A Method of Protecting Relational Databases Copyright with Cloud Watermark

Yong ZHANG, Xiamu NIU, Dongning ZHAO

Abstract**—**With the development of Internet and databases application techniques, the demand that lots of databases in the Internet are permitted to remote query and access for authorized users becomes common, and the problem that how to protect the copyright of relational databases arises. This paper simply introduces the knowledge of cloud model firstly, includes cloud generators and similar cloud. And then combined with the property of the cloud, a method of protecting relational databases copyright with cloud watermark is proposed according to the idea of digital watermark and the property of relational databases. Meanwhile, the corresponding watermark algorithms such as cloud watermark embedding algorithm and detection algorithm are proposed. Then, some experiments are run and the results are analyzed to validate the correctness and feasibility of the watermark scheme. In the end, the foreground of watermarking relational database and its research direction are prospected.

Keywords**—**cloud watermark, copyright protection, digital watermark, relational database

I. Introduction

The progress in multimedia storage and transmission technology has allowed to store and transmit an ever increasing amount of information in digital format. This possibility has greatly expanded by the advent of the Internet. However, the ease of copying and reproducing digital data is likely to encourage Intellectual Property Rights (IPR) violation. The digital watermark technique represents a valid solution to the above problem, since it makes possible to identify the source, author, creator, owner, distributor or authorized consumer of digital image, video, audio and text, etc.

Now, with the development of Internet and databases application techniques, the demand that lots of databases in the Internet are permitted to remote query and access for authorized users becomes common. Meanwhile in that the copyright of the data may not be protected effectively, the data providers are worried about the data being burgled, illegal copy more and more. So we need find a mechanism to indicate the invading and pirating of the databases. We can solve this problem through embedding digital watermark into the relational databases, but there is little correlative work on it [5, 6, 8, 9].

The rest of the paper is organized as follows. Section II specifies the cloud model, includes forward cloud generator, backward cloud generator, similar cloud and the corresponding algorithms. Section III gives our scheme of cloud watermark relational databases and the corresponding algorithms. Section IV analyzes the results of the experiments. And we conclude with a summary and directions for future work in section V.

II. Cloud Model

The cloud model [1, 2] is a model of the uncertainty transition between a linguistic term of a qualitative concept and its numerical representation. In short, it is the model of the uncertainty transition between qualitative concept and quantitative description. The cloud model has three digital characteristics, Expected value (*Ex*), Entropy (*En*) and Hyper Entropy (*He*), which well integrates the fuzziness and randomness of spatial concepts in a unified way [2]. Expected value (*Ex*) represents the most representative numerical value of a linguistic term from fuzzy logic point of view. *Ex* is the position corresponding to the center of the cloud gravity. Entropy (*En*) is a measure of the fuzziness of the concept in the quantitative universe showing how many elements and to what extent in the universe could be accepted by the linguistic term. Meanwhile the Entropy (*En*) defined here is probabilistic, determining the probability distribution function of all the elements in the quantitative universe. Hyper Entropy (*He*) is the entropy of Entropy (*En*), All cloud drops would be much convergent if *He* is very small. When the ratio of Entropy (*En*) and Hyper Entropy (*He*) is very small, or the Hyper Entropy (*He*) relative to Entropy (*En*) is very big, the cloud will behave as the shape of fog on the whole, and the cloud is called fog. A single cloud drop is no real meaning, only the whole shape of all cloud drops can have some meaning.

A. Forward Cloud generator

Given three digital characteristics *Ex*, *En* and *He*, the forward cloud generator can produce as many drops of the cloud as you would like. The forward cloud generator algorithm FCG(*Ex*, *En*, *He*, *N*) in details [1, 2] is:

Input:

Manuscript received November 17, 2004. This work was supported by the National Natural Science Foundation of China (Project Number: 60372052, 60496323), the Foundation for the Author of National Excellent Doctoral Dissertation of P. R. China (Project Number: FANEDD-200238), the Multidiscipline Scientific Research Foundation of Harbin Institute of Technology. P. R. China (Project Number: HIT.MD-2002.11), and the Foundation for the Excellent Youth of Heilongjiang Province.

Yong ZHANG is with Harbin Institute of Technology Shenzhen Graduate School, Shenzhen University Town, 518055 China and State Key Laboratory of Software Engineering (Wuhan University), 430072 China (phone: +86-755-26033461; e-mail: zhang076@163.com).

Xiamu NIU is with Harbin Institute of Technology Shenzhen Graduate School, Xili University Town, Shenzhen, 518055 China(phone: +86-755-26033461; e-mail: xiamu.niu@hit.edu.cn)

Dongning ZHAO is a professional, she is now in Shenzhen, China (e-mail: zhaodongning1979@163.com).

three digital characteristics of a linguistic term (*Ex*, *En*, *He*), and the required number of cloud drops *N*

Output:

N cloud drops with their positions *x* in the quantitative universe and the compatibility degree *y*, and each cloud drop can represent the linguistic term.

Algorithm:

(1) Produce a random number *En'* which satisfies with the normal distribution probability of mean *En*, and standard error *He*;

(2) Produce a random number *x* which satisfies with the normal distribution probability of mean *Ex*, and standard error *En'*;

(3) Calculate: $y = e^{\frac{-(x-Ex)^2}{2(En')^2}}$;

(4) Let (*x*, *y*) be a drop of the cloud in the universe of discourse;

(5) Repeat step 1-4 until *N* cloud drops are generated.

B. *Backward Cloud generator*

Simultaneously, the input of the backward cloud generator is the quantitative positions of *N* cloud drops x_i, while the output is three digital characteristics *Ex*, *En*, *He*, of the linguistic term represented by *N* cloud drops. The backward cloud generator algorithm BCG(x_i, *N*) in details [3] is:

Input:

N cloud drops x_i (i=1,…,*N*),

Output:

three digital characteristics (*Ex*, *En*, *He*) which can represent the linguistic term.

Algorithm:

(1) Calculate: $V = \sum_{i=1}^{N} x_i / N$, $M1 = \sum_{i=1}^{N} |x_i - V| / N$, $M2 = \sum_{i=1}^{N} (x_i - V)^2 / N$;

(2) $Ex = V$;

(3) Calculate: $En = M1 \times \sqrt{\pi / 2}$;

(4) Calculate: $He = \sqrt{M2 - En^2}$.

C. *Similar cloud*

How to deal with the relationship among the different clouds represent a linguistic term of a qualitative concept? This is a problem how to measure the similarity degree among the different clouds. In that every cloud can be represented by three parameters (*Ex*, *En*, *He*), we can measure the similarity of the different clouds according to three parameters of the cloud. If there is a certain similarity among some clouds which represent the same qualitative concept, then the clouds are called similar clouds, or equipollence clouds. Here the certain similarity means that the similarity degree among the clouds is bigger than the similarity threshold given before.

In theory, the cloud is composed of infinity cloud drops. However, in our experiments, the whole shape of the cloud is represented by the finity cloud drops actually. Even two clouds are generated by the same three parameters (*Ex*, *En*, *He*), the two clouds are only similar, and are not same. Of course, the number of generated cloud drops is more, the similarity degree of the two clouds is bigger. When we measured the similarity degree of the clouds, we should notice that the similarity threshold is related with the number of the generated cloud drops.

Based on the particularity of the cloud itself, we proposed a method of measuring the similarity of some clouds based on the distance between the generated cloud drops. The corresponding similar cloud algorithm SC (*Ex1*, *En1*, *He1*, *Ex2*, *En2*, *He2*, *n*, *b*) in details [4] is:

Input:

n; //the number of cloud drops generating
Ex1, *En1*, *He1*; // the digital character of the firs cloud
Ex2, *En2*, *He2*; // the digital character of the second cloud
b; // the threshold of the cloud similarity degree

Output:

Similar or not?

Algorithm:

(1) $Drop1_i$=FCG(*Ex1*,*En1*,*He1*,*n*); //$i \in [1 \ldots n]$

(2) $Drop2_i$=FCG(*Ex2*,*En2*,*He2*,*n*); //$i \in [1 \ldots n]$

(3) Bubble Sort(*Drop1*);

(4) Bubble Sort(*Drop2*);

(5) select the cloud drops which falled into the range of *Ex*-3*En* and *Ex*+3*En* from sorted *Drop1* and sorted *Drop2*, and recorded into the *Drop1'* and *Drop2'* respectively;

(6) let *n1* equal the number of cloud drops in *Drop1'*;

(7) let *n2* equal the number of cloud drops in *Drop2'*;

(8) suppose *n1*≤*n2*, then there are C_{n2}^{n1} kinds combinations of the cloud drops in the *Drop2'*, noted as $Drop2'_j$ respectively which size equal that of *Drop1'*, $j \in [1, 2 \ldots C_{n2}^{n1}]$, if *n1*>*n2*, the same to *Drop1'*;

(9) Calculate the sum of the difference square of the corresponding cloud drops between *Drop1'* and $Drop2'_j$, noted as $Distance_j$, $j \in [1, 2 \ldots C_{n2}^{n1}]$;

(10) Calculate: $Similar = \sqrt{\sum_{j=1}^{n1} Dis\tan ce_j / C_{n2}^{n1}}$;

(11) If *Similar* <*b* then the clouds are similar

(12) Else the clouds are not similar.

For the algorithm, the key problem is how to set the similarity degree threshold *b*, the threshold *b* given is related with the number of the cloud drops. Commonly, the threshold can be set 30/*n*, the variable *n* is the number of the generating cloud drops.

With the given algorithms of forward and backward cloud generators, it is easy to build the mapping relationship inseparably and interdependently between qualitative concept and quantitative data. At the same time, the similar cloud algorithm is narrated and can be used to analyze the similarity of the embedded and extracted cloud watermark. Based on these algorithms of cloud model, a scheme of watermarking relational databases with cloud watermark is proposed and introduced in section III and section IV.

III. Relational Databases Cloud Watermark Technique

In the image watermarking algorithms, the idea of the patchwork algorithm [6, 7] is to choose random pairs of points (x_i, y_i) of an image in the spectral domain, and increase the brightness at x_i by 1 unit while correspondingly decreasing the brightness at y_i. Random changes to relational data can potentially introduce large errors. It is also not clear how to handle incremental updates and how to protect the watermark from various forms of attacks if one was to apply the patchwork algorithm to relational data [6, 7]. So we proposed a cloud watermarking relational databases algorithm according to the idea of the patchwork algorithm [5, 8, 9] and the character of some numerical attributes values which may tolerate certain errors.

A. Cloud watermark technique

Because of the property of cloud model, the most of the cloud drops generated by forward cloud generator will fall into the range of Ex-3En and Ex+3En. Supposing the error range of the numerical attribute value be -D and D, we can set the parameter Ex of forward cloud generator to be 0, the parameter En to be D/3, and the parameter of He to be arbitrary number given by the user. Thus, the cloud drops generated by three parameters (Ex, En, He) will fall into the range of -D and D mostly. From the distribution of the generated cloud drops and the whole shape of cloud, the cloud drops which will be embedded into the selected numerical attribute of some tuples would not bring more change to the mean of the numerical attribute value. Indeed, more the cloud drops were embedded, smaller the change of the mean would be. This is consistent with the property of cloud that a single cloud drop is no real meaning, only the whole shape of the cloud drops can have some meaning. So, the proposed method of watermarking relational databases with cloud model according to the patchwork algorithm idea of image watermark is feasibility. In this paper, the method is aimed at the history relational data to mining knowledge or other data need not being updated [5, 9].

To use cloud models in watermarking relational databases should utilize the properties of the database. That is to say, some numerical attribute of the databases should tolerate the little changes, and the changes do not affect the usability of the data, we can modify the value of the changeable bits with the cloud drops. We call this precondition the Rigorous Precondition.

In addition, there are some targets to evaluate the effect of the watermarking embedded to databases, that is:

- Imperceptible, the cloud watermark can not be sensed by users of the data.
- Usability, the cloud watermark do not weaken the usability of the protected data.
- Robustness, the watermark must survive the rational operations.
- Security, the watermark can not be removed by the users without the secret key private to the owner.
- Detectability, the watermark can be detected from a suspect database by the owners or the third party.
- Updatability, the watermark can be embedded into the update of the data.
- Transfer Capability, the watermark should transfer with the copy of the data.
- Watermarking overhead, the overhead of the watermark should be under a receivable range.

B. Cloud watermark embedding algorithm

Figure 1 is the framework of relational databases cloud watermark embedding. In the figure, the parameters of Ex, En and He is the key of cloud watermark embedding algorithm for the owners. RDB denotes original relational databases, and CWMRDB denotes watermarked relational databases. The relational databases cloud watermark embedding algorithm in details is:

Input:

Key: (Ex, En, He), RDB, *Scale*;

Output:

CWMRDB;

Algorithm:

```
(1) Select the numerical attribute A in the RDB;
(2) Let $N$ equal the number of tuples in the RDB, $N$ is the number of generating cloud drops;
(3) Call FCG($Ex$, $En$, $He$, $N$), and generate N cloud drops ($x_i$, $y_i$), $i \in [1 \ldots n]$;
(4) RDB.first;
(5) $i = 1$;
(6) While not RDB.EOF do
(7)   begin
(8)     if A.value is not null and random(0..1) is less than Scale then
(9)       A.value=A.value+$x_i$;
(10)    $i = i+1$;
(11)    RDB.next;
(12)  end.
```

The input parameter *Scale* represents the embedding scale.

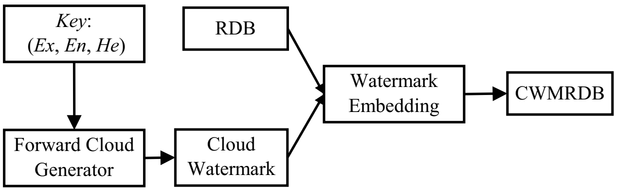

Figure1 Framework of relational databases cloud watermark embedding

The line 4 represents making the pointer of RDB point to the first tuple, and the line 6 represents that the loop conditions are if the attribute value of the tuple which is pointed by the pointer is null and if the generated random in the range of 0 and 1 is less than the parameter scale given, and the line 11 represents the pointer of RDB being moved to the next tuple.

C. *Cloud watermark detection algorithm*

Figure 2 is the framework of relational databases cloud watermark detection. In the figure, the parameters of *Ex*, *En* and *He* are the key of cloud watermark detection algorithm. RDB denotes original relational databases, and Suspicious RDB denotes suspicious relational databases, and the “similarity analysis” represents that to call the similar cloud function (see section II.C) and to analyze the similarity, then the result if have watermark or not is validated. The relational databases cloud watermark detection algorithm in details is:

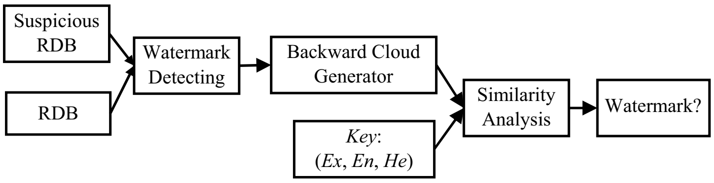

Figure2 Framework of relational databases cloud watermark detection

Input:

Ex, *En*, *He*, RDB, Suspicious RDB, *b*;

Output:

Has watermark or not ?

Algorithm:

(1) Select the numerical attribute A in the Suspicious RDB;

(2) According to the matched tuples between RDB and Suspicious RDB, extract the marks from the attribute A, note x_i, and the number of marks is noted N, $i \in [1\ldots N]$;

(3) Call BCG(x_i, N), get the parameters of *WmEx*, *WmEn*, *WmHe*;

(4) Call SC(*Ex*, *En*, *He*, *WmEx*, *WmEn*, *WmHe*, *N*, *b*), validate if has watermark or not.

The parameter of *b* represents a similarity threshold which is related with the number of cloud drops (see section II.C).

IV. Experiment Result And Analysis

Now we can validate the method of watermarking relational databases our proposed through running some experiments and analyzing the results. We ran the experiments on SQL server version 7.0 using ODBC connectivity on a Windows XP PC with 2.8GHz CPU, 256 MB of RAM, and a 80G hard disk driver. And the programming language is Delphi version 5.0. The data source is a forest cover dataset of America. In this dataset, there are 581,012 tuples and 54 attributes, among the attributes there are ten numerical attributes.

In the experiment, we only select approximately 5,000 tuples for cloud watermark embedding and detecting, so the embedded scale can be set only approximately 1 percent (5,000/581,012). Thus, the cloud drops can be spread into the relational databases and are hard to be deleted. In that the measurement data are the mean data of the geography pane 30 meters multiply 30 meters, the slight change will be constraint to lie within the measurements tolerance and do not affect the usability of the data. We can watermark relational databases with the cloud watermark algorithm. We selected one of the numerical attribute, for example, elevation. The tolerance error range of the attribute is between -10 and 10, then we set the key that *Ex* equal 0, *En* equal 3, and *He* equal 1. Let the embedded scale equal 0.01, and then we randomly selected approximately 5,000 tuples from the relational databases which size is 581,012 to embed the cloud drops. Figure 3 shows the generated cloud drops in the experiment. Figure 4 shows the extracted cloud drop marks from relational databases. The extracted cloud drop marks are transformed to three parameters (*WmEx* equals 0.001, *WmEn* equals 2.999, *WmHe* equals 0.998) by backward cloud generator, and figure 5 shows the cloud drops generated by three transformed parameters (*WmEx*, *WmEn* and *WmHe*) through backward cloud generator.

Figure 3 illustrates that the cloud drops generated through forward cloud generator falled into the rang of -10 and 10, which may be satisfied with the error demand of the attribute named elevation, and do not affect the usability of the relational data.

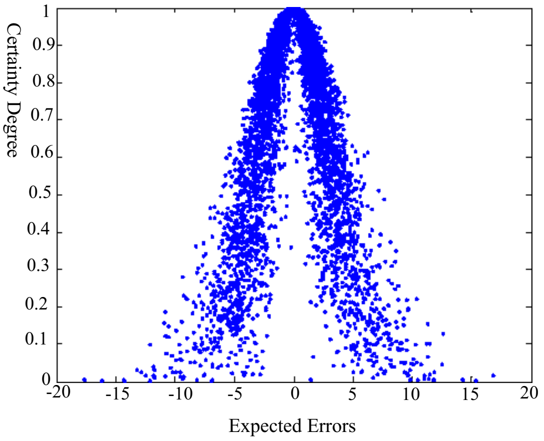

Figure3 Embedded cloud watermark

Then we can make the cloud similarity analysis between the cloud character parameters of Figure 3 (*Ex*, *En* and *He*) and those of figure 5 (*WmEx*, *WmEn* and *WmHe*) through the similar cloud algorithm (see section II.C). The result of similarity analysis shows that the similarity degree is bigger than the threshold *b* given before. That is to say, the relational databases have watermark, and figure 5 shows the extracted

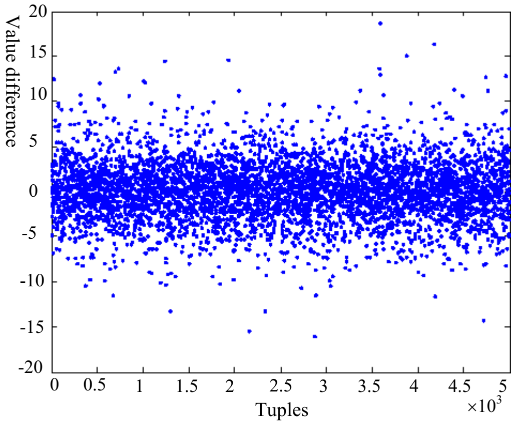

Figure4 Extracted cloud drop marks

cloud drops watermark.

The experiment shows that the method of protecting relational databases copyright with cloud watermark can adapt to the particularity of the relational databases, and the cloud drops have randomicity and the embedded cloud drops are irrespective of the tuples' order. In that three parameters of the cloud can only be transformed by the backward cloud generator with some cloud drops, relational databases cloud watermark technique has detectability and robustness. But there is a limitation for the scheme, the cloud watermark detection algorithm need the original RDB participating, and has not implemented blind detection. So the scheme should be reinforced unceasingly, and some attacked experiments should be studied.

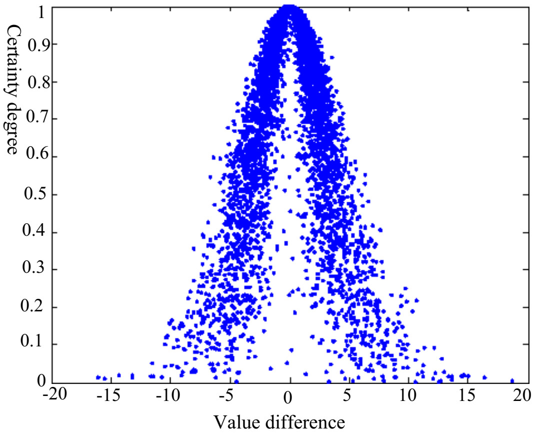

Figure5 Extracted cloud watermark

In addition, there are some other ways to improve the effect of the watermarking. For primary keys are the most important part of the relational data and the most difficult part to omit, embedding the cloud watermark on primary keys can improve the watermark's Robustness and Transfer Capability. In that there are many attributes and tables in the relational databases, we can improve the robustness of the database by strengthening the association among these attributes and tables using the cloud generator, to realize that with the cloud watermark information of one attribute can not resume the watermark on any other attributes.

V. Conclusions

With the need of databases security, it is urgent to solve the problem of relational databases copyright. For some centuries, it is noteworthy that the publishers of books of mathematical tables (e.g. logarithm tables and astronomical ephemerides) have been introducing small errors in their tables for centuries to identify pirated copies [5, 6, 8, 9]. So we can study on watermarking relational databases according to the idea.

In this paper, the method of protecting relational databases copyright with cloud watermark is studied only on the numerical attributes. In the future, we should expand our studies to non-numerical attributes fields [5, 6, 8]. Because of the particularity of the relational data, there are some difficulties to research on watermarking relational databases in depth. But the right management of the relational databases copyright with watermark should be an important topic for database research. And this research direction will be deserved serious attention more and more [5, 6, 8, 9].

Acknowledgment

Yong ZHANG thanks his Ph.D. supervisor professor Deyi LI, his colleagues Dr. Shuliang WANG, Luying LIU, Changyu LIUetc.

References

[1] LI Deyi, MENG Haijun, SHI Xuemei, 1995, Membership clouds and membership clouds generator. Journal of Computer Research and Development, 32(6): 15-20

[2] WANG SHULIANG, LI Deren, SHI Wenzhong, LI Deyi, WANG Xinzhou. 2003, Cloud Model-Based Spatial Data Mining. Geographical Information Science, 9(2):67-78.

[3] CHEN Gang, 2003, Handwriting Identification Based on Data Field. Master dissertation, PLA University of Science and Technology, Nanjing, PRC.(in Chinese)

[4] ZHANG Yong, ZHAO Dongning, LI Deyi, 2004, The Similar Cloud and the Measurement Method. Information and Control, 33(2):129-132.(in Chinese)

[5] ZHANG Yong, ZHAO Dongning, LI Deyi, 2003, Watermarking Relational Databases. Journal of PLA University of Science and Technology(Nature Science), 4(5):1-4.(in Chinese)

[6] Agrawal R, Kiernan J. Watermarking Relational Databases. In: Proceeding of the 28th VLDB Conference. Hong Kong: University of Science & Technology, Hong Kong, 2002: 155-166.

[7] Bender W, Gruhl D, Morimoto N, Lu A. Techniques for Data Hiding. IBM Systems Journal, 1996, 35(3,4):313-336.

[8] Sion R, Atallah M, Prabhakarl S. Watermarking Relational Databases. Technical Report. Indiana: the Center for Education and Research in Information Assurance and Security of Purdue University, 2002.

[9] ZHANG Yong, ZHAO Dongning, LI Deyi, 2003, Digital Watermarking for Relational Databases. Computer Engineering and Application. 39(25):193-195.(in Chinese)

Yong ZHANG, was born in China, in December, 1976, received the B.S. degree, M.S. degree in System Engineering from Institute of Communication Engineering, Nanjing, China in 1997 and 2001 respectively, and received the Ph.D degree in Communication Engineering from Institute of Communication Engineering, Nanjing, China, in 2004.

Currently, he is working in information security technique research center, Shenzhen graduate school, Harbin Institute of Technology, China. His interest includes databases security, digital watermarking, System Integration and Optimization.

Xiamu Niu was born in China, in May 1961, received the B.S. degree and M.S. degree in Communication and Electronic Engineering from Harbin Institute of Technology (HIT), Harbin, P. R. China in 1982 and 1989 respectively, and received the Ph.D degree in Instrument Science and Technology in 2000. He was an invited scientist and staff member in Department of Security Technology for Graphics and Communication System, Fruanhofer Institute for Computer Graphics, Germany, from 2000 to 2002. He was awarded the Excellent Ph.D Dissertation of China in 2002. He now is the Professor (doctoral advisor) and Superintendent of Information Countermeasure Technique Institute HIT, Director of Information Security Technique Research Center, HIT-ShenZhen. He is SPIE member, ACM member, IEEE member, and the advanced CIE member. He has published 3 works and more than 70 papers, and about 20 papers were cited by SCI and EI. His current research fields include computer information security, hiding communication, cryptography, digital watermarking, signal processing and image processing etc.