PAPER Special Section on Data Mining and Statistical Science

Privacy Preserving Frequency Mining in 2-Part Fully Distributed Setting

The Dung LUONG[†a)], *Nonmember and* Tu Bao HO[††b)], *Member*

SUMMARY Recently, privacy preservation has become one of the key issues in data mining. In many data mining applications, computing frequencies of values or tuples of values in a data set is a fundamental operation repeatedly used. Within the context of privacy preserving data mining, several privacy preserving frequency mining solutions have been proposed. These solutions are crucial steps in many privacy preserving data mining tasks. Each solution was provided for a particular distributed data scenario. In this paper, we consider privacy preserving frequency mining in a so-called 2-part fully distributed setting. In this scenario, the dataset is distributed across a large number of users in which each record is owned by two different users, one user only knows the values for a subset of attributes, while the other knows the values for the remaining attributes. A miner aims to compute the frequencies of values or tuples of values while preserving each user's privacy. Some solutions based on randomization techniques can address this problem, but suffer from the tradeoff between privacy and accuracy. We develop a cryptographic protocol for privacy preserving frequency mining, which ensures each user's privacy without loss of accuracy. The experimental results show that our protocol is efficient as well.

key words: *privacy preserving frequency mining, 2-part fully distributed setting, cryptography*

1. Introduction

Data mining plays an important role in the current world, and provides us a powerful tool for discovering knowledge from huge amounts of data. However the process of mining data can result in violations of privacy. So, issues of privacy preservation in data mining are receiving more and more attention from the public [5] and many privacy preserving data mining approaches have been proposed for tackling the problem of privacy violation [17], [22], [23].

Generally, there are mainly two kinds of privacy preserving data mining approaches: the perturbation-based approach and the cryptography-based approach. The methods based on perturbation (e.g., [1], [3], [27]) are very efficient, but have a tradeoff between privacy and accuracy. The methods based on cryptography (e.g., [11], [23], [25]) can safely preserve privacy without loss of accuracy, but have high complexity and communication cost. These privacy preserving data mining methods have been presented for various scenarios in which the general idea is to allow mining datasets distributed across multiple parties, without disclosing each party's private data [2].

Computing the frequencies of values or tuples of values in a dataset is a fundamental operation that is repeatedly used in various data mining methods. Within the context of privacy preserving data mining, several privacy preserving frequency mining solutions have been proposed. The goal is for one of the participants to obtain the global frequencies of values or tuples of values from the joint data set of all parties, with the requirement that no information about private data, except that which can be deduced from the frequency values, will be disclosed.

Each privacy preserving frequency mining solution can be applied in a particular privacy preserving data mining scenario. In [12], [13] and [28] they developed a private frequency computation solution from the vertically distributed data based on secure scalar product protocols, where the final goal was to design privacy preserving protocols for learning naive Bayes classification, association rules and decision trees. In [15], private frequency computation was addressed for horizontally distributed data by computing the secure sum of all local frequencies of participating parties. Much more complicated solutions have been proposed for the fully distributed setting [6], [18] and [20], where the goal is to allow a miner to compute the frequencies of values or tuples of values from a data set distributed across a large number of users, while preserving privacy of each user's private data.

In this paper, we study privacy preserving frequency mining in yet another scenario that exists in various practical applications but has not been investigated. In this scenario, the data set is distributed across a large number of users, and each record is owned by two different users, one user only knows the values for a subset of attributes, while the other knows the values for the remaining attributes. We call this *2-part fully distributed setting* (2PFD, for short). Computing the privacy preserving frequencies of the tuples in 2PFD setting is more complicated than in other settings, because a tuple of values here may belong to two different users.

Let us take some examples of 2PFD. Consider the scenario in which a sociologist wants to find out the depersonalization behavior of children depending on the parenting style of their parents [29]. The sociologist provides the sample survey to collect information about the parenting style from parents and behavior from their children. Clearly, the information is quite sensitive, parents do not want to objec-

Manuscript received December 1, 2009.
Manuscript revised April 2, 2010.

[†]The author is with Vietnam Government Information Security Commission, 105-Nguyen Chi Thanh, Hanoi, Vietnam.

[††]The author is with Japan Advanced Institute of Science and Technology, Nomi-shi, 923–1292 Japan, and also with Institute of Information Technology, 18 Hoang Quoc Viet, Hanoi, Vietnam.

a) E-mail: thedung@bcy.gov.vn
b) E-mail: bao@jaist.ac.jp

DOI: 10.1587/transinf.E93.D.2702

Copyright © 2010 The Institute of Electronics, Information and Communication Engineers

tively reveal their limitations in educating children, while it is also difficult to ask the children to answer honestly and truthfully about their depersonalization behavior. Therefore, in order to get accurate information, the researcher must ensure the confidentiality principle of information for each subject. In this case, each data record is privately owned by both the parents and their children.

Another example is the scenario where a medical researcher needs to study the relationship between living habits, clinical information and a certain disease [7], [14]. A hospital has a clinical data set of the patients that can be used for research purposes and the information of living habits can be collected by a survey of patients, though, neither the hospital nor the patients are willing to share their data with the miner because of privacy. This scenario meets the 2PFD setting, where each data object consists of two parts: one part consisting of living habits belongs to a patient, the remaining part consisting of clinical data of this patient is kept by the hospital. Furthermore, we can see that the 2PFD setting is quite popular in practice, and that privacy preserving frequency mining protocols in 2PFD are significant and can be applied to many other similar distributed data scenarios.

The main contribution of this work is to develop a cryptographic protocol for privacy preserving frequency mining in 2PFD setting. The proposed protocol ensures each user's privacy without loss of accuracy. In addition, it is efficient, requiring only 1 or 2 of interactions between each user and the miner, while the users do not have to communicate with each other. Experimental evaluation of computation cost of the protocol has shown that it is efficient and practical.

2. Related Works

A variety of privacy preserving data mining solutions have been proposed that relate to our task.

Some randomization-based solutions proposed in [1], [3], [18], [19] can be applied to 2PFD setting. The basic idea of these solutions is that every user perturbs its data, before sending it to the miner. The miner then can reconstruct the original data to obtain the mining results with some bounded error. These solutions allow each user to operate independently, and the perturbed value of a data element does not depend on those of the other data elements, but only on its initial value. Therefore, they can be used in various distributed data scenarios. Although these solutions are very efficient, their use generally involves a tradeoff between privacy and accuracy, i.e. if we require the more privacy, the miner loses more accuracy in the data mining results, and vice-versa.

In contrast, the cryptographic approaches proposed in [6], [25] provided strong privacy without loss of accuracy. The key idea of these approaches is a private frequency computation method in the fully distributed setting that allows the miner to compute frequencies of values or tuples in the data set, while preserving privacy of each user's data. To compute the frequency of a tuple of values, each user outputs a boolean value (either 1 or 0) indicating whether the data it holds matches the pattern or not, and the miner uses private frequency computation method to privately compute the sum of boolean values from all users. The result of the private frequency computation is then used for various privacy preserving data mining tasks such as naive Bayes learning, decision tree learning, association rule mining etc. Here we aim at solving the privacy-preserving frequency mining problem in 2PFD setting. Note that in this setting, each user may only know some values of the tuple but not all. Therefore, the above mentioned cryptographic approaches can not be used in 2PFD setting. For more details, see the following section.

Some other solutions based on k-anonymization of user's data have been proposed in [20], [21]. The advantage of these solutions is that they do not depend on the underlying data mining tasks, because the anonymous data can be used for various data mining tasks without disclosing private information. However, these solutions are inapplicable in 2PFD setting, because the miner can not link two anonymous parts of one object with each other.

One of the requirements in our computation model is the connection of two different parts of the partitioned records to obtain the desired computation results without disclosing any attribute information. It is similar with the problem of secure scalar product [13] and the problem of computing the intersection of private datasets in two-party vertically partitioned model [16]. Indeed, we consider the problem of computing the intersection of private datasets of two parties. This problem requires combination of two values belonged two different parts of the two-party partitioned records to obtain the matching results while preserving each party's privacy. To solve this problem based on the proposed protocol in [16], we follow the basic structure: one party defines a polynomial whose roots are her inputs, and then encrypts the coefficients of this polynomial by homomorphic encryption. Thus, other party can use the homomorphic properties of the encryption system to evaluate the polynomial at each of his inputs. He then multiplies each result by a random number and adds to it an encryption of the value of his input. The result allows the party with the encrypted polynomial to find the values in the intersection of the two parties' inputs while protecting privacy of the remaining values. Here, we note that the evaluation party owns the value one of each combined values pair, thus it can easily combine its values with other party's corresponding values by evaluating the encrypted polynomial. In our problem, the miner plays a role as a combiner; however the miner does not know any values in each partitioned record. Therefore, our problem is clearly more difficult than the similar problems in the vertically portioned data model.

3. Preliminaries

3.1 Privacy-Preserving Frequency Mining Formulation in 2PFD Setting

In 2PFD setting, a data set (a data table) consists of *n* records, and each record is described by values of nominal attributes. The data set is distributed across two sets of users $U = \{U_1, U_2, \dots, U_n\}$ and $V = \{V_1, V_2, \dots, V_n\}$. Each pair of users ($U_i$, V_i) owns a record in which user U_i knows values for a proper subset of attributes, and user V_i knows the values for the remaining attributes. Note that in this setting, the set of attributes whose values known by each U_i is equal, and so for each user V_i.

The miner aims to mine the frequency of a tuple of values in the data set. Assume that each user's data includes the sensitive attribute values. To protect users' privacy and also enable learning frequency, our purpose is to design a protocol that enable the miner to learn frequency from all users' data without learning each individual's sensitive values. Assume that the tuple consists of two parts, the first part consists of values for some attributes belong to U_i, and the second part consists of the remaining values for some attributes belong V_i. In this case, each U_i outputs a boolean value u_i (either 1 or 0) to indicate whether or not the data it holds matched the first part, and each V_i outputs a boolean value v_i to indicate whether or not the data it holds matched the second part. Therefore, our purpose is to design a protocol that allows the miner to obtain the sum $f = \sum u_i v_i$ without revealing u_i and v_i.

Our formula is still appropriate when the tuple consisting of values for some attributes only belongs to U_i (or V_i). For example, when the tuple consists of values for some attributes only belonging to U_i, U_i outputs a boolean value u_i to indicate whether the data it holds matches all values in the tuple and V_i outputs $v_i = 1$. Therefore, clearly the sum $f = \sum u_i = \sum u_i v_i$ is the frequency value which needs be computed. However, to compute $\sum u_i$, we can use the privacy-preserving frequency mining protocol for the fully distributed setting proposed in [6].

To be applicable, we require that the protocol can ensure users' privacy in an environment that doesn't have any secure communication channel between the user and the miner, as well as it should not require any communication among the users. In addition, it should minimize the number of interactions between the user and the miner. Particularly, the user U_i must not interact with the miner more than twice, and the user V_i must interact with the miner exactly once. Those requirements make our protocol more applicable. For example, considering a real scenario when a miner uses a web-application to investigate a large number of users for his research, a user only needs to use his browser to communicate with the server one or two times, while he does not have to communicate with the others.

3.2 Definition of Privacy

The privacy preservation of the proposed protocol is based on the semi-honest security model. In this model, each party participating in the protocol has to follow rules using correct input, and cannot use what it sees during execution of the protocol to compromise security. A general definition of secure multi-party computation in the semi-honest model is stated in [8]. This definition was derived to make a simplified definition in the semi-honest model for privacy preserving data mining in the fully distributed setting scenario [6], [25]. This scenario is similar to 2PFD setting, so here we consider the possibility that some corrupted users share their data with the miner to derive the private data of the honest users, we assume that all users are semi-honest, thus any user can be corrupted. One requirement is that no other private information about the honest users be revealed, except a multivariate linear equation in which each variable presents a value of an honest user. In our model, information known by users is no more than information known by the miner, so we do not have to consider the problem in which users share information with each other.

Definition. *Assume that each user U_i has a private set of keys $D_i^{(u)}$ and a public set of keys $E_i^{(u)}$, and each user V_i has a private set of keys $D_i^{(v)}$ and a public set of keys $E_i^{(v)}$. A protocol for the above defined frequency mining problem protects each user's privacy against the miner along with t_1 corrupted users U_i and t_2 corrupted users V_i in the semi-honest model if, for all I_1, $I_2 \subseteq \{1, \dots, n\}$ such that $|I_1| = t_1$ and $|I_2| = t_2$, there exists a probabilistic polynomial-time algorithm M such that*

$$\{M(f, [u_i, D_i^{(u)}]_{i\in I_1}, [E_j^{(u)}]_{j\notin I_1}, [v_k, D_k^{(v)}]_{k\in I_2}, [E_l^{(v)}]_{l\notin I_2})\}$$
$$\overset{c}{\equiv} \{View_{miner,\{U_i\}_{i\in I_1},\{V_k\}_{k\in I_2}}[u_i, D_i^{(u)}, v_i, D_i^{(v)}]_{i=1}^{n}\}$$

where $\overset{c}{\equiv}$ denotes computational indistinguishability.

Basically, the definition states that the computation is secure if the joint view of the miner and the corrupted users (the t_1 users U_i and the t_2 users V_i) during the execution of the protocol can be effectively simulated by a simulator, based on what the miner and the corrupted users have observed in the protocol using only the result f, the corrupted users' knowledge, and the public keys. Therefore, the miner and the corrupted users can not learn anything from f. By the definition, in order to prove the privacy of a protocol, it suffices to show that there exists a simulator that satisfies the above equation.

3.3 ElGamal Encryption Scheme

In this section, we briefly review ElGamal encryption scheme [24] to be used later.

Let G be a cyclic group of order q in which the discrete logarithms are hard. Let g be a generator of G, and x be uniformly chosen from $\{0, 1, \dots, q-1\}$. In ElGamal encryption schema, x is a private key and the public key is $h = g^x$. Each user securely keeps their own private keys, otherwise public keys are publicly known.

To encrypt a message M using the public key h, one randomly chooses k from $\{0, \dots, q-1\}$ and then computes the ciphertext $C = (C_1 = Mh^k, C_2 = g^k)$. The decryption of the ciphertext C with the private key x can be executed by computing $M = C_1(C_2^x)^{-1}$.

ElGamal encryption is semantically secure under the Decisional Diffie-Hellman (DDH) Assumption [4]. In ElGamal encryption scheme, one cleartext has many possible encryptions, since the random number k can take many different values. ElGamal encryption has a randomization property in which it allows computing a different encryption of M from a given encryption of M.

4. Privacy Preserving Frequency Mining Protocol in 2PFD Setting

4.1 Protocol

In this section, we use ELGamal encryption scheme and the joint decryption technique to build a privacy-preserving frequency mining protocol. This idea has been extensively used in previous works, e.g., [9], [25], [26].

In the proposed protocol, we assume that each user U_i has private keys x_i, y_i, z_i and public keys $X_i = g^{x_i}$, $Y_i = g^{y_i}$, $Z_i = g^{z_i}$, and each user V_i has private keys p_i, q_i, s_i and public keys $P_i = g^{p_i}$, $Q_i = g^{q_i}$, $S_i = g^{s_i}$. We note that computations in this paper take place in the group G. We define

$$X = \prod_{i=1}^{n} X_i P_i = g^x$$
$$Y = \prod_{i=1}^{n} Y_i Q_i = g^y$$

where

$$x = \sum_{i=1}^{n} (x_i + p_i)$$
$$y = \sum_{i=1}^{n} (y_i + q_i)$$

In our protocol, the values X and Y are known by all users, each user use X and Y as the public keys to encrypt its data. Thus, decrypting its encryptions requires the use of the private keys x and y, where no individual user known these values.

As presented in Sect. 3.1, our purpose is to allow the miner to privately obtain the sum $f = \sum_{i=1}^{n} u_i v_i$. The privacy preserving protocol for the miner to compute f consists of the following phases:

- Phase 1. Each user U_i does the following:
  - Choose randomly c_i from $\{0, 1, \ldots, q-1\}$,
  - Compute $C_1^{(i)} = g^{u_i} Z_i^{c_i}$ and $C_2^{(i)} = g^{c_i}$,
  - Send $C_1^{(i)}$ and $C_2^{(i)}$ to the miner.
- Phase 2. Each user V_i does the following:
  - Get $C_1^{(i)}$ and $C_2^{(i)}$ from the miner,
  - Choose randomly r_i from $\{0, 1, \ldots, q-1\}$,
  - if $v_i = 0$ then compute $R_1^{(i)} = X^{q_i}$, $R_2^{(i)} = (C_2^{(i)})^{s_i r_i} Y^{p_i}$ and $R_3^{(i)} = S_i^{r_i}$,
  - if $v_i = 1$ then compute $R_1^{(i)} = (C_1^{(i)})^{v_i} X^{q_i}$, $R_2^{(i)} = (C_2^{(i)})^{s_i r_i} Y^{p_i}$ and $R_3^{(i)} = (Z_i)^{-1} S_i^{r_i}$,
  - Send $R_1^{(i)}$, $R_2^{(i)}$ and $R_3^{(i)}$ to the miner.
- Phase 3. Each user U_i does the following:
  - Get $R_1^{(i)}, R_2^{(i)}$ and $R_3^{(i)}$ from the miner,
  - Compute $K_1^{(i)} = R_1^{(i)} (R_3^{(i)})^{c_i} X^{y_i}$, $K_2^{(i)} = R_2^{(i)} Y^{x_i}$,
  - Send $K_1^{(i)}$ and $K_2^{(i)}$ to the miner.
- Phase 4. The miner does the following:
  - Compute $d = \prod_{i=1}^{n} \frac{K_1^{(i)}}{K_2^{(i)}}$,
  - Find f from $\{0, 1, \ldots, q-1\}$ that satisfies $g^f = d$,
  - Output f.

4.2 Proof of Correctness

Theorem 1. *The above presented protocol correctly computes the frequency value $f = \sum_{i=1}^{n} u_i v_i$ as defined in Sect. 3.1.*

Proof. We show that the miner can compute the desired value f by using the above protocol. Indeed,

$$\begin{aligned} d &= \prod_{i=1}^{n} \frac{K_1^{(i)}}{K_2^{(i)}} \\ &= \prod_{i=1}^{n} \frac{R_1^{(i)} (R_3^{(i)})^{c_i} X^{y_i}}{R_2^{(i)} Y^{x_i}} \end{aligned}$$

If $v_i = 0$ then $g^{u_i v_i} = 1$, therefore

$$\begin{aligned} K_1^{(i)} &= X^{q_i} g^{s_i r_i c_i} X^{y_i} \\ &= g^{u_i v_i} g^{s_i r_i c_i} X^{y_i + q_i} \end{aligned}$$

If $v_i = 1$, we have

$$\begin{aligned} K_1^{(i)} &= (C_1^{(i)})^{v_i} X^{q_i} (Z_i^{-1} S_i^{r_i})^{c_i} X^{y_i} \\ &= g^{u_i v_i} g^{z_i c_i v_i} X^{q_i} g^{-z_i c_i} g^{s_i r_i c_i} X^{y_i} \\ &= g^{u_i v_i} g^{s_i r_i c_i} X^{y_i + q_i} \end{aligned}$$

In both cases, we also have

$$\begin{aligned} K_2^{(i)} &= (C_2^{(i)})^{s_i r_i} Y^{p_i} Y^{x_i} \\ &= g^{s_i c_i r_i} Y^{x_i + p_i} \end{aligned}$$

Finally, we obtain

$$\begin{aligned} d &= \prod_{i=1}^{n} \frac{K_1^{(i)}}{K_2^{(i)}} \\ &= \prod_{i=1}^{n} \frac{g^{u_i v_i} g^{s_i r_i c_i} X^{y_i + q_i}}{g^{s_i r_i c_i} Y^{x_i + p_i}} \\ &= \prod_{i=1}^{n} g^{u_i v_i} \prod_{i=1}^{n} \frac{X^{y_i + q_i}}{Y^{x_i + p_i}} \\ &= g^{\sum_{i=1}^{n} u_i v_i} \prod_{i=1}^{n} \frac{(g^{\sum_{j=1}^{n} (x_j + p_j)})^{(y_i + q_i)}}{(g^{\sum_{j=1}^{n} (y_j + q_j)})^{(x_i + p_i)}} \end{aligned}$$

$$= g^{\sum_{i=1}^n u_i v_i} \frac{g^{\sum_{i=1}^n \sum_{j=1}^n (x_j+p_j)(y_i+q_i)}}{g^{\sum_{i=1}^n \sum_{j=1}^n (y_j+q_j)(x_i+p_i)}}$$
$$= g^{\sum_{i=1}^n u_i v_i}$$

Therefore, we can obtain f from the equation $d = g^f = g^{\sum_{i=1}^n u_i v_i}$. □

Note that, in practice, the value of f is not too large, so that the discrete logarithms can be successfully taken (for example $f = 10^5$).

4.3 Proof of Privacy

In this section, we first show that under the DDH assumption, our protocol preserves each user's privacy in the semi-honest model. Then, we show that in the case of collusion of some corrupted users with the miner, the protocol still preserves the privacy of each honest user.

In our model, the communication only occurs between each user and the miner, thus the miner receives the messages of all users. Assume that each user can get the messages of the remaining users via the miner, then the information known by the miner and each user are the same during the execution of the protocol. Therefore, it is sufficient to only consider the view of the miner, as follow:

In Phase 1, the miner receives the messages $C_1^{(i)}$ and $C_2^{(i)}$ of each U_i. Here $C = (C_1^{(i)}, C_2^{(i)})$ is an ElGamal encryption of the value g^{u_i} under the private key z_i, the public key $Z_i = g^{z_i}$, and the value c_i is randomly chosen from $\{1, 2, \dots, q-1\}$.

In Phase 2, the messages $R_1^{(i)}, R_2^{(i)}$ and $R_3^{(i)}$ sent by each V_i are equivalent to the first part of ElGamal encryptions $R_1 = (\alpha X^{q_i}, g^{q_i})$, $R_2 = (\beta Y^{p_i}, g^{p_i})$ and $R_3 = (\gamma S_i^{r_i}, g^{r_i})$, respectively ($\alpha = 1$ or $(C_1^{(i)})^{v_i}$; β=$(C_2^{(i)})^{s_i r_i}$; $\gamma = 1$ or $(Z_i)^{-1}$). Here x, y and s_i are the private keys, and q_i, p_i and r_i are randomly chosen from $\{1, 2, \dots, q-1\}$.

Similarly, in Phase 3, the messages $K_1^{(i)}$ and $K_2^{(i)}$ sent by each U_i can be represented as the first part of ElGamal encryptions $K_1 = (\alpha' X^{y_i}, g^{y_i})$ and $K_2 = (\beta' Y^{x_i}, g^{x_i})$.

As well known, the ElGamal encryption is semantically secure under the DDH assumption. So, the view of the miner can be efficiently simulated by a simulator for ElGamal encryptions.

Now, we show that the protocol preserves the privacy of the honest users against the collusion of the corrupted users with the miner, even up to $2n-2$ corrupted users. We have the following theorem

Theorem 2. *The protocol in Sect. 3.3 preserves the privacy of the honest users against the miner and up to* $2n-2$ *corrupted users. In cases with only two honest users, it remains correct as long as two honest users do not own the attribute values of the same record.*

Proof. In the proposed protocol, the information known by each user is the same, thus we need to only consider the case where a user U_i and a user V_j ($i \neq j$) are honest. The remaining cases can be proved similarly. Without loss of generality, we assume that $I = \{2, 3, 4, \dots, n\}$ and $J = \{1, 3, 4, \dots, n\}$.

Now we need to design a simulator M that simulates the joint view of the miner and the corrupted users by a probabilistic polynomial-time algorithm, and then this simulator is combined with a simulator for the ElGamal ciphertexts to obtain a completed simulator. To do so, basically we show a polynomial-time algorithm for computing the joint view of the miner and the corrupted users. The computation of the algorithm is based on what the miner and the corrupted users have observed in the protocol using only the result f, the corrupted users' information, and the public keys. The algorithm outputs the simulated values for the encryptions generated by a simulator of ElGamal encryptions.

- M simulates $C_1^{(1)}$, $C_2^{(1)}$ using two random ElGamal ciphertexts.
- M takes the following encryptions as its input
  $$(a_1, a_1') = (\alpha g^{(x_1+p_2)q_2}, g^{q_2})$$
  $$(a_2, a_2') = (\beta g^{(y_1+q_2)p_2}, g^{p_2})$$
  where $\alpha = 1$ or $C_1^{(2)}, \beta = (C_2^{(2)})^{s_2 r_2}$, and it computes the following values
  $$R_1'^{(2)} = a_1 Q_2^{\sum_{i \in I} x_i + \sum_{j \in J} p_j} / g^\delta$$
  $$R_2'^{(2)} = a_2 P_2^{\sum_{i \in I} y_i + \sum_{j \in J} q_j}$$
  where $\delta = f - \sum_{l=3}^n u_l v_l - \lambda v_1 - \theta u_2$, and $\lambda, \theta \in \{0, 1\}$. Next, M simulates $R_3^{(2)}$ using a random ElGamal ciphertext.
- M takes the two following encryptions as its input
  $$(b_1, b_1') = (R_1^{(1)} (R_3^{(1)})^{c_1} g^{(x_1+p_2)y_1}, g^{y_1})$$
  $$(b_2, b_2') = (R_2^{(1)} g^{(y_1+q_2)x_1}, g^{x_1})$$
  and computes
  $$K_1'^{(1)} = b_1 . Y_1^{\sum_{i \in I} x_i + \sum_{j \in J} p_j}$$
  $$K_2'^{(1)} = b_2 . X_1^{\sum_{i \in I} y_i + \sum_{j \in J} q_j}$$

This finishes the simulation algorithm. □

4.4 Efficiency Evaluation

In this section, we show results of the complexity estimation of the protocol and the efficiency measurement of the protocol in practice

In the proposed protocol, the computational cost of each user U_i in the first phase and in the third phase are 2 and 3 modular exponentiations, respectively. The computational cost of each user V_i in the second phase is at most 4 modular exponentiations. The miner uses $2n$ modular multiplications and at most n comparisons. We note that these computational costs do not include the overhead of key generation and computing two parameters X and Y. This looks quite expensive (as shown in the following experimental results), however generating these parameters belongs to the

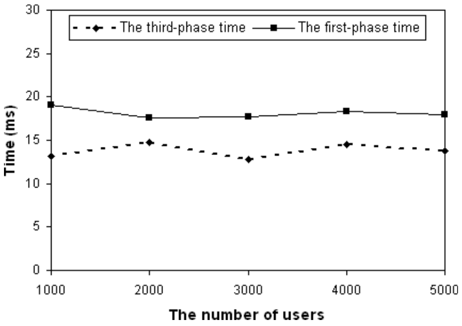

Fig. 1 The average time used by each U_i for computing the messages in the first phase and the third phase.

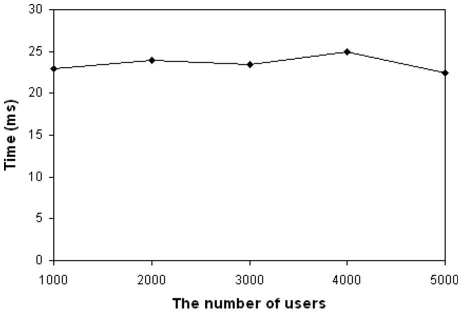

Fig. 2 The average time used by each V_i for computing the messages in the second phase.

preparation period of the mining process, so it can be implemented before the protocol is executed without affecting the computation time of the protocol.

For evaluating the efficiency of the protocol in practice, we build an experiment on the privacy preserving frequency mining in C environment, which runs on a laptop with CPU Pentium M 1.8 GHz and 1 GB memory. The used cryptographic functions are derived from Open SSL Library.

We measure the computation cost of the frequency mining protocol for different numbers of users, from 1000 to 5000. Before executing the protocol, we generate three pairs of keys for each user, with the size of public keys set at 512 bits, and then compute parameters X and Y. The results show that it takes 3.2 s to generate three key pairs for each user and 281 ms to compute two parameters X and Y for 5000 users.

As shown in Fig. 1, the average time used by each U_i for computing the first-phase messages and the third-phase messages are about 20 ms and 13 ms, respectively. Figure 2 shows that each V_i needs about an average 24 ms to compute her messages. For the miner, Fig. 3 shows that the computation time is very efficient and nearly linearly related to the number of users.

5. Conclusion

In this paper, we have proposed a protocol for privacy preserving frequency miming in 2PFD setting, which has not been investigated previously. Basically, the proposed protocol is based on ElGamal encryption scheme, and it provided strong privacy without loss of accuracy. We conducted an experiment to evaluate the complexity of the protocol, and the results showed that, the protocol is efficient and practical as well.

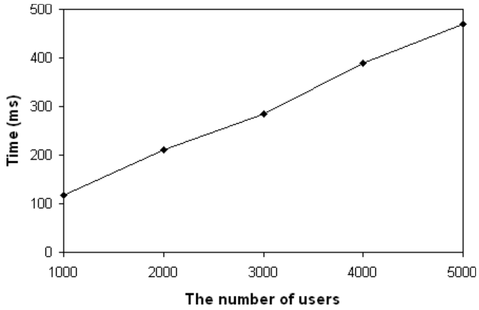

Fig. 3 The computation time of the frequency value f.

At the present, the proposed protocol only solves the problem of privacy preserving frequency mining, which can be the key component of privacy preserving protocols for several data mining tasks such as naive Bayes classification, decision tree learning, association rules mining, linear regression analysis and correlation analysis. There may be many other tasks of privacy preserving data mining in 2PFD setting, such as privacy preserving model selection, privacy preserving clustering, etc., that would be of interest for future work.

Also, although our approach is technically mature enough to be used in the privacy preserving frequency mining scenario with 2PFD setting, there are still issues we need to tackle to enhance the efficiency of the protocol. For example, in the proposed protocol, half of the users need two interactions with the miner, so a natural question is whether we can design a privacy preserving frequency mining protocol in which each user needs only one interaction with the miner.

References

[1] A. Evmievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," Proc. Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.217–228, ACM Press, 2002.

[2] C.C. Aggarwal and P.S. Yu (Eds.), Privacy-Preserving Data Mining: Models and Algorithms, Series: Advances in Database Systems, vol.34, Springer, 2008.

[3] D. Agrawal and C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," Proc. ACM SIGMOD, pp.247–255, 2001.

[4] D. Boneh, "The decision Diffe-Hellman problem," ANTS-III, vol.1423, pp.48–63, LNCS, 1998.

[5] EU Directive 95/46/EC of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official J. European Communities, p.31, 1995.

[6] F. Wu, J. Liu, and S. Zhong, "An efficient protocol for private and accurate mining of support counts," Pattern Recognit. Lett., vol.30, Issue 1, pp.80–86, 2009.

[7] G. Joachim, "The relationship between habits of food consumption and reported reactions to food in people with inflammatory bowel

disease-testing the limits," J. Nutrition and Health, vol.13, pp.69–83, 1999.

[8] O. Goldreich, Foundations of Cryptography, Basic Tools, vol.1, Cambridge University Press, 2001.

[9] H. Martin and S. Kazue, "Efficient receipt-free voting based on homomorphic encryption," Proc. Advances in Cryptology-Eurocrypt, 2000.

[10] J. Benaloh and D. Tuinstra, "Receipt-free secret- ballot elections (extended abstract)," Proc. 26th Annual ACM Symposium on Theory of Computing, pp.544–553, ACM Press, 1994.

[11] J. Sakuma and R.N. Wright, "Privacy-preserving evaluation of generalization error and its application to model and attribute selection," Proc. 1st Asian Conference on Machine Learning (ACML'09), pp.338–353, 2009.

[12] J. Vaidya and C. Clifton, "Privacy preserving naive Bayes classifier for vertically partitioned data," Proc. 2004 SIAM Conference on Data Mining, 2004.

[13] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," Proc. Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.639–644, 2002.

[14] K.J. Bjarne and S.T. Dag, "The Tromson heart study: The relationship between food habits and the body mass index," J. Chronic Diseases, vol.40, no.8, pp.795–800, 1987.

[15] M. Kantarcoglu and J. Vaidya, "Privacy preserving naive Bayes classifier for horizontally partitioned data," IEEE ICDM Workshop on Privacy Preserving Data Mining, pp.3–9, 2003.

[16] M. Freedman, K. Nissim, and B. Pinkas, "Efficient private matching and set intersection," Proc. Eurocrypt, vol.3027, pp.1–19, LNCS, Springer-Verlag, 2004.

[17] R. Agrawal and R. Srikant, "Privacy preserving data mining," Proc. ACM SIGMOD Conference on Management of Data, pp.439–450, 2000.

[18] R. Agrawal, R. Srikant, and D. Thomas, "Privacy preserving OLAP," Proc. 2005 ACM SIGMOD International Conference on Management of Data, SIGMOD'05, pp.251–262, ACM, 2005.

[19] R. Agrawal and R. Srikant, "Privacy-preserving data mining," Proc. ACM SIGMOD Conference on Management of Data, pp.439–450, ACM Press, 2000.

[20] S. Zhong, Z. Yang, and R.N. Wright, "Privacy-enhancing k-anonymization of customer data," Proc. Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp.139–147, 2005.

[21] S. Zhong, Z. Yang, and T. Chen, "k-Anonymous data collection," J. Information Sciences, vol.179, no.17, pp.2948–2963, 2009.

[22] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis, State-of-the-art in privacy preserving data mining, ACM SIGMOD Record, vol.3, no.1, pp.50–57, 2004.

[23] Y. Lindell and B. Pinkas, "Privacy preserving data mining," Advances in Cryptology Crypto2000, vol.1880, pp.36–53, LNCS, Springer-Verlag, 2000.

[24] Y. Tsiounis and M. Yung, "On the security of ElGamal-based encryption," Public Key Cryptography'98, vol.1431, pp.117–134, LNCS, 1998.

[25] Z. Yang, S. Zhong, and R.N. Wright, "Privacy-preserving classification of customer data without loss of accuracy," Proc. 2005 SIAM International Conference on Data Mining (SDM), pp.21–23, 2005.

[26] Z. Yang, S. Zhong, and R.N. Wright, "Anonymity-preserving data collection," Proc. Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD'05), pp.334–343, ACM, 2005.

[27] W. Du and Z. Zhan, "Using randomized response techniques for privacy preserving data mining," Proc. Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.505–510, ACM Press, 2003.

[28] W. Du and Z. Zhan, "Building decision tree classifier on private data," Proc. IEEE International Confonference on Privacy, Security, and Data Mining, pp.1–8, 2002.

[29] W. Uwe, H. Susanne, and N.V. Miles Jeremy, "Perceived parenting styles, depersonalisation, anxiety and coping behaviour in adolescents," J. Adolescence, vol.34, no.3, pp.521–532, 2003.

The Dung Luong is a Ph.D student, He received Bachelor of Information Technology from Le Quy Don Technical University in 2001. His research interests include privacy preserving data mining and computer security.

Tu Bao Ho is a professor at School of Knowledge Science, Japan Advanced Institute of Science and Technology. He received M.S. and Ph.D. degrees in Computer Science from Pierre and Marie Curie University, Paris (1984, 1987). His research interests include knowledge-based systems, machine learning, data mining, medical informatics and bioinformatics.