A new anonymity-based protocol preserving privacy based cloud environment

Jian Wang[1*], Le Wang[2]

[1]*College of Computer and Information Engineering, Henan University of Economics and Law, China*

[2]*SIAS International University Zhengzhou, China*

Received 1 March 2014, www.cmnt.lv

Abstract

With the development of cloud computing application, more and more people would like to do business under this environment. But more attention should be paid to the disclosure of privacy during the transaction. Customers will reject to do business on the cloud platform if the cloud environment cannot avoid disclosing their private data. Nowadays, little work has been done about how to prevent sensitive attributes leaking between service providers. Therefore, this paper proposed a new anonymity-based protocol to protect privacy.

Keywords: anonymity, cloud computing, privacy preserving, sensitive attributes, protocol

1 Introduction

Cloud computing raises a range of important privacy issues as acknowledged by a number of recent work [1]. Such issues are due to the fact that, in the cloud, users' data and applications reside on the cloud cluster which is owned and maintained by a third party [2]. Concerns arise since in the cloud it is not clear to individuals why their personal information is requested or how it will be used or passed on to other parties [3, 4].

Despite increased awareness of the privacy issues in the cloud, little work has been done in this space. Recently, Pearson et al. has proposed accountability mechanisms to address privacy concerns of end users [5] and then develop a simple solution, a privacy manager, relying on obfuscation techniques [6]. Their basic idea is that the user's private data is sent to the cloud in an encrypted form, and the processing is done on the encrypted data. The output of the processing is de-obfuscated by the privacy manager to reveal the correct result. However, the privacy manager provides only limited features in that it does not guarantee protection once the data is being disclosed [7-9].

There are many service providers in the cloud, we can call each service as a cloud, each cloud service will exchange data with other cloud, when the data is exchanged between the clouds, and there exists the problem of disclosure of privacy [10]. So my research aims at avoiding the disclosure of the sensitive attributes of users' when user ask for service from the service provider in cloud computing.

2 Related work

There are some privacy problems we want to address in the cloud computing [11]: The first problem is the disclosure of sensitive private information when exchanging data through the cloud service. And the sensitive private information includes: personally identifiable information, usage data, unique device identities and so on [12].

The second problem is that people getting inappropriate or unauthorized access to personal data in the cloud by taking advantage of certain vulnerabilities, such as lack of access control enforcement, security holes and so on [13].

The third problem is that: because the feature of cloud computing is that it is a dynamic environment, in that service interactions can be created in a more dynamic way than traditional e-commerce scenarios [14]. Services can potentially be aggregated and changed dynamically by service providers can change the provisioning of services. In such scenarios, personal sensitive data may move around within an organization or across organizational boundaries, so adequate protection of this information must be maintained despite the changes. So design the method to protect the privacy in cloud computing must meet the dynamical exchange of data.

Nowadays, some researchers focus on cloud data storage security in cloud computing [15]. And other researchers focus on the disclosure of sensitive private information when exchanging data through the cloud computing [16].

Some researchers use identity technology to solve these privacy problems in cloud computing [17]. They propose some requirements for identity services [18]. For

[*] ***Corresponding author*** e-mail: goodjian121@126.com

example, the identity services should be independent of devices; should permit a single sign-on to thousands of different online services; should allow pseudonyms and multiple discrete identities to protect user privacy; should be interoperable, based on open standards, and should be transparent and auditable.

When create an identity management infrastructure in the cloud computing, we should consider:

1) There are already a number of identity management systems in place on a wide variety of platforms. These need to be supported by the identity management infrastructure. The infrastructure must support cross-system interaction as well as interoperation and delegation between them. This is only possible if the infrastructure and the individual systems are based on open standards, available on all platforms.

2) Identity management systems will support a wide variety of privacy and security properties, ranging from low-security password-based one-factor authentication to high-end, attribute-based systems deploying state-of-the-art privacy-enhancing certificates. While the infrastructure needs to support all of these systems, users should understand the implications of using one system over the other [19].

3 Anonymity-based method

Our paper proposed a new anonymity protocol for the cloud computing services. Before the micro data are published, this anonymity algorithm will process these data. And then send these anonymous data to service providers in the cloud. Then the service provider can integrate the auxiliary information (also called external knowledge, background knowledge, or side information that the service provider can get from other channels such as the web, public records, or domain knowledge) to analyse the anonymous data in order to mine the knowledge they want.

For example, the traffic management board of a region collects records of road accidents for research and analysis. Suppose each record has five attributes, namely occupation, age, vehicle-type, postcode, and faulty. Consider the tuples in Table 1.

The traffic management board anonymised the traffic accident records on attributes age, vehicle-type, and postcode before the data can be released to the service provider in the cloud. Suppose 2-anonymity is required. Table 2 shows a 2-anonymous release of the records with respect to quasi-identifier (age, vehicle-type, postcode).

And then the traffic management board anonymised the traffic accident records on attributes occupation, age, and postcode. Again, suppose 2-anonymity is required. Table 3 shows a 2-anonymous release of the records with respect to quasi-identifier (occupation, age, postcode).

Suppose one service provider in the cloud obtains both releases in Tables 2 and 3. By comparing the two tables, the service provider immediately knows that a family doctor of age 30 driving a white Sedan living in area 31043 was faulty in an accident. The victim may be easily re-identified by both these anonymous tables.

TABLE 1 A set of traffic accident record

Occupation	Age	Vehicle	Postcode	Faulty
Dentist	30	Red Truck	31043	No
Family doctor	30	White Sedan	31043	Yes
Banker	30	Green Sedan	31043	No
Mortgage broker	30	Black Truck	31043	No

TABLE 2 2-anonymous release of Table 1 with respect to quasi-attributes (age, vehicle-type, postcode)

Occupation	Age	Vehicle	Postcode	Faulty
Dentist	30	Truck	31043	No
Family doctor	30	Sedan	31043	Yes
Banker	30	Sedan	31043	No
Mortgage broker	30	Truck	31043	No

TABLE 3 A 2-anonymous release of Table 1 with respect to quasi-attributes (occupation, age, postcode)

Occupation	Age	Vehicle	Postcode	Faulty
Medical	30	Red Truck	31043	No
Medical	30	White Sedan	31043	Yes
Finance	30	Green Sedan	31043	No
Finance	30	Black Truck	31043	No

Besides, when publish anonymous data, we should consider multiple quasi-identifiers (QI) attributes for different service providers carrying different background knowledge [20]. So the data publishing side should anonymise different quasi-identifiers (QI) attributes of the records for different service providers carrying different background knowledge.

We can take an example to explain. Suppose the traffic management board have records of road accidents for research and analysis. Suppose each record has five attributes, namely occupation, age, vehicle-type, postcode, and faulty. And the traffic management board wants to release these records to different service providers in the cloud.

Suppose there are two service provider named "auto insurance companies" and "human resource department" in the cloud service. Both of them want to use the anonymous records of road accidents.

Importantly, different service providers may carry different background knowledge. For example, the auto insurance company may join the traffic accident records with the vehicle registration records on attributes age, vehicle-type, and postcode to find out whether its customers were at fault in some accidents. Typically, the company does not have the occupation information of its customers, as such information is not required in applying for auto insurance. Therefore, to protect privacy, the traffic management board has to anonymise the traffic accident records on attributes age, vehicle-type, and postcode before the data can be released to the auto insurance company.

Simultaneously, the human resource department may join the traffic accident records with the resident records on attributes occupation, age, and postcode to find out which residents were faulty in some accidents. Therefore, to protect privacy, the traffic management board needs to anonymise the traffic accident records on attributes occupation, age, and postcode. Note that vehicle-type is

not part of the anonymisation, because the human resource department typically does not have information about residents' vehicle types.

Our method is different from the traditional cryptography technology to protect individuals' privacy in the cloud computing services. If we use the cryptography technology. Then we should use cryptography technology to process the data, and send the processed data to service provider in the cloud. Then this service provider can't use these data if it didn't get the key of cryptography. So the service provider in the cloud should use the key to restore the data firstly and then can use these data. However, if we use anonymity technology to process these data and send these anonymous data to service providers in the cloud. Then the service provider can directly use these data without any key and without restoring these data. So this will be more flexible and safe to protect individuals' privacy in the cloud computing services.

4 Private matching protocol

The main purpose for our research is to use private matching technology [21] to intersect user's data and datasets of service provider without accessing each other's data, so this can let the user check whether his anonymous data meet k-anonymity to the datasets of service provider, meanwhile avoiding the disclosure of each other's data.

Our proposed private matching protocol:

1) Both the Client and the SP serialize the attributes and concatenate them to convert their data set into a string of concatenated attribute values, and produce set Vs and set Vc. Which Vc is the set of elements (tuples) in Client, corresponding to the attributes requested by the SP in the current service request, while Vs is the set of elements (tuples) in SP, corresponding to the attributes requested by the SP in the current service request.

Notion 1: Before the user in client adds his tuple to the table of SP, the user should check whether the one new tuple meets k-anonymity to the table of SP. So |Vc|=1. Because before adding one new tuple to the table of SP, the set of Vc only has one tuple to check. But |Vs|>=1. Because the set of Vs may have many tuples before checking.

2) Both Client (named C, for short) and SP (named S, for short) apply hash function h to their sets.

Xc=h (Vc),

Xs=h (Vs).

Each party randomly chooses a secret key. And ec is the key for C, es is the key for S.

3) Both parties encrypt their hashed sets:

Yc=Fec (Xc)=Fec (h (Vc)),

Ys=Fes (Xs)=Fes (h (Vs)).

4) C sends to S its encrypted set

Yc=Fec (h (Vc)).

5a) S ships to C its set Ys=Fes (h (Vs)).

5b) S encrypts each y Yc, with S' key es and sends back to C the pairs

<y, Fes (y)>=<Fec (h (v)), Fes (Fec (h (v)))>.

6) C encrypts each y Ys, with C's key ec, obtaining Zs=Fec (y)=Fec (Fes (h (v))), here the v Vs.

Also, from pairs <Fec (h (v)), Fes (Fec (h (v)))> obtained in Step 5b for the v Vc, It creats pairs <v, Fes (Fec (h (v)))> by replacing Fec (h (v)) with the corresponding v.

7) Because the Vc has only one element v (that is to say Vc has only one tuple). So if this element v Vc meets (Fes (Fec (h (v)))) Zs, then this v Vs. then Vs and Vc has one identical element, else Vs and Vc has no identical element.

Vs and Vc has one identical element is equal to

$$|V_s \cap V_c|=1.$$

Vs and Vc has no identical element is equal to

$$|V_s \cap V_c|=0.$$

5 Case studies

Despite increased awareness of the privacy issues in the cloud, little work has been done in this space. When users order service in the cloud, lots of data including users' attributes need to be transmitted between users and service providers. How to protect users' sensitive attributes and avoid the identification by adversary is still an important problem to solve. Most existing methods are that the user's private data is sent to the cloud in an encrypted form [22], and the processing is done on the encrypted data. However, in this method, the service provider needs to decode these data before they access them. So this will be a waste of time and will be very inconvenient to service providers. Besides, once the data is decrypted the user's privacy may be at risk, since the SP has full control of it. So our proposed approach is different from traditional methods to avoid disclosure of individuals' privacy.

In our proposed approach, the datasets of service provider meet k-anonymity. If a user wants to request a service in the cloud, the user should anonymised his attributes corresponding to the attributes requested by the SP in the current service request, then if this user's anonymised data meet k-anonymity to datasets of this service provider, this user can send his anonymised data to this service provider. Then the service provider will prepare service for this right user. Our proposed approach has two advantages, the first one is the data transmitted in the cloud are anonymised data, so if the adversary get these anonymised data, he cannot identify users. The second advantage is that when service provider gets these anonymised data, service provider can directly use these data without restoring these data. So this will be more flexible and safe to protect individuals' privacy in the cloud.

6 Avoiding privacy indexing

In order to avoid the disclosure of individuals' privacy in cloud, our research address that the user should send the anonymous data to service provider (SP) in the cloud, and the datasets in SP should meet k-anonymity.

The k-anonymity approach publishes a table T' which changes the values on the quasi-identifier attributes so that every tuple in T is published in a group-by of at least k tuples on the quasi-identifier. Publishing T' instead of T can protect privacy effectively against privacy indexing, since the attacker cannot re-identify any individual with a confidence more than 1/k.

In fact the k-anonymity approach meets the second condition of privacy-preserving index (PPI)'s definition in paper [23]. Because the datasets in SP meet k-anonymity, so the datasets in SP also meet the second condition of privacy-preserving index (PPI)'s definition. If an adversary wants to identify who satisfy the query, this adversary executes query on the datasets of SP, then the returned results contain k users (both true positives users and false positives users are in the k users), therefore the probability that the adversary can identify the real user is 1/k. So our based on k-anonymity approach like the method in paper [23], can avoid privacy indexing issue.

Let us take an example.

If an adversary executes a query on the datasets of SP, and this adversary wants to identify user's phone number (sensitive attribute) through this query. Assume table is dataset of one service provider, and table meets 3-anonymity. Then assume the adversary's query is on table.

The query is (select phone number from table where Age=27 & Zip code=555345 & gender=M). So the returned results will be (8142346789, 8123456789, 2346789090, 8456789090). In these results, 8456789090 is true positives and the other three phone numbers are false positives. So the adversary gets four different phone numbers and this adversary can't identify the user whose other attributes satisfy this query. So executing query on these datasets which meet k-anonymity will avoid privacy indexing issue.

Another reason is that the datasets in SP meet 3-anonymity, so this query will return at least 3 different results. All of these results will contain both true positives and false positives. According to the definition of privacy-preserving index (PPI) in paper [23], our proposed approach can avoid privacy indexing issue.

7 Security analysis

Now we use a formal proof method proposed by [24] to prove the security of proposed protocol.

Firstly we give some definitions of the signs used in the proof process.

PD : Private data.

$PDDM$: Privacy-preserving data mining protocol.

PD_{P_i} : Private data of P_i.

EXT_{P_i} : Extra information P_i can obtain through the underlying protocol.

$GAIN_{P_i}$: Advantage of P_i to get access to any other party's private data using a protocol.

$GAIN_{SEC}$: Advantage of P_i to get access to any other party's private data using a protocol by looking at a semantically secure ciphertext which is negligible when using an RSA [25] type of encryption.

$\Pr(PD)$: Probability of disclosing the private data PD without using privacy preserving protocol.

ε : Level of security.

As stated in [24] that if you want to prove whether our proposed protocol named *PRPT* is privacy preserving, according to any private data, you can find a ε such that:

$$|\Pr(PD \mid PDDM) - \Pr(PD)| \le \varepsilon .$$

So if we want to prove whether the proposed *PRPT* protocol is privacy preserving, we only need to find ε such that:

$$|\Pr(PD \mid PRPT) - \Pr(PD)| \le \varepsilon .$$

Because the advantage of each party P_j by the protocol can access another party's private data can be shown as:

$$GAIN_{P_j} = \Pr(PD_{P_k} \mid EXT_{P_j}, PRPT) - \Pr(PD_{P_k} \mid EXT_{P_j}),$$

$$(k \ne j) .$$

Because each party P_j only runs secure dot product protocol with P_j by using his own randomly generated vector. And party P_j runs secure multi-party addition protocol [26] with other parties by using his private output share. Therefore:

$$GAIN_{P_j} = GAIN_{SEC}, \ (j \ne i) .$$

Because $GAIN_{SEC}$ means that advantage of P_j to get access to any other party's private data, using a protocol by looking at a semantically secure [27] ciphertext which is negligible, when using an RSA type of encryption. So $GAIN_{P_i}$ is also negligible.

Participant P_j by decrypting the message received from other parties and decrypting the signs of their private output, can only know his private value of the final weight vector, which is his own final output. So we can conclude: $\varepsilon = \max(GAIN_{Pi}, GAIN_{P_j}) = GAIN_{P_j}$.

Therefore for each $k, j \in \{1,...,n\}$, $k \ne j$, we can conclude:

$$\Pr(PD_{P_k} \mid EXT_{P_j}, PRPT) - \Pr(PD_{P_k} \mid EXT_{P_j}) \le GAIN_{P_i} = \varepsilon .$$

So at last we can find $\varepsilon = GAIN_{P_i}$, such that $|\Pr(PD \mid PRRT) - \Pr(PD)| \le \varepsilon$. Therefore, the proposed *PRPT* protocol is privacy preserving.

8 Experimental evaluation

In this section we design experiments to show the application and performance of our proposed protocol named *PRPT*. We have used Java language to implement our experiments. The experiments are carried out on Windows XP operating system with 2.13 GHz Intel Core i3 processor and 4 GB of memory. In the following experiments we use IBM Quest Synthetic Data Generator to generate the experimental data.

Figure 1 illustrates that distortion ratio changes depending on the variation of l-value, k-value and i-value. When l-value increases, privacy protection degree increases. The number of different sensitive attributes within the same equivalent group increases, resulting in the generalization level of identifier attributes increased, then leading to substantial information loss from the original data increased, ultimately causing distortion ratio of multiple sensitive attributes set increased. When other conditions are identical, k-value increased, the group size becomes larger. Because the grouped data should meet diverse, attributes in equivalent group increase. Then generalization levels increase and data distortion rate also increases. When the number of data sets and the parameters are equal, the dimension of sensitive attribute i increases, the distortion rate of multiple sensitive attributes is higher, which is caused by sensitive property diversity in each dimension. By contrasting the following figure, we can conclude that *PRPT* protocol by using multiple sensitive attributes generalization can avoid excessive generalization of the identifier attributes, so the total distortion rate is less than the other method. Therefore the total distortion ratio of *PRPT* protocol is smaller than other algorithm. And compared to the l-diversity, k-l-sensitive rules has a relatively low level of data loss. This experiment suggests that *PRPT* protocol has lower information loss than l-diversity algorithm and k-l algorithm. According to the above experimental results analysis, we can conclude that *PRPT* protocol can get better performance than other algorithm under the same conditions.

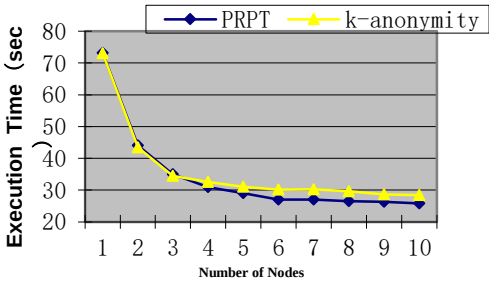

FIGURE 1 Distortion ratio changes depending on l-value changes (k=40, i=3)

9 Conclusions

Cloud computing has been envisioned as the next generation architecture of IT Enterprise. This technology not only gave us more convenience, but also exposed some security problem. When people access cloud service or receive cloud service, they should provide many attributes to ensure this service accomplish successfully. But much private data are included in these attributes. If these sensitive attributes are disclosed, it will bring suffering to people. This paper illustrated the importance of protecting customers' private data in cloud computing. We have argued that it is very important to take privacy into account and we proposed a novel anonymity-based protocol preserving privacy based cloud environment.

Acknowledgements

This research is supported by the Education Science and Technology Project of Henan Province under the grant No. 13A520026 and by the Technology Bureau Project of Zhengzhou City under the grant No. 131PPTGG423-1.

References

[1] Wang J, , Zhao Y, Shuo J, Le J 2009 Providing Privacy Preserving in Cloud Computing *in ICTM Proceedings IEEE* 213-6
[2] Almorsy M 2011 Collaboration-Based Cloud Computing Security Management Framework, *in CLOUD Proceedings IEEE* 364-71
[3] Kretzschmar M, Golling M 2011 Security management spectrum in future multi-provider Inter-Cloud environments-Method to highlight necessary further development *in SVM Proceedings IEEE* 1-8
[4] Bouayad A, Blilat A, El Houda Mejhed N, El Ghazi M 2012 Cloud computing: Security challenges *in CIST Proceedings IEEE* 26-31
[5] Khalil I, Khreishah A, Bouktif S, Ahmad A 2013 Security Concerns in Cloud Computing *in ITNG Proceedings IEEE* 411-6
[6] Kandukuri B, Paturi V, Rakshit A 2009 Cloud Security Issues *in SCC Proceedings IEEE* 517-20
[7] Hamdi M 2012 Security of cloud computing, storage, and networking *in CTS Proceedings IEEE* 1-5
[8] Behl A, Behl K 2012 Security Paradigms for Cloud Computing *in CICSYN Proceedings IEEE* 200-5
[9] Al Awadhi E, Salah K, Martin T 2013 Assessing the security of the cloud environment *in GCC Proceedings IEEE* 251-6
[10] Wang J, Le J A New Privacy Preserving Approach Used in Cloud Computing *Journal of Key Engineering Materials* **439-440**(1) 1318-23 *(in Chinese)*
[11] Francis T, Vadivel S 2012 Cloud computing security: Concerns, strategies and best practices *in ICCCTAM Proceedings IEEE* 205-7
[12] Wang J, Novel 2012 A Anonymity Algorithm for Privacy Preserving in Publishing Multiple Sensitive Attributes *Research Journal of Applied Sciences, Engineering and Technology* **4**(22) 4923-7
[13] Mell P 2012 What's Special about Cloud Security *Journal of IT Professional* **14**(4) 6-8
[14] Viega J 2012 Cloud Security: Not a Problem *Journal of Security & Privacy* **10**(4) 3
[15] Goth G 2011 Public Sector Clouds Beginning to Blossom: Efficiency, New Culture Trumping Security Fears *Journal of Internet Computing* **15**(6) 7-9
[16] Kaufman L M 2009 Data security in the World of Cloud Computing *in Security and Privacy Proceedings IEEE* 61-4
[17] Behl A, Behl K 2012 An analysis of cloud computing security issues *in WICT Proceedings IEEE* 109-14

[18] Wang J, Le J 2010 Novel Privacy Preserving Classification Mining Approach Based Cloud Environment *in BTMC Proceedings IEEE* 236-9
[19] Wang J, Luo Y, Zhao Y, Le J 2009 A Survey on Privacy Preserving Data Mining *in DBTA Proceedings IEEE* 111-4
[20] Zhang Q, Koudas N, Srivastava D, Yu T 2007 Aggregate query answering on anonymized tables *in ICDE Proceedings IEEE* 116-25
[21] Li Y, Tygar JD, Hellerstein J M 2005 Private Matching, in Computer Security *Proceedings Springer* 25-50
[22] Siani P, Yun S, Miranda M 2009 A privacy manager for cloud computing *In CloudCom Proceedings IEEE* 90-106
[23] Mayank B, Rakesh A, Jaideep V 2009 Privacy-preserving indexing of documents on the network *The VLDB Journal* 837-56 *(in Chinese)*
[24] Zhan J, Matwin S 2006 A Crypto-Based Approach to Privacy-Preserving Collaborative Data Mining *in Data Mining Proceedings IEEE* 546-50
[25] Wang S, Liu G 2011 File Encryption and Decryption System Based on RSA Algorithm *in Computational and Information Sciences Proceedings IEEE* 797-800
[26] Liu W, Wang Y, Cao Y 2011 Research on Secure Multi-Party Closest String Problem *in Wireless Communications, Networking and Mobile Computing Proceedings IEEE* 1-4
[27] Lee J, Chang J 2007 Semantically Secure Authenticated Encryption Scheme and the Same Scheme for Ad-hoc Group Called a Ring *in Information Technology Proceedings IEEE*825-30

Authors

Jian Wang, born in 1981, November 26th, China

Current position: an instructor in Henan University of Economics and Law.
University studies: Ph.D. degree from Donghua University, Shanghai, China in June 2011.
Scientific interest: privacy preserving in cloud computing, information security.
Publications: 10 papers.

Le Wang, born in 1983, October 29th, China

Current position: an instructor in SIAS International University.
University studies: Bachelor degree from Zhengzhou University, Zhengzhou, China in June 2005.
Scientific interest: computer science, database application.