This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2015.2435697, IEEE Transactions on Information Forensics and Security

Security Analysis on One-to-Many Order Preserving Encryption Based Cloud data Search

Ke Li, Weiming Zhang, Ce Yang, and Nenghai Yu

***Abstract*—For ranked search in encrypted cloud data, order preserving encryption (OPE) is an efficient tool to encrypt relevance scores of the inverted index. When using deterministic OPE, the ciphertexts will reveal the distribution of relevance scores. Therefore, Wang et al. [16] proposed a probabilistic OPE, called One-to-Many OPE, for applications of searchable encryption, which can flatten the distribution of the plaintexts. In this paper, we proposed a differential attack on One-to-Many OPE by exploiting the differences of the ordered ciphertexts. The experimental results show that the cloud server can get a good estimate of the distribution of relevance scores by a differential attack. Furthermore, when having some background information on the outsourced documents, the cloud server can accurately infer the encrypted keywords by using the estimated distributions.**

***Keywords*—searchable encryption, order preserving encryption, privacy, cloud computing**

I. Introduction

Nowadays users connected to the Internet may store their data on cloud servers and let the servers manage or process their data. They can enjoy convenient and efficient service without paying too much money and energy, as one of the most attractive feature of cloud computing is its low cost [1].

However, no matter how advantageous cloud computing may sound, large number of people still worry about the safety of this technology. If cloud server get direct access to all these users' data, it may try to analyse the documents to get private information. The initial purpose of this action may be kind. The server wants to provide better service by digging into these data and then displaying customer-oriented advertisement, which could be convenient but also annoying. Besides, when we consider sensitive data such as personal health records and secret chemical ingredients, the situation becomes even more serious [2]. Theoretically, the server is not supposed to have access to sensitive data at all; therefore we should ensure the server has no access to leaking these data to an untrusted third party. Thus, sensitive data have to be encrypted before being outsourced to a commercial public cloud [3].

This work was supported in part by the Natural Science Foundation of China under Grant 61170234 and Grant 60803155, by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDA06030601, and by the Funding of Science and Technology on Information Assurance Laboratory under Grant KJ-13-02.

Copyright (c) 2013 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org

K. Li, W. Zhang, C. Yang, R. Liu, and N. Yu are with CAS Key Laboratory of Electromagnetic Space Information, University of Science and Technology of China, Hefei, 230026, China.

However, encryption on sensitive data presents obstacles to the processing of the data. Information retrieval becomes difficult in the encrypted domain because the amount of outsourced files can be very large and traditional search patterns can not be deployed to ciphertext retrieval directly. Users need to download all the data, decrypt it all, and then search keywords like plaintext retrieval. To overcome this, Searchable Encryption (SE) [4] was proposed to make query in the encrypted domain possible while still preserving users' privacy. There are several problems in searchable encryption: fuzzy search, ranked search, multi-keyword search and so on. D. Song et al. [5] first proposed a search scheme only supporting single Boolean keyword search. After that plenty of searchable encryption methods [6] [7] [8] [9] arose to improve efficiency and reduce communication overhead.

Applying order preserving encryption (OPE) [10] is one practical way of supporting fast ranked search. This algorithm was first proposed in 2004 to solve encrypted query problems in database systems. OPE is a symmetric cryptosystem, therefore it is also called order-preserving symmetric encryption (OPSE). The order-preserving property means that if the plaintexts $x_1 < x_2$, then the corresponding ciphertexts $E(x_1)$ and $E(x_2)$ satisfy $E(x_1) < E(x_2)$.

Boldyreva et al. initiated the cryptographic study of OPE schemes [11], [12], in which they defined the security of OPE and proposed a provably secure OPE scheme. However, the security definition and the constructions of OPE in [11], [12] are based on the assumption that OPE is a deterministic encryption scheme which means that a given plaintext will always be encrypted as a fixed ciphertext. However, deterministic encryption leaks the distribution of the plaintexts, so it cannot ensure data privacy in most applications. For instance, in privacy-preserving keywords search, OPE is used to encrypt relevance scores in the inverted index [16]. As noted by Wang et al. [16], when using a deterministic OPE, the resulting ciphertext shares exactly the same distribution as the relevance score, by which the server can specify the keywords [14] [15]. Therefore, Wang et al. [16] improved the OPE in [11] and proposed a "One-to-Many OPE" in their secure keyword search scheme, where they tried to construct a probabilistic encryption scheme and conceal the distribution of the plaintexts.

However, we discover that the One-to-Many OPE [16] cannot ensure the expected security. In fact, although the ciphertexts of One-to-Many OPE conceals the distribution of the plaintexts, an adversary may estimate the distribution from the differences of the ciphertexts. So in this paper, we propose a differential attack on the One-to-Many OPE. Our

1556-6013 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2015.2435697, IEEE Transactions on Information Forensics and Security

experimental results show that, when applying this attack to the secure keyword search scheme of [16], the cloud server can get an estimation of the distribution of the the relevance scores, and furthermore accurately reveal the encrypted keywords.

The rest of this paper is organized as follows. We first describe the plaintext search model and ciphertext search model in Section II. Then, in Section III, the basic OPE, One-to-Many OPE, and privacy requirement in cloud computing are briefly reviewed. We elaborate on differential attack on One-to-Many OPE and further attack with background information of outsourced data in Section IV and Section V respectively. Finally the conclusion is given in Section VI.

II. Searching Model

A. Plaintext Searching Model

In practice, to realize effective data retrieval on large amount of documents, it is necessary to perform relevance ranking on the results. Ranked search can also significantly reduce network traffic by sending back only the most relevant data. In ranked search, the ranking function plays an important role in calculating the relevance between files and the given searching query. The most popular relevance score is defined based on the model of $TF \times IDF$, where term frequency (TF) is the number of times a term (keyword) appears in a file and inverse document frequency (IDF) is the ratio of the total number of files to the number of files containing the term. There are many variations of $TF \times IDF$-based ranking functions, and in [16], the following one is adopted.

$$Score(w, F_d) = \frac{1}{|F_d|} \cdot (1 + \ln f_{d,w}) \cdot \ln\left(1 + \frac{N_d}{f_w}\right) \quad (1)$$

Herein, w denotes the keyword and $f_{d,w}$ denotes the TF of term w in file F_d; N/f_w denotes IDF where f_w is the number of files that contain term w and N_d is the total number of documents in the collection; and $|F_d|$ is the number of indexed terms containing in file F_d, i.e., the length of F_d.

To realize fast search, the keywords, IDs of files, and the relevance scores are usually organized as an index structure named "Inverted Index". An example on posting list of the Inverted Index is shown in TABLE I. With a complete Inverted Index, the server can complete retrieval task by simply comparing the relevance scores stored in the index which represent the importance level of each file for a certain keyword.

TABLE I
Example of Posting List of the Inverted Index

Keyword	w			
File ID	F_1	F_2	...	F_{f_w}
Relevance Score	8.6	6.1	...	7.3

B. Ciphertext Searching Model

Due to the special background of cloud computing, unlike traditional plaintext information retrieval, there are usually three entities in cloud data retrieval as shown in Fig. 1:

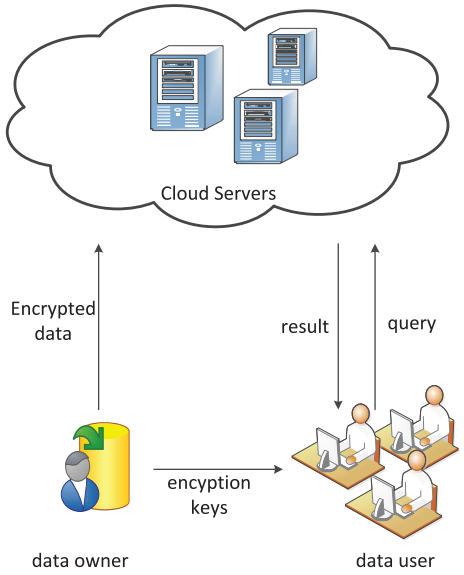


Fig. 1. Framework of retrieval over encrypted cloud data

data owner, remote cloud server and users. A data owner can be an individual or a corporation, i.e., it is the entity that owns a collection of documents $\mathcal{D}_c = \{D_1, D_2 \ldots D_{N_d}\}$ that it wants to share with trusted users. The keyword set is marked as $\mathcal{W} = \{w_1, w_2 \ldots w_{N_w}\}$. For security and privacy concerns, documents have to be encrypted into $\xi = \{E(D_1), E(D_2) \ldots E(D_{N_d})\}$ before being uploaded to the cloud server. Additionally, the plaintext index has to be encrypted into $\mathcal{I}$ to prevent information leakage.

The encrypted form of the example of the posting list of the Inverted Index is shown in TABLE II, in which the keyword w_i is protected by a Hash function $hash()$, and the relevance scores are encrypted by a encryption scheme $E'()$.

TABLE II
Example of Encrypted Posting List of the Inverted Index

Keyword	$hash(w)$			
File ID	F_1	F_2	...	F_{f_w}
Relevance Score	$E'(8.6)$	$E'(6.1)$	...	$E'(7.3)$

We use TABLE II as an example to see how a cloud server conducts a secure search based on an encrypted index. In the search procedure, a user first generates a search request in a secret form — a trapdoor $\mathcal{T}(w)$. In this example, the trapdoor is just the hash values of the keyword of interest.

Once the cloud server receives the trapdoor $\mathcal{T}(w)$, it compares it with the hash values of all keywords in the index $\mathcal{I}$, then the desired documents which are corresponding to keyword w are found. Next, the server returns the matched file IDs: F_1, F_2, ... , F_{f_w} to the user. Finally, the user can download all the encrypted documents based on the given IDs and decrypt them. A desirable system is supposed to return the documents in a ranked order by their relevance with the

1556-6013 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2015.2435697, IEEE Transactions on Information Forensics and Security

queried keyword, but using traditional encryption schemes will disorder relevance scores. Therefore, in [16] Order Preserving Encryption (OPE) is applied to encrypt the relevance scores, which enables the server to quickly perform ranked search without knowing the plain relevance scores.

III. OPE VS. ONE-TO-MANY OPE

A. OPE

OPE is a symmetric cryptosystem, so it is also called order-preserving symmetric encryption (OPSE). The order-preserving property means that if the plaintexts have such a relationship as $x_1 < x_2$, then the corresponding ciphertexts $E(x_1)$ and $E(x_2)$ satisfy $E(x_1) < E(x_2)$.

Boldyreva et al. [11] initiated the cryptographic study of OPE schemes, and they defined the security of an OPE scheme using the ideal object. Note that any order-preserving function g from domain $\mathcal{D} = \{1, 2, \cdots, M\}$ to range $\mathcal{R} = \{1, 2, \cdots, N\}$ can be uniquely defined by a combination of M out of N ordered items. The ideal object is just a function that is randomly selected from all order-preserving functions, which is called a random order-preserving function (ROPF). Thus, with the spirit of pseudorandom functions, an OPE scheme is defined to be secure if the adversary cannot distinguish the OPE from the ROPF. In [11], the authors also constructed an efficient OPE scheme satisfying this secure criterion. The construction is based on the relation between the random order-preserving function and the hyper-geometric probability distribution (HGD), and a HGD simpler is used to select an order-preserving function in a pseudorandom manner.

In the OPE scheme of [11], the range $\mathcal{R}$ is divided into some nonoverlapping interval buckets with random sizes. The random-sized bucket is determined by a binary search based on a random HGD sampler. In [16], the procedure of binary search is described as Algorithm 1, where $TapeGen()$ is a random coin generator.

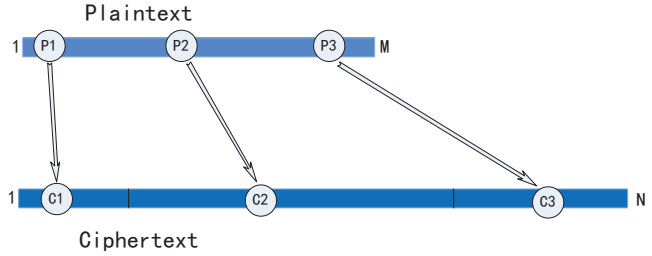
Algorithm 1 BinarySearch

Input: $\{K, \mathcal{D}, \mathcal{R}, m\}$

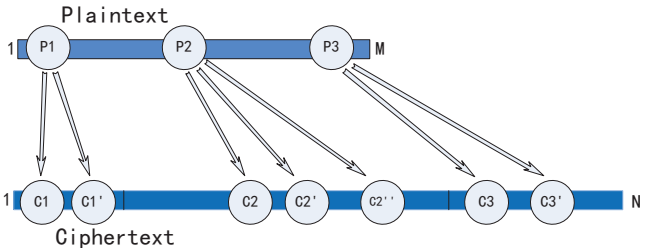
$M \leftarrow length(\mathcal{D})$; $N \leftarrow length(\mathcal{R})$
$d \leftarrow min(\mathcal{D}) - 1$; $r \leftarrow min(\mathcal{R}) - 1$
$y \leftarrow r + ceil(N/2)$
$coin \xleftarrow{R} TapeGen(K, (\mathcal{D}, \mathcal{R}, y||0))$
$x \xleftarrow{R} d + HGD(coin, M, N, y - r)$
$x = d + f$
if $m \leq x$ **then**
  $\mathcal{D} \leftarrow \{d + 1, \cdots, x\}$
  $\mathcal{R} \leftarrow \{r + 1, \cdots, y\}$
else
  $\mathcal{D} \leftarrow \{x + 1, \cdots, d + M\}$
  $\mathcal{R} \leftarrow \{y + 1, \cdots, r + N\}$
end if

Output: $\{\mathcal{D}, \mathcal{R}\}$

After the binary search, a plaintext m is mapped into a bucket in the range $\mathcal{R}$, and then the OPE algorithm assigns a fixed value in the bucket as the encrypted value of m. The encryption process of Algorithm 1 is illustrated in Fig. 2(a), which shows that a given plaintext m_i will always be mapped to a fixed ciphertext c_i belonging to a bucket selected by the binary search procedure, therefore it is a deterministic encryption.



(a) Deterministic OPE



(b) One-to-Many OPE

Fig. 2. Comparison between deterministic OPE and One-to-Many OPE

B. One-to-Many OPE

Wang et al. [16] noted that, in applications of privacy-preserving keyword search, if a deterministic OPE is used to encrypt relevance scores, the ciphertexts will share exactly the same distribution as its plain counterpart, by which the server can specify the keywords.

Therefore, Wang et al. [16] modified the original OPE [11] to a probabilistic one, called "One-to-Many OPE". For a given plaintext m, i.e., a relevance score, the "One-to-Many OPE" first employs Algorithm 1 to select a bucket for m, and then randomly chooses a value in the bucket as the ciphertext. The randomly choosing procedure in the bucket is seeded by the unique file IDs together with the plaintext m, and thus the same relevance score in the Inverted Index will be encrypted as different ciphertexts. The encryption process of "One-to-Many OPE" is described in Algorithm. 2 [16], which is also illustrated in Fig. 2(b).

Example 1: To compare the encrypted results of OPE and One-to-Many OPE, we take the posting list of keyword "weather" as example that is generated from the the TREC data [17]. The relevance scores are encoded into integers, from which we get the plaintext distribution shown in Fig. 3(a). The distributions of the encrypted results obtained by deterministic OPE and One-to-Many OPE are shown in Fig. 3(b) and Fig. 3(c) respectively.

Comparing Fig. 3(b) with Fig. 3(a), we can see that deterministic OPE makes the plaintexts and the ciphertexts share

1556-6013 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2015.2435697, IEEE Transactions on Information Forensics and Security

Algorithm 2 One-to-Many order preserving encryption

Input: $\{K, \mathcal{D}, \mathcal{R}, m, id(F)\}$
  while $|\mathcal{D}|! = 1$ **do**
    $\{\mathcal{D}, \mathcal{R}\} = binarysearch(K, \mathcal{D}, \mathcal{R}, m)$
  end while
  $coin \xleftarrow{R} TapeGen(K, (\mathcal{D}, \mathcal{R}, 1||m, id(F)))$
  $c \xleftarrow{coin} \mathcal{R}$
  $c = round(coin)$
Output: c

the same distribution, which would make it too easy for an attacker to get the exact keyword's information. As shown in Fig. 3(c), when the size of the ciphertext domain is large enough (such as 2×10^6), One-to-Many OPE can flatten the distribution of plaintexts. In fact, with One-to-Many OPE, almost all encrypted values appear only once except for a few appearing twice or more.

C. Privacy threat models

The purpose of both OPE and One-to-Many OPE is to prevent information leakage to the cloud server. The cloud server is considered as "semi-honest", also called "honest but curious" [18]. Specifically, the cloud server will not attempt to remove encrypted data files or index from the storage, and it will also correctly follow the designed protocol specification and execute the procedure faithfully. However, it is curious to handle the stored data and tries to analyze the data to learn additional information.

When talking about the "honest but curious" model, usually there are two attack models "**Known Ciphertext Model**" and "**Known Background Model**" [19].

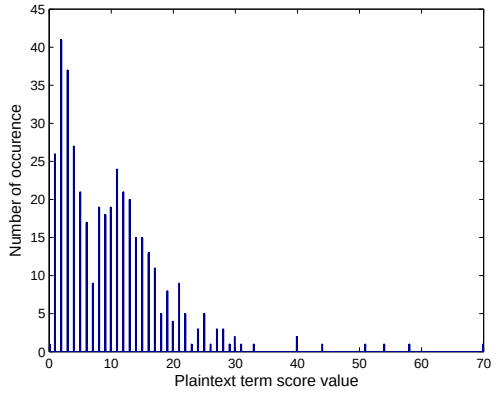
"**Known Ciphertext Model**" assumes that the cloud server can only get access to the encrypted files and the encrypted index. In this model the server can only dig into the ciphertexts without any other background information, and thus security means that the keywords and documents information are strictly protected and there is no indirect way to speculate these information.

"**Known Background Model**" is closer to the real-world situation in the cloud application. The cloud server is supposed to possess more knowledge than what can be accessed in the Known Ciphertext Model. It may intentionally collect related statistical information about the outsourced documents, and with this information the server can infer more sensitive information.

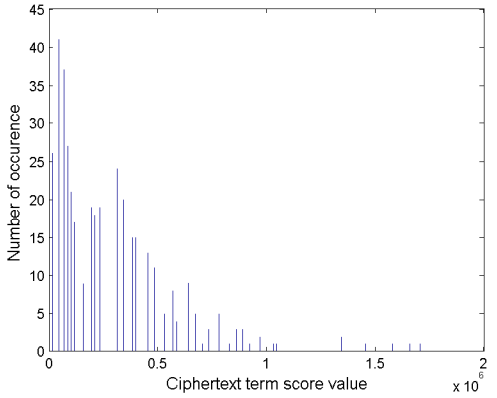
Next, we will propose attacking methods on One-to-Many OPE under these two threat models respectively.

IV. Differential Attack on One-to-Many OPE under Known-Ciphertext-Model

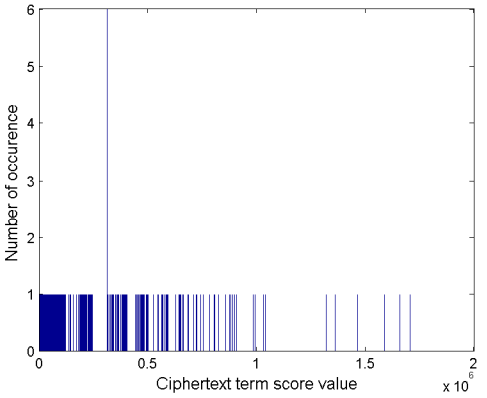
In Fig. 3(c), it can be seen that One-to-Many OPE has successfully hidden the distribution of the plaintexts, but the security of One-to-Many OPE has not endured strict crypt-analysis. In this section, we will show that, by analyzing the differences between the ciphertexts, the cloud server can get an estimation on the distribution of the plaintexts.



(a) Plaintext distribution



(b) Ciphertext Distribution of deterministic OPE



(c) Ciphertext distribution of One-to-Many OPE

Fig. 3. Comparison between plaintext distribution and ciphertext distribution obtained by two kinds of OPE

As shown in Fig. 2(b), each plaintext value m is mapped into many possible ciphertexts belonging to a fixed bucket, and the ciphertext is randomly selected in the bucket. Therefore, the scatter of ciphertexts in a bucket will be dense for a plaintext value with high frequency, but will be sparse for a plaintext value with low frequency. Although the sizes of the buckets are randomly determined, the density of ciphertexts in each bucket will vary according to the frequency of the corresponding plaintext, and thus the profile of the plaintexts'

1556-6013 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2015.2435697, IEEE Transactions on Information Forensics and Security

frequency can be portrayed by the density of ciphertexts. Note that the density of ciphertexts can be revealed by the differences between the neighboring ciphertexts that we call "differential ciphertexts". In other words, if we can locate the change points of the distribution of the differential ciphertexts, we can determine the boundaries of the buckets in the ciphertext range $\mathcal{R} = \{1, 2, \cdots, N\}$. With these boundaries, the histogram of the plaintexts can be easily estimated by counting the number of ciphertexts belonging to each bucket. Therefore, the cloud server may reconstruct the distribution of plaintexts from the differential ciphertexts, which we call "differential attack".

The key issue in "differential attack" is locating the change point in the differential sequence of the ciphertexts. There are many statistical methods to realize such Change Point Analysis (CPA), and we use the cumulative sum (CUSUM) based CPA [21] to describe the procedure of "differential attack", which consists of six steps.

1. Sort the encrypted values:

Suppose that the original ciphertext sequence is c_1, c_2, ..., c_L. Sort the ciphertext sequence in ascending order, and get $c_{i_1} \leq c_{i_2} \leq \cdots \leq c_{i_L}$.

2. Generate the differential sequence:

The differential sequence $\{d_i, 1 \leq i \leq L-1\}$ of the ordered ciphertexts is obtained by calculating $d_1 = c_{i_2} - c_{i_1}$, $d_2 = c_{i_3} - c_{i_2}$, ..., $d_{L-1} = c_{i_L} - c_{i_{L-1}}$.

3. Generate CUSMU sequence:

To get the CUSUM of d_i $(1 \leq i \leq L-1)$, we first compute their average value:

$$\bar{d} = \frac{1}{L-1} \sum_{i=1}^{L-1} d_i. \tag{2}$$

Set the initial value of cumulative sum as $S_0 = 0$. The other cumulative sum values are calculated in a recursion way such that

$$S_i = S_{i-1} + (d_i - \bar{d}), \quad i = 1, 2, ..., L-1. \tag{3}$$

The CUSUM is defined as the cumulative sum of each data minus the average value, so the final value should always be zero, i.e., $S_{L-1} = 0$. A CUSUM chart can be obtained by drawing the cumulative sum S_i in order for $0 \leq i \leq L-1$. If there is a period of data which is greater than the average value, an ascending curve will occur on the chart; otherwise, a descending curve will occur on the chart. A change point on the chart refers to a sudden change in the curve. In Fig. 4, we depict the CUSUM chart of the differential sequence of ciphertexts obtained by One-to-Many OPE in Example 1, which shows that a change point took place.

4. Locating one change point

To be sure that a change indeed took place, we determine a confidence level by performing a bootstrap analysis. Define

$$S_{\max} = \max_{0 \leqslant i \leqslant L-1} S_i, \tag{4}$$

$$S_{\min} = \min_{0 \leqslant i \leqslant L-1} S_i, \tag{5}$$

$$S_{diff} = S_{max} - S_{min}. \tag{6}$$

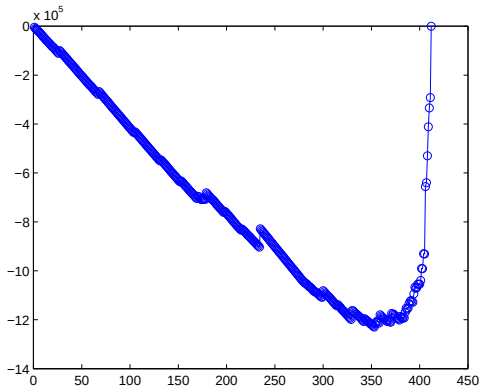


Fig. 4. CUSUM graph of differential sequence of One-to-Many OPE

S_{diff} is an estimator of the changing magnitude, with which a single bootstrap is performed by following four steps:

1) Generate a bootstrap sample of $L-1$ units, denoted as $x_1, x_2, \cdots, x_{L-1}$, by randomly reordering the original $L-1$ differential values $d_1, d_2, \cdots, d_{L-1}$.

2) Calculate the CUSUM of the bootstrap sample, denoted as $S_0^0, S_1^0, \cdots, S_{L-1}^0$.

3) Calculate the changing magnitude of the bootstrap sample S_{diff}^0.

4) Determine whether S_{diff}^0 is less than the original changing magnitude S_{diff}.

The bootstrap sample mimics the behavior if no change has occurred. Perform N_b bootstraps, where N_b is large enough, and let U be number of bootstraps for which $S_{diff}^0 < S_{diff}$. Then the confidence level that a change occurred is defined as follows:

$$ConfidenceLevel = \frac{U}{N_b}. \tag{7}$$

If the $ConfidenceLevel$ is larger than a pre-set threshold β (typically $\beta = 0.9$ or 0.95), we determine that a change has occurred. After detecting a change, the change point v_1 is determined by

$$v_1 = \arg\max_{0 \leqslant i \leqslant L-1} |S_i|. \tag{8}$$

5. Locating other change points

Assume that Step **4** outputs one change point v_1, which thus divides the sequence $d_1, d_2, \cdots, d_{L-1}$ into two subsequences: d_1, d_2, ... , d_{v_1} and d_{v_1+1}, d_{v_1+2}, ... , d_{L-1}. Then we take change point analysis, i.e., Steps **3** and **4**, on these two subsequences respectively. Assume that change points, v_2 and v_3, are detected in the two subsequences respectively. Obviously, $v_2 < v_1 < v_3$, which can divide the original sequence into four subsequences. Take change point analysis on these four subsequences and output eight change points...and so on. This process will end when we cannot detect change points in any subsequence. Assume that B change points in total are found, denoted by $v_1, v_2, \cdots, v_B$.

6. Generating the estimated histogram of plaintexts

Sort the change points $v_1, v_2, \cdots, v_B$ in ascending order, and get $v_{i_1} < v_{i_2} < \cdots < v_{i_B}$. Let $v_{i_0} = 0$ and $v_{i_{B+1}} = N$, and then the ciphertext range $\mathcal{R} = \{1, 2, \cdots, N\}$ can be

1556-6013 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

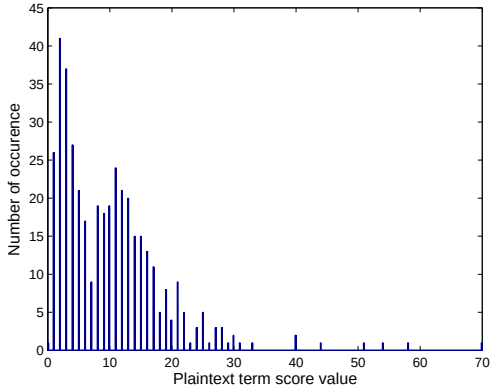
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2015.2435697, IEEE Transactions on Information Forensics and Security

divided into $B+1$ disjointed intervals: $[v_{i_0}+1, v_{i_1}], [v_{i_1}+1, v_{i_2}], \cdots, [v_{i_B}+1, v_{i_{B+1}}]$. Count the number of ciphertexts dropped in each interval:

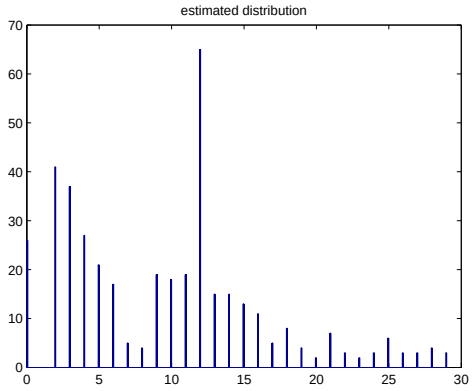
$$h_k = \left|\left\{c_j | v_{i_{k-1}} + 1 \leqslant c_j \leqslant v_{i_k}, 1 \leqslant j \leqslant L\right\}\right|, 1 \leq k \leq B+1. \tag{9}$$

Let h_k be the height of the kth bin, and we get a histogram with $B+1$ bins, which is just the estimation of the distribution of the plaintexts.

Fig. 5(b) is the estimated distribution of the relevance scores of keyword "weather" in Example 1. Comparing with Fig. 5(a), we can see that, by differential attack, the cloud server can find a similar profile of the distribution of the relevance scores with limited errors. Note that this attack is executed in "**Known Ciphertext Model**", because the cloud server only needs to observe the ciphertexts.



(a) Original distribution



(b) Estimated distribution

Fig. 5. Comparison of Original and Estimated distribution

V. Further Attack under Known-Background-Model

A. Identifying the keywords

In this subsection, we will show that, if the cloud server has some background knowledge of the stored data, it can even infer what the keyword is based on the estimated distribution of the relevance scores.

If the curious server knows what the encrypted documents are roughly about, it can collect many relative documents using a tool such as a web crawler, and get a mimic document collection.

For instance, suppose that a server wants to attack an encrypted dataset whose documents are from website $english.cntv.cn/news/sports$, and the attacker has priori knowledge that these documents are about sports news. Then it can conduct a document mining work on another similar website $www.chinadaily.com.cn/spoorts$ to get a mimic document set. As sports news in a short period share high similarity, we can assume that the distributions of keywords from two data sets are remarkably similar and this imitation has high accuracy.

Based on these, the cloud server can then generate an Inverted Index for the mimic document collection. Assume that there are N_w keywords of interest in this Inverted Index. For the ith keyword $\bar{w}_i$, the cloud server can calculate the histogram of the relevance scores in the corresponding posting list, denoted as H_i. Take H_i as the feature of the keyword $\bar{w}_i$.

On the other hand, for the encrypted keyword $hash(w)$ in the encrypted Inverted Index, the cloud server can get an estimated histogram of the relevance scores by using differential attack, denoted as H_w. Then the cloud server can guess what $hash(w)$ is by comparing H_w with H_i for $1 \leq i \leq N_w$.

If the top $k\%$ most similar features to H_w are H_{i_1}, H_{i_2}, $\cdots$, H_{i_J}, the cloud server can be confident that the keyword w belongs to the set $\{\bar{w}_{i_1}, \bar{w}_{i_2}, \cdots, \bar{w}_{i_J}\}$. When such inference is correct, the smaller k or J is, the more information on w is leaked to the cloud server. In fact, if the cloud server has enough background knowledge, it can accurately identify what the keyword w is, i.e., $J = 1$.

We adopt relative entropy here to evaluate the similarity between H_w and H_i. Relative entropy is also called K-L divergence, which is a measurement of the similarity of two probability distributions. For discrete probabilistic distribution P and Q, their relative entropy is defined as follows:

$$D_{KL}(P||Q) = \sum_i P(i) ln \frac{P(i)}{Q(i)}, \tag{10}$$

where all $Q(i)$ and $P(i)$ should be larger than zero and the default value of $0ln0$ is 0.

Because change point analysis on ciphertext differences contains inevitable deviation, the estimated distribution includes location errors and height errors, we should not directly use histograms of H_w and H_i to compute relative entropy. Instead, we should regularize both H_w and H_i to z bins according to the procedure of CPA, and then compare the regularized histograms. Herein we call z as *BINsize*. The detailed regularization procedure is as follows:

From the recursive change point analysis on ciphertext differences we can get first $z-1$ change points: $v'_1, v'_2, \cdots, v'_{z-1}$. Sort the change points $v'_1, v'_2, \cdots, v'_{z-1}$ in ascending order, and get $v'_{i_1} < v'_{i_2} < \cdots < v'_{i_{z-1}}$. Let $v'_{i_0} = 0$ and $v'_{i_z} = N$, and then the ciphertext range $\mathcal{R} = \{1, 2, \cdots, N\}$ can be divided into z disjointed intervals: $[v'_{i_0}+1, v'_{i_1}], [v'_{i_1}+1, v'_{i_2}], \cdots, [v'_{i_{z-1}}+1, v'_{i_z}]$. Count the number of ciphertexts

1556-6013 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2015.2435697, IEEE Transactions on Information Forensics and Security

dropped in each interval, we can get a histogram with z bins as ep_1, ep_2,..., ep_z, which is the regularized distribution of H_w.

To regularize H_i, we denote $H_i = \{b_1, b_2, \ldots, b_M\}$, where b_j is the occurrence of value j for $1 \leq j \leq M$. We do change point analysis on the sequence $\{b_1, b_2, \ldots, b_M\}$ and record the first $z-1$ change points as $u'_1, u'_2, \cdots, u'_{z-1}$. Sort the change points $u'_1, u'_2, \cdots, u'_{z-1}$ in ascending order, and get $u'_{i_1} < u'_{i_2} < \cdots < u'_{i_{z-1}}$. These $z-1$ change points will divide H_i into z regularized bins as e_1, e_2,..., e_z, which is the regularized distribution of H_i.

Normalize these bins as $P_i(i) = ep_i/(ep_1 + ep_2 + ... + ep_z)$ and $Q_w(i) = e_i/(e_1 + e_2 + ... + e_z)$ for $1 \leq i \leq z$, and the relative entropy of H_w and H_i is estimated as:

$$D_{KL}(H_w||H_i) = \sum_{i=1}^{z} P_i(i) ln \frac{P_i(i)}{Q_w(i)}, \tag{11}$$

For a given keyword w whose estimated histogram by using differential attack is $H(w)$, and mimic histogram by the cloud server is $H(\bar{w})$, we define the identifying accuracy of keyword w as

$$IA_w = \frac{\sum_{i=1}^{N_w} F(D_{KL}(H_w||H_i), D_{KL}(H_w||H_{\bar{w}}))}{N_w}, \tag{12}$$

where

$$F(x, y) = \begin{cases} 1 & y < x \\ 0 & y \geq x \end{cases}. \tag{13}$$

Note that the maximum value of $IA(w)$ is $(N_w - 1)/N_w$, and, when $IA(w)$ reaches the maximum value, the keyword w can be uniquely identified by the cloud server. Furthermore, we define the average identifying accuracy (AIA) for N_t keywords as follows.

$$AIA = \frac{\sum_{j=1}^{N_t} IA_{w_j}}{N_t}. \tag{14}$$

B. Experimental Results

We demonstrate a thorough experimental evaluation on the TREC data [17], which consists of 7594 documents and 18238 distinct keywords, from which we select $N_w = 2061$ keywords of interest. In other words, the Inverted Index consists of 2061 post listings in our experiments.

To imitate the background obtained by the cloud server, we randomly select a subset of 100α percent of the whole document set. The keywords $\bar{w}_i$ and corresponding feature H_i are generated from this subset for $1 \leq i \leq N_w$. Herein, we use α as a parameter to describe the similarity of the background acquired by the cloud server to the outsourced document collection. We call α the background strength. A large background strength means that the server has a distribution close to the real distribution of the relevance scores that have been encrypted, and vice versa. In this experiment, we choose α = 0.90, 0.66 and 0.50.

First, the $BINsize$ z in Eq. (11) is an important parameter that has to determined, and obviously choosing different $BINsizes$ will affect the identifying accuracy. We perform a series of experiments to see which range of z would be the best for the attacker to get the most accurate result. In this experiment, we set α = 0.66 for consistency, and select $N_t = 100$ representative keywords from the total N_w keywords to estimate AIA. The AIA trend chart for different $BINsizes$ is depicted in Fig. 6, which shows that choosing z=8 would reach the best identifying accuracy. Therefore, in the following experiments we set $z = 8$. We should note that the choice of z should vary with term frequency, otherwise setting z too small or too big will cause measurement deviation. For keywords with higher term frequencies we should set bigger z values, and vice versa.

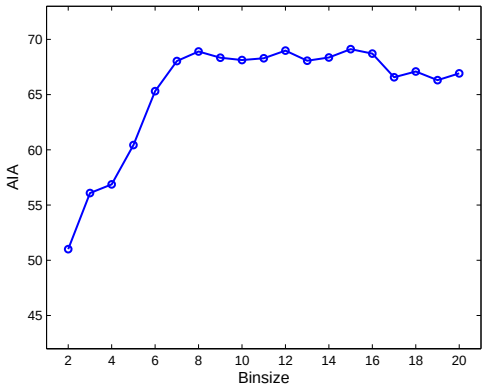


Fig. 6. Average identifying accuracy (AIA) for different BINsize.

Next, we try to recognize the keyword "w =weather". The relative entropies between H_w and H_i for $1 \leq i \leq N_w$ are depicted in Fig. 7(a), Fig. 7(b), and Fig. 7(c) for $\alpha = 0.90, 0.66$ and 0.5 respectively. In these figures, for the "correct" keyword "$\bar{w}_i$=weather", the relative entropy is marked by a red circle. As shown in Fig. 7(a), the minimum relative entropy appears at "$\bar{w}_i$=weather", that is, the server can successfully identify that w is just "weather". Fig. 7(b) shows the relative entropy of "$\bar{w}_i$=weather" is among the top 2.28% results, which means when α = 0.66, the server can successfully identify w as "weather" in a small range as 2.28% of all. When α comes to 0.50, the server can successfully identify w as "weather" in a relatively bigger range as 7.67%.

Furthermore, we conduct extensional experiments on representative $N_t = 100$ keywords selected from the total N_w keywords including "weather" to make our results more convincing. The calculation of AIA is repeated ten times with different encryption keys to avoid accidental error, and the average AIA results are shown in Fig. 8. The AIA curve is drawn to show the trend of AIA values with different background strengths α from 0.1 to 0.9.

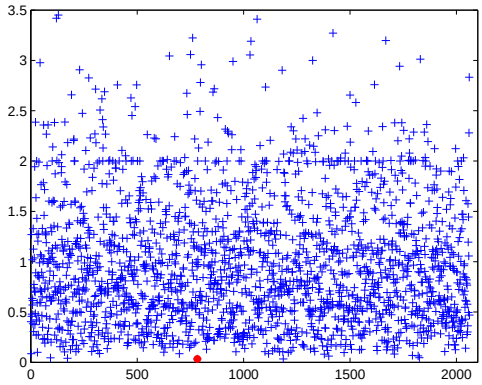
TABLE III
NUMBER OF SUCCESSFULLY IDENTIFIED KEYWORDS OUT OF 100 TEST ONES FOR DIFFERENT BACKGROUND STRENGTH α

α	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Count	2	14	16	24	29	36	40

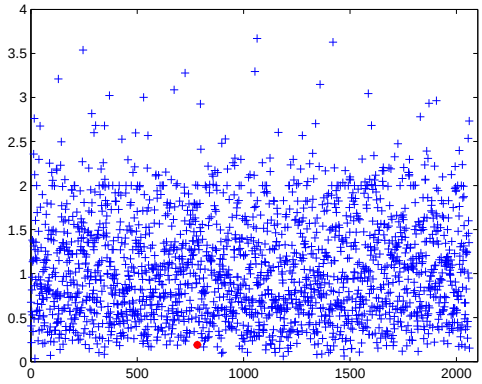
In addition, in TABLE III, we introduce the notion of "successfully identified". An attacked keyword is considered

1556-6013 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

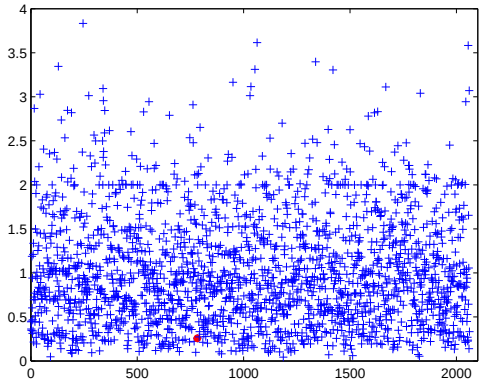
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2015.2435697, IEEE Transactions on Information Forensics and Security



(a) α=0.90



(b) α=0.66



(c) α=0.50

Fig. 7. Identifying keyword "weather" for different background strength α.

"successfully identified" if the relative entropy of the keyword falls in the top 5% results. We list the number of "successfully identified" keywords out of the 100 test ones for different background strengths α from 0.3 to 0.9.

Fig. 8 and TABLE III show that when there is more background knowledge, i.e., a larger α, the server can infer the keyword information more accurately.

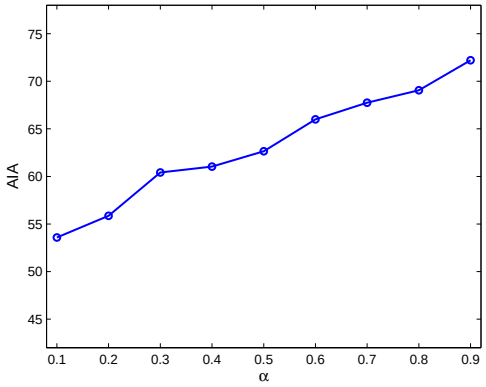


Fig. 8. AIA for different background strength α.

VI. Conclusion

In ranked search of encrypted cloud data, probabilistic OPE is needed to preserve the order of relevance scores and conceal their distributions at the same time. One-to-Many OPE [16] is a scheme designed for such a purpose. However, in this paper, we demonstrate that the cloud server can estimate the distribution of relevance scores by change point analysis on the differences of ciphertexts of One-to-Many OPE. Furthermore, the cloud server may identify what the encrypted keywords are by using the estimated distributions and some background knowledge.

On the other hand, some methods can be used to resist the proposed attack. One is to improve the One-to-Many OPE itself. For instance, we can divide plaintexts having the same value into several sets and divide the corresponding bucket into several sub-buckets. By mapping each plaintext set into one sub-bucket, some new change points will appear in the differential attack, which will cover up the original distribution of plaintexts. Another possible method is to add noise into the inverted index by adding some dummy documents IDs and keywords, and forging corresponding relevance scores.

In our future work, we will elaborate these ideas to design secure methods of probabilistic OPE and schemes for search in encrypted data.

References

[1] P. Mell and T. Grance, "The NIST definition of cloud computing," *NIST special publication*, 800(145): 7, 2011.

[2] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing," *Journal of Network and Computer Applications*, 34(1): 1-11, 2011.

[3] B. Krebs, "Payment processor breach may be largest ever," http://voices. washingtonpost. com/securityfix/2009/01/payment processor breach may b. html, 2009.

[4] M. Abdalla, M. Bellare and D. Catalano, "Searchable encryption revisited: Consistency properties, relation to anonymous IBE, and extensions," Advances in *Cryptology–CRYPTO*, 2005. *Springer Berlin Heidelberg*, pp. 205-222, 2005.

[5] D. Song, D. Wagner and A. Perrig, "Practical techniques for searches on encrypted data," *Security and Privacy*, 2000. *Proceedings. 2000 IEEE Symposium on. IEEE*, pp. 44-55, 2000.

[6] E.-J. Goh, "Secure Indexes," *Technical Report* 2003/216, *Cryptology ePrint Archive*, http://eprint.iacr.org/, 2003

1556-6013 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2015.2435697, IEEE Transactions on Information Forensics and Security

[7] D. Boneh, G. Di Crescenzo and R. Ostrovsky, "Public key encryption with keyword search," Advances in *Cryptology-Eurocrypt*, 2004. *Springer Berlin Heidelberg*, pp. 506-522, 2004.

[8] Y.-C. Chang and M. Mitzenmacher, "Privacy preserving keyword searches on remote encrypted data," *Applied Cryptography and Network Security*. *Springer Berlin Heidelberg*, pp. 442-455, 2005.

[9] R. Curtmola, J. Garay and S. Kamara, "Searchable symmetric encryption: improved definitions and efficient constructions," *Proceedings of the 13th ACM conference on Computer and communications security*. ACM, pp. 79-88, 2006.

[10] R. Agrawal, J. Kiernan and R. Srikant, "Order preserving encryption for numeric data," *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. ACM, pp. 563-574, 2004.

[11] A. Boldyreva, N. Chenette and Y. Lee , "Order-preserving symmetric encryption," Advances in *Cryptology-EUROCRYPT*, 2009. *Springer Berlin Heidelberg*, pp. 224-241, 2009.

[12] A. Boldyreva, N. Chenette and A. ONeill, "Order-preserving encryption revisited: improved security analysis and alternative solutions," Advances in *CryptologyCCRYPTO*, 2011. *Springer Berlin Heidelberg*, pp. 578-595, 2011.

[13] L. Xiao, I.-L Yen, "Security analysis for order preserving encryption schemes," *Proc. of 46th Annual Conference on Information Sciences and System*, pp. 1-6, 2012.

[14] A. Swaminathan, Y. Mao and G.-M Su, "Confidentiality-preserving rank-ordered search," *Proceedings of the 2007 ACM workshop on Storage security and survivability*. ACM, pp. 7-12, 2007.

[15] S. Zerr, D. Olmedilla and W. Nejdl, "Zerber+ r: Top-k retrieval from a confidential index," P*roceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*. ACM, pp. 439-449, 2009.

[16] C. Wang, N. Cao and K. Ren, "Enabling secure and efficient ranked keyword search over outsourced cloud data," *Parallel and Distributed Systems, IEEE Transactions* 23(8), pp. 1467-1479, 2012.

[17] P. Bailey, N. Craswell, I. Soboroff and A.-P. Vries, "The CSIRO Enterprise Search Test Collection," *SIGIR Forum* 41(2), pp. 42-45, 2007.

[18] S. Yu, C. Wang and K. Ren, "Achieving secure, scalable, and fine-grained data access control in cloud computing," *INFOCOM, 2010 Proceedings IEEE*. IEEE, pp. 1-9, 2010.

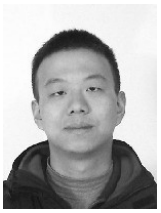
[19] N. Cao, C. Wang and M. Li, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," *INFOCOM, 2011 Proceedings IEEE*. IEEE, pp. 829-837, 2011.

[20] S. Bttcher and C.-L. Clarke, "A Security Model for Full-Text File System Search in Multi-User Environments," *FAST*, 2005.

[21] W.-A. Taylor, "Change-point analysis: a powerful new tool for detecting changes," http://www.variation.com/cpa/tech/changepoint.html, 2000.

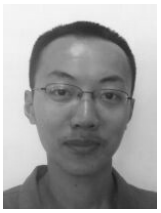
[22] G.Box and T. Kramer, "Statistical process monitoring and feedback adjustmentła discussion," *Technometrics*, 34(3), pp. 251-267, 1992.

[23] G.-K. Zipf, "The psycho-biology of language," 1935.



Ke Li received his B.S. degree in 2013 from University of Science and Technology of China, where he is now pursuing the M.S. degree in University of Science and Technology of China. His research interests include security, privacy and reliability in cloud computing.

Weiming Zhang received his M.S. degree and PH.D. degree in 2002 and 2005 respectively from the Zhengzhou Information Science and Technology Institute, P.R. China. Currently, he is an associate professor with the School of Information Science and Technology, University of Science and Technology of China. His research interests include multimedia security, information hiding, and privacy protection.



Ce Yang received his B.S. degree in 2011 from University of Science and Technology of China, where he is now pursuing the Ph.D. degree in University of Science and Technology of China. His research interests include security, privacy and reliability in cloud computing.

Nenghai Yu received his B.S. degree in 1987 from Nanjing University of Posts and Telecommunications, M.E. degree in 1992 from Tsinghua University and Ph.D. degree in 2004 from University of Science and Technology of China, where he is currently a professor. His research interests include multimedia security, multimedia information retrieval, video processing, information hiding and security, privacy and reliability in cloud computing.

1556-6013 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.