Threshold Estimation in Self-Destructing Scheme Using Regression Analysis

Young Ki Kim, Choong Seon Hong
Department of Computer Science and Engineering
Kyung Hee University
South Korea
{qoo0144, cshong}@khu.ac.kr

***Abstract*—As technologies and services that leverage cloud computing evolve, more and more businesses or individuals are using them. However, as users increasingly use and store personal information in cloud storage, research on privacy protection models in the cloud environment is becoming more important. A self-destructing scheme has been proposed to prevent the decryption of encrypted user data after a certain period of time using a DHT network. However, the existing privacy protection model does not mention the method of setting the threshold value considering the availability and security of the data. Therefore, in this paper, we propose an optimal threshold finding method considering both data availability and security of privacy protection model by applying regression analysis.**

***Keywords*—Self-Destructing Scheme, DHT Network, Privacy Protection, Regression Analysis**

I. Introduction

Cloud computing means entrusting data to information systems that are managed by external parties on remote servers "in the cloud". Cloud computing technology is considered to be more popular because of its convenience, scalability, and availability. In fact, many companies are taking significant economic benefits by using cloud computing for data related to the services they provide.

As users increasingly use and store personal information in cloud storage, research on privacy protection models in the cloud environment is becoming more important[1]. To do this, Shamir Secret Sharing[2] is used to divide the keys required for encryption into several pieces and use them to decrypt the encrypted user data after a certain period of time using a distributed hash table network Self-Destructing Scheme.

However, the existing Self-Destructing Scheme does not mention a method for setting a threshold value considering the availability and security of the encrypted data. Therefore, in this paper, we propose a method to find optimal threshold value considering the availability and security of data in Self-Destructing Scheme by applying regression analysis.

In Section 2, the existing Self-Destructing Scheme and regression analysis are discussed. In Section 3, the problems and suggestions of existing research are described. In Section 4, experimental results and performance evaluation are analyzed. Finally, in Section 5, conclusion of this paper and future research plans are discussed.

II. Related Work

A. Self-Destructing Scheme

In 2009, Geambasu et al. proposed Self-Destructing Scheme to protect privacy of electronic data.

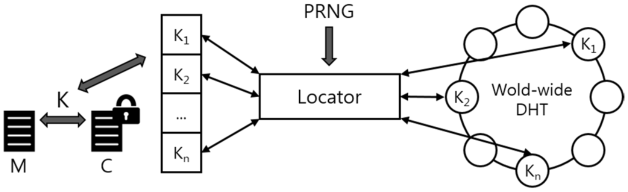


Fig. 1. Vanish: Self-Destructing Scheme

The Self-Destructing Scheme divides the key used to encrypt the data into several shares and distributes it to the distributed hash table network where the data disappears after a certain period of time.

The proposed Self-Destructing Scheme in [3] and [4] do not require the user to perform any delete operation or a third party to delete the data. Also, if the attacker obtains a copy of the data or the key used to encrypt, the data can not be decrypted after a user-specified period of time.

B. Regression Analysis

Regression analysis, which is one of the supervised learning, is a method of estimating the relationships among variables[5]. A typical method of measuring fitness in regression analysis is residual analysis. In this paper, we use the method of finding the residuals of the predicted and actual results and minimize them.

We apply the nonlinear regression analysis model to reduce the error of the distribution of input values and result values. Unlike the linear regression analysis model, the nonlinear regression model has a feature that its structure is not linear with respect to the coefficients. The nonlinear regression model is more complicated than the linear regression model, but also the least squares technique is used in estimating the parameters. More details on this will be discussed later in Section 3.

When dividing a key into shares using Shamir Secret Sharing in the Self-Destructing Scheme, the total number of key shares and the number required for decryption should be determined by considering data availability and security.

978-1-5386-1101-2/17/$31.00 ©2017 IEEE

III. Proposed scheme

In this section, we discuss our proposed threshold estimation based on regression analysis in Self-Destructing Scheme and learning phase consisting of training phase and testing phase. As discussed in more detail in this section, previous works do not mention at all how to determine the optimal threshold for data availability and security. Therefore, we propose a solution to this problem.

A. System Architecture and Assumptions

As mentioned above, in the Self-Destructing Scheme, when dividing the key shares, the number of total shares and the minimum number necessary for decryption must be determined. And according to [3], the total number of divided key shares and the minimum number required to decrypt are related to the availability and security of the data. However, the method of determining the optimal threshold ratio considering availability and security is not mentioned.

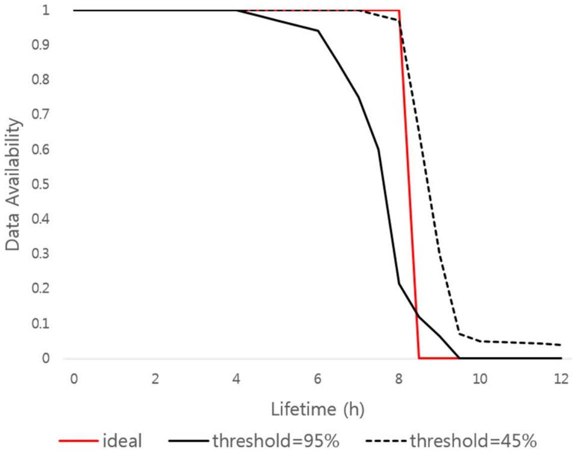


Fig. 2. Availability of Data According to Threshold Ratio

Figure 2 shows the graph of the same total number of N key shares with different thresholds. If the ratio of the threshold value is 95%, the user-specified time required to maintain the data is not satisfied. If the threshold value is 45%, the key shares are not completely lost even after a certain period of time, which may cause a security problem. This problem is caused by the peculiarity of distributed hash table network, and studies for solving this problem are also actively proceeding [6].

In order to consider both the availability and security of data, we need to find thresholds with the most similar results to the ideal graph. Therefore, in this paper, we propose a method to find the optimal threshold value by using regression analysis based on the similarity and the threshold value of the graph.

$$\mathrm{Y} = \alpha + \beta_1 X_1^{r_1} \beta_1 X_2^{r_2} + \beta_1 X_i + \ ... + \beta_k X_k^{r_k} + \varepsilon \qquad (1)$$

(1) is a nonlinear regression function for measuring the availability of data according to a threshold value. *Y* means the similarity of the graph based on the ideal graph and the actual threshold value and X_i means threshold value.

The similarity between the graphs is measured using the Gromov-Hausdorff distance using the length, curvature, and area divided by the graph [7]. As the degree of similarity between the graph drawn with the specific threshold value and the ideal graph is higher, the user can be guaranteed the access time desired by the user. And also if the period expires, not only the user but also the attacker can not obtain the key even if the user does not perform any other delete operation.

The nonlinear regression model for finding the optimal threshold is mentioned in equation (1). Based on the training set and the results, the least squares method is used to minimize the residual sum of squares in order to generate a graph to infer the optimal threshold.

When developing a least squares estimator for a nonlinear model, we encounter complex problems that were not seen in linear models. Given a training set, the estimates of α and β are obtained by minimizing the following equation.

$$SS_{Res} = \sum_{i=1}^{n} (Y_i - \alpha e^{\beta X_i})^2 \qquad (2)$$

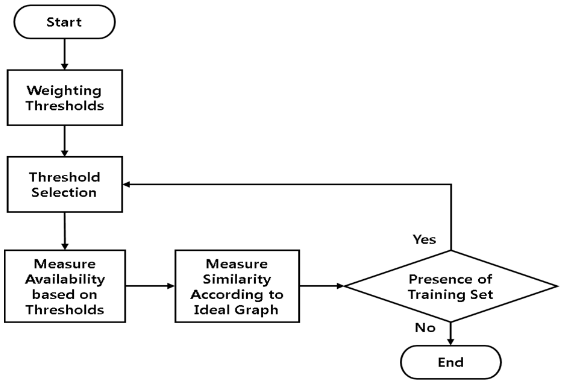
SS_{Res} means residual sum of squares and (2) is differentiated with respect to α and β and each derivative is set to 0, the following equations can be obtained.

$$\sum_{i=1}^{n} (Y_i - \alpha e^{\widehat{\beta} X_i})(-e^{\widehat{\beta} X_i}) = 0 \qquad (3)$$

$$\sum_{i=1}^{n} (Y_i - \alpha e^{\widehat{\beta} X_i})(-\hat{\alpha} e^{\widehat{\beta} X_i} X_i) = 0 \qquad (4)$$

Unlike the least squares estimation equation of the linear model, (3) and (4) are nonlinear with respect to the parameter estimators $\hat{\alpha}$ and $\hat{\beta}$. Therefore, the estimator can not be calculated for the elementary matrix algebra, and an iterative process should be used. Many methods have been introduced to minimize the residual sum of squares in nonlinear models. In this paper, we apply the Gaussian and Newton method to find the least squares estimator.

Applying the regression analysis model in the proposed environment is important because it is very hard to find the optimal value by analyzing the results of all the thresholds in the Self-Destructing Scheme. Also, due to the characteristics of the distributed hash table network(churn), the optimal threshold value for determining the availability and security of the data to be protected is different every time the nodes are added or deleted.



Start
Weighting Thresholds
Threshold Selection
Measure Availability based on Thresholds
Measure Similarity According to Ideal Graph
Presence of Training Set
Yes
No
End

Fig. 3. Flowchart of Learning Process for Regression Analysis

Figure 3 is a flowchart of the learning process for regression analysis. When the learning process starts, we assign weights to thresholds that are likely to be similar to the ideal graph and select thresholds. If a graph based on a threshold value is observed as shown in Fig. 2, the similarity with the ideal graph is measured and the presence or absence of an additional training set is checked.

B. Training Phase for Regression Analysis

The learning process for calculating the linear regression equation consists of a training phase and a testing phase. Data sets used as learning groups are divided into training set and testing set at a ratio of 7: 3.

Algorithm 1 Training Phase

```
if there is training set then
    extract data information from training set
    N = total number of key shares
    T = threshold ratio
    set current time
    measure data availability graph
    if the user-specified time is expires then
        F = similarity of availability graph with ideal
        save the result of training set
    else
        wait until the user-specified time is expires
    end if
else
    calculate nonlinear regression equation
end if
```

Algorithm 1 is the pseudocode for the training phase. The training phase measures the availability graph of the data based on the data set for learning and calculates the similarity with the ideal graph. The main purpose of the training phase is to calculate the nonlinear regression equation based on the input value(threshold ratio) and the output value(similarity of availability graph).

When the training phase begins, firstly, if training set exists, information such as total number of key shares, threshold ratio, size of data and user-specified time are extracted. Then, initialize the necessary information into variables and measure the availability and form a graph. If the user-specified time expires, measure the similarity to the ideal graph and store the result. The training phase repeats the above steps until there are no more training sets, and the nonlinear regression equation is calculated after all the training sets have been learned.

C. Testing Phase for Regression Analysis

Algorithm 2 is the pseudocode for the testing phase. The testing phase is a process basically similar to the training phase and is the process of verifying the fitness of the nonlinear regression equation which calculated as a result of the training phase.

Algorithm 2 Testing Phase

```
if there is testing set then
    extract data information from testing set
    N = total number of key shares
    T = threshold ratio
    NE = closest N from training set
    TE = estimated threshold
    set current time
    measure data availability graph
    if the user-specified time is expires then
        F = similarity of availability graph with ideal
        FE = similarity of estimated availability graph
        save the result of testing set
    else
        wait until the user-specified time is expires
    end if
else
    update nonlinear regression equation
end if
```

Similar to the training phase, when the testing phase begins, it first determines the presence or absence of a data set. Then, information such as total number of key shares, threshold ratio, size of data and user-specified time is extracted from the testing set. In the process of initializing the variables based on the information obtained from the data set, we initialize the closest N value among the training sets and the estimated optimal threshold ratio to verify the fitness of the nonlinear regression equation. After the above procedure, we update the nonlinear regression equation based on the actual results and estimated values. The testing phase can minimize the error of the nonlinear regression equation.

IV. EVALUATION

Table 1: Experiment Environment

DHT Network Type	Vuze(Azureus)
Num. of Key Shares	100
Num. of Training Set	1000

Table 1 shows the experimental environment for performance evaluation in this paper. In the Vuze distributed hash table network environment, which is widely used as P2P application, we constructed 100 key shares and 1000 training set for the learning process. More specifically, it is divided into 700 training sets and 300 fitness test groups. At this time, one training set has a key K for encryption and 100 key shares $K_1, K_2, \ldots, K_{100}$.

The scenario for performance evaluation in this paper is as follows. First, we generate a key K for encrypting the data and generate 100 key shares using Shamir Secret Sharing. The initial learning group does not have a threshold set for obtaining the full key from the key shares. Therefore, we prioritize the thresholds that the measured availability graph is expected to be similar to the ideal graph, and randomly select the results.

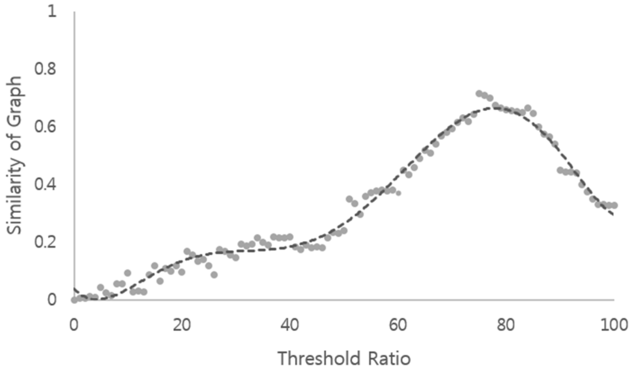


Fig. 4. Regression Analysis According to Similarity Graph

Figure 4 shows the result of analyzing the similarity between the data availability graph according to the threshold value and the ideal graph mentioned in figure 2. Based on the output value(similarity of availability graph) of the input value(threshold ratio), we derive function and graph for determining optimal threshold value using nonlinear regression function. And then, the threshold value at the point where the similarity of the graph is highest is calculated and applied to the actual Self-Destructing Scheme.

The prediction of the threshold value through regression analysis has an important meaning, because the shape of the graph continuously changes as the learning progresses, and that the optimum threshold value differs depending on the size of the data and the number of key shares.

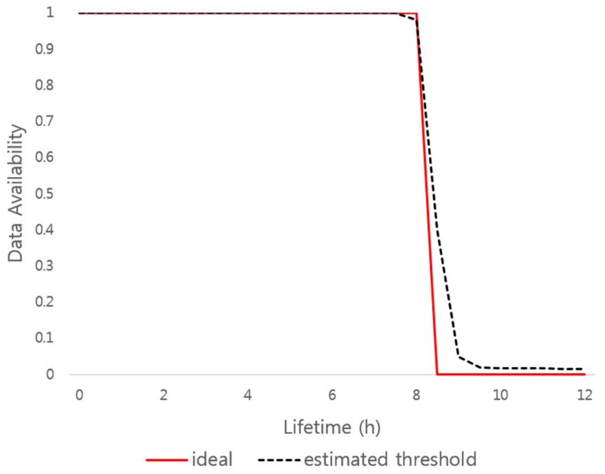


Fig. 5. Availability Graph According to Estimated Threshold

Figure 5 compares the availability graph of the data with the ideal graph based on the thresholds estimated through regression analysis. Comparing the results of the other thresholds in figure 2 above, when the key shares were distributed with the estimated thresholds, the key shares were retained in the distributed hash table environment for 8 hours(user-specified time), and after user-specified time, almost all key shares disappeared .

V. Conclusion

Self-Destructing Scheme is proposed as a model to protect the privacy of personal information in the cloud computing environment. However, the existing Self-Destructing Scheme does not mention the method of determining the threshold considering data availability and security. Therefore, in this paper, we propose a method to find optimal threshold considering both availability and security of data by applying nonlinear regression analysis. In addition, the availability of the data with the estimated threshold value by the regression analysis and the result are sufficiently reliable.

However, there is a limitation that the threshold estimation proposed in this paper takes much time in the learning process. This is because the time spent in the learning process is dependent on the user-specified time. Therefore, in the future, we will study methods to minimize the time required for the learning process by using approaches such as clustering or parallel computation.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2015-0-00274,Development of Smart Mediator for Mashup Service and Information Sharing among ICBMS Platform) *Dr. CS Hong is the corresponding author

References

[1] Mark D. Ryan, “Cloud computing privacy concerns on out doorstep,” Communications of the ACM, vol. 54, no. 1, January 2011, pp. 36-38.

[2] A. Shamir, “How to share a secret,” Communications Magazine, ACM, vol. 22, no. 11, November 1979, pp. 612-613.

[3] Roxana Geambasu, Tadayoshi Kohno, Amit A. Levy and Henry M.Levy, “Vanish: Increasing data privacy with self-destructing data,” USENIX Security Symposium, June 2009, pp.299-316.

[4] Guojun Wang, Fengshun Yue, Qin Liu, “A secure self-destructing scheme for electronic data,” Journal of Computer and System Sciences, vol. 79, no. 2, March 2013, pp.279-290.

[5] Gene Golub, Victor Pereyra, “Seperable nonlinear least squares: the variable projection method and its applications,” Inverse Problems, vol. 19, no. 2, February 2003, R1.

[6] S. Rhea, D. Geels, T. Roscoe and J. Kubiatowicz, “Handling churn in a DHT,” USENIX Annual Technical Conference, vol. 6, December 2004, pp.127-140.

[7] Alexander M. Bronstein, Michael M. Bronstein, Ron Kimmel, Mona Mahmoudi and Guillermo Sapiro, “A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching,” International Journal of Computer Vision, vol. 89, no. 2, September 2010, pp.266-286.