2011 Third International Conference on Multimedia Information Networking and Security

Privacy Preserving Multiple Keyword Search for Confidential Investigation of Remote Forensics

Shuhui Hou
Dept. of Information and Computer Science
University of Science and Technology Beijing
Beijing, China
Email: shuhui@ustb.edu.cn

Tetsutaro Uehara
Academic Center for Computing and Media Studies
Kyoto University
Kyoto, Japan
Email: uehara@media.kyoto-u.ac.jp

S.M. Yiu
Dept. of Computer Science
The University of Hong Kong
Hong Kong
Email: smyiu@cs.hku.hk

Lucas C.K. Hui
Dept. of Computer Science
The University of Hong Kong
Hong Kong
Email: hui@cs.hku.hk

K.P. Chow
Dept. of Computer Science
The University of Hong Kong
Hong Kong
Email: chow@cs.hku.hk

***Abstract*—Remote forensics can help investigators perform investigation without need to ship hard drives or travel to a remote location. With increased use of cloud computing technologies, it is becoming more and more difficult to perform post-event forensic investigation. The difficulty consists in that thousands upon thousands of disparate data from different data owners may be stored on a single storage device (e.g., a remote server). To clone a copy of data from the storage device is a costly and time consuming task and may not be easy due to the huge volume of data. Even if it is possible to make a clone, investigating all the data one by one will inevitably result in exposing irrelevant data to the investigators while data owners may be unwilling to expose it because it may involve their privacy information. The other alternative is to let the server administrator search the relevant information and retrieve the data for the investigators provided a warrant can be provided. However, sometimes, the investigators need to keep the investigation subject confidential due to the confidentiality of the crime or the server administrator may be one of the suspects. In this paper, we address how to solve this problem by multiple keyword search over encrypted data, so that the investigators can obtain the necessary evidence while keeping the investigation subject confidential and at the same time, the irrelevant data can be protected from exposing to the investigators.**

***Keywords*-remote forensics; privacy-preserving multiple keyword search; confidential forensic investigation; homomorphic encryption**

I. INTRODUCTION

In recent years, computer crimes are growing and becoming a serious problem for businesses, the public, and government. How to capture digital evidence is critical for counteracting against computer crimes. Remote forensics can help investigators capture evidence without need to ship hard drives or travel to a remote location. With increased use of cloud computing technologies, it is becoming more and more difficult to perform post-event forensic investigation. The difficulty consists in that thousands upon thousands of disparate data from different data owners may be stored on a single storage device (e.g., a remote server), where some of data are relevant to the crime but some are irrelevant. The storage device may only contain a small portion of evidential data and it may not be easy to clone a copy of data from the storage device due to the huge volume of data. Even if it is possible to make a clone, investigating all the data one by one will inevitably result in exposing irrelevant data to the investigators while data owners may be unwilling to expose it, especially in the case that the irrelevant data involves confidential information or privacy information. The other alternative is to let the server administrator search the relevant information and retrieve the data for the investigators provided a warrant can be provided. However, sometimes, the investigators need to keep the investigation subject confidential, that is, the investigators may not want the server administrator to know what information they are looking for due to the confidentiality of the crime or the server administrator may be one of the suspects.

In this paper, we address how to solve this problem by multiple keyword search over encrypted data, so that the investigators can obtain the necessary evidence while keeping the investigation subject confidential and at the same time, the irrelevant data can be protected from exposing to the investigators.

We assume that the evidence required by the investigator is stored together with a huge amount of irrelevant data on a remote server or a distributed set of storage devices. It is not possible to make a clone of all data. The server administrator is willing to cooperate and search the relevant information for the investigator. However, they want to make sure that only relevant information will be given to the investigator, no other information of other users will be disclosed to the investigator. At the same time, the investigator does not want the server administrator to know what information they are searching. We further assume that the server administrator is trustable in the sense that he will not hide any information if it satisfies the searching criteria of the investigator. In other words, the server administrator will give out all the information located.

978-0-7695-4559-2/11 $26.00 © 2011 IEEE
DOI 10.1109/MINES.2011.90

Our main idea to solve this problem is as follows. The server administrator will encrypt all the data stored on the server for preventing the investigator from learning the irrelevant data; the investigator will provide the administrator keywords (which are in an encrypted form for preventing the administrator from learning the investigation subject) and the "trapdoor" so that the administrator can search for the relevant data from the encrypted data; the administrator will only return the relevant data to the investigator and the investigator will only decrypt and perform investigation on such relevant data for capturing the evidence. In our work [1], we proposed two forensically sound schemes by utilizing homomorphic encryption and commutative encryption. However, a limitation common to both schemes is that they only allow the server administrator or the Trusted Third Party to identify the subset of documents that match a certain keyword rather than simultaneous multiple keywords. To obtain fine search results and improve investigation efficiency, it is essential to perform multiple keyword searches since single keyword searches often yield coarse results.

There are a number of studies on searching over encrypted data by keywords. Traditional schemes ([2] [3] [4] [5] [6] [7]) only support single keyword search over encrypted data. To enrich search functionalities, the schemes ([8] [9] [10]) are proposed to support conjunctive keyword search, in which the scheme [10] can also support disjunctive keyword search. Conjunctive keyword search returns "all-or-nothing", which means it only returns those documents in which all the keywords specified by the search query appear; disjunctive keyword search returns undifferentiated results, which means it returns every document that contains a subset of the specific keywords, even only one keyword of interest. In addition, improvements for ranked search over encrypted data are given by Ning cao et al. [11] recently. The common to all these schemes is that they attempt to tackle the same problem: a user wishes to outsource his data to a remote server while preventing the untrusted server from learning the data content. The user can manage his data in a convenient way before outsourcing the data since he is usually the data owner. That is, the data owner can build an encrypted searchable index and provide the server a "trapdoor" such that the server can perform search without learning the data content. Our problem is quite different from this problem because that the investigator is not data owner so he cannot manage the data in any convenient way. On the other hand, both the confidentiality of investigation subject and the privacy of irrelevant server data need to be protected. Briefly, the above schemes with multiple keyword search cannot be used to solve our problem at least cannot be directly used. Our work [1] proposed two forensically sound solutions but it only supports single keyword search, so now we will explore the solutions which support multiple keyword search over encrypted data.

The remainder of the paper is organized as follows. In Section II, we make assumptions to formulate our problem and clarify its requirements. To keep the investigation subject confidential and protect the privacy of irrelevant server data, Section III presents a scheme to support both conjunctive and disjunctive keyword search. How to enhance the security of this scheme is stated in Section IV. Finally, discussions are conducted and conclusions are drawn in Section V.

II. Problem Formulation

We make the following assumptions.

1) The investigator and the administrator do not trust each other. To prevent the administrator (who may be a potential suspect) from learning the investigation subject, the investigator will provide the administrator keywords which are in an encrypted form. To prevent the investigator from obtaining irrelevant data, the administrator will verify what keywords are used later. For example, during the evidence presentation in a court of law, the investigator can be required to show what evidence is collected based on what keywords, so the administrator can check whether the investigator cheated for obtaining other information from the server.
2) Evidential data is stored alongside the irrelevant data on a remote server in non-encrypted form. For simplicity, we view the data as a set of documents and every document W is a series of word blocks which has fixed length as follows:

w_1	$\cdots$	w_i	$\cdots$	w_v

   We assume that every keyword specified by the investigator has the same length as w_i.
3) It is difficult to distinguish the relevant data from the irrelevant ones. We view the documents involving the specified keywords as relevant data and those without involving the specified keywords as irrelevant ones. Then, the investigator is only allowed to perform investigation on documents which involve the specified keywords.
4) Both the keywords specified by the investigator and the data stored on the server are encrypted with the cryptographic scheme, which is assumed to be provably secure in the sense that the server administrator cannot learn anything about the specified keywords when they are encrypted and the investigator cannot learn more than the search result. The search result must involve the specified keywords, so the investigator can treat them as potential evidence.

We formulate our problem in TABLE I, where a scheme supporting multiple keyword search is desired.

A scheme which satisfies the following properties is desirable for our problem.

Inputs	*Investigator*	w^*: set of specified keywords
	Server administrator	D: whole set of server-side documents
Outputs	*Investigator*	Nothing
	Server administrator	$W(\in D)$: document involving w^*
Privacy	*Investigator*	Server administrator cannot learn w^*
	Server administrator	Investigator cannot learn more than W

TABLE I

To keep the investigation subject confidential and protect the privacy of irrelevant data, the set of specified keywords w^* and the set of documents D need to be encrypted. This will lead to a non-index, sequential search on the entire server. Besides, public key encryption is required. Both the investigator and the server administrator can perform encryption but only the one who owns private key can perform decryption.

III. A Scheme on Multiple Keyword Search

After an event involving computer crime has occurred, the investigator or the police usually search for evidence over all the documents stored on the server. However, as the data is irrelevant to the crimes and involves confidential information or privacy information, data owners may be unwilling to reveal it to the investigator. Data owners usually trust the administrator who is responsible for managing the data in a secure manner. Hence, the alternative is to let the administrator perform the searching and only return the relevant data to the investigator. Take the company server as an example, if there are only a few employees suspected, the administrator usually provides the investigator their data rather than all the employees' data. Here, we assume that the administrator honestly returns all the searching results without holding some of them.

For the brevity of description, we adopt the following notation: The set of keywords specified by the investigator is denoted as w^*=$\{w_1^*$, w_2^*, …, $w_u^*\}$, where the length of every keyword w_i^* is l-bit; The whole set of documents stored on the server is denoted as D=$\{W^1$, W^2, …, $W^L\}$ and any document W∈D is denoted as W=$\{w_1$, w_2, …, $w_v\}$, where the length of every word block w_i is l-bit. Obviously, both w^* and W come from the same domain; Encrypting a set means encrypting every element of the set. For example, the encryption of the set w^* and W can be denoted as $E(w^*)$=$\{E(w_1^*)$, $E(w_2^*)$, …, $E(w_u^*)\}$ and $E(W)$=$\{E(w_1)$, $E(w_2)$, …, $E(w_v)\}$, where E is the encryption function.

Using the above notation, we will describe how to realize disjunctive and conjunctive keyword search below.

A. Disjunctive Keyword Search

The administrator realizes disjunctive keyword search over encrypted data by checking if the intersection of $E(w^*)$ and $E(W)$ is nonempty. In other words, the document W involves one or more than one keyword of w^* is equivalent to the intersection of $E(w^*)$ and $E(W)$ is nonempty, i.e., $E(w^*)$∩$E(W)$≠ϕ holds. The procedures are detailed as follows.

1) The investigator performs the following.
   a) After specifying a couple of keywords w^*, he generates a key pair for a homomorphic public key system and sends the public key to the administrator (the public key is only known to the investigator and the administrator). The corresponding encryption is denoted as $E(\cdot)$;
   b) To prevent the administrator from learning the investigation subject, he encrypts the set w^*=$\{w_1^*$, w_2^*, …, $w_u^*\}$ using his public key, takes the logarithm $log_2(\cdot)$ of $E(w^*)$=$\{E(w_1^*)$, $E(w_2^*)$, …, $E(w_u^*)\}$ and denotes it as X=$\{x_1$, x_2, …, $x_u\}$;
   c) To hide the X=$\{x_1$, x_2, …, $x_u\}$, he constructs a polynomial function of degree u, $f(x)$=$(x - x_1)(x - x_2)\cdots(x - x_u)$=$a_0$+$a_1x$+…+$a_ux^u$. It is obvious that $f(x_i) = 0$ (i = 1,2,…,u) holds. Technically, the polynomial function meets the requirement that $f(x) = 0$ if and only if $x \in X$;
   d) He uses the coefficients a_0, a_1, …, a_u to form a private vector and sends this vector α=(a_0, a_1, …, a_u) to the administrator.
2) To prevent the investigator from learning the irrelevant data, the administrator encrypts the data set D with the public key and the resulting data is denoted by $E(D)$=$\{E(W^1)$, $E(W^2)$, …, $E(W^L)\}$. Similary, the encryption of any document W∈D can be denoted by $E(W)$=$\{E(w_1)$, $E(w_2)$, …, $E(w_v)\}$. He takes the logarithm $log_2(\cdot)$ of $E(W)$ and denotes it as Y=$\{y_1$, y_2, …, $y_v\}$. For every y_i∈ Y(i = 1,2,…,v), the administrator performs the following.
   a) He constructs a private vector β_i=(r_i, r_iy_i, …, $r_iy_i^u$), where r_i is a non-zero number;
   b) He computes δ=α·β_i+λ, where "·" means the scalar product of vectors and λ is a number specified by the administrator. Taking λ=0 here, it is obvious y_i∈X if and only if $\delta = 0$. That is, $E(w^*)$∩$E(W)$≠ϕ holds if and only if $\delta = 0$;
   c) He retrieves the encrypted document $E(W)$ and starts searching on next encrypted document if δ = 0 holds. Otherwise, he continues to check y_{i+1}.

   The administrator collects all the encrypted document $E(W)$ in which $\delta = 0$ holds for some β_i and sends them to the investigator.
3) The investigator decrypts the encrypted documents involving one or more than one specified keyword and performs investigation on such decrypted data for capturing evidence.

B. Conjunctive Keyword Search

That the document W contains all the keywords of w^* is equivalent to $E(w^*)$⊆$E(W)$. It is easy to see that $E(w^*)$⊆$E(W)$ if and only if the elements in $E(w^*)$∩$E(W)$ are different and $|E(w^*)$∩$E(W)|$=$|E(w^*)|$, where $|\cdot|$ denotes the cardinality of a set. Since there may exist same word block in a document, we add the condition "the elements belonging to the intersection need to be different" to the above set inclusion relation.

Based on the above procedures of disjunctive keyword search, the administrator can realize conjunctive keyword search by retrieving and collecting the encrypted document $E(W)$ where the β_is which make $\delta = 0$ true are different and the number of such β_is is u. That is, the administrator sends the investigator the encrypted document $E(W)$ which contains all the keywords of $E(w^*)$ ($E(w^*)$⊆$E(W)$).

IV. Improvement of Security

To improve the security, the investigator can encrypt each of the $u+1$ coefficients (a_0, a_1, …, a_u) with the semantically secure homomorphic encryption scheme and send to the administrator the resulting vector of ciphertext, ($E(a_0)$, $E(a_1)$, …, $E(a_u)$).

The administrator realizes disjunctive keyword search over encrypted data by checking if the intersection of $E(w^*)$ and $E(W)$ is nonempty. Given an $E(W) \in E(D)$, the administrator performs the following for every $y_i \in Y (i = 1,2,\ldots,v)$.

1) He uses the homomorphic properties to evaluate the encrypted δ, i.e., $E(\delta)$. Here, $\delta = \alpha \cdot \beta_i$(Taking $\lambda = 0$);
2) He chooses a random value $\widetilde{r}$ and computes $E(\widetilde{r}\delta + y_j)$ ($\forall y_j \in Y$);
3) He retrieves the encrypted document $E(W)$ and starts searching on next encrypted document if $E(\widetilde{r}\delta + y_j) = E(y_j)$ holds for some j (j=1,2,…,v). Otherwise, he continues to check y_{i+1}.

The rest of procedures is as the same as ones in Section III.

Proof of Correctness: we show the correctness of this improvement based on a semantically secure public key encryption scheme that preserves the group homomorphism of addition and allows multiplication by a constant. This property is obtained by Paillier's cryptosystem [12]. Without knowledge of the private key, the Paillier's cryptosystem supports the following operations: (i) Given two encryptions $E(m_1)$ and $E(m_2)$, we can efficiently compute $E(m_1 + m_2)$; (ii) Given some constant k belonging to the same group, we can compute $E(km)$. We will show the correctness by using Paillier's cryptosystem shown in Fig. 1. Without special remarks, the notation in [13] is used directly below.

The investigator encrypts coefficients (a_0, a_1, …, a_u) with the Paillier's cryptosystem (the random number $r = 1$) by computing

$$\begin{aligned} E(a_0) = g^{a_0} \cdot r^n = g^{a_0} \ mod \ n^2 \\ \cdots \ \cdots \ \cdots \\ E(a_u) = g^{a_u} \cdot r^n = g^{a_u} \ mod \ n^2. \end{aligned} \tag{1}$$

Then, he sends the resulting coefficients ($E(a_0)$, $E(a_1)$, …, $E(a_u)$) to the administrator. The administrator evaluates an encryption of δ by computing

$$\begin{aligned} E(\delta) &= g^{\delta} = g^{\alpha \cdot \beta_i} = g^{a_0 r_i + a_1 r_i y_i + \ldots + a_u r_i y_i^u} \ mod \ n^2 \\ &= [(g^{a_0})(g^{a_1})^{y_i} \cdots (g^{a_u})^{y_i^u}]^{r_i} \ mod \ n^2 \\ &= [E(a_0)E(a_1)^{y_i} \cdots E(a_u)^{y_i^u}]^{r_i} \end{aligned} \tag{2}$$

The administrator knows ($E(a_0)$, $E(a_1)$, …, $E(a_u)$), y_i and r_i, so he can evaluate $E(\delta)$ by computing $[E(a_0)E(a_1)^{y_i} \cdots E(a_u)^{y_i^u}]^{r_i}$.

By the homomrophic property of Paillier's cryptosystem, the administrator computes $E(\widetilde{r}\delta + y_j)$ as follows.

$$\begin{aligned} E(\widetilde{r}\delta + y_j) &= E(\widetilde{r}\delta)E(y_j) \\ &= (g^{\delta})^{\widetilde{r}} E(y_j) \\ &= E(y_j) \qquad if \ \delta = 0, \ i.e., \ y_i \in X \end{aligned} \tag{3}$$

Thus, the administrator can check if $y_i \in X$ (i.e., $E(w^*) \cap E(W) \neq \phi$) even in an encrypted form.

The encryption processes are applied twice here, so the security is improved. Similar to the Section III, such improved disjunctive keyword search can be easily generalized to the case of conjunctive keyword search.

V. Discussion and Conclusion

A. Discussion

To improve the efficiency in forensic investigation, the investigator is supposed to capture evidence only from the relevant data in our solutions. Through multiple keyword (specified by investigation subject) search on encrypted data (stored on the server), we realized that the investigator can search for evidence without learning any information of irrelevant data and the server administrator cannot learn the investigation subject. Obviously, whether the confidentiality of the investigation subject and the privacy of irrelevant server data can be completely protected relies on the security of cryptosystem.

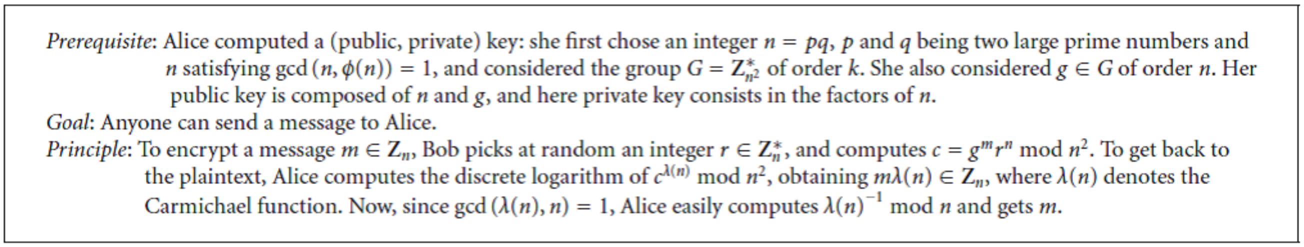
In the above schemes, we assumed that the document can be easily broken into a sequence of words of a fixed length. However, this assumption might not be true in a normal file. To deal with variable-length words, we can pick a fixed-size block that is long enough to contain most words like the work [2], where words that are too short or too long may be padded to a multiple of the block size with some pre-determined padding format.

B. Conclusions

In this paper, we formulated the problem: in the remote forensic investigation, the investigator may need to keep the investigation confidential and have no right to access the irrelevant data, especially in the case that this data involves privacy information or confidential information. Based on multiple keyword search over encrypted data, we addressed solutions to this problem so that the investigator can obtain the necessary evidence while keeping the investigation subject confidential and at the same time, the irrelevant data can be protected from exposing to the investigator. For future work, we will consider how to implement them and verify their feasibility.

Acknowledgment

The work described in this paper was partially supported by the Fundamental Research Funds for the Central Universities (No. 06108041) and partially supported by the General Research Fund from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. RGC GRF HKU 713009E), the NSFC/RGC Joint Research Scheme (Project No. N_HKU 722/09), and HKU Seed Fundings for Basic Research 200811159155 and 200911159149.



Prerequisite: Alice computed a (public, private) key: she first chose an integer $n = pq$, p and q being two large prime numbers and n satisfying $\gcd(n, \phi(n)) = 1$, and considered the group $G = \mathbb{Z}^*_{n^2}$ of order k. She also considered $g \in G$ of order n. Her public key is composed of n and g, and here private key consists in the factors of n.

Goal: Anyone can send a message to Alice.

Principle: To encrypt a message $m \in \mathbb{Z}_n$, Bob picks at random an integer $r \in \mathbb{Z}^*_n$, and computes $c = g^m r^n \bmod n^2$. To get back to the plaintext, Alice computes the discrete logarithm of $c^{\lambda(n)} \bmod n^2$, obtaining $m\lambda(n) \in \mathbb{Z}_n$, where $\lambda(n)$ denotes the Carmichael function. Now, since $\gcd(\lambda(n), n) = 1$, Alice easily computes $\lambda(n)^{-1} \bmod n$ and gets m.

Fig. 1. Paillier cryptosystem in [13]

References

[1] Shuhui Hou, Tetsutaro Uehara, S.M. Yiu, Lucas C.K. Hui, and K.P. Chow, “Privacy Preserving Confidential Forensic Investigation for Shared or Remote Servers”, 2011 Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2011), in press.

[2] D. Song, D. Wagner, and A. Perrig, “Practical Techniques for Searches on Encrypted Data”, in Proceedings of IEEE Symposium on Security and Privacy 2000, pp.44-55.

[3] Y.C. Chang, and M. Mitzenmacher, “Privacy Preserving Keyword Searches on Remote Encrypted Data”, Cryptology ePrint Archive, Report 2004/051, Feb 2004.

[4] Reza Curtmola, Juan Garay, Seny Kamara and Rafail Ostrovsky, “Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions”, in CCS, 2006, pp.79-88.

[5] Eu-jin Goh, “Secure Indexes”, in the Cryptology ePrint Archive, Report 2003/216, March 2003.

[6] Dan Boneh, Giovanni Di Crescenzo, Rafail Ostrovsky, and Giuseppe Persiano, “Public Key Encryption with Keyword Search”, in Proceedings of Eurocrypt 2004 (Lecture Notes in Computer Science 3027), pp.506-522.

[7] Dan Boneh, Eyal Kushilvitz, Rafail Ostrovsky, and William E. Skeith III, “Public Key Encryption That Allows PIR Queries”, CRYPTO 2007, pp.50-67.

[8] P. Golle, J. Staddon, and B. R. Waters, “Secure Conjunctive Keyword Search over Encrypted Data”, in Proc. of ACNS, 2004, pp.31-45.

[9] Dan Boneh, and Brent Waters, “Conjunctive, Subset, and Range Queries on Encrypted Data”, in Proc. of TCC, 2007, pp.535-554.

[10] Emily Shen, Elaine Shi, and Brent Waters, “Predicate Privacy in Encryption Systems”, Lecture Notes in Computer Science, Volume 5444, 2009, pp.457-473.

[11] Ning Cao, Cong Wang, Ming Li, Kui Ren, and Wenjing Lou, “Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data”, IEEE INFOCOM 2011, pp.10-15.

[12] Pascal Paillier, “Public-Key Cryptosystems Based on Composite Degree Residuosity Classes”, Advances in Cryptology EUROCRYPT 99, vol.1592, 1999, pp.223-238.

[13] Caroline Fontaine, and Fabien Galand, “A Survey of Homomorphic Encryption for Nonspecialists”, EURASIP Journal on Information Security, vol.2007, 2007, pp.1-10.