Hindawi
Security and Communication Networks
Volume 2017, Article ID 8026787, 8 pages
https://doi.org/10.1155/2017/8026787

Research Article

Privacy Preserved Self-Awareness on the Community via Crowd Sensing

Huiting Fan, Kai Xing, Lei Tan, Weikang Rui, Zhonghu Xu, Shuo Zhang, and Jing Xu

School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, China

Correspondence should be addressed to Kai Xing; kxing@ustc.edu.cn

Received 26 February 2017; Accepted 16 April 2017; Published 14 June 2017

Academic Editor: Qing Yang

In social activities, people are interested in some statistical data, such as purchase records, monthly consumption, and health data, which are usually utilized in recommendation systems. And it is seductive for them to acquire the ranking of these data among friends or other communities. In the meantime, they want their privacy data to be confidential. Therefore, a strategy is presented to allow users to obtain the result of calculating their privacy data while preserving these data. In this method, firstly a polynomial approximation function model is set up for each user. Afterwards, "fragment" the coefficients of each model into pieces. Eventually "blend" all scraps to build the global model of all users. Users can use the global model to gain their corresponding ranking results after a special computing. Security analyses of three aspects elaborate the validity of proposed privacy method, even if some spiteful attackers try to steal private data of users, no matter who they are (users or someone outside the community). Experiments results manifest that the global model competently fits all users data and all privacy data are protected.

1. Introduction

People are increasingly involved in various online/offline social activities. Especially with smart mobile terminals (mobile phones, tablet PCs, wearing equipment, etc.) continuing invading people's lives, the time people spend on social software is also in escalation. Of course, an increase in social activities related data is also observed. For example, when people use applications (like Amazon, Uber, and Twitter), the amount of purchase records, route, and registration information constantly augments.

For Internet service providers a large number of data are propitious to provide better service to people by existing state-of-the art data mining technology. Take recommendation systems, for example. Recommendation system can provide a recommendation list from thousands of items for users when they want to select one or more items, which avoids plenty of time in choosing. The service should not only meet curiosity of users (data providers) about ranking or other knowledge of themselves, but also ensure the security of users' private data. For instance, users among the same social community are curious about the rankings of their income while they do not want their income data to be leaked to any service provider, even if to their social friends. After all, even though two users are in the same social community, they may be not familiar with each other. Particularly, even between two good friends income data also should not be disclosed. The aforementioned case leads us to find a strategy to both meet the user's ranking calculation requirements moreover and protect the user's privacy data, that is, a social ranking strategy under privacy data protected.

In this paper, we, respectively, establish a polynomial function approximation model for each user's data, since there exists a unique optimal uniform approximation polynomial function for any function in the polynomial space. For a user's local model, all coefficients of the function model are sliced, which occurred in this user's local site. We set a separate site as the global data analyst to integrate all users sliced coefficients data. The data analyst is responsible for blending all fragments sent by users so as to establish the global model. Notice that one piece of coefficients of each user is kept by the user himself while others are delivered to the data analyst. After these processes, in the condition of user's privacy data being protected, the global model of all users is established. All users can get their ranking results from a calculation of the global model. In our previous work [1]

a privacy preserved method for a community was presented, which did not consider the social relationship among users. The social link can impel more curiosity of users, which can lead to more data attributed by users.

The rest of the paper is structured as follows: Section 2 introduces some existing works on social ranking and privacy-preserving methods. Section 3 presents some assumptions, our goals, and definitions. In Section 4, how to get the global model with our privacy-preserving algorithm is presented as well as obtaining a ranking among a user's social friends. In Section 5 we analyse the security of our privacy-preserving algorithm. Section 6 demonstrates the accuracy of our scheme with experiment results. Then some conclusions are shown in Section 7.

2. Related Work

In our privacy-preserving scheme, all users (data providers) cooperatively learn a global model while no data and parameters of model of all users are disclosed. Meanwhile, a data analyst who holds the global model will provide a computation result of ranking of any user's private data. Reference [2] also lets participants jointly learn a model while no input data sets of participants are revealed, but in training process participants should share small subaggregates of their models' key parameters. In the case of massive data stored in an untrusted server by a client, when the client would like to calculate a function on some part of its outsourced data, it could require the server. Reference [3] presents a protocol that the client can efficiently verify the results provided by the server. Rial and Danezis [4] allow grid users to prove the accuracy of computations according to the readings on their device with a privacy-preserving protocol. Ahn et al. [5] proposed a generic framework for kinds of concepts of computing on authenticated data, such as arithmetic, transitive, and homomorphic signatures. In order to protect the privacy of the data, there exist varieties of methods. Dwork et al. [6] implemented distributed protocols to achieve privacy-preserving by generating shares of random noise. Li et al. [7] applied a homomorphic encryption to their privacy-preserving demand response (EPPDR) scheme to realize demand response efficiently in a privacy-preserving way. Acs and Castelluccia [8] protect call-data-record (CDR) data sets using an anonymization scheme with differential privacy.

In this paper, we concentrate on how to model users and compute what users need (their social ranking) while how to get the social relationship among users is not presented. Such related work can be found in [9–13]. The first privacy-preserving data mining algorithm was introduced by Agrawal and Srikant [14], which allows parties to cooperate without revealing personal data of any party.

3. Assumptions and Goals

In order to concentrate on how to achieve the protection of user privacy data, we have the following assumptions:

(i) Each user is a local site where there is a data processing program to set up the local model.

(ii) Those users who are willing to gain the model of a group where they are provide their own data to model establishment process, and their data is stored in local site and others cannot get it.

(iii) The social relations of a user are authorized to the processing program and we do not need to mine them.

(iv) The data analyst is distinguished from any user but can communicate with all users and is equipped to handle complex data.

(v) There is a semihonest model among the data analyst and all users.

(vi) The attacker can be any one, even the data analyst. However, the number of attackers is limited to less than n.

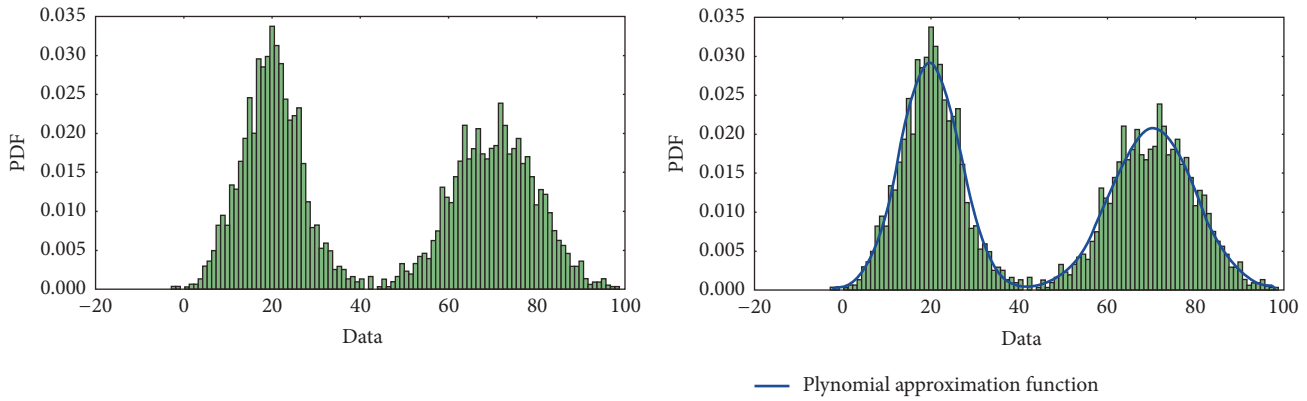
In the semihonest model, each participator (users and the data analyst) evolved in these protocols has to follow the rules using correct input, and all of them will not gather the middle temporary results during the construction of the global model, which can endanger the security. This semihonest model is reasonable in many situations because any participator who wants to mine data will follow these protocols to ensure the final correct result. Even one participator colludes with some others, which is called the collusion attack; no one can obtain anything about others' privacy data.

In our scheme, we consider that in a social community there are n users $1, 2, \ldots, n$. X^j is the private data of user i and it has p_i observed data $\{x_1^{(j)}, x_2^{(j)}, \ldots, x_{p_i}^{(j)}\}$. We set $y_i^{(j)}$ as the probability that $x_i^{(j)}$ appears in X^j. Each user's data is independently identically distributed. Unless the user divulges his privacy data to other people, his data is safe.

We want to attain a global model for all users at the data analyst site while users do not need to send private data to the data analyst. If a user would like to know his ranking among friends, he is able to gain the result utilizing the global model. As we have assumed, our privacy protection is based on a semihonest security model with the proposed protocols. It is applicable because any user must conform our protocols to get the eventual valid result; even a user desires to filch other user's data. Furthermore, if some users conspire to one user or more users data, that is, a collusion attack, it is still fruitless.

4. Social Group Modeling and Social Ranking with Privacy-Preserving

Our procedures about modeling social group under privacy-preserving are propounded in this section. There are mainly three steps to accomplish modeling procedure for a group. Foremost, for each user's own data a polynomial approximation function is built as his local model and the parameters of the model thereupon are created. Subsequently, a probability coefficient is allocated to each user in line with the quantity of the user's data. Thenceforth, the probability coefficient of each user is multiplied by his function model and the product is sliced into several parts in a specific approach. One part is kept by the user himself and others are sent



(a) Private data probability distribution

(b) Polynomial approximation of the data probability distribution

Figure 1: Data probability distribution and polynomial approximation function.

to other users randomly. Afterwards, all users dispatch all received information to the data analyst. Finally, the data analyst merges all received fragments and construct the global model.

4.1. Function Modeling. In user j's local site, his data X^j is private and possessed securely, which means that no one can infer and infringe it. Every user's data X^j is converted into (x_i^j, y_i^j), $i = 1, 2, \ldots, p_i$. We have come to know that there exists a unique optimal uniform approximation polynomial function for any function in the polynomial space; therefore each user's data model can be built by the polynomial approximation function algorithm. Then each user has his data model in a form of

$$f^j(x) = \sum_{i=0}^{m} A_i^j x^i, \tag{1}$$

where $f^j(x) = y^j = A^j X$, $A^j = (A_0^j, A_1^j, \ldots, A_m^j)$, and $X = (1, x, \ldots, x^m)^T$.

An experiment is conducted to verify the accuracy of our method. In Figure 1(a), we demonstrate the Probability Distribution Model of a user's privacy data, as well as indicating the distribution features. The polynomial approximation function model generated on the basis of the same data is denoted in Figure 1(b). Apparently, the polynomial approximation function could suffice the distribution features of a set of data, in that it fits the data very well, nearly having the same characteristics as the data. It also does not reveal the real data, which accommodates our privacy requirement.

Due to the outstanding properties of the polynomial approximation function, we undoubtedly select it as the global model. In addition, we take the amount of data contributed by each user into account when modeling for all users. The data analyst assigns a weighted coefficient π^j to each user based on the amount of his data. It is the rate between the number of users j's data and the number of all users. π^j is multiplied by every user j ($j = 1, 2, \ldots, n$). Thus, in user j's site he has $\pi^j f^j(x)$.

$$\pi^j f^j(x) \begin{bmatrix} \pi^j A_0^j & \pi^j A_1^j & \pi^j A_2^j & \cdots & \pi^j A_m^j \end{bmatrix} \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^m \end{bmatrix}. \tag{2}$$

4.2. Fragmenting User Model and Blending Users Models

Fragmenting. For the sake of protecting private data and acquiring the global model, first of all, every coefficient $\pi^j A_i^j$ of user j ($j = 1, 2, \ldots, n$) is sliced into c parts: $A_{1i}^j, A_{2i}^j, \ldots, A_{ci}^j$, and $\sum_{k=1}^{c} A_{ki}^j = \pi^j A_i^j$, where c is a constant and $n/2 \le c \le n$, $i = 0, 1, \ldots, m$. $A_{1i}^j, A_{2i}^j, \ldots, A_{ci}^j$ can be negative or positive.

Consequently user j's model function $\pi^j f^j(x)$ can be transformed into a new functional form:

$$\begin{aligned} \pi^j f^j(x) &= \mathbb{I}_c \begin{bmatrix} \begin{bmatrix} A_{10}^j & A_{11}^j & A_{12}^j & \cdots & A_{1m}^j \\ A_{20}^j & A_{21}^j & A_{22}^j & \cdots & A_{2m}^j \\ \vdots & \vdots & \vdots & & \vdots \\ A_{c0}^j & A_{c1}^j & A_{c2}^j & \cdots & A_{cm}^j \end{bmatrix} \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^m \end{bmatrix} \end{bmatrix} \\ &= \mathbb{I}_c B^j X, \end{aligned} \tag{3}$$

where $\mathbb{I}_c = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \end{bmatrix}_c$ is a c-dimensional unit vector. Denote the ith row of the matrix B^j as B_i^j.

$$B_i^j = \begin{bmatrix} A_{i0}^j & A_{i1}^j & A_{i2}^j & \cdots & A_{im}^j \end{bmatrix}. \tag{4}$$

Input: A user's data set with p_j observational data $\{x_1^{(j)}, x_2^{(j)}, \ldots, x_{p_j}^{(j)}\}$, user's data number probability coefficient π^j.
Output: The user's polynomial approximation function model $\pi^j f^j(x)$, and the coefficient matrix B^j.

(1) Each user transforms his data into the form (x_i^j, y_i^j), where y_i^j is the probability of x_i^j in the user's data set;
(2) All users build their own data models with the polynomial approximation function algorithm in a distributed system;
(3) Each user slices his data distribution model by Eq. (3) to obtain the coefficient matrix B^j;
(4) For each user, one of the coefficient matrix rows is kept by himself and the remaining row pieces are sent to other users randomly;
(5) Each user collects all the received matrix rows, mixes them and send the mixed result to the data analyst, in the same way, the data analyst can reconstruct the final model in the community by Eq. (5).

Algorithm 1: Privacy preserved community modeling.

Blending. If an attacker amasses all pieces of the coefficient matrix or the matrix B of a user, the user's local model may be deduced. To avoid this case, each user j retains one row of the matrix B^j, sending all the other ones to other users randomly. B_i^{jk} is a piece of matrix row form user j to user k. If user k does not receive any matrix row from user j, $B^{jk} = \mathbf{0}$.

4.3. Global Modeling. To make sure that every user can receive all the pieces sent from others, there is a period of exclusive time for all users to transmit pieces. When all pieces sent to a user j arrive, he aggregates and mixes them as $U^j = B^{1j} + B^{2j} + B^{3j} + \cdots + B^{nj}$. Ultimately, U^j is sent to the data analyst.

In the aforementioned exclusive time, the data analyst can gather all the U^j $(j = 1, 2, \ldots, n)$ which can be applied in modeling all users' data. Hence, the final model $F(x)$ is generated.

$$F(x) = \mathbb{I}_{1\times n}\left[\begin{bmatrix} U^1 \\ U^2 \\ \vdots \\ U^n \end{bmatrix}_{n\times(m+1)} \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^m \end{bmatrix}_{(m+1)\times 1}\right] = \sum_{i=0}^{m} b_i x^i. \quad (5)$$

How our privacy preserved social group modeling works is exemplified in Figure 2 with $n = 4$ users and slicing size $c = 4$.

Our model scheme is instructed by Algorithm 1.

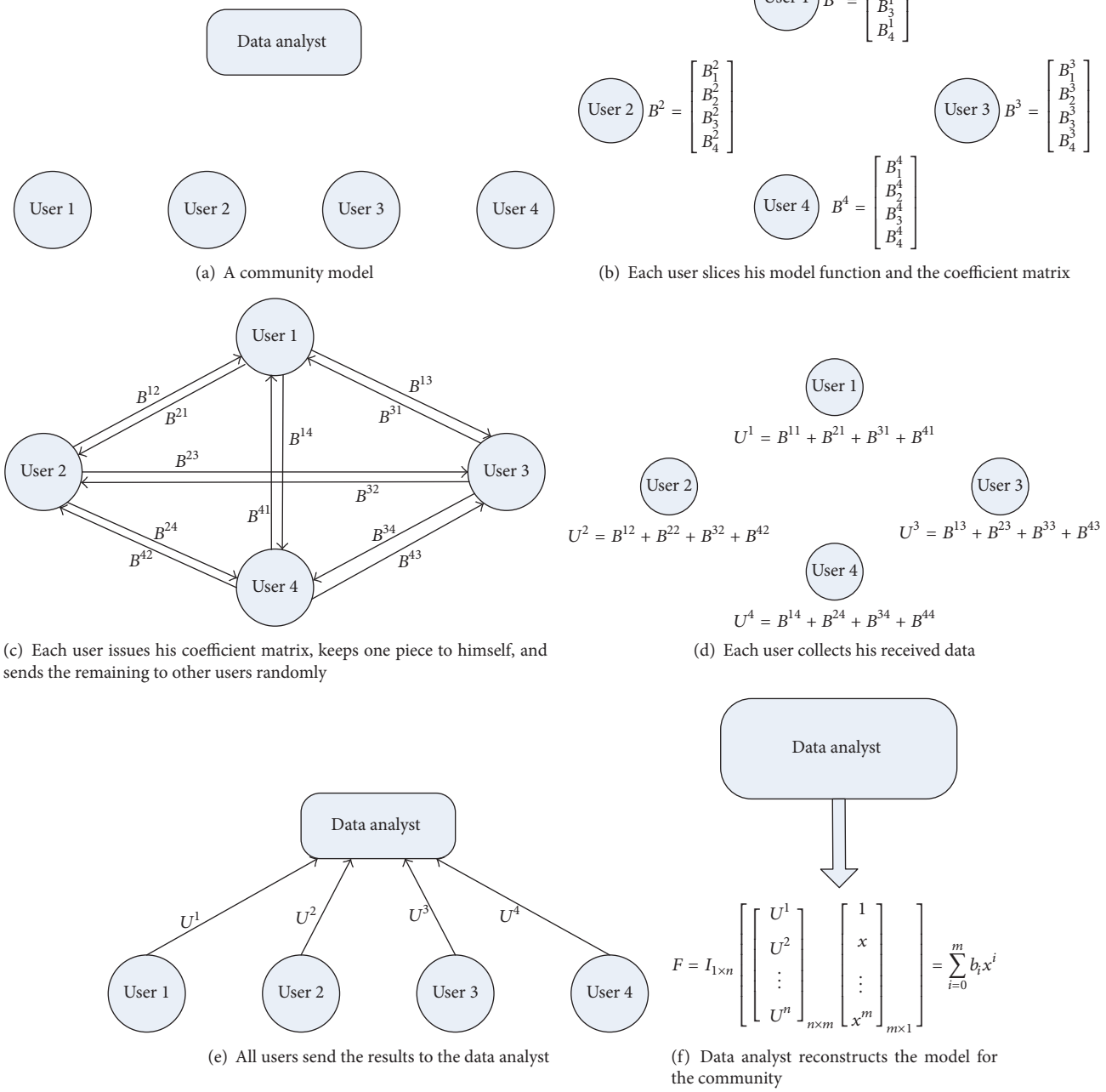
4.4. Getting My Ranking in a Social Group. Now for a social group the distribution model of all users is in the possession of the data analyst and all users have knowledge of it; the ranking of a user is available. According to previous knowledge, $F(x)$ characterizes the Probability Distribution Model of all users' data. Therefore, no private data is disclosed to anyone, even the data analyst. A user can calculate $F(x)$ to get his ranking, a probability value, with his data x.

5. Security Analysis

We should state the aims of social ranking under privacy protecting before security analysis for all steps abovementioned in Section 4.

(i) Only the user himself can hold his model, which means others cannot obtain it in any ways, even the data analyst.
(ii) Even if some users collude together and share all they have received, extracting the model of someone else is beyond their abilities.
(iii) The data analyst is responsible for constructing the global model of all users while it should be absolutely ignorant of any user's data and local data model.

5.1. Security Analysis at Each User. In Section 4.1 the local model of a user is generated by the polynomial approximation function with his own private data. We have demonstrated that making use of polynomial approximation function will not leak out the real accurate data. Moreover, the whole process of local modeling thoroughly occurred in user's local site and no one else is involved in. As a consequence, the private data and all the coefficients are only held by the user himself, which elucidates that these are unavailable for other users and the data analyst.



(a) A community model

(b) Each user slices his model function and the coefficient matrix

(c) Each user issues his coefficient matrix, keeps one piece to himself, and sends the remaining to other users randomly

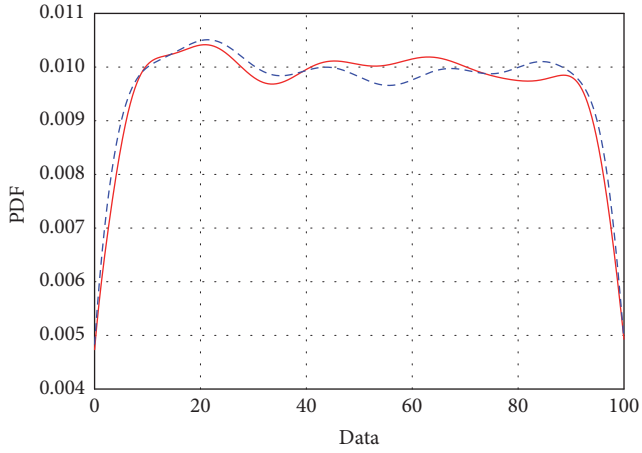
(d) Each user collects his received data

(e) All users send the results to the data analyst

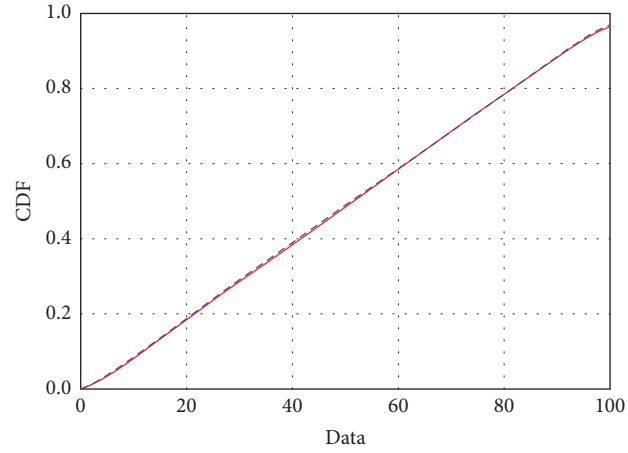
(f) Data analyst reconstructs the model for the community

Figure 2: The process of privacy preserved community modeling.

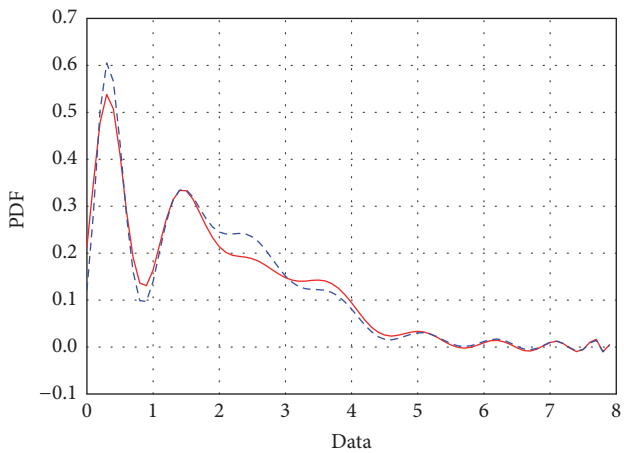
5.2. Security Analysis for Fragmenting and Blending for All Users. Fragmenting and blending parts presented in Section 4.2 strengthen the privacy protection efforts. First of all, fragmenting drives the coefficients of each user to be more complicated, which turns coefficients of a polynomial approximation function into a matrix, and the user himself decides the row number of this matrix. As we assumed, the number of the attackers is limited to less than n. That is because the probability of all users being attackers is extremely low. In addition, what is sent to others is one row of a variant of the coefficients matrix. As a final point, send pieces randomly instead of designating stationary recipients and it is perplexing for others to collect all pieces, not to mention one piece always kept by the user. Therefore, even $n-1$ attackers among n users cannot do harm to the private data of the remaining one. That is why the number of attackers is just less than n. In conclusion, all we have done is to prevent any user, even numerous users attacking together, from acquiring entire model coefficients. In other words, we protect the local model of each user.



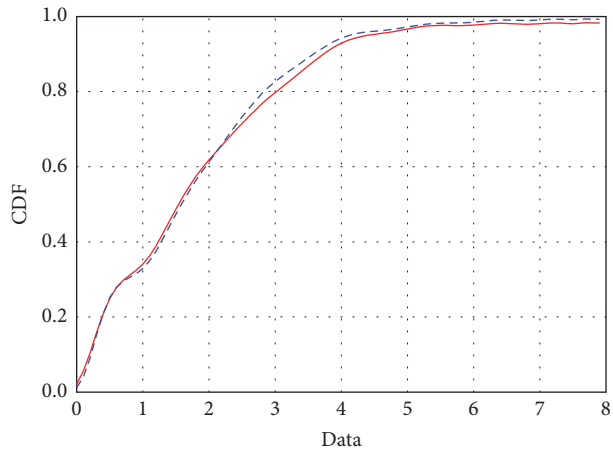
(a) The PDFs comparison on random data set



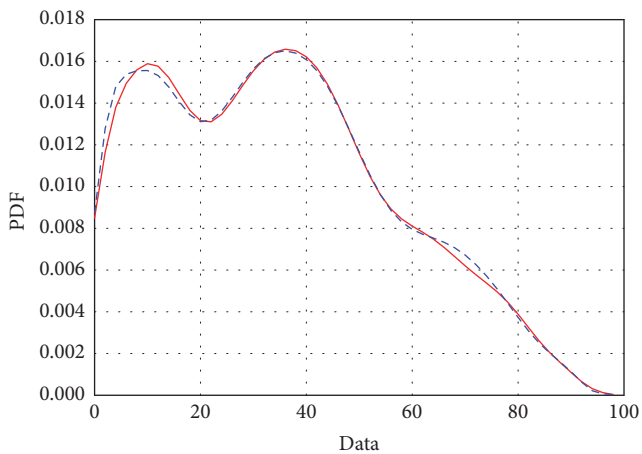
(b) The CDFs comparison on random data set



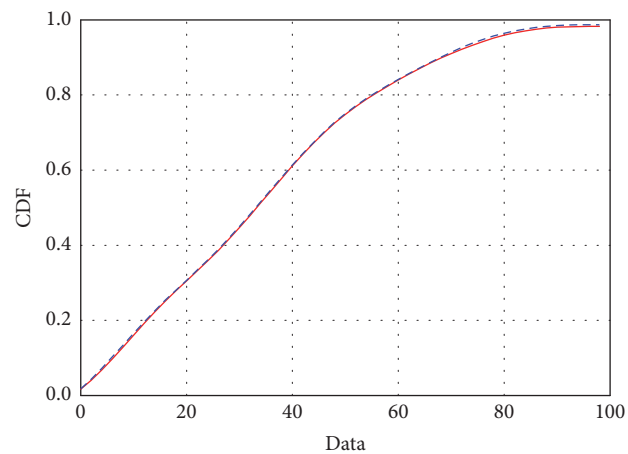
(c) The PDFs comparison on individual household electric power consumption data set



(d) The CDFs comparison on individual household electric power consumption data set

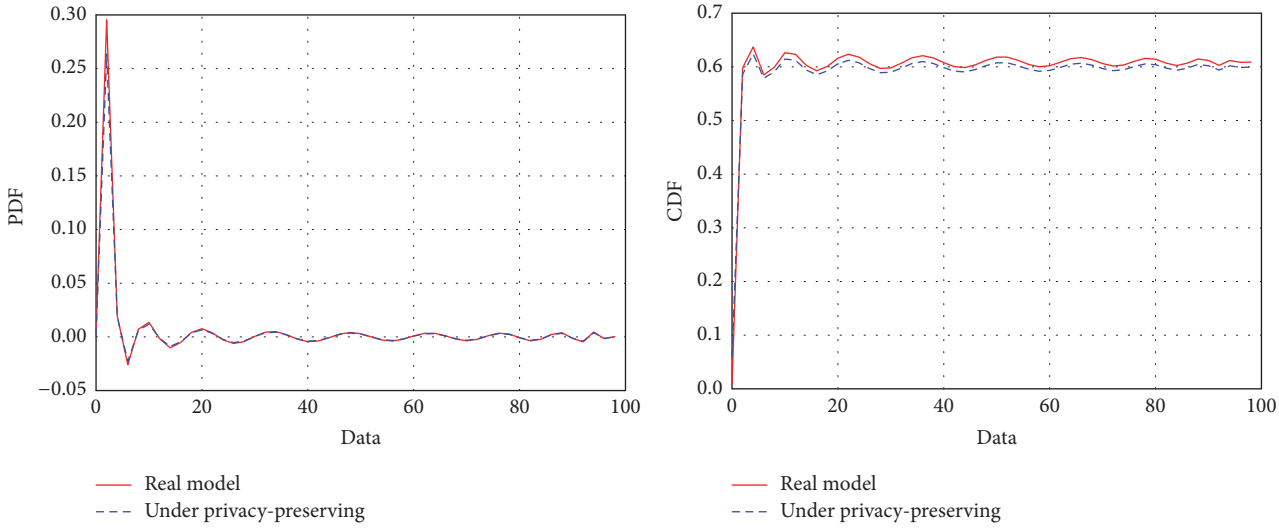


(e) The PDFs comparison on census-income (KDD) data set



(f) The CDFs comparison on census-income (KDD) data set

Figure 3: Continued.



(g) The PDFs comparison on cuff-less blood pressure estimation data set

(h) The CDFs comparison on cuff-less blood pressure estimation data set

FIGURE 3: The PDFs and CDFs comparisons of four experiments on different data sets between real model and privacy-preserving constructed model.

5.3. Security Analysis for Social Ranking. In Section 4.4, the data analyst obtains the global model. Nevertheless, all local models are protected because what the data analyst receives from each user is more than nothing but a sum of pieces that go through a series of variants. For this reason the data analyst is absolutely ignorant of any user's data and local data model.

Briefly, we have achieved all the proposed goals and our method is capable of obtaining a user's social ranking without leaking any private data and user local model.

6. Experiment Study

For appraising our methods proposed in this paper, we designed several experiments. In our experiments, there are four data sets. The first one consists of 100000 randomly generated floating numbers among $(0, 100)$. The remaining three data sets are all from real life and private, which can be found in the UCI Machine Learning Repository. The second one is the age data extracted from census-income (KDD) data set. The third data set is a part of photoplethysmograph (PPG) signal data of cuff-less blood pressure estimation data set and its number is 132000. In the last experiment, we utilize the household global minute-averaged active power data from individual household electric power consumption data set to build the whole community model.

In this paper the problem that needs solving is to get a ranking for a user among his social friends. For example, a user is willing to know the ranking of his salary of a year assumed as s. To achieve his goal, we firstly model the salary data of his social friends and him in our proposed privacy-preserving schemes. Then, we get the PDF and CDF of this constructed model. The value of PDF of s can be used as his ranking.

All the results of our experiments which verify the validity of our proposed schemes are illustrated in Figure 3. For the randomly generated data, Figures 3(a) and 3(b) show, respectively, the PDFs and the CDFs comparisons. Under our privacy protecting scheme, the PDF of our constructed model basically fits that of real data model and the CDF of our constructed model perfectly fits that of real data model. In order to be more persuasive, we implement the rest of the experiments with real life data sets. Figures 3(c) and 3(d) are the results on the individual household electric power consumption data set and the description of the PDF and CDF of household global minute-averaged active power, respectively, which substantiates that our scheme is able to acquire a high accuracy in the real life. The whole data set is separated into two parts in a ratio of $7:3$. The larger part is used as training data and the other one as test data. Figures 3(e)–3(h) are more evidences of the accuracy of our methods. Therefore, we can affirm that our presented solution has perfect performance not only in randomly generated data but also in real life data and is capable of ranking users among their social friends without revealing their private data.

7. Conclusion

We provide a method for those who are willing to get the ranking of the group of their friends and not leak out their own private data in this paper. We construct a polynomial approximation function model for each user and then fragment and blend the function model to protect the user privacy data. Eventually, we establish an overall model for all users, which can help users get their ranking. Our experiments, based on both randomly generated data and real life data set, strongly support our proposed schemes. In the future work, we will apply our privacy-preserving scheme to recommendation system.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (no. 61332004).

References

[1] L. Tan, H. Fan, W. Rui et al., "Mining myself in the community: privacy preserved crowd sensing and computing," in *Proceedings of the International Conference on Wireless Algorithms, Systems, and Applications*, pp. 272–282, Springer, 2016.

[2] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS 2015*, pp. 1310–1321, October 2015.

[3] M. Backes, D. Fiore, and R. M. Reischuk, "Verifiable delegation of computation on outsourced data," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp. 863–874, ACM, 2013.

[4] A. Rial and G. Danezis, "Privacy-preserving smart metering," in *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society (WPES '11)*, Y. Chen and J. Vaidya, Eds., pp. 49–60, Chicago, Ill, USA, October 2011.

[5] J. H. Ahn, S. Hohenberger, D. Boneh, J. Camenisch, A. Shelat, and B. Waters, "Computing on authenticated data," *Journal of Cryptology. The Journal of the International Association for Cryptologic Research*, vol. 28, no. 2, pp. 351–395, 2015.

[6] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: privacy via distributed noise generation," in *Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503, Springer, 2006.

[7] H. Li, X. Lin, H. Yang, X. Liang, R. Lu, and X. Shen, "EPPDR: an efficient privacy-preserving demand response scheme with adaptive key evolution in smart grid," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 8, pp. 2053–2064, 2014.

[8] G. Acs and C. Castelluccia, "A case study: privacy preserving release of spatio-temporal density in Paris," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2014*, pp. 1679–1688, ACM, August 2014.

[9] G. Mezzour, A. Perrig, V. Gligor, and P. Papadimitratos, "Privacy-preserving relationship path discovery in social networks," in *Proceedings of the International Conference on Cryptology and Network Security*, pp. 189–208, Springer, 2009.

[10] M. J. Freedman and A. Nicolosi, "Efficient private techniques for verifying social proximity," *IPTPS*, vol. 5, p. 1, 2007.

[11] S. Garriss, M. Kaminsky, M. J. Freedman, B. Karp, D. Mazières, and H. Yu, "Re: reliable email," *NSDI*, vol. 6, p. 22, 2006.

[12] B. Carminati, E. Ferrari, and A. Perego, "Private relationships in social networks," in *Proceedings of the Workshops in Conjunction with the 23rd International Conference on Data Engineering - ICDE 2007*, pp. 163–171, April 2007.

[13] J. Domingo-Ferrer, "A public-key protocol for social networks with private relationships," in *Proceedings of the International Conference on Modeling Decisions for Artificial Intelligence*, pp. 373–379, Springer, 2007.

[14] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *ACM Sigmod Record*, vol. 29, no. 2, pp. 439–450, 2000.