IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON SECURITY AND PRIVACY IN EMERGING DECENTRALIZED COMMUNICATION ENVIRONMENTS

Received November 5, 2019, accepted November 16, 2019, date of publication November 26, 2019,
date of current version December 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2956019

User Behavior Clustering Scheme With Automatic Tagging Over Encrypted Data

MINGHUI GAO[1], BO LI[1], CHEN WANG[2], LI MA[1], AND JIAN XU[2]
[1]China NARI Group Corporation (State Grid Electronic Power Research Institute), Nanjing 210003, China
[2]Software College, Northeastern University, Shenyang 110004, China

Corresponding author: Chen Wang (w@domoe.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant N171704005, and in part by the Shenyang Science and Technology Plan Projects under Grant 18-013-0-01.

ABSTRACT User behavior clustering analysis has a wide range of applications in business intelligence, information retrieval, and image pattern recognition and fault diagnosis. Most of existing methods of user behavior have some problems such as weak generality and the lack of tags of clustering. With the increasing awareness of privacy protection, user behavior analysis also needs to support for ciphertext to protect user data. Based on clustering algorithm, homomorphic encryption technology and information security, in this paper, we propose a user behavior clustering scheme that supports automatic tags on ciphertext. Firstly, design a security protocol corresponding to the basic operations such as addition, multiplication and comparison and apply to the scheme. Then, the relevant features of the user behavior are merged with the clustering process, the latent factor model, and matrix decomposition. We have implemented our method and evaluated its performance using K-means and K-means++ clustering. The results show that the scheme can auto tags over encrypted data, and the tag also meets the actual situation, which proves the validity and generality of the scheme.

INDEX TERMS User behavior clustering, encrypted data, clustering with tagging.

I. INTRODUCTION

With the increasing maturity of mobile Internet technology, people use various mobile devices and wireless communication networks to browse the web, read news and carry out social activities at any time and any place, and information exchange is more and more convenient. Massive data is constantly generated in various fields, which makes the Internet data and resources show massive characteristics. How to get useful information and knowledge from redundant data to help us make more objective and effective decisions has become an important problem. User behavior analysis can solve the above problems, which refers to the statistics and analysis of user interest. Clustering algorithm is a common means to achieve it, which is widely used in data statistical analysis fields such as business intelligence, information retrieval, image pattern recognition and fault diagnosis [1]. At present, most clustering algorithms still exit two problems: the number of clustering and the tags after clustering is unknown. Without iterating through the data in each group, the category represented by the group cannot be known. After clustering user behavior data, there is no suitable method to mark each group directly. For example, shopping websites usually record members to buy the product information or comments, and also has a product category. They wants to get each group of tags after clustering to combine them, so they can obtain information about what kind of products the users in the group like. Company can offer different marketing plans to different groups.

While user behavior clustering is widely applied, it also causes serious privacy disclosure, which will bring harm to the data owner [2], [3]. For example, when using clustering for stock analysis, if the behavior information of individual stock is leaked in the process of clustering, it will bring chaos to the stock market. Criminals steal user behavior data, which often reflects the user's interests and hobbies, criminals for this fraud. By analyzing the information and privacy protection is often considered as contradictory, actually user behavior clustering and privacy protection can coexist, it can construct privacy protection user behavior clustering scheme, which combined with clustering and differential privacy, security multi-party computation and

The associate editor coordinating the review of this manuscript and approving it for publication was Zheli Liu.

homomorphic encryption. Differential privacy can only process plaintext data disturbed by noise, so it has semantic security [4]. The security multi-party computation requires all participants to join the process, the data of each party will not be leaked to other parties and only be known by itself [5]. However, the intermediate computing tasks are based over non-encrypted data, and the data is also unencrypted during transmission, which is easy to leak information. Generally, it is much faster than homomorphic encryption, but from a customer economics perspective, secure multi-party computing requires to computing so lots of online data to generate bandwidth, and homomorphic encryption is more convenient and economical. However, homomorphic encryption only supports data of integer type and total homomorphic encryption does not support comparison and maximum operation, and there are still some shortages in practical application [6].

A. RELATED WORKS

In recent years, the user behavior analysis method based on clustering algorithm has been widely studied, and become a common technical means in the field of statistical analysis such as business intelligence, information retrieval, image pattern recognition and fault diagnosis. The K-means algorithm is representative of the clustering algorithm, and most user behavior clustering analysis uses the K-means algorithm. Xue and Luan [7] analyzed microblog online behavior data to grasp the user's habits and potential relationships, and use the improved K-means algorithm to cluster behaviors. Considering the user elements, compared with the distance between the two records. It makes more sense to group two different records of the same user at the edge of two neighborhood groups into one group. Phan *et al.* [8], the hierarchical clustering algorithm is combined with the famous independent waterfall model *IC* to analyze the user's sports behavior characteristics and construct a physical exercise communication model *TaCPP* to obtain the relationship between users and physical exercise. In [9]–[11], the user's browsing behavior is analyzed to predict the webpage access. Wang *et al.* [9] mainly uses the clickstream data to capture the user's behavior, and uses the similarity between the clickstreams to construct the similarity graph. The hierarchical clustering algorithm clusters the users to predict the user's future behavior; Kumar *et al.* [10] uses the improved Levenshtein distance to measure the similarity, and uses the hierarchical clustering algorithm and the Markov model to analyze the user's usage behavior to predict webpages; Cavusoglu and Zengin [11] solves the above two problems by using the K-means++ algorithm for the traditional K-means algorithm, which has high dependence on the initial clustering center and needs to input the number of clusters in advance. Zhao *et al.* [12] and Hui *et al.* [13] applied user behavior analysis to the system anomaly detection. K-means is not conducive to data analysis algorithm because of take a long time, the K-means++ algorithm is used to cluster the electricity data and detect abnormal user behavior. Because of the K-means algorithm belongs to unsupervised learning, it is not possible to automatically generate cluster tags. It is necessary to traverse the data of each group to know the group representation. Hu and Ogihara [14] proposed a framework to identify the social tags of songs, first cleaned and filtered the noise; then applied the improved hierarchical clustering algorithm to group the tags to construct the tag categories; finally, according to the categories, the lyrics were clustered and used. The centroid of the corresponding cluster represents the lyrics, and the possibility of assigning lyrics to a specific category is predicted based on the naive Bayesian method. The framework uses cluster center to represent cluster, so there can only be one cluster tag. In practice, there are often many tags, which cannot guarantee the accuracy of the tags. Haiyan *et al.* [15] mainly studies the automatic generation of Weibo user tags based on cluster analysis. According to the analysis of content, the keywords or phrases are extracted as tags, and the cluster tag problem cannot be generated after clustering. Yang and Wang [16] combined with latent semantic analysis (LSA), using the minimum-maximum similarity (MMS) to establish the initial clustering center to improve the selection of initial clustering center, and combined the three to propose label clustering. It can apply social tags to personalized searches.

The existing privacy protection technology cannot be directly applied to the user behavior clustering. They exit shortages in practical applications as following: 1) the current technology cannot guarantee the semantic security of the data, the clustering result is inaccurate; 2) the data is transmitted in plaintext. There is no guarantee that the data will not be stolen during the transmission process, resulting in a privacy leak; 3) Full homomorphic encryption can satisfy arbitrary operations but is inefficient, and does not support comparison and seeking the most value. In order to solve the above problems, a large number of scholars have carried out research work. In [17], [18] proposed the distributed privacy protection K-means algorithm, and Baby and Chandra [18] uses the code-based threshold encryption sharing scheme as a privacy protection mechanism, which is processed separately on different servers and iterated fewer times compared with existing protocols. In [19]–[22] mainly studies selective clustering that supports privacy protection, encrypts sensitive data of users to prevent privacy leakage from external analysts and cluster service providers, and fully supports the selection of online user behavior analysis. Class features while ensuring differential privacy. Su *et al.* [23] also applies differential privacy to the privacy-protected K-means algorithm. However, too many cluster iterations mean less privacy budget for each iteration. Encrypted data calculation is a major difficulty. The homomorphic encryption scheme can support a series of arithmetic operations applied to ciphertext data [24]–[28]. In [24]–[26] proposed a security protocol to support comparison operations. The Paillier cryptosystem is used to encrypt the plaintext data, and then the plaintext operation is replaced with the ciphertext security protocol, but the computational cost is too large. Jaschke and Armknecht [27] solves the division problem in ciphertext operation, and does not allow

direct division of two ciphertext, but can divide a ciphertext data by a constant. This constant represents the sum of data, and even if exposed, it will not reveal the key information. Cheon *et al.* [28] replaces the non-polynomial kernel with a polynomial kernel so that it can be efficiently computed under homomorphic encryption. In [29], [30] considers that existing methods require the participation of all data owners, and the data involved is too large. Therefore, a secure third party is introduced, and the calculation is given to a third party, which saves computational cost and ensures mutual privacy.

B. CONTRIBUTIONS

The main purpose of this paper is to propose a user behavior clustering scheme with automatic tagging over encrypted data, which using homomorphic encryption technology and combined two parties. User's behavior data can be applied to different situations and clustering algorithm can be switched with other clustering algorithm according to the actual circumstance or data types.

In order to ensure the privacy of user data and tag is not leaked, the clustering process is studied in detail, the basic operations such as addition, multiplication and comparison are proposed, and the security protocols corresponding to the plaintext operation is designed, so that the operation result of ciphertext is consistent with the same plaintext operation after decryption.

Our major contributions are presented as follows:

1) We present a scheme that allows privacy-preserving clustering with automatic tagging over encrypted data. Clusters can be labeled at the same time without checking in each group.

2) Tags represent as a link between users and information resources, and privacy issues loom large. This experiment encrypts the tag generation process of cluster class to prevent privacy disclosure of tag information.

3) The k-means algorithm will reveal privacy while calculating the distance between the sample point and the center point, so it can hide the cluster center to prevent the attacker from inferring the cluster grouping to which the user belongs. In this experiment, homomorphic encryption is used to solve the encrypted distance calculation problem and encrypted comparison problem, and these security protocols are applied to the k-means algorithm framework to realize privacy protection.

II. PRELIMINARIES

A. CLUSTERING ALGORITHM

The process of dividing a physical or abstract set into multiple classes consisting of objects that are similar to each other is called clustering. Clustering is an important mining method. Unlike the classification algorithm, the sample objects are not marked and need to be automatically determined by the clustering algorithm. It belongs to unsupervised learning, which people will not provide any before classification. The K-means algorithm is a typical distance-based clustering algorithm. The distance is used as the similarity evaluation index, that is, the closer the distance between two objects is, the greater the similarity is. Euclidean or cosine angles are often used when calculating distances. K represents the number of clusters of target clusters and K-means is an algorithm that clusters data points by mean.

The K-means algorithm is divided into two steps: cluster and moving cluster center.

1) Select k objects randomly, each object represents the average of the cluster. For each of the remaining objects, it is assigned to the nearest cluster based on its distance from each cluster center.

2) Recalculate the average of each cluster. This process is repeated until the criterion function E converges, the cluster center no longer undergoes significant changes. The error squared criterion function E is usually used as a performance metric, which represents the sum of the distances of all sample points to the mean vector of the respective cluster. The smaller the E value, the higher the similarity of the sample values within the cluster. The minimization criterion function E is an NP problem, and the clustering algorithm can be regarded as a coordinate ascending algorithm, that is, by fixing one variable, adjusting another variable, and continuously adjusting through an iterative process, and finally obtaining a local optimal solution.

The K-means algorithm has the advantages of simplicity, easy understanding and implementation, and low time complexity, but it still has the following four shortcomings.

1) Sensitive to the initial cluster center and the number of cluster and k needs to be given in advance. The K-means++ algorithm and the binary K-means clustering algorithm can make up for the above shortcomings.

2) It is sensitive to noise and outliers. The K-Mediods algorithm uses a median representation of each cluster to avoid sensitivity to outlier data.

3) Belongs to hard clustering, that is, each sample belongs to only one category, and Gaussian hybrid clustering allows soft clustering.

4) Only spherical clusters can be found, and non-convex shapes cannot be found. The spectral clustering algorithm can find clusters of arbitrary shapes.

B. ENCRYPTION METHOD

Homomorphic encryption allows any data to remain encrypted during processing and operation, enabling third parties to apply functionality on encrypted data without revealing the value of the data. A homomorphic cryptosystem, like other forms of public encryption, uses a public key to encrypt data and only allows individuals with matching private keys to access their unencrypted data, although there are also examples of symmetric key homomorphic encryption. However, it differs from other forms of encryption in that it uses an algebraic system to allow others to perform various calculations or operations over encrypted data.

The homomorphic encryption is divided into an additive homomorphism and a multiplicative homomorphism algorithm, and the descriptions are as follows:

Let R and S are integers, where R is the plaintext space and S is the encrypted space. $a, b \in R$, E is the encryption function on $R \rightarrow S$, if there are algorithms ADD and$MULT$:

If it satisfies $E(a+b) = ADD(E(a), E(b))$, it is called an additive homomorphic algorithm;

If it satisfies $E(a \times b) = MULT(E(a), E(b))$, it is called a multiplicative homomorphic algorithm.

If an encryption function only satisfies the addition and homomorphism, only the addition and subtraction operations over the encrypted can be performed; f an encryption function only satisfies the multiplicative homomorphism, only the multiplication and division operations on the encrypted can be performed; When the state is the same as the multiplication, the encryption function is called a fully homomorphic algorithm and has full homomorphism. This paper uses the Pailler encryption method to encrypt the data to satisfy the additive homomorphism. The multiplication operation on the encrypted is given in next section.

III. DESIGN

A. SCHEME DESIGN

The scheme adopts a model to analyze the user's behavior data, cluster users with similar behaviors into the same cluster, and automatically assign appropriate tags to each cluster, without checking the inside of the cluster to obtain the tags. Because the tag plays a role in contacting users and behavioral information, it can directly reflect the user's preferences to a certain extent. To prevent privacy leakage, the user behavior data is encrypted. The whole scheme under encrypted, which combined with the two parties to ensure safety. Furthermore, the user behavior data may have a problem of missing values. The user has not done this behavior, and does not mean that he does not like to do it. The behaviors that have not been done here are regarded as missing values in the original data. The potential factor model is established by the *NMF* matrix decomposition to deal with the missing value problem. In addition, if the initial data is directly clustered directly, the dimension is too high, which will increase the difficulty of clustering and the effect of clustering. The experiment transforms the user behavior data into a matrix form, combines the behavior with the label data, and uses the principle of matrix multiplication to obtain each user cluster and corresponding label. The schematic diagram of the model work is shown in Fig.1.

B. SECURITY PROTOCOL

Analyze the operation of clustering model and get the basic operations included in the process such as addition, dot product, multiplication, and comparison. For the above basic operations, a corresponding secure communication protocol is designed. This section gives the construction method of the communication protocol. The protocol consists of two parties, denoted as A and B. The encryption scheme involved is the Paillier encryption scheme and the QR encryption scheme, both of which satisfy the addition and homomorphism operation.

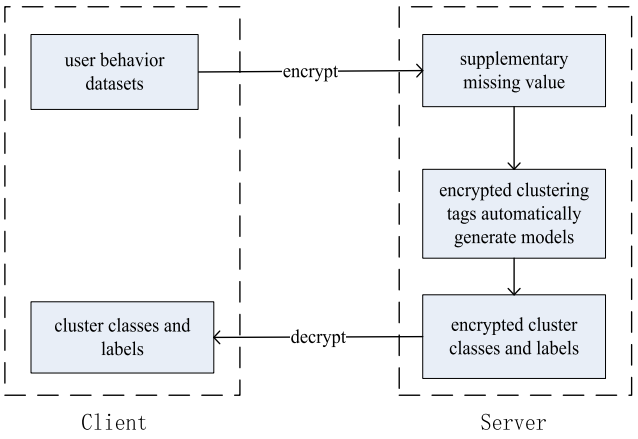


FIGURE 1. Schematic diagram of the model.

Protocol 1 Secure Dot Product Protocol

Input A: $x = (x_1, \ldots, x_d)$, public keys pk_p
Input B: $y = (y_1, \ldots, y_d)$, secret keys sk_p
Output: $E_{pk}(\langle x, y \rangle)$
1. B: encrypt $y = (y_1, \ldots, y_d)$, send $E_{pk}(y_i)$to A
2. A: compute $E_{pk}(v) = \prod_i E_{pk}(y_i)^{x_i} \bmod N^2$
3. A: output $E_{pk}(v)$

1) SECURE DOT PRODUCT PROTOCOL

Since the entire scheme is performed under encrypted, a secure dot product protocol is used to solve the encryption matrix multiplication, as shown in *Protocol 1*. It is attended by both A and B. A represents the client, enters the test sample and records it as x; B represents the server, enters the training sample, and records it as y.

2) SECURE MULTIPLICATION PROTOCOL

The secure multiplication protocol mainly realizes the multiplication through attributes of homomorphic encryption, so that their results can be obtained from two encrypted data. Specifically, A has two encrypted data such as $E_{pk}(x)$ and $E_{pk}(y)$. The goal is to get $E_{pk}(xy)$ through interaction with B and ensure the privacy of x and y. B has the private key encrypted by Paillier and the public key is public. The basic idea of secure multiplication protocol is based on this formula:

$$x^*y = (x + r_x)^*(y + r_y) - x^*r_x - y^*r_y, \quad r_x, r_y \in Z_n \tag{1}$$

The proof of the correctness of the agreement is given below:

The purpose of the secure multiplication protocol is to obtain $E_{pk}(xy)$, and the value of x^*y can be derived from the following:

$$x^*y = (x + r_x)^*(y + r_y) - x^*r_x - y^*r_y \tag{2}$$

According to the nature of Paillier homomorphic encryption, the value of $E_{pk}(xy)$ can be derived from the

Protocol 2 Secure Multiplication Protocol

Input A: $E_{pk}(x)$ and $E_{pk}(y)$, public keys pk_p
Input B: Secret keys sk_p
Output: $E_{pk}(xy)$
4. A: randomly select two numbers $r_x, r_y \in Z_n$
5. A: compute $x' \leftarrow E_{pk}(x)E_{pk}(r_x)$
6. A: compute $y' \leftarrow E_{pk}(y)E_{pk}(r_y)$
7. A: send x' and y' to B
8. B: decrypt x' and y', compute $h_x \leftarrow D_{pk}(x'), h_y \leftarrow D_{pk}(y')$, $h \leftarrow h_x h_y \bmod N$ and $h' \leftarrow E_{pk}(h)$
9. B: send h' to A
10. A: compute $s \leftarrow h'E_{pk}(x)^{N-r}$, $s' \leftarrow sE_{pk}(x)^{N-r}$ and $E_{pk}(xy) \leftarrow s'E_{pk}(r_x r_y)^{N-1}$
11. A: output $E_{pk}(xy)$

following formula:

$$E_{pk}(xy) = E_{pk}\left((x+r_x)^*\left(y+r_y\right)\right)^* E_{pk}(x)^{N-r_y} * E_{pk}(y)^{N-r_x}$$

where $(x + r_x)^*(y + r_y)$ is calculated by A after decryption according to $x' \leftarrow E_{pk}(x)E_{pk}(r_x)$ and $y' \leftarrow E_{pk}(y)E_{pk}(r_y)$, and then encrypted again to obtain $E_{pk}\left((x + r_x)^*\left(y + r_y\right)\right)$, and then $E_{pk}\left((x + r_x)^*\left(y + r_y\right)\right)$ is sent to B. B owns data $E_{pk}(x)$, $E_{pk}(y)$, r_x, and r_y. According to the nature of Paillier homomorphic encryption, $E_{pk}(x)^{N-r_y}$ and $E_{pk}(y)^{N-r_x}$ can be obtained, and $E_{pk}(xy)$ is calculated according to the formula.

3) SECURE DISTANCE COMPUTING PROTOCOL

The secure distance protocol implements the Euclidean distance calculation between two encrypted vectors. The basic idea is based on the following equation:

$$\left(|x-y|^2\right) = \sum_{i=1}^{l}(x_i - y_i)^2 \tag{3}$$

Firstly, for all $1 \le i \le l$, A calculates by the properties of $E_{pk}(x_i - y_i) = E_{pk}(x_i)E_{pk}(y_i)^{N-1}$ Paillier homomorphic encryption, and then $E_{pk}\left((x_i - y_i)^2\right)$ is calculated by multiplying security protocol M and B. Finally, A uses the properties of homomorphic encryption to sum $E_{pk}\left((x_i - y_i)^2\right)$ as:

$$E_{pk}\left(|x-y|^2\right) = \prod_{i=1}^{l} E_{pk}\left((x_i - y_i)^2\right) \tag{4}$$

Specific security protocols are shown in Protocol 3.

The proof of the correctness of the agreement is given below:

The purpose of the safety distance calculation protocol is to calculate the Euclidean distance of the encrypted form of two encrypted vectors. The Euclidean distance of the plaintext is calculated as $\left(|x-y|^2\right) = \sum_{i=1}^{l}(x_i - y_i)^2$, so $E_{pk}\left(|x-y|^2\right) = \prod_{i=1}^{l} E_{pk}\left((x_i - y_i)^2\right)$ can be obtained according to the nature of Paillier homomorphic encryption,

Protocol 3 Safety Distance Calculation Protocol

Input A: $E_{pk}(x)$ and $E_{pk}(y)$, the bit length l of x and y, public keys pk_p
Input B: Secret keys sk_p, the bit length l
Output: $E_{pk}\left(|x-y|^2\right)$
1. A: for i=1 to l do
2. compute $E_{pk}(x_i - y_i) \leftarrow E_{pk}(x_i)E_{pk}(y_i)^{N-1}$
3. A and B: for i=1 to l do
4. compute $E_{pk}\left((x_i - y_i)^2\right) \leftarrow$ M $\left(E_{pk}(x_i - y_i), E_{pk}(x_i - y_i)\right)$
5. A: compute $E_{pk}\left(|x-y|^2\right) \leftarrow \prod_{i=1}^{l} E_{pk}\left((x_i - y_i)^2\right)$
6. A: output $E_{pk}\left(|x-y|^2\right)$

Protocol 4 Safety Comparison Protocol

Input A: $E_{pk}(x)$, $E_{pk}(y)$, the bit length l of x and y, public keys pk_p
Input B: Secret keys sk_p, the bit length l
Output A: 1 or 0
7. A: compute $x' \leftarrow E_{pk}(y)^* 2^{l} * E_{pk}(x)^{-1} \bmod N^2$
8. A: randomly select a number r from $\left(0, 2^{l+1}\right) \cap \mathbb{Z}$
9. A: adding noisy r to the encrypted data x' makes it impossible for the B party to know the real data x': $z \leftarrow x'E_{pk}(r) \bmod N^2$
10. A: send z to B
11. B: compute decrypted data z': $z' \leftarrow D_{pk}(z)$
12. A: compute $c \leftarrow r \bmod 2^l$
13. B: compute $d \leftarrow r \bmod 2^l$
14. transfer DGK comparison protocol, A, B as input, c, d as input data, B gets comparison result $E_{pk}(t')$, where t=(d<c)
15. A: sent $E_{pk}(r_{l+1})$ to B, where r_{l+1} is the $l+1$th bit of r
16. B: encrypt the $l+1$ th bit of z to get $E_{pk}(z_{l+1})$
17. B: compute $t' \leftarrow E_{pk}(t')E_{pk}(r_{l+1})E_{pk}(z_{l+1})$
18. B: send t' to A
19. A: compute decrypted data t, $t \leftarrow D_{pk}(t')$
20. output t

so as long as $E_{pk}\left(|x-y|^2\right)$ are obtained. The value is fine. It can be known from the secure multiplication protocol that $E_{pk}(x_i - y_i)^2$ is introduced through $E_{pk}(x_i - y_i)$, and A knows all the encrypted information x_i and y_i. As long as the Paillier homomorphic cryptographic property and the calculation formula $E_{pk}(x_i - y_i) = E_{pk}(x_i)E_{pk}(y_i)^{N-1}$ are used again, $E_{pk}(x_i - y_i)$ can be obtained, so that A can obtain encrypted of two encrypted vectors. The European distance of the form.

4) SECURITY COMPARISON PROTOCOL

The main idea of the comparison protocol is to compute encrypted data $2^l + y - x$, and then refer to bit $l+1$ which corresponds to bit 2^l. If the result is 1, then $y \ge x$; otherwise $y < x$. This paper assumes that the encryption scheme is additive homomorphism, N denotes encrypted modulus.

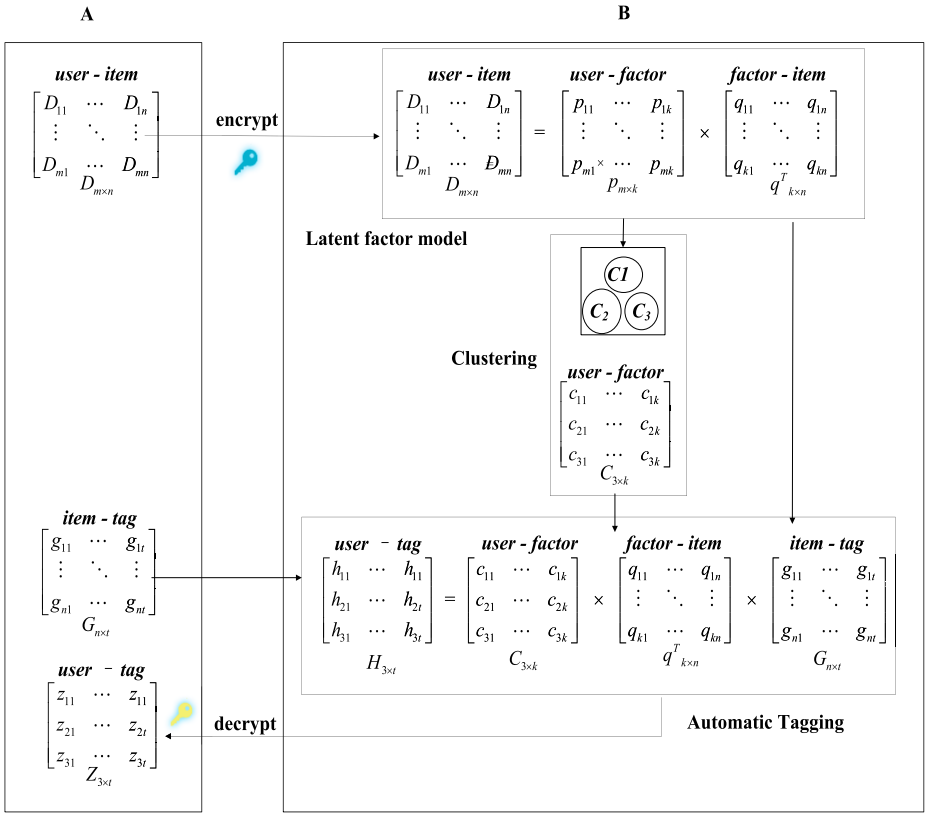


FIGURE 2. Encrypted clustering tag automatic generation process.

C. AUTOMATIC TAG GENERATION PROCESS

This section uses the communication protocol in last Section to construct an encrypted clustering label auto-generation model. The construction process is shown in Fig.2.

Step 1: A adopts Paillier algorithm to encrypt data, and simultaneously constructs *user-item* matrix D and *item-tag* matrix G, where each row of matrix D represents user *user* and each column represents user behavior *item*. As shown in the figure, there are m users and n behaviors. A sends the public key pk_p together with the encrypted matrix to party B, and the private key is kept in its own hand for decryption to prevent data leakage.

Step 2: B uses matrix decomposition technology to establish a potential factor model and decompose D into matrix p, q^T. However, due to the high dimension of this matrix, the clustering effect is not ideal at this time. Therefore, this paper regards p as *user-latent factor* matrix, conducts clustering on p, calculates the cluster class center C_i, $i = 1, \ldots, k$ of each group after clustering, and obtains*user-factor* matrix C.

Step 3: The *user-factor* matrix C is obtained after clustering the matrix p, first find out all the data of individual corresponding labels, because C regards user as latent factor, on the relationship between the label will be *user-factor* matrix C, *factor - item* matrix q^T and *item-tag* matrix G multiplication matrix H, each row represents a group, each column represents the label, so H can be regarded as the user's relationship to the tag, use this matrix can be statistical group within the tag number, prioritize after former n, you can get the most representative tag of each group in the encrypted state.

Step 4: B sends the encrypted *user-tag* matrix H to A, and A decrypts it using the corresponding private key sk_p to obtain the cluster class label under plaintext.

1) LATENT FACTOR MODEL

Latent factor model (LFM) is a technology to find out potential factors. It is usually applied in recommendation system to find out potential factors and their relationship between users and commodities. Clustering can be done automatically based on user's behavior, and namely the granularity of clustering is completely controllable. The main idea of this technique is to obtain the category matrix of users and goods respectively by assuming an implicit factor space, and then multiply the two matrices to get the final result. Whether a certain commodity belongs to a class or not is completely determined by the user's behavior. If two commodities are liked by many users at the same time, there is a great probability that these two commodities belong to the same class.

Since the data set D of users and behaviors that are generally collected is not very completely, behaviors include the purchased goods, rated items and listening to music and other data. Items that the user rated highly or listened to more often could be assumed to be liked by the user, but items that the user hadn't purchased, rated, or listened didn't mean the user

Protocol 5 Clustering Protocol

Input C: sample $y = (y_1, \ldots, y_n)$, secret keys sk_p
Input S: public keys pk_p
Output C: encrypted clustering result c=*(c1,. . .,ci)*
1. C: encrypt $E_{pk}(y_n)$ and send it to S
2. S: for $1 \leq i \leq n$:
3. randomly selected the initial clustering center x
4. compute $E_{pk}\left(|x - y_i|^2\right) = \prod_{i=1}^{n} E_{pk}\left((x - y_i)^2\right)$
5. the results were stored in the array *dis_p*
6. find the minimum *Min* in the array *dis_p*
7. find the maximum *Max* in the array *dis_p*
8. check this point which the sample farthest from its clustering center is the new clustering center
9. end
10. S: Returns $E_{pk}(c_i)$ to C
11. C: Decrypt $E_{pk}(c_i)$ and get result $c = (c1 \ldots ci)$

didn't like them. Considering that the obtained data is incomplete, matrix decomposition is required to reconstruct the potential factor matrix model between users and behaviors, and the calculation is shown in equation.

$$D \approx p \times q^T \tag{5}$$

In the part of data processing, the relationship matrix p between users and potential factors and the relationship matrix A between potential factors and behavior are obtained. According to formula as follow, the relationship matrix H between potential factors and labels is calculated.

$$H = q^T \times G \tag{6}$$

The next step is clustering and tagging.

2) CLUSTERING

The clustering process is jointly completed by server S and client C. The data processed is encrypted by Paillier. In essence, the basic calculation of plaintext data in the clustering algorithm is replaced by the communication protocols, so that the clustering can be operated on encrypted data and finally the clustering result can be obtained.

To prevent privacy disclosure while calculating the distance between the sample point and the center point, the cluster center is hidden to prevent the attacker from inferring the group of class clusters to which the user belongs, so as to ensure the security and homomorphism of the data in the clustering process.

Protocol 5 is a description of clustering protocols.

3) AUTOMATICALLY TAGS

P is regarded as the relationship matrix between users and potential factors, and the clustering of users is converted to the clustering of p. After grouping p, there are two ways to automatically mark the group. One is to take the average value of the data in the group to get the group center of each group, as shown in equations (4-3) and (4-4).

$$S_i = \{x \in clusteri | p_x\} \tag{7}$$

$$C_i = means(s_i) \tag{8}$$

A large value in each column of C_i indicates that the group has strong characteristics in this dimension, that is, the corresponding potential factor is easy to be observed. Therefore, as long as the group is multiplied by the relationship matrix between the corresponding potential factor and the tag, the relationship matrix Q between the group and the corresponding label can be obtained. The calculation formula is shown in equation.

$$Q = C \times H = \begin{bmatrix} q_{11} & \cdots & q_{1t} \\ \vdots & \ddots & \vdots \\ q_{k1} & \cdots & q_{kt} \end{bmatrix} \tag{9}$$

Each column in Q represents the label of a group. For example, to give n tags to the first group, you can take the largest n from $t_{11} \ldots t_{1t}$, as the tag of this group, or give a threshold, if it beyond which the tag of this group can be seen.

The second is to find out all the data of individual corresponding tag, because the matrix p represents the relationship between the user and the potential factor. The matrix H represents the relationship between potential factor and the tags. So $p * H$ can be regarded as the relationship between the user and the tag, which statistics with the matrix group within the tag numbers. Find nth as this group of tags.

IV. EXPERIMENT AND ANALYSIS

In order to investigate the user behavior analysis scheme with automatic tagging over encrypted data, we proposed in this paper that the experiment uses the user behavior data set last.fm for testing. It combined with the data of users, singers and singer types, including the number of times each user listens to each singer's music and the singer types that the user has tagged. Different users may tag different types of the same singer, among which there are 1892 users, 17,632 singers and 11946 singer types. The last.fm dataset contains five files, the details of which are shown in Table.1.

If the types marked by the same singer are summed up, the type marked more times can be regarded as the label of the singer's preferred type. The experiment requires two types of data, namely, user-singer relationship data set and singer and type data set. The number of times users have listened to the music is converted into 1∼5 points of liking degree through data standardization, and those who have not listened to the music are not rated as missing values, and the data set user_artists.dat of the relationship between users and singers is used to construct D matrix. At the same time, the data set *user_taggedar. dat* is used to construct G matrix.

In the experiment, k-means and k-means++ two classical clustering algorithms were integrated into the scheme, and the effectiveness and universality of the scheme were verified through Silhouette Coefficient and Label Coefficient.

TABLE 1. Last.fm dataset.

dataset name	meaning
artists.dat	the user tags the singer
tags.dat	tags available in the dataset
user_artists.dat	the number of times the user listens to each singer
user_taggedartists.dat	user tags for each singer
user_taggedartists-timest amps.dat	the user's timestamp for each singer
user_friends.dat	relationship between users in datasets

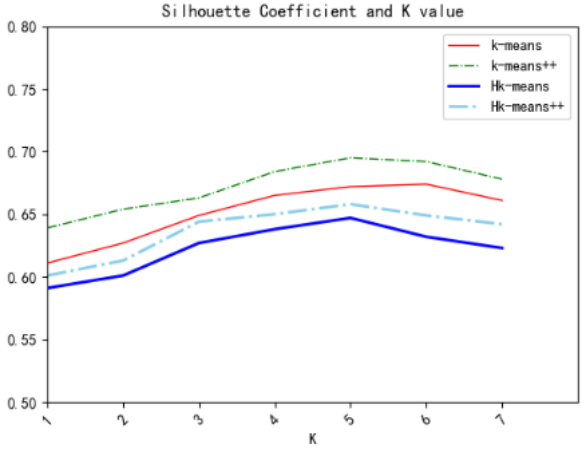


FIGURE 3. Relationship between the Silhouette Coefficient and K value.

A. SILHOUETTE COEFFICIENT

A popular method to measure the performance of clustering algorithms is to calculate the Silhouette Coefficient. The goal of clustering algorithm is to make the clustering results have small inter-class similarity and large intra-class similarity, and the larger the Silhouette Coefficient is, the better the clustering performance will be. The original k-means algorithm, k-means++ algorithm, the k-means algorithm with privacy protection (Hk-means) and the k-means++ algorithm (Hk-means++) are compared. The influence of k value on clustering results was observed through experiments, and the experimental results were shown in FIG.3.

As shown in FIG.3, with the increase of k value, the average Silhouette Coefficient of the four algorithms generally shows an upward trend. When k=5, the contour coefficient value is larger, and the clustering effect is better. K-means++ algorithm is better than k-means algorithm on the whole, while the clustering effect of the Hk-means++ algorithm under encrypted is better than that of the Hk-means algorithm on the whole, and the clustering performance of the Hk-means++ algorithm is closer to that of the non-privacy protected k-means++. It is proved that this scheme is effective in encrypted clustering, and supports that most clustering algorithms are universal.

Cluster Tag Scoring Algorithm

Input: the clustering result C of p_{train}, p_{test}, p_{train} and its label matrix T, and the *latent factory-label* matrix H
Output: *F1score*
1. calculate the label represented by the user in the test set $S \leftarrow p_{test} \times H$
2. p_{test} is allocated to C in p_{train} existing group by using the same clustering algorithm
3. for $i \leftarrow 1$ to N do:
4. for $j \in cluster_i$ do:
5. calculate $TP \leftarrow count\left(TP + S_j \cap T_i\right)$
6. calculate $FP \leftarrow count\left(FP + S_j \backslash T_i\right)$
7. calculate $FN \leftarrow count\left(FN + T_i \backslash S_j\right)$
8. end
9. end
10. calculate *Precision*
11. calculate *Recall*
12. calculate *F1score*

B. LABEL COEFFICIENT

In this scheme, multiple tags are generated for each cluster, so it is necessary to propose a tag evaluation method to evaluate the fitness between tags and clusters.

First, the matrix p is divided into p_{test} and p_{train}. Select 80% as the training data set to do clustering and label, and the other 20% as the test data set. According to the clustering results p_{train}, the same clustering algorithm will be used to assign p_{test} to the existing cluster p_{train}, and the label represented by the user in the test set will be obtained from $p_{test} \times H$. Compare the tags obtained by users with cluster tags, and *F1 score* is calculate as:

$$F1score = \frac{2 \times precision \times recall}{precision + recall} \tag{10}$$

Precision and *Recall* is verified by the following equation:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

Suppose x belongs to a data sample of p_{test}, and x belongs to cluster C. TP represents the number of labels owned by x and C, FP represents the number of labels owned by C without x, and FN represents the number of labels owned by x without c.

The scoring process of cluster labels is as follows:

Variable S represents the tag set p_{test} in the data set. Compare S with the obtained tag set p_{train} to calculate the *F1 score*.

The score of the tag was calculated by F1score, and the closer it was to 1, the better. The original k-means algorithm, k-means++ algorithm, the k-means algorithm with privacy

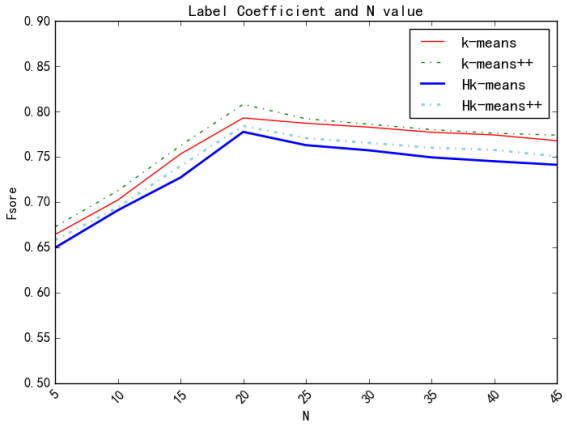


FIGURE 4. Relationship between the Label Coefficient and K value.

protection (Hk-means) and the k-means++ algorithm (Hk-means++) are compared. With different number of tags, the calculated F1score is also different, as shown in FIG.4.

When the number of tags is small, more common tags are usually taken, but it is easy to obscure the meaning of group. On the contrary, the more the number of tags, the higher the ability to distinguish the tags, so that the meaning of cluster more vivid. When the number of tags is 20, the tag score is the largest, and after 20, the tag score starts to decline. At the same time, k-means++ algorithm has the best effect and higher accuracy of automatic tags. Because k-means++ algorithm does not need to input the cluster number k value in advance, it is difficult to know the cluster number in advance in practical application, so the effect of k-means algorithm is not as good as k-means++ algorithm.

At the same time, the effect of plaintext and encrypted is similar, so as to verify that this scheme can realize clustering of encrypted user behavior data and generate cluster labels automatically.

V. CONCLUSION

In this work, we have proposed a user behavior clustering scheme with automatic tagging over encrypted data. In our scheme, nothing of personal and user behavior data is leaked to either of the service providers. Moreover, we construct secure basic protocols with secure multi-party computing technology and homomorphic encryption to achieve a secure user behavior clustering model. Then, combine user behavior data and tags data with user data to realize automatic tagging. Finally, we realized K-means clustering and K-means++ clustering algorithm on the user behavior dataset. Evaluate from Silhouette Coefficient and Label Coefficient evidences that the scheme is feasible and versatile, and achieves privacy protection. It can be applied to user behavior data in different situations, and the clustering algorithm can be replaced with other clustering algorithms.

REFERENCES

[1] C. Zhong, J. Shao, F. Zheng, K. Zhang, H. Lv, and K. Li, "Research on electricity consumption behavior of electric power users based on tag technology and clustering algorithm," in *Proc. 5th Int. Conf. Inf. Sci. Control Eng.*, Zhengzhou, China, Jul. 2018, pp. 459–462.

[2] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, "A comprehensive review on privacy preserving data mining," *SpringerPlus*, vol. 4, Nov. 2015, Art. no. 694.

[3] A. H. Celdrán, G. D. Tormo, F. G. Mármol, M. G. Pérez, and G. M. Pérez, "Resolving privacy-preserving relationships over outsourced encrypted data storages," *Int. J. Inf. Secur.*, vol. 15, no. 2, pp. 195–209, Apr. 2016.

[4] Z. Gheid and Y. Challal, "Efficient and privacy-preserving K-means clustering for big data mining," in *Proc. IEEE/Trustcom/BigDataSE/ISPA*, Tianjin, China, Aug. 2016, pp. 791–798.

[5] J. Yuan and Y. Tian, "Practical privacy-preserving mapreduce based k-means clustering over large-scale dataset," *IEEE Trans. Cloud Comput.*, vol. 7, no. 2, pp. 568–579, Apr./Jun. 2019.

[6] H. Yin, J. Zhang, Y. Xiong, X. Huang, and T. Deng, "PPK-Means: Achieving privacy-preserving clustering over encrypted multi-dimensional cloud data," *Electronics*, vol. 7, no. 11, p. 310, Nov. 2018.

[7] L. Xue and W. Luan, "Improved K-means algorithm in user behavior analysis," in *Proc. 9th Int. Conf. Frontier Comput. Sci. Technol.*, Dalian, China, Aug. 2015, pp. 339–342.

[8] N. Phan, J. Ebrahimi, D. Kil, B. Piniewski, and D. Dou, "Topic-aware physical activity propagation in a health social network," *IEEE Intell. Syst.*, vol. 31, no. 1, pp. 5–14, Jan./Feb. 2016.

[9] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao, "Unsupervised clickstream clustering for user behavior analysis," in *Proc. Conf. Hum. Factors Comput. Syst.*, May 2016, vol. 7, no. 12, pp. 225–236.

[10] B. T. H. Kumar, L. Vibha, and K. R. Venugopal, "Web page access prediction using hierarchical clustering based on modified levenshtein distance and higher order Markov model," in *Proc. TENSYMP*, Bali, Indonesia, May 2016, pp. 1–6.

[11] M. M. Öztürk, U. Cavusoglu, and A. Zengin, "A novel defect prediction method for Web pages using K-means++," *Expert Syst. Appl.*, vol. 42, no. 19, pp. 6496–6506, Nov. 2015.

[12] Z. Zhao, J. Wang, and Y. Liu, "User electricity behavior analysis based on K-means plus clustering algorithm," in *Proc. Int. Conf. Comput. Technol., Electron. Commun.*, Dalian, China, Dec. 2017, pp. 484–487.

[13] S. Hu, Z. Xiao, Q. Rao, and R. Liao, "An anomaly detection model of user behavior based on similarity clustering," in *Proc. 4th Inf. Technol. Mechatronics Eng. Conf.*, Chongqing, China, Dec. 2018, pp. 835–838.

[14] Y. Hu and M. Ogihara, "Identifying accuracy of social tags by using clustering representations of song lyrics," in *Proc. 11th Int. Conf. Mach. Learn. Appl.*, Boca Raton, FL, USA, Dec. 2012, pp. 582–585.

[15] L. Haiyan, C. Xiaowei, and R. Ying, "The research on the automatic generation of micro-blog user tags based on clustering analysis," in *Proc. 5th Int. Conf. Softw. Eng. Service Sci.*, Beijing, China, Jun. 2014, pp. 633–636.

[16] J. Yang and J. Wang, "Tag clustering algorithm LMMSK: Improved K-means algorithm based on latent semantic analysis," *J. Syst. Eng. Electron.*, vol. 28, no. 2, pp. 374–384, Apr. 2017.

[17] N. B. Jinwala and G. B. Jethava, "Privacy preserving using distributed K-means clustering for arbitrarily partitioned data," *Int. J. Eng. Develop. Res.*, vol. 2, no. 2, pp. 2291–2295, Nov. 2014.

[18] V. Baby and N. S. Chandra, "Distributed threshold K-means clustering for privacy preserving data mining," in *Proc. Int. Conf. Adv. Comput., Commun. Informat.*, Jaipur, India, Sep. 2016, pp. 2286–2289.

[19] M. A. Mustafa, N. Zhang, G. Kalogridis, and Z. Fan, "DEP2SA: A decentralized efficient privacy-preserving and selective aggregation scheme in advanced metering infrastructure," *IEEE Access*, vol. 3, pp. 2828–2846, 2015.

[20] J. Qian, F. Qiu, F. Wu, N. Ruan, G. Chen, and S. Tang, "A differentially private selective aggregation scheme for online user behavior analysis," in *Proc. IEEE Global Commun. Conf.*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[21] J. Qian, F. Qiu, F. Wu, N. Ruan, G. Chen, and S. Tang, "Privacy-preserving selective aggregation of online user behavior data," *IEEE Trans. Comput.*, vol. 66, no. 2, pp. 326–338, Feb. 2017.

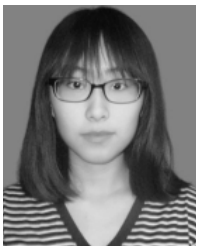
[22] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, New York, NY, USA, 2017, pp. 1175–1191.

[23] D. Su, J. Cao, N. Li, E. Bertino, M. Lyu, and H. Jin, "Differentially private K-means clustering and a hybrid approach to private optimization," in *Proc. 6th ACM Conf. Data Appl. Secur. Privacy*, New Orleans, LA, USA, 2016, pp. 26–37.

[24] N. Almutairi, F. Coenen, and K. Dures, "K-means clustering using homomorphic encryption and an updatable distance matrix: Secure third party data clustering with limited data owner interaction," in *Proc. Int. Conf. Big Data Anal. Knowl. Discovery*, Lyon, France, 2017, pp. 274–285.
[25] H.-J. Kim and J.-W. Chang, "A privacy-preserving k-means clustering algorithm using secure comparison protocol and density-based center point selection," in *Proc. 11th Int. Conf. Cloud Comput.*, San Francisco, CA, USA, Jul. 2018, pp. 928–931.
[26] Y. Chen, J.-F. Martínez-Ortega, P. Castillejo, and L. López, "A homomorphic-based multiple data aggregation scheme for smart grid," *IEEE Sensors J.*, vol. 19, no. 10, pp. 3921–3929, May 2019.
[27] A. Jäschke and F. Armknecht, "Unsupervised machine learning on encrypted data," in *Proc. Int. Conf. Sel. Areas Cryptogr.*, Calgary, AB, Canada, 2018, pp. 453–478.
[28] J. H. Cheon, D. Kim, and J. H. Park, "Towards a practical cluster analysis over encrypted data," in *Proc. Int. Conf. Sel. Areas Cryptogr.*, Calgary, AB, Canada, 2019, pp. 123–147.
[29] K. Xing, C. Hu, J. Yu, X. Cheng, and F. Zhang, "Mutual privacy preserving *k*-means clustering in social participatory sensing," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2066–2076, Aug. 2017.
[30] N. Almutairi, F. Coenen, and K. Dures, "Third party data clustering over encrypted data without data owner participation: Introducing the encrypted distance matrix," in *Proc. Int. Conf. Big Data Anal. Knowl. Discovery*, Regensburg, Germany, 2018, pp. 163–173.

MINGHUI GAO received the B.E. degree in information security from Northeastern University, in 2009. He is currently working with NARI Group Corporation, China. His research interests include computer networks and information secuirty.

BO LI received the B.E. degree in computer science and technology from Shenyang Aerospace University, in 2003. He is currently working with NARI Group Corporation, China. His research interests include computer networks and information security.



CHEN WANG received the B.E. degree in software engineering from Shenyang Technology University, in 2018. She is currently pursuing the M.E. degree with Northeastern University. Her research interests include cryptography and machine learning.

LI MA received the B.E. degree in computer science and technology from the Kunming University of Science and Technology, in 2008. He is currently working with NARI Group Corporation, China. His research interests include computer networks and information security.

JIAN XU received the Ph.D. degree in computer application technology from Northeastern University, in 2013. He is currently an Associate Professor with Northeastern University. His research interests include cryptography and network security.