The application of Randomized HITS algorithm in the fund trading network

Xingyu Xu[1], Zhen Wang[1] ,Chunhe Tao[1],Haifeng He[1]

[1]The Third Research Institute of Ministry of Public Security,China

Abstract. For the economic crime investigation, it is very important for the rapid detection of the case to determine the key suspects efficiently and accurately in the fund trading network. In this paper, the design idea of Randomized HITS algorithm, combined with the fund flow characteristics of the personnel in the fund trading network , is used to study the key suspects. Based on the characteristics of the amount of funds and the frequency of transactions in the personnel relationship of the fund trading network, this paper expounds in detail how to construct the probability transfer matrix. In the end, the method is applied to a really case. From the final result, the algorithm can find the key suspects quickly from the large amount of fund trading data, improve the efficiency of information analysis. It provides a new alternative method of fund network analysis.

1 Introduction

In the economic crime investigation, the analysis of the fund trading network plays a very important role. After a few suspects have been determined, investigators will collect the data related to the suspects in accordance with the law (including bank transaction data, POS transaction data, third party payment platform transaction data, etc.). We can associate each account with a specific person (the ID number is the only person), the transaction data can build a person-to-person fund flow network. Each node in the network represents a person, and the directed edge between the nodes represents the fund flow, while the weight of the side can be defined flexibly by using the information of the transaction data (such as the amount, frequency, average and so on) according to the analytical requirements. The effective analysis of such a fund flow network can find out possible suspects or even criminal groups, so as to quickly point out the direction of investigation for investigators.

Data statistics is the most common method of fund network analysis. The investigators search out one who direct with the suspected criminal suspects from the funds data, which used to calculate the total amount, frequency, the average amount of the transaction and so on. After obtaining statistical information, the investigators infer the criminal suspects with their own experience and provide the results as a reference. This method ignores the structure information of the fund trading network, and requires the investigators to make artificial judgment based on the experience, which is weak in facing large scale data.

With the development of data visualization, there are many network visualization tools, such as Gephi[1], I2 [2], and so on. These tools can use raw data to build a fund trading network and display the network on the computer screen intuitively. Before the final presentation, these tools tend to cluster the nodes in the network based on some algorithms found by the community and determine the layout of the nodes in the two-dimensional plane according to the results of the clustering. As a result, when the final visualization results are seen, investigators can easily identify which personnel are closely related to the identified suspects, thus identifying possible suspects. Compared with the traditional data statistics method, these visualization tools can make good use of the structural information of the network and display the entire network intuitively. However, because of the large consumption of data visualization to the computer hardware resources, in the case of more network nodes, the performance and time consumption of the machine have high requirements. Even in the case of drawing the result diagram, because the nodes are more dense, it is not conducive to the relationship discovery through artificial observation. The key information is passed to the decision maker.

In recent years, with the development of Internet technology, all kinds of mobile payments and online transactions have become very common. This has directly led to the rapid growth of the scale of financial transaction data in the economic crime investigation. It also brought new challenges to the investigators who analyzed these transaction data to determine the key suspects efficiently and accurately. In this paper, the design idea of Randomized HITS algorithm is combined with the fund flow characteristics of the personnel in the fund trading network, and the Randomized HITS algorithm is introduced to solve above problem efficiently.

[a] Corresponding author: xuxy@139.com

PageRank algorithm and HITS algorithm are all based on network link structure sorting algorithm, and successfully applied to commercial search engine. The PageRank algorithm has been successfully applied to the Google search engine, and the HITS algorithm is also successfully applied to the Clever system of IBM, and all of them have achieved a good search effect [4].

The Randomized HITS algorithm is improved by HITS. The core ideas of these three algorithms are network link analysis, which are used to solve the search ranking problem. The ranking recommended for investigators is according to the important parameter parameters (PR, Authority, Hub) of the calculated nodes.

In this paper, the network is built between people through the relationship of fund transaction. The design idea of Randomized HITS algorithm is combined with the fund flow characteristics of the personnel in the fund trading network. The transfer probability matrix used in Randomized HITS algorithm is defined by introducing the total amount of funds transfer or transferring frequency information among the personnel. So as to realize the application of Randomized HITS algorithm in economic cases investigation and analysis.

2 Randomized HITS

Randomized HITS algorithm is improved based on PageRank algorithm and HITS algorithm. PageRank algorithm was proposed by Sergey Brin and Larry Page in 1998 to solve the problem of web page ranking in link analysis [5]. By defining the behavior of random walk in the web link network, the algorithm creates a Markov chain and uses the access probability in a stable distribution to characterize the importance of each web page. HITS algorithm is proposed by Jon Kleinberg in 1997, and is part of research project entitled "CLEVER" at IBM Almaden Research Center. It defines two different types of feature values for each web page, representing the authority and importance of a web page, and is different from PageRank algorithm, which defines a access probability for each web page. The core idea of the whole algorithm is based on the assumption that if a web page is linked to a lot of important pages, the authority of the page will be correspondingly large; if a web page is linked by a lot of very authoritative web pages, the importance of the web page itself will be correspondingly large. Although PageRank and HITS algorithms have a good performance in addressing the importance of web page ranking, they still have some shortcomings, it is unstable when a small amount of data missing. In a collection of web pages, when a few pages lost, the sorting results of PageRank and HITS algorithm will have a greater impact. In order to solve this instability, Andrew Y. Ng et al. Put forward Randomized HITS algorithm[8] in combination with PageRank[3] and HITS algorithms[4].

2.1 PageRank algorithm

Given a set of web pages containing N pages and the reference relationship between them, PageRank algorithm constructs an adjacency matrix A. For each element $A_{i,j}\{i,j \in \{1,2,\dots,n\}$ n is a web page number) in A, there is:

$$A_{i,j} = \begin{cases} 1 & \text{i refer to j} \\ 0 & \text{i don't refer to j} \end{cases} \quad (1)$$

By normalization of each row of the matrix A (each element divided by the sum of all elements of the row), the algorithm gets a probability transfer matrix M, representing a user's probability of transferring from a page i to a page j when browsing a web page. At the same time, PageRank algorithm defines a random jump probability, considering that the user is not necessarily clicking on a link on the web page, and may also enter a web site directly in the browser's address bar to jump to a new web page. For each web page, the user has a certain probability $(1-\epsilon)$ to select the next page according to the probability transfer matrix M, and has a certain probability(ϵ) to select one from the whole web page. As a result, each page is regarded as a state, and the jump between the web pages is regarded as a state transfer. PageRank algorithm actually defines a Markov chain, which is the state transfer matrix of the chain.

$$\epsilon U + (1-\epsilon)M \quad (2)$$

Among them, $U_{i,j} = 1/n$(for all i,j) indicates that each web page has the same probability of being selected when the user chooses randomly. After completing the Markov chain according to the above definition, PageRank algorithm initializes the probability that all web pages are accessed to $1/n$, and iterates according to formula (3).

$$p^{t+1} = [\epsilon U + (1-\epsilon)M]^T p^t \#(3)$$

$p^t = [p_1^t, p_2^t, p_3^t, \dots, p_n^t]^T$ is the probability distribution of all web pages at step t After the final iteration is stable ($||p^{t+1} - p^t|| = 0$), the algorithm can get the probability of each web page being accessed. These probabilities are quantified for the important to web page.

2.2 HITS algorithm

As for a set of web pages containing N pages, HITS algorithm, similarly to PageRank algorithm will first construct a adjacency matrix A (such as formula (1)) based on the reference relationship between web pages, and then HITS algorithm will iteratively calculate the weight (h_i) and importance (a_i) of each web page according to the following way:

$$a_i^{t+1} = \sum_{j:A_{j,i}=1} h_j^t, \qquad h_j^{t+1} = \sum_{i:A_{j,i}=1} a_i^{t+1} \quad (4)$$

$a^t = [a_1^t, a_2^t, \dots, a_n^t]^T, h^t = [h_1^t, h_2^t, \dots, h_n^t]$, the formula (4) can be written as the form of the following matrix:

$$a^{t+1} = A^T h^t = A^T A a^t \quad (5)$$

Before entering the iteration, a^0 and h^0 are initialized to$[1,1,\dots,1]^T$. Through continuous iteration to achieve the final stable state, we can get the authority and importance of each web page.

2.3 Randomized HITS algorithm

Randomized HITS algorithm assumes that the user is browsing the web based on such a probability process: every time a user wants to jump to a new web page, he first throws a coin with uneven weight distribution (the probability of upside up is the probability ϵ). If it is toward the front, he will judge the current step that is an odd step, and the odd step will be randomly selected from the web page that is linked by the current web page, and the even number of steps will be selected randomly from the links to all the pages of the current web page. If the coin is upside down, the user randomly selects a web page from the entire web page collection.

The above assumption is actually a random walk process defined in the graph formed by Web links, which is similar to that of PageRank algorithm. In odd and even steps, the probability distribution of users on different web pages is different. Analogical HITS algorithm, Randomized HITS algorithm defines the probability distribution of odd steps as the importance of the web page, defines the probability distribution of even number steps as the authority of the web page, and there is a state transfer process as follows:

$$a^{t+1} = \epsilon\vec{1} + (1-\epsilon)A_r^T h^t,$$
$$h^{t+1} = \epsilon\vec{1} + (1-\epsilon)A_c a^{t+1}$$

Among them, $\vec{1}$ is a column vector with a total element of 1, A is the adjacency matrix of the web link graph, A_r and A_c are the transfer probability matrices between the nodes obtained from the rows and columns of the A respectively. At the beginning, Randomized HITS algorithm initializes the value of importance and authority of all web pages to be 1, and then iterates until stable to obtain the importance and authority of the final page.

2.4 Implementation

In the web rankings problem, the web page in the foundation root is a web page related to search queries. On the basis of the root set root, the HITS algorithm extends the set of the set of web pages, which is extended to the collection base, including the links pointing to the root set, or the root set with a link. Pointing, they are all extended into the collection base. The HITS algorithm searches for the good "Hub" page and the good "Authority" page in this expanded webpage collection.

In the process of economic case investigation, the investigators usually take the records of the related personnel according to the suspected evidence of the existing crimes. The transaction data of the account transactions with the suspects are included in the investigation data, and the different personnel are labeled as network nodes. The characteristics of the data set selected are similar to the base data set of HITS algorithm.

In Randomized HITS algorithm design, a good "Authority" page will be directed by a lot of good "Hub" pages, and a good "Hub" page points to a lot of good "Authority" pages. Similarly, in the economic case investigation, there is often a lot of frequent exchange of funds, with a large amount of money or more frequent fund flow of the account node, including fund inflow nodes and fund outflow nodes. these nodes are usually considered significant suspect accounts. This is consistent with the high "Authority" value node in Randomized HITS algorithm and the high "Hub" value node. Based on the above consideration, Randomized HITS algorithm used to look out the key suspicions in the fund trading network.

This method needs to focus on the construction of the transfer probability matrix of Randomized HITS in the fund trading network. In the traditional HITS algorithm, the weight of the links between the web nodes is 1, that is to say, all the web nodes have no difference in the impact on the other nodes. It can't be used directly in the fund trading network. The amount of money trading and the frequency of the transaction are very important characteristic amount in the fund transaction data. That is to say, the degree of influence between the different account nodes is different. We use this characteristic to construct the transfer probability matrix. As shown in Figure 1.

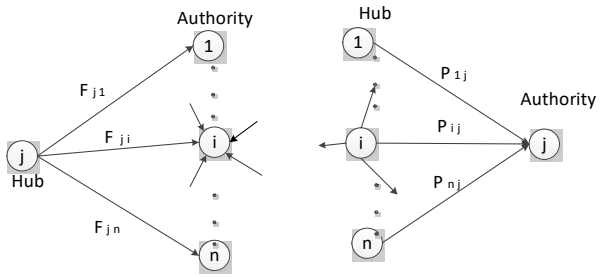

Fig. 1 Diagram of transfer probability matrix of fund network node

$$a_j = \sum_i P_{ij} h_i \qquad P_{ij} = \frac{Out_{i\to j}}{\sum Out_i}$$
$$h_j = \sum_i F_{ji} a_i \qquad F_{ji} = \frac{In_{j\to i}}{\sum In_i}$$

Before entering the iteration, a_0 and h_0 are initialized to$[1,1,\ldots,1]^T$. Among them, $Out_{i\to j}$ represents the total amount of transfer funds of account node i to account node j. Out_i represents all the total amount of borrowed funds of the account node i, P_{ij} is defined as a positive transfer probability, representing the influence degree of the account node i on the "Authority" value of the account node j; On the other hand, $In_{j\to i}$ represents the total amount of the account node j to the account node i, and In_i represents all the total loan amount of the account node i, and F_{ji} is defined as the reverse transfer probability, representing the influence degree of the account node i on the "Hub" value of the account node j.

For example ,there are three nodes j,k,l, we can build a matrix with their transaction funds.

	j	k	l
j	0	1000	4000
k	2000	0	0
l	1000	1000	0

$$P_{jk} = \frac{Out_{j\to k}}{\sum Out_j} = \frac{1000}{1000+4000} = 0.2$$

Through above matrix and calculation P_{jk}, we can construct an adjacency matrix to calculate "Authority" value.

	j	k	l
j	0	0.2	0.8
k	1	0	0
l	0.5	0.5	0

This idea combines the characteristics of the funds transaction in economic case investigation: the account number which has a close connection (with a high total amount of fund transaction) to the key suspect account (high "Authority" value node or the high "Hub" value node) has high suspicion.

At the same time, due to the definition of "Authority" value and "Hub" value, when the network is stable, the high "Authority" account node represents the focus of the inflow of funds, and the high "Hub" value account node represents the focus of the outflow of funds. This feature can be used to combine the experience of the investigators to the different roles in the criminal gang. Role labels are given for different suspect accounts in the case, providing information for the role of suspects.

In addition, as a result of quantitative calculation, when the final results are read, the importance of the two suspects can be compared according to the value of "Authority" and "Hub". Finally, the number of the account nodes is sorted according to this value. The higher ranking, the higher degree of suspicion.

3 Experiment

3.1 Setting

In this experiment, we use the data of a case, which includes bank transaction information and third party payment transactions. The number of transaction information is 7055590, including the account information, the opponent's account information, the amount of exchange and the direction of the transaction.

Based on the above information, the corresponding transfer probability matrix is constructed according to the flow of funds between the account nodes. Through the iteration of Randomized HITS algorithm, the "Authority" value and the "Hub" value of each account node are obtained.

Finally, the suspect is ordered by the "Authority" value, and the result is submitted to the investigators. The investigators check the result set through the information of the suspect and verify the effectiveness of the algorithm.

3.2 Result

The result is shown in Table 1. According to the first line check, we list the number of suspects in the rankings recommended by the algorithm.

Table 1. Number & Ranking of suspects

data sources	The Number (determined by investigators)	The Ranking （In the position of the last suspect）
Bank transactions	15	Top 20
Third party payment platform transactions	16	Top 30

Compared with the results of bank data, the results of the third party transaction data have a certain weakening effect, because the current online payment and online shopping are very common, so the degree of suspicion of some very frequent traders will also rise. This phenomenon is called the "Closely linked community phenomenon" in HITS algorithm. In view of this problem, we can further use commodity trading information to preprocess the base set of data before calculating it.

The whole calculation process costs ten seconds. From the final result, Randomized HITS algorithm can find the key suspects quickly from the large-scale fund transaction data. It can effectively improve the detection efficiency In the investigation. The rapid discovery of key suspects and the initial positioning of hierarchical roles provide a new way of analysis for the rapid detection of Internet economic crime cases, which can be used as a useful supplement to the current economic case investigation means.

4 Summary

For economic crime investigation, it is important to determind the key suspect accounts effectively and accurately in the fund trading network. In this paper, Randomized HITS algorithm is introduced to analyze the key suspects in the analysis of funds transaction network. We construct a probability transfer matrix is elaborated in detail according to the characteristics of the amount of funds and the frequency of transaction in the relationship between the personnel of the fund trading network. At last, the method is applied to a real case. From the final verification data, it can be seen that the algorithm is available. It can quickly discover the key suspects from the large amount of funds transaction data, improve the efficiency of information analysis, and strengthen the fast response ability and operational capability of information detection. Randomized HITS algorithm provides effective decision support for the case investigation.

References

1. https://gephi.org/
2. https://www.ibm.com/us-en/marketplace/analysts-notebook
3. https://en.wikipedia.org/wiki/PageRank
4. https://en.wikipedia.org/wiki/HITS_algorithm

5. http://www.almaden.ibm.com/cs/k53/clever.html
6. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report. Stanford InfoLab.
7. Jon M Kleinberg. 1999. Authoritative sources in a hyperlinked environment. J. ACM 46, 5 (1999), 604–632.
8. Andrew Y Ng, Alice X Zheng, and Michael I Jordan. 2001. Stable algorithms for link analysis. In Proceedings of the 24th annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 258–266.