2015 IEEE 16th International Symposium on High Assurance Systems Engineering

Signal Processing Oriented Approach for Big Data Privacy

Xiaohua Li
Dept of Electrical and Computer Engineering
State University of New York at Binghamton
Binghamton, NY 13902
Email: xli@binghamton.edu

Thomas Yang
Dept of ECSSE
Embry-Riddle Aeronautical University
Datona Beach, FL 32114
Email: tianyu.yang@erau.edu

***Abstract*—This paper addresses the challenge of big data security by exploiting signal processing theories. We propose a new big data privacy protocol that scrambles data via artificial noise and secret transform matrices. The utility of the scrambled data is maintained, as demonstrated by a cyber-physical system application. We further outline the proof of the proposed protocol's privacy by considering the limitations of blind source separation and compressive sensing.**

***Keywords*-big data privacy, signal processing, cyber-physical systems**

I. Introduction

Big Data refers to the explosive amount of data generated in today's society. Compared with conventional databases, big data has new features in terms of volume, variety and velocity. One of the major hurdles for the application of big data is the challenge of data privacy.

It is well known that it is not sufficient to preserve privacy by simply removing the identities of data owners. With some external knowledge, it is often possible to discover a substantial amount of private information through analyzing the published data, even after the removal of data owner identities. To guarantee privacy, data must be processed by more advanced anonymization techniques, such as perturbation and k-anonymity. However, these conventional privacy techniques are not very effective for big data [1].

In this paper, we propose a big data privacy scheme that guarantees privacy by exploiting signal processing theories in blind source separation and compressive sensing. Our method is computationally more efficient over cryptography (encryption)-based approaches which usually have prohibitively high computational complexity for big data. Our method can also overcome the security weaknesses encountered by most existing noise perturbation techniques.

The remaining of the paper is organized as follows. In Section II, we develop the new data privacy technique and present an example application to illustrate the data utility. In Section III, we analyze privacy based on signal processing theories. Conclusions are drawn in Section IV.

II. New data privacy protocol with application in power demand forecast

Consider a data set consisting of N data records, each of which has M attributes (i.e., dimensions or values). The set of data is organized into an $N \times M$ matrix $\mathbf{D}$. We assume that there is a trusted data broker that collects the data $\mathbf{D}$, anonymizes it, and makes it publicly available to all the data users. To preserve privacy, we propose that the data broker applies a $N \times I$ artificial noise matrix $\mathbf{W}$ and a $(M+I) \times (M+I)$ transformation matrix $\mathbf{H}$ to scramble the data matrix $\mathbf{D}$, and publishes the scrambled data

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

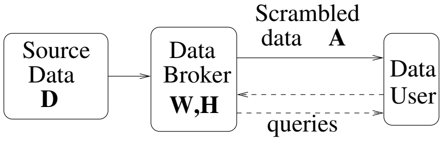

Figure 1. The proposed big data privacy preserving protocol.

Note that $\mathbf{D}$, $\mathbf{W}$, and $\mathbf{H}$ are kept unknown to data users. This scheme can be adapted easily to various big data tools such as MapReduce. Note that we just need a matrix multiplication, which can be implemented in block-style, and is more efficient than full data encryption.

All data users then conduct their own data analysis based on the scrambled data $\mathbf{A}$. After the data analysis, to remove the scrambling and obtain the information of interest to data users, a query procedure is implemented, in which data users submit their scrambled results to the data broker, who then returns the unscrambled information to data users. This protocol is shown in Fig. 1.

Although (1) limit certain data utility, many important big data analytic techniques, especially those based on the geometric properties among the rows of the data matrix $\mathbf{D}$, can still be conducted. This includes a large amount of big data techniques that rely on matrix factorization and dimension reduction [2]. Many important machine learning methods, such as data clustering and classification, can be supported as well. In the sequel, we consider specifically a cyber-physical system application to demonstrate the utility of the scrambled data, i.e., the problem of forecasting the charging demand of electrical vehicle (EV) or plug-in hybrid electrical vehicle (PHEV).

The majority of charging stations take long hours to fully charge an EV. For charging stations, the prediction of overall charging demand of all EVs or PHEVs is important, because such information is needed to adjust the stations' inventory. For EV or PHEV users, the prediction of the charging time of each charging station is helpful for planning their charging needs. Here we consider a special task, i.e., exploiting the historical power consumption data of a charging station to predict its power consumption in the near future. The prediction can be conducted by the popular k-Nearest Neighbor (kNN) algorithm [3]. Making the source data public may leak user's privacy, e.g., some users' activities may be tracked. To preserve privacy, we propose to use our protocol to scramble the

978-1-4799-8111-3/15 $31.00 © 2015 IEEE
DOI 10.1109/HASE.2015.23

source data. We will show that the task of prediction can be realized on the scrambled data.

First, consider the unscrambled data $\mathbf{D}$, where each row of $\mathbf{D}$ consists of the power consumption data sampled during several days. Let $\mathbf{D}(i, j)$, $j = 1, \cdots, T$, be the first T power consumption data of the ith record. For example, they can be the first day's data samples. Let $\mathbf{D}(i, j)$, $j = T + 1, \cdots, M$, be the power consumption data samples in the days before the first day. For example, they can be the data samples of the K days before the first day. Now, with the most recent K days' data sample $\mathbf{x}$, which is a $1 \times (M - T)$ vector, we need to predict tomorrow's power consumption data $\mathbf{y}$, which is a $1 \times T$ vector.

With the kNN algorithm [3], we first calculate the distance between $\mathbf{x}$ and each of the rows of $\mathbf{D}$ as $d_i = \mathbf{x}\mathbf{D}(i, T+1 : M)'$. Next, we select k rows that have the smallest distances among all N distances. The row index can be written as $\alpha(j)$, where $0 \leq j \leq k-1$, $d_{\alpha(j)} \leq d_i$, $0 \leq i \leq N-1$, $0 \leq \alpha(j) \leq N-1$, and $i \neq \alpha(j)$ for any i and any j. Then, the output $\mathbf{y}$ can be calculated as a weighted sum $\mathbf{y} = \sum_{j=0}^{k-1} \gamma_j \mathbf{D}(\alpha(j), 1 : T)$, where the weighting coefficients satisfy $\gamma_j \in [0, 1]$ and $\sum_{j=0}^{k-1} \gamma_j = 1$.

In order to preserve data privacy, we consider the case where EV or PHEV users are provided with the scrambled data $\mathbf{A}$ only, rather than the original data $\mathbf{D}$. Due to the special structure of the power consumption data record, we let $\mathbf{H}$ be block diagonal and rearranged (1) into

$$\begin{aligned}\mathbf{A} = [&\mathbf{D}(:, 1 : T), \mathbf{W}(:, 1 : J),\\ &\mathbf{D}(:, T+1 : M), \mathbf{W}(:, J+1 : I)] \operatorname{diag}\{\mathbf{H}_1, \mathbf{H}_2\}, \qquad (2)\end{aligned}$$

where $\mathbf{H}_1$ and $\mathbf{H}_2$ are $(T+J) \times (T+J)$ and $(M-T+I-J) \times (M-T+I-J)$ orthogonal matrices. Similarly, the data vector $\mathbf{x}$ is scrambled into $\tilde{\mathbf{x}} = [\mathbf{x} \;\; \mathbf{w}]\mathbf{H}_2$ with an artificial noise vector $\mathbf{w}$.

Applying the kNN algorithm to the scrambled data $\mathbf{A}$ and $\tilde{\mathbf{x}}$, it is easy to verify that the distance $\tilde{d_i} = \tilde{\mathbf{x}}\mathbf{A}(i, T+J+1 : M+I)' = d_i + \mathbf{w}\mathbf{W}(i, J+1 : I) \approx d_i + c$. Obviously, the selection of k smallest distances leads to the same results $\alpha(j)$ as using the original data. The weighted sum $\tilde{\mathbf{y}} = \sum_{j=0}^{k-1} \gamma_j \mathbf{A}(\alpha(j), 1 : T+J)$ is scrambled by $\mathbf{H}_1$. The user can send $\tilde{\mathbf{y}}$ to the data broker, which descrambles the vector and sends back $\tilde{\mathbf{y}}\mathbf{H}_1^{-1}$ after removing the artificial noise vector.

III. Privacy analysis

We consider a passive attacker who attempts to recover the data matrix $\mathbf{D}$ from the published scrambled data $\mathbf{A}$.

Blind source separation (BSS) [4] based attacks are a severe concern to data security. Referring to (1), even if the adversary does not know any elements of $[\mathbf{D} \;\; \mathbf{W}]$, based on statistical properties of $[\mathbf{D} \;\; \mathbf{W}]$, attacks can still be launched using BSS methods to estimate columns of $[\mathbf{D} \;\; \mathbf{W}]$ (attributes). However, we argue that this type of attacks will not be successful in our case. First, for many applications, the attributes are generally not independent from each other. For example, for a specific time of the day, the consumption data of different days are almost certain to be correlated. Therefore, the essential BSS assumption that the components to be estimated are independent is violated. Second, the inherent order and variance uncertainties of BSS estimation require the adversary to have some additional knowledge about the original data set in order to resolve the ambiguities in the estimated attribute data. Many times, this type of additional knowledge is not available. Third, even if the attributes in $\mathbf{D}$ are independent and the adversary has enough information to resolve the order and variance uncertainties, the artificial noise $\mathbf{W}$ can be easily designed to prevent BSS estimation. For example, to preserve the distance among the rows of $\mathbf{D}$, rows of $\mathbf{W}$ are typically made orthogonal, which means they are not independent. Therefore, if BSS estimation is still attempted in this case, $\mathbf{W}$ has to be treated as additive noise in the signal model, and the estimation will be impossible as long as the variance of the transformed noise (multiplication of $\mathbf{W}$ with the corresponding submatrix in $\mathbf{H}$) is sufficiently large. This can be easily illustrated from the well-known channel identification problem in wireless communications, since information theory dictates that correct symbol recovery is not possible when the signal to noise ratio is lower than some threshold value.

Another major risk to privacy comes from the query procedure, where any data user can send a number of query data vectors to collect enough data to estimate the scrambling matrices. The data broker can limit the number of queries per data user to mitigate this risk. A more serious problem, however, is that even with a small number of queries, the attacker may still apply compressive sensing to estimate $\mathbf{D}$, especially if $\mathbf{D}$ is sparse. Compressive sensing-based attack is a severe challenge to privacy-preserving data sharing [5]. Fortunately, in our case, the data broker can mitigate this attack successfully by replying the queries with a generalized $\hat{\mathbf{y}}_i$ (e.g., by quantizing it to just a few levels). Our preliminary simulations have demonstrated this effective mitigation against compressive sensing attacks (not included due to space limitation).

IV. Conclusions

In order to preserve big data privacy, we propose a signal processing oriented approach, which uses both artificial noise and secret transform matrices to scramble the data. The application of the proposed approach in power demand forecast is presented to demonstrate the preservation of data utility. Data privacy is analyzing by blind source separation and compressive sensing theories.

References

[1] C. C. Aggarwal and P. S. Yu (editors), *Privacy-preserving data mining: Models and Algorithms*, Springer, 2008.

[2] Y. Koren, R. Bell, et al, "Matrix factorization techniques for recommender systems," *Computer*, 2009.

[3] M. Majidpour, et al, "Fast demand forecast of electric vehicle charging stations for cell phone application," *Proc. IEEE/PES General Meeting*, Washington, DC, July 2014.

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

[5] C. Dwork, F. McSherry and K. Talwar, "The price of privacy and the limits of LP decoding," *ACM STOC'07*, San Diego, CA, June 2007.