ZIBELINE INTERNATIONAL™ PUBLISHING

ISSN: 2773-6202 (Online)
CODEN: JTIOCC

Journal of Technology & Innovation (JTIN)

DOI: http://doi.org/10.26480/jtin.02.2021.58.60

RESEARCH ARTICLE

PRESERVING PRIVACY FOR PAGERANK ALGORITHM

Tho Thi Ngoc Le*

Faculty of Information Technology, Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam
Corresponding Author Email: ltn.tho@hutech.edu.vn

ARTICLE DETAILS

Article History:

Received 23 September 2021
Accepted 26 October 2021
Available online 08 November 2021

ABSTRACT

Data mining has been emergingly applied in many fields to discover the knowledge from the huge data. To do that, information has been sent forward and backward among data owners, users, and maybe third parties. In this situation, it is necessary to design systems to exchange the data between the data owner, the client and third parties during data mining process without scarifying the sensitiveness of data. Hence, we need a privacy preserving mechanism while mining to protect the data as in the situation of sophisticated cyber-attack. In this work, we describe a model for ensuring the privacy in ranking on the graph using PageRank and Shamir Secure Sharing scheme. Specifically, Shamir Secure Sharing scheme has been applied to share the information of graph from the data owner to many servers (i.e. third party). Then, the share of graph on each server will be ranked separately. When the users need the results of ranking and make a request, the information from servers will be combined for the users. Doing this way, the third party doesn't know the meaning of data but still run analyzing the data. Hence, data owner preserves the privacy of his data while users still retrieve a piece of the information as needed.

KEYWORDS

graph ranking, PageRank, privacy preserving data mining, Shamir Secret Sharing scheme

1. INTRODUCTION

We have been seeing an explosion of information processing in all aspects which are generated from many resources, such as business activities, social networks, or daily life activities. As the necessary of knowledge discovery, data mining and machine learning have been considered as paths to exploit and analyze the useful knowledge for life. Currently, data mining and machine learning have been applied for discovering hidden patterns in large data sets in various fields such as medicine, education, health care, transportation, weather forecast, agriculture, etc. Two examples are addressed as in work by (Le and Phuong, 2020) are:

(1) Credit scoring: By building historical customer data, the bank can determine loan quota or credit limit which customers are affordable.

(2) Health care: Providers find the best practices and the most effective treatments. They apply tools to compare symptoms, causes, treatments and effects to analyze which actions are efficient for patients.

Despite of many practical advantages, data mining is potentially suffering from the privacy issues (Wakabayashi, 2019). It would give people troubles if their personal information is disclosed without consent. In another scenario, when a data owner outsources data to a third party, such as a data analyst, third party might have a way to learn from the data. For this reason, there is a requirement to secure data before sending to third party but still ensuring that the third party can run mining on data.

In this work, we report a preliminary result when considering a popular ranking algorithm, PageRank, and try to preserve the privacy of graph information while ranking on third party side (Page et al., 1998). We proposed a scheme to secure ranking process using Shamir Secret Sharing. First, graph data is shared to different servers of third party. Second, each server computes the graph vertex ranks separately. Third, when the user retrieves data, rank scores from (some) servers are collected and recover the right score for user.

2. PRELIMINARIES

2.1 PageRank Algorithm

PageRank algorithm is a popular approach to rate the relative importance of web pages (Page et al., 1998). PageRank is known as an efficient algorithm to measure the human interest and attention devoted to web pages. Since it works on graph, PageRank's spirit has inspired a series of approaches in other disciplines, such as biology, chemistry, neuroscience, and physics (Gleich, 2015), whose data is also presentable as graphs naturally.

Given a graph $G = \{V, E\}$, where V is a set of n vertices $V = \{v_1, v_2, \cdots, v_n\}$, and E is a set of edges $E = \{e | e = (v_i, v_j), \forall v_i, v_j \in V, i \neq j\}$. PageRank analyzes the importance of vertices based on their links. Specifically, the important vertices are likely to have more links. Mathematically, the importance of a vertex v_i is expressed as its rank $PR(v_i)$:

$$PR(v_i) = \frac{1-d}{n} + d \sum_{v_j \in M(v_i)} \frac{PR(v_j)}{L(v_j)} \quad (1)$$

where, $M(v_i)$ is the set of all neighbor vertices of v_i and $L(v_i)$ is the size of $M(v_i)$. The rank PR of every vertex will be computed for pre-specified loops or until convergence.

2.2 Shamir Secret Sharing Scheme

Quick Response Code	Access this article online	
	Website: www.jtin.com.my	DOI: 10.26480/jtin.02.2021.58.60

In this section, we describe the way to share a secret with Shamir Secret Sharing (Shamir, 1979). Given that there are n parties who wish to share a secret s and the number of parties required to recover secret must be at least t. In cryptography, this is a kind $(t+1)$-out-of-n secret sharing scheme. When a trust dealer wishes to share a secret s in $\boldsymbol{Z}_p$ (p is a chosen prime number), he needs to generates a secret polynomial $f(X)$ of degree t with $f(0) = s$. Then, he generates random integers f_i in $\boldsymbol{Z}_p$ for $i = 1, \cdots, t$ and sets:

$$f(X) = s + f_1 X + \cdots + f_t X^t \tag{2}$$

The trusted dealer establishes each of the n players by an element in a set $X \subset \boldsymbol{Z_p}\{0\}$, X can be setup as $X = \{1,2,\cdots,n\}$. Then, if $i \in X$, party P_i is given the share $s_i = f(i)$. We then have a share vector $(s_1;\cdots,s_n)$.

If $(t+1)$ parties come together, they can recover the secret s, and the original polynomial via Lagrange interpolation using the belong equation:

$$s = f(0) = \sum_{i=1}^{n} s_i \delta_i(0) \tag{3}$$

Furthermore, we define a set $Y \subset X$, the vector $r_Y = \left(r_{x_i,Y},\cdots,r_{x_i,Y}\right)_{x_i \in Y}$ to be the public recombination vector where:

$$r_{x_i,Y} = \prod_{x_j \in Y, x_j \neq x_i} \frac{-x_j}{x_i - x_j} \tag{4}$$

Then, if we obtain a set of shares back from a subset $Y \subset X$ with $\#Y > t$, we can recover s via the summation:

$$s = \sum_{x_i \in Y} r_{x_i,Y} \cdot s_i \tag{5}$$

3. Securing Pagerank Using Shamir Secret Sharing Scheme (SPS)

We describe our proposed scheme SPS, which is outlined in Figure 1, for securing the graph information during ranking process with PageRank algorithm. Three main stakeholders in the scheme are database owner, servers on cloud and users. Given that, the database owner possesses graph data and he need to protect the information about graph including vertices and edges, which are required by ranking process of PageRank algorithm. Additionally, we need to protect the rank scores of graph vertices after ranking. When a user making a request to the cloud, e.g. retrieving the rank score of a vertex, each server on the cloud itself must not know the exact answer.

Our proposal includes the following three steps:

(1) Sharing: To preserve privacy of data, the database owner shares his data to K servers on cloud.

(2) Ranking: On each server, the graph will be ranked for vertex scores separately.

(3) Retrieving: When a user requests for the information about rank scores of vertices, the cloud combines corresponding elements on different servers to get results.

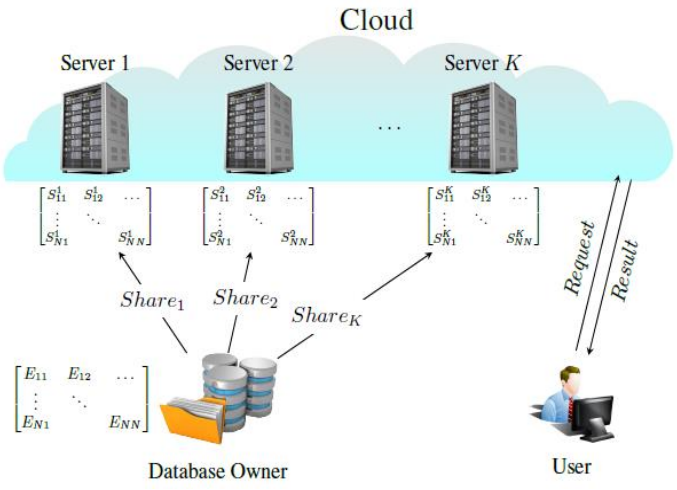

Figure 1: Securing PageRank Using Shamir Secret Sharing Scheme (SPS).

We outline the notations and data representation of a graph $G = \{V, E\}$. V is the set of N vertices, represented as a vector $V = \{v_1, v_2, \cdots, v_N\}$. E is the set of edges expressing the relations among vertices, conveniently represented as an adjacent matrix E:

$$E = \begin{bmatrix} e_{11} & \cdots & e_{1N} \\ \vdots & \ddots & \vdots \\ e_{N1} & \cdots & e_{NN} \end{bmatrix}$$

where, E_{ij} is either 0 or 1, indicating the existing of connection between vertex v_i and vertex v_j. The rank scores of all graph vertices are represented as a vector $PR = \{pr_1, pr_2, \cdots, pr_N\}$. Note that, the values of all elements at initialization is 1.

As in the formula of PageRank in equation (1), we need the information about current rank scores, vertex adjacencies and vertex out-degrees. For convenience, we also count out-degrees in advance and store in a vector $L = \{\frac{1}{l_1}, \frac{1}{l_2}, \cdots, \frac{1}{l_N}\}$. Then, each ranking iteration is a multiplication of matrices:

$$R^{new} = \frac{1-d}{N} + d \times E \times PR \times L \tag{6}$$

$$P = \frac{1-d}{N} + \begin{bmatrix} e_{11} & \cdots & e_{1N} \\ \vdots & \ddots & \vdots \\ e_{N1} & \cdots & e_{NN} \end{bmatrix} \times \begin{bmatrix} pr_1 \\ \vdots \\ pr_N \end{bmatrix} \times \begin{bmatrix} d/l_1 \\ \vdots \\ d/l_N \end{bmatrix} \tag{7}$$

Note that, for convenience in computing, we multiply d to vector L since damping factor d in equation (7) is a scalar.

3.1 Step 1: Sharing

We assume the cloud servers are "honest but curious." Owner accesses to his graph data and shares to the servers three parts of information: adjacent matrix, vertex ranks and outdegrees. Regarding to the Shamir secret sharing scheme, the secret to share should be integer. Hence, we normalize all values of matrix and vectors to integers before sharing.

For each normalized value to share v, database owner shares it to K values and sends to K cloud servers using equation (2). At the end of this step, each k^{th} cloud server has a share including an adjacent matrix E^k, a vertex rank score vector PR^k and an out-degree vector L^k.

3.2 Step 2: Ranking

Each server k separately ranks its shared data, E^k, PR^k and L^k using equation (7). In this formula, the values of elements in vertex rank vector PR^k will be updated every iteration. Other parts, i.e. adjacent matrix E^k and out-degree vector L^k are kept constant.

3.3 Step 3: Retrieving

When a user makes a request, the cloud will collect the corresponding information from servers to recover secret and result answer to user. For an example, user requests the rank score of the i^{th} vertex (with $i \in [1, \cdots, N]$), cloud collects K corresponding rank scores from k servers at the position i of vertex rank score vectors, i.e. $\{PR_i^1, PR_i^2, \cdots, PR_i^K\}$. Then, the rank score of vertices i^{th} will be recovered using equation (5).

4. Conclusions

In this work, we introduce a scheme to ensure the privacy of information during ranking with PageRank algorithm using Shamir Secret Sharing scheme. The graph information is shared to different servers.

In the future, we are going to experiment this scheme on a graph data such as Gnutella (Ripeanu et al., 2002) to measure performance of graph ranking with and without applying Shamir Secret Sharing scheme. The performance going to be measured includes accuracy and duration.

Acknowledgement

The author would like to thank Dr. Tran Viet Xuan Phuong for many discussions and advising Shamir Secret Sharing scheme being used in this work.

REFERENCES

Gleich, D. F. 2015. PageRank beyond the Web. SIAM Review, 57(3), 321-363.

Le, T. T., Phuong, T. V. 2020. Privacy Preserving Jaccard Similarity by Cloud-Assisted for Classification. Wireless Personal Communications, volume 112, 1875-1892.

Page, L., Brin, S., Motwani, R., Winograd, T. 1998. The PageRank citation ranking: Bringing order to the Web. Proceedings of the 7th International World Wide Web Conference, pp. 161-172. Brisbane, Australia.

Ripeanu, M., Foster, I., Iamnitchi, A. 2002. Mapping the Gnutella Network: Macroscopic Properties of Large-Scale Peer-to-Peer Systems. IEEE Internet Computing Journal, volume 6, 85-93.

Shamir, A. 1979. How to Share a Secret. Communications of the ACM, volume 22, 612-613.

Wakabayashi, D. 2019. Google and the University of Chicago Are Sued Over Data Sharing. The New York Times.