is often possible to discover a lot of private information from the published data. Despite significant accomplishments of privacy research in areas like statistical database [3], privacy-preserving data mining [4] and privacy-preserving data publishing [5], privacy remains one of the major challenges for big data.

To guarantee privacy, source data must be processed for anonymization. Many data anonymization techniques have been developed, including cryptographic techniques such as homomorphic encryption or secure multi-party computation [6], and non-cryptographic techniques such as perturbation or k-anonymity [3][4]. In perturbation, the original data is modified by adding noise, or generalized to less accurate values, etc. In k-anonymity, the original data is modified such that a given data is not distinguishable from at least k other data.

Unfortunately, each of these techniques has some problems when applied to big data [4] [7]. The popular k-anonymity methods can no longer anonymize the data without losing an unacceptable amount of information. The algorithms become impractical because the underlying problem is NP-hard. The noise-perturbation methods become less effective because it is possible to estimate the original data from the perturbed data when the data volume becomes large. The encryption based approaches are computationally prohibitive.

Considering the sheer scale of big data, techniques based on matrix transform looks promising because of their efficiency in both computation and data scrambling. Within the conventional database research, many of such techniques have been studied, such as matrix rotation [8][9], matrix multiplication [10][11], sketches [7], cancellable biometrics [12], etc. Unfortunately, they have been challenged with a number of inverse-transform attacks [10][13], and has since been looked as insecure. Note that most of these works did not take the special characteristics of big data into consideration.

In this paper we show that, with the help of the sheer scale of big data, the matrix transform technique can be strengthened into an efficient and secure big data privacy preserving methodology. We propose an innovative scheme that exploits both secret matrix transform and distance-preserving artificial noise, and the latter plays a critical role to guarantee both the required scale and the privacy.

As another major contribution, we conduct a comprehensive privacy analysis based on advanced communication signal processing theories, in particular non-blind and blind channel identification and source separation theories. The analysis shows that quantifiable metrics such as SNR or estimation accuracy used in signal processing can be applied to study the level of privacy.

The organization of this paper is as follows. In Section 2, we give the big data publishing model and develop the proposed privacy preserving scheme. In Section 3, we give a comprehensive privacy analysis. In Section 4, we conduct simulations based on a typical big data application scenario. Finally, a conclusion is given in Section 5.

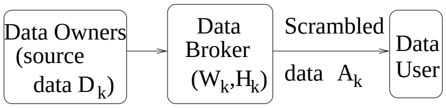

Fig. 1. Privacy-preserving data publishing model.

2. JOINT ARTIFICIAL NOISE AND SECRET MATRIX TRANSFORM

2.1. Data scrambling in big data publishing

Consider a set of data consisting of N data records. Each of the data records has M attributes, i.e., dimensions or values. This data set can be described by an $N \times M$ matrix $\mathbf{D}$. For example, in collaborative filtering or recommender generating, each entry $\mathbf{D}(i,j)$ denotes the user i's preference of the item j [14] [15]. The sheer scale of big data means that N and M can be very large, e.g., in millions [14]. Such a large data set may consist of several separated data blocks $\mathbf{D}_k$ in practice, which are stored physically in different places.

We consider a privacy-preserving data publishing model [5], as illustrated in Fig. 1. We assume that there is a trustable data broker that collects the data $\mathbf{D}$ from the data owners, anonymizes them, and makes them publicly available to all the data users. Data privacy in our case means that no data user can estimate the source data with sufficient accuracy from the scrambled data and possibly some outside knowledge about the source data and the scrambling scheme. To preserve privacy, the data broker applies both artificial noise and secret matrix transform to scramble the data matrix $\mathbf{D}_k$, and publishes the scrambled data matrix $\mathbf{A}_k$.

For the data publishing model, what makes privacy challenge is to keep desirable data utility to any data user, even the privacy attacker. Both of them should be allowed to conduct normal data analytic functions with the scrambled data and obtain similar results as using the unscrambled source data. Alternatively, our proposed techniques can be applied to guarantee the privacy of data stored in untrusted cloud servers. What makes privacy challenge in this case is that the data owners have to use the cloud to process the data for big data analytics.

We propose the following data scrambling scheme. The data broker transforms the source data matrix $\mathbf{D}_k$ into the scrambled data matrix

$$\mathbf{A}_k = \begin{bmatrix} \mathbf{D}_k & \mathbf{W}_k \end{bmatrix} \mathbf{H}_k, \tag{1}$$

where $(\cdot)_k$ denotes the k^{th} block of the huge big data set $\mathbf{D}$. Each block $\mathbf{D}_k$ has N_k rows, and $\sum_k N_k = N$. Therefore, the dimension of $\mathbf{D}_k$ is $N_k \times M$. Processing each block separately fits nicely with the big data processing platforms like MapReduce. $\mathbf{W}_k$ is an $N_k \times K$ matrix of artificial noise, and $\mathbf{H}_k$ is the $(M+K) \times J$ secret transform matrix. The dimension of $\mathbf{A}_k$ is $N_k \times J$. We can adjust $M \leq J \leq M+K$ to control the published data size.

While somewhat similar to conventional matrix transform methods such as [8][9], the inclusion of the extra artificial noise matrix $\mathbf{W}_k$ is new and makes our scheme fundamentally different. The purpose of $\mathbf{W}_k$ is to both add intentional noise and guarantee the sheer scale of the matrices. Without it, the transform matrix along may not guarantee privacy in some special cases.

To support data utility, similar to conventional matrix transform methods, our proposed scheme preserves some important geometric properties among the data records during scrambling, such as distance and manifold. Therefore, many important data analytic functions such as dimension reduction, classification and clustering can be conducted with the scrambled data and produce the same results as using the unscrambled source data. Note that many big data analytics functions are based on matrix factorization and dimension reduction techniques such as singular value decomposition (SVD), principle component analysis (PCA) and non-negative matrix factorization [16][17].

To preserve the distance among the rows of $\mathbf{D}_k$, we need to find $\mathbf{W}_k$ and $\mathbf{H}_k$ such that $[\mathbf{D}_k \quad \mathbf{W}_k]\mathbf{H}_k\mathbf{H}_k'[\mathbf{D}_k \quad \mathbf{W}_k]' - \mathbf{D}_k\mathbf{D}_k'$ is diagonal or diagonal matrix plus certain constant, where $(\cdot)'$ denotes matrix transform. A simpler way is to find $\mathbf{W}_k$ such that $\mathbf{D}_k(\mathbf{W}_k\mathbf{W}_k' - \mathbf{I}_M)\mathbf{D}_k'$ is diagonal or diagonal plus certain constant. A special way is to make $\mathbf{W}_k$ orthogonal. Similarly, the secret transform matrix $\mathbf{H}_k$ can be set as the J columns of an $(M+K)\times(M+K)$ orthogonal matrix $\mathbf{T}_k$, where $\mathbf{T}_k'\mathbf{T}_k = \mathbf{I}_{M+K}$.

2.2. Collaborative filtering with the scrambled data

Privacy-preserving techniques make tradeoff between utility and privacy [18]. In our case, although not as flexible as using the unscrambled source data, many important big data analytic functions, especially those based on dimension reduction, clustering or classification, can still be conducted. In the sequel, we demonstrate this by the collaborative filter based recommender generation.

Let us consider collaborative filtering for preference prediction [14]. Taking the movie rating as example, we need to use existing ratings to predict a user's preference (rating) on an item that he has not rated. In this case, a data user's own data is within $\mathbf{D}_k$, i.e., this data user is also a data owner. He can use all the published scrambled data, but should not be able to get access to the unscrambled data except his own data.

One of the well known ways is to use SVD to find the reduced-dimensional reconstruction of $\mathbf{D}_k$, which can then be used to generate preference prediction. With the scrambled data, we first find the SVD of $\mathbf{A}_k$

$$\mathbf{A}_k = \mathbf{U}_a\mathbf{\Sigma}_a\mathbf{V}_a' \tag{2}$$

where $\mathbf{U}_a$ and $\mathbf{V}_a$ are $N_k \times N_k$ and $J \times J$ orthogonal matrices, respectively, and $\mathbf{\Sigma}_a$ is the $N_k \times J$ diagonal singular-value matrix. Then, we consider the reduced dimension $L \leq \min\{N_k, J\}$. Let the matrices $\tilde{\mathbf{U}}_a$ and $\tilde{\mathbf{V}}_a$ consist of the first L columns of the matrices $\mathbf{U}_a$ and $\mathbf{V}_a$, respectively. Let $\tilde{\mathbf{\Sigma}}_a$ be the $L \times L$ left-top submatrix of $\mathbf{\Sigma}_a$. The reconstructed L-dimensional subspace matrix

$$\tilde{\mathbf{A}}_k = \tilde{\mathbf{U}}_a\tilde{\mathbf{\Sigma}}_a\tilde{\mathbf{V}}_a' \tag{3}$$

can be used to generate the preference predictions. Specifically, the $(i,j)^{\text{th}}$ entry $\tilde{\mathbf{A}}_k(i,j)$ is a prediction of the user i's preference (or rating) on the item j.

To see the relationship between the preference predictions made with the scrambled data, i.e., using (3), and those made using the original data matrix $\mathbf{D}_k$, let the SVD of $\mathbf{D}_k$ be

$$\mathbf{D}_k = \mathbf{U}_d\mathbf{\Sigma}_d\mathbf{V}_d'. \tag{4}$$

Without loss of generality, we consider the case when $\mathbf{W}_k$ and $\mathbf{H}_k$ are orthogonal matrices, which also means $K = N_k$ and $J = M + K$. From (1), we have

$$\begin{aligned}
\mathbf{U}_d'\mathbf{A}_k &= \begin{bmatrix} \mathbf{U}_d'\mathbf{D}_k & \mathbf{U}_d'\mathbf{W}_k \end{bmatrix}\mathbf{H}_k \qquad (5)\\
&= \begin{bmatrix} \mathbf{\Sigma}_d & \mathbf{U}_d'\mathbf{W}_k \end{bmatrix}\begin{bmatrix} \mathbf{V}_d' & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_K \end{bmatrix}\mathbf{H}_k \\
&= \begin{bmatrix} \mathbf{\Sigma}_d & \mathbf{I}_K \end{bmatrix}\begin{bmatrix} \mathbf{V}_d' & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_d'\mathbf{W}_k \end{bmatrix}\mathbf{H}_k \qquad (6)\\
&= \begin{bmatrix} [\mathbf{\Sigma}_d \;\; \mathbf{0}] + \mathbf{I}_K & \mathbf{0} \end{bmatrix}\mathbf{Q}\begin{bmatrix} \mathbf{V}_d' & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_d'\mathbf{W}_k \end{bmatrix}\mathbf{H}_k,
\end{aligned}$$

where $\mathbf{Q}$ is an orthogonal matrix that combines $\mathbf{\Sigma}_d$ and $\mathbf{I}_M$ together (which can be realized by a sequence of right Givens rotations). Defining

$$\hat{\mathbf{V}} = \mathbf{Q} \left[\begin{array}{cc} \mathbf{V}'_d & \mathbf{0} \\ \mathbf{0} & \mathbf{U}'_d \mathbf{W}_k \end{array} \right] \mathbf{H}_k, \tag{7}$$

which is an orthogonal matrix, we have

$$\mathbf{A}_k = \mathbf{U}_d \left[\begin{array}{cc} [\mathbf{\Sigma}_d \ \ \mathbf{0}] + \mathbf{I}_K & \mathbf{0} \end{array} \right] \hat{\mathbf{V}}. \tag{8}$$

Comparing (8) with (2) and (4), we can see that the matrix $\mathbf{A}_k$ shares the same left orthogonal matrix $\mathbf{U}_d$ with $\mathbf{D}_k$. They have the same row vector space. The singular values of $\mathbf{A}_k$ equal the singular values of $\mathbf{D}_k$ plus 1. The column vector space (i.e., orthogonal matrix $\hat{\mathbf{V}}$) is scrambled by the secret transform matrix $\mathbf{H}_k$ and the artificial noise matrix $\mathbf{W}_k$.

The reduced dimension L in (3) is equivalently as determined according to the singular values (diagonal elements) in $[\mathbf{\Sigma}_d \ \ \mathbf{0}]+\mathbf{I}_M$. Obviously, this is the same as determining L according to $\mathbf{\Sigma}_d$ of (4). From (6), we can verify that the L-dimensional subspace matrix $\tilde{\mathbf{A}}_k$ in (3) also equals

$$\tilde{\mathbf{A}}_k = \left[\begin{array}{cc} \tilde{\mathbf{D}}_k & \tilde{\mathbf{W}}_k \end{array} \right] \mathbf{H}_k, \tag{9}$$

where $\tilde{\mathbf{D}}_k$ is the L-dimensional subspace reconstruction of $\mathbf{D}_k$. Each i^{th} row of the matrix $\tilde{\mathbf{A}}_k$ is a preference prediction for the data owner i scrambled by $\tilde{\mathbf{W}}_k$ and $\mathbf{H}_k$. If a user i can provide a row of source data $\mathbf{D}_k$ to the data broker (to verify that he is the owner of the data, or he knows this user's data anyway) and the corresponding row $\tilde{\mathbf{A}}_k(i)$ in $\tilde{\mathbf{A}}_k$, then the data broker can descramble the prediction results simply via $\tilde{\mathbf{A}}_k(i)\mathbf{H}'_k$ and return the descrambled results to the user i. Note that there is no loss of privacy in this procedure.

If the proposed scheme is applied to preserve the privacy of the data outsourced to untrusted cloud, then the cloud conducts the above big data analytics based on the scrambled data, and feedbacks the results $\tilde{\mathbf{A}}_k(i)$ to the data owner. The data owner can remove the scrambling via $\tilde{\mathbf{A}}_k(i)\mathbf{H}'_k$. This procedure does not lose any data privacy.

Another popular big data analytic task is to generate a list of recommendations that best fit a user's preference [14] [17]. An example is to recommend a list of movies to a user. This can be conducted by finding a set of users whose preferences are most similar to that of this user, and then looking for the items that this set of users rated highest. Since the SVD of the scrambled data matrix $\mathbf{A}_k$ shares the same row space as the source data matrix $\mathbf{D}_k$, the selection can be conducted simply based on (2), and the result is the same as that based on (4). In other words, since our scheme preserves distance, distance calculation and neighborhood selection can be conducted by just using the scrambled data.

3. PRIVACY ANALYSIS OF THE PROPOSED SCHEME

The privacy in this paper means that the adversary is unable to estimate the source data $\mathbf{D}_k$ from the published data $\mathbf{A}_k$ with certain accuracy. Equivalently, the adversary should not be able to estimate the secret transform matrix $\mathbf{H}_k$ as well. Since the estimation of $\mathbf{D}_k$ or $\mathbf{H}_k$ is similar to channel identification or source separation, extensive results in blind or non-blind channel identification can be applied to guide the comprehensive privacy analysis. Considering the severe space limit of this paper, we can only outline the major ideas, with an emphasis on showing the critical role of the sheer scale of big data for guaranteeing privacy. Detailed formulation and quantitative analysis will be reported elsewhere.

In general, channel identification depends on the knowledge available to the receiver. Similarly, in our case, we need to consider different levels of knowledge the adversary may have. Without loss of generality, we consider the problem of estimating one column of the secret matrix $\mathbf{H}_k$, which we denote as $\mathbf{h}$, from

$$\mathbf{y} = \left[\begin{array}{cc} \mathbf{D}_k & \mathbf{W}_k \end{array} \right] \mathbf{h}, \tag{10}$$

where $\mathbf{y}$ is the corresponding column in $\mathbf{A}_k$. Obviously, we can use $\mathbf{W}_k$ to make the matrix $[\mathbf{D}_k \ \ \mathbf{W}_k]$ to be a wide matrix by setting $M + K > N_k$. Then there are infinite number of sets of $\mathbf{D}_k$, $\mathbf{W}_k$ and $\mathbf{h}$ that satisfy (10). Therefore, even with full knowledge about the matrix $[\mathbf{D}_k \ \ \mathbf{W}_k]$, the adversary still can not find the correct secret transform matrix $\mathbf{H}_k$. With less information about this matrix, the estimation of $\mathbf{H}_k$ will only become more difficult. This scale issue limits fundamentally the capability of the adversary's system inversion attacks. Note that existing works do not have this advantage because of the lack of $\mathbf{W}_k$.

More practically, the adversary may know some elements of the source data $\mathbf{D}_k$ only. The adversary can not know $\mathbf{W}_k$ or $\mathbf{H}_k$ *a priori*. Then the problem becomes training-based channel identification. Similar to [19], we can rewrite (10) into

$$\mathbf{y} = \mathbf{X}_1\mathbf{h} + \mathbf{X}_2\mathbf{h}, \tag{11}$$

where $\mathbf{X}_1$ contains all the known elements in $\mathbf{D}_k$ and $\mathbf{X}_2$ denotes the rest of $[\mathbf{D}_k \ \ \mathbf{W}_k]$. Denote $\mathbf{C}$ as the covariance matrix of $\mathbf{X}_2\mathbf{h}$ and assume $\mathbf{y}$ has Gaussian distribution $\mathcal{N}(\mathbf{X}_1\mathbf{h}, \mathbf{C})$. Then we can formulate the estimation of $\mathbf{h}$ into maximum likelihood optimization

$$\hat{\mathbf{h}} = \arg \min_{\mathbf{h}} \quad \ln|\mathbf{C}| + E\left[(\mathbf{y} - \mathbf{X}_1\mathbf{h})'\mathbf{C}^{-1}(\mathbf{y} - \mathbf{X}_1\mathbf{h})\right]. \tag{12}$$

The estimation accuracy is described by the Cramer-Rao Lower Bound (CRLB) $(\mathbf{X}'_1\mathbf{C}^{-1}\mathbf{X}_1)^{-1}$. Thanks to $\mathbf{W}_k$ and the large dimensions, the matrix $\mathbf{X}_1$ is singular almost surely. Therefore, the estimation accuracy is extremely worse. Even if the majority of $\mathbf{D}_k$ is known, the $\mathbf{W}_k$ still prevents any accurate estimation of $\mathbf{h}$.

If the adversary does not know any elements of $[\mathbf{D}_k \ \ \mathbf{W}_k]$, attacks can still be launched based on statistical information about $[\mathbf{D}_k \ \ \mathbf{W}_k]$. The problem changes to blind channel identification. More generally, considering the matrix $\mathbf{H}_k$ in (1), we have the blind source separation (BSS) problem [20].

If $\mathbf{D}_k$ and $\mathbf{W}_k$ are independent, it may be possible for the adversary to use BSS to separate $\mathbf{D}_k$ and $\mathbf{W}_k$ from the mixture $\mathbf{A}_k$. Specifically, the adversary may try to find a $(M + K) \times (M + K)$ matrix $\mathbf{G}$ such that

$$\mathbf{G}\mathbf{H}'_k = \left[\begin{array}{cc} \mathbf{G}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_2 \end{array} \right], \tag{13}$$

where $\mathbf{G}_1$ and $\mathbf{G}_2$ are $M \times M$ and $K \times K$ matrices, respectively. Applying $\mathbf{G}$ to $\mathbf{A}_k$ can separate the data matrix $\mathbf{D}_k$ from the noise $\mathbf{W}_k$, albeit with certain permutation and scaling. Note that the permutation and scaling can be removed based on some existing knowledge about the source data. If $\mathbf{D}_k$ and $\mathbf{W}_k$ are independent, it looks possible for the adversary to find $\mathbf{G}$ so as to separate $\mathbf{D}_k$ from $\mathbf{W}_k$. From BSS theory, we know that the independent source signals can be separated and estimated unless they are Gaussian distributed with proportional covariances.

Existing matrix transform methods suffer greatly from the attacks based on blind channel identification and BSS. In particular, [10] reported the attack that used the correlations of the transformed data, in a way similar to blind channel identification. BSS-based attacks were discussed in [9], and some unrealistic assumptions, such as the source data are dependent rather than independent, had to be applied for privacy. As a matter of fact, these results cast great doubts on the validity of the matrix-transform based methods.

In contrast to the pessimistic results of the existing works, our scheme can successfully prevent this class of attacks thanks to the artificial noise matrix $\mathbf{W}_k$ and the data scale. The key point is that they make the accuracy of source separation extremely low. The adversary has to estimate an $(M+K) \times (M+K)$ matrix $\mathbf{G}$ from just $N_k \times (M+K)$ samples, where $N_k < M+K$. The number of parameters to be estimated is even larger than the number of source data. An upper bound of the estimation accuracy can be derived from the accuracy of estimating the $(M+K) \times (M+K)$ correlation matrix $\mathbf{A}_k'\mathbf{A}_k$ with a limited number of N_k data samples in $\mathbf{A}_k$. Obviously, the rank of $\mathbf{A}_k'\mathbf{A}_k$ is no larger than N_k when it should be $M+K$ ideally.

This idea can be explained more clearly based on a simpler blind channel identification model. Consider the n^{th} row of the matrix $[\mathbf{D}_k \quad \mathbf{W}_k]$, which we denote as $\mathbf{x}(n)$. From (1) we have signal model

$$\mathbf{y}(n) = \mathbf{H}_k'\mathbf{x}'(n), \quad n = 1, \cdots, N_k. \tag{14}$$

Note that $\mathbf{x}'(n)$ is an $M+K$ dimensional vector. Each of its elements can be looked as a source with N_k samples. The first M sources are from the data matrix $\mathbf{D}_k$, while the rest K sources are from the artificial noise matrix $\mathbf{W}_k$. The artificial noise may be independent, and have different distributions, from the source data. The adversary may try to extract one source in $\mathbf{x}'(n)$ based on $\mathbf{y}(n)$. The problem becomes to find a single equalizer vector $\mathbf{g}$ such that

$$\mathbf{g}\mathbf{H}_k' = \mathbf{e}_d \triangleq \left[\ 0, \cdots, 0, 1, 0, \cdots, 0\ \right], \tag{15}$$

where $\mathbf{e}_d$ is a unit vector with all zero entries except a 1 in the d^{th} entry. With only N_k sample vectors $\mathbf{y}(n)$, it is almost impossible to estimate the $1 \times J$, where $N_k < J \leq M+K$, vector $\mathbf{g}$ with sufficient accuracy. For example, according to the Wiener filter theory, the optimal MMSE (minimum mean square error) estimation is

$$\mathbf{g}' = \mathbf{R}^{-1}\mathbf{H}_k'(d), \tag{16}$$

where $\mathbf{R} = E[\mathbf{y}(n)\mathbf{y}'(n)]$ is the $J \times J$ correlation matrix, and $\mathbf{H}_k'(d)$ is the d^{th} column of the matrix $\mathbf{H}_k'$. We can design the size of $\mathbf{W}_k$ so that the correlation matrix $\mathbf{R}$ can not be estimated accurately enough. This effectively prevents the estimation of $\mathbf{g}$.

Without being able to separate the source data $\mathbf{D}_k$ from the artificial noise $\mathbf{W}_k$, the adversary has to consider all the artificial noise as unknown additive noise. Then the model (14) becomes

$$\mathbf{y}(n) = \tilde{\mathbf{H}}_k\mathbf{x}_d'(n) + \mathbf{v}'(n), \quad n = 1, \cdots, N_k, \tag{17}$$

where $\mathbf{x}_d(n)$ is the n^{th} row of the source data matrix $\mathbf{D}_k$, and $\mathbf{v}(n)$ is the n^{th} row of the multiplication of the artificial noise matrix $\mathbf{W}_k$ with the corresponding submatrix in the secret transform matrix $\mathbf{H}_k$. The matrix $\tilde{\mathbf{H}}_k$ is the submatrix of $\mathbf{H}_k$ corresponding to the source data $\mathbf{D}_k$. Compared with (14), besides the insufficient number of sample vectors, the channel identification is further suffered from a low signal-to-noise ratio (SNR), thanks to the artificial noise. Specifically, the SNR of $\mathbf{y}(n)$ can be derived as

$$\gamma = \frac{E[\|\tilde{\mathbf{H}}_k\mathbf{x}_d'(n)\|^2]}{E[\|\mathbf{v}'(n)\|^2]} = \frac{M}{K}. \tag{18}$$

We can select large enough K to reduce the SNR γ. Fundamentally, information theory specifies that correct symbol recovery are not available when the SNR is lower than some threshold value. In this sense, even if the adversary knows $\tilde{\mathbf{H}}_k$ *a priori*, it still can not estimate $\mathbf{x}_d'(n)$ from $\mathbf{y}(n)$ with sufficient accuracy.

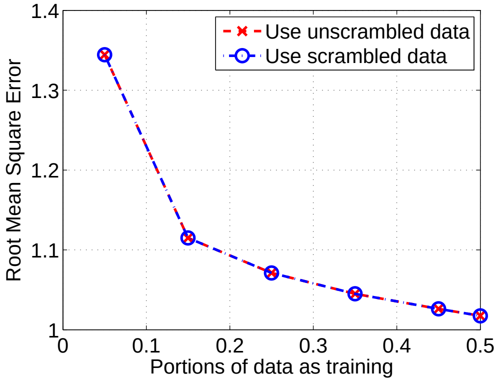

Fig. 2. Comparison of SVD-based rating predictions based on unscrambled data or the scrambled data with the proposed scheme.

In summary, the artificial noise matrix $\mathbf{W}_k$ plays a critical role in preventing the estimation of the secret transform matrix $\mathbf{H}_k$ as well as the source data $\mathbf{D}_k$ from the scrambled data $\mathbf{A}_k$ and some *a priori* information. The privacy of the source data can be guaranteed.

4. SIMULATIONS

We used the movie rating data in [14] to verify the data utility of our proposed scheme. We compared our approach with the scrambled data against the conventional SVD-based approach with unscrambled data [14]. We randomly generated a 1682×1682 orthogonal matrix $\mathbf{H}_k$ and random artificial noise to scramble the data. We used different portions of the rating data as training to generate rating predictions, and calculated the root mean square error (RMSE) between the predicted rating and the actual rating. The simulation results are shown in Fig. 2, which suggests that our approach could keep the same data utility as the conventional SVD with the original data.

We also simulated some blind or nonblind channel identification based attacks. The accuracy of the estimation of the source data $\mathbf{D}_k$ was extremely low. For the movie rating data, the RMSE of the adversary were generally around 3. Note that the rating data were distributed as integers from 1 to 5. This demonstrates the extremely low estimation accuracy for the adversary.

5. CONCLUSIONS

For privacy-preserving big data publishing, we propose a scheme which uses jointly artificial noise and secret transform matrix to scramble the data. This scheme preserves desirable data utility by preserving the geometric properties such as distance among the data records so that many existing big data analytics can be conducted based on the scrambled data. This scheme preserves privacy thanks to the artificial noise and the sheer scale of big data. Privacy is analyzed based on blind and non-blind channel identification and source separation theories. Simulations are conducted to demonstrate the proposed scheme.

6. REFERENCES

[1] B. Chandramouli, J. Goldstein and S. Duan, “Temporal analytics on big data for web advertising,” *IEEE 28th International Conf. on Data Engineering*, 2012.

[2] X. Wu, X. Zhu, G.-Q. Wu and W. Ding, “Data mining with big data,” *IEEE Trans. Knowledge and Data Engineering*, vol. 26, no. 1 pp. 1041-4347, Jan. 2013.

[3] N. R. Adam and J. C. Wortmann, “Security-control methods for statistical databases: A comparative study,” *ACM Computing Surveys*, vol. 21, no. 4, pp. 515-556, Dec. 1989.

[4] C. C. Aggarwal and P. S. Yu (editors), *Privacy-Preserving Data Mining: Models and Algorithms*, Springer, 2008.

[5] B. C. M. Fung, K. Wang, R. Chen and P. S. Yu, “Privacy-preserving data publishing: A survey of recent developments,” *ACM Computing Surveys*, vol. 42, no. 4, article 14, June 2010.

[6] C. Aguilar-Melchor, S. Fau, C. Fontaine, G. Gogniat and R. Sirdey, “Recent advances in homomorphic encryption: A possible future for signal processing in the encrypted domain,” *IEEE Signal Processing Magazine*, vol. 30, no. 2, pp. 108-117, March 2013.

[7] C. C. Aggarwal and P. S. Yu, “On privacy-preservation of text and sparse binary data with sketches,” *Proceedings of the Seventh SIAM International Conference on Data Mining*, April 26-28, 2007, Minneapolis, Minnesota, USA.

[8] S. R. M. Oliveira and O. Zaane, “Privacy preserving clustering by data transformation,” *Proc. 18th Brazilian Symp. Databases*, pp. 304-318, Oct. 2003.

[9] K. Chen and L. Liu, “Towards attack-resilient geometric data perturbation,” *SIAM Data Mining Conference*, 2007.

[10] K. Liu, C. Giannella and H. Kargupta, “An attacker’s view of distance preserving maps for privacy preserving data mining,” *Knowledge Discovery in Databases: PKDD 2006*, Lecture Notes in Computer Science, vol. 4213, pp. 297-308, 2006.

[11] W. Lu, A. L. Varna, A. Swaminathan and M. Wu, “Secure image retrieval through feature protection,” *ICASSP*, 2009.

[12] N. K. Ratha, J. H. Connell and R. M. Bolle, “Enhancing security and privacy in biometrics-based authentication systems,” *IBM Systems. Journal*, vol. 40, no. 3, pp. 614-634, 2001.

[13] S. P. Kasiviswanathan, M. Rudelson and A. Smith, “The power of linear reconstruction attacks,” *arXiv preprint, arXiv:1210.2381*, 2012.

[14] B. M. Sarwar, G. Karypis, J. A. Konstan and J. T. Riedl, “Application of dimensionality reduction in recommendation system - A case study,” In *ACM WEBKDD Workshop*, 2000.

[15] J. Canny, “Collaborative filtering with privacy,” *IEEE Symp. Security and Privacy*, pp. 45-57, Oakland, CA, May 2002.

[16] G. Takacs, I. Pilaszy, B. Nemeth and D. Tikk, “Investigation of various matrix factorization methods for large recommender systems,” *ICDM*, 2008.

[17] Y. Koren, R. Bell, et al, “Matrix factorization techniques for recommender systems,” *Computer*, 2009.

[18] L. Sankar, S. R. Rajagopalan and H. V. Poor, “Utility-privacy tradeoffs in databases: An information-theoretic approach,” *IEEE Trans. Info. Forensics and Security*, vol. 8, no. 6, pp. 838-852, June 2013.

[19] O. Rousseaux, G. Leus, P. Stoica and M. Moonen, “Gaussian maximum likelihood channel estimation with short training sequences,” *IEEE Trans. Wirel. Commun.*, vol. 4, no. 6, pp. 2945-2955, Nov. 2005.

[20] P. Comon and C. Jutten (editors), *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 2010.