Identification of key nodes in abnormal fund trading network based on improved pagerank algorithm

Jianying Xiong¹, Wen Xiao³

¹,²Security Technology Department of Jiangxi Police College, Jiangxi Police Institute Nanchang, 330013, China

Corresponding author and e-mail: Jianying Xiong, splcq@ncu.edu.cn

Abstract. The traditional network theory is unsuitable for money laundering path analysis because of the complexity of financial transaction. The abnormal accounts are identified only by effectively mining the characteristics of fund flow. The importance of transaction node in the whole illegal capital flow is equivalent to that of the connection among these nodes. Through combining the page rank algorithm with the characteristics of fund flow, an improved weighted and iterative initial value mechanism is designed to calculate the transaction heat value of the account, thus the abnormal capital transaction account is screened out. When the method is applied to a money laundering case, the results show the improved page rank algorithm can effectively screen out the key account nodes in the money laundering network and provide clues for the monitoring of money laundering crimes.

1. Introduction

With the rapid development of society and financial, there are many types of economic crimes. Economic crimes are characterized by multiple characteristics such as stakeholder, high-tech, concealment and virtuality, which seriously endanger the national financial order and people's property security. In the era of big data, economic crime is also highly integrated with information technology. In the process of violation, a large number of data traces will be produced, such as capital transaction data, communication data, traffic data, etc. Using big data to promote fund data research and judgment has become a trend of illegal behavior monitoring and investigation. However, the huge amount of capital transaction data and its complexity greatly increase the staff's cognitive and thinking burden [1]. The identification of suspicious financial transactions is basis of anti-money laundering, and the screening of abnormal accounts is the key to analyze the money laundering path. The most common fund network analysis method is the data statistics. Technicians can directly search out the accounts with direct fund transactions from the collected fund transaction data, and accurately calculate the total amount, frequency and average amount of a single transaction between them. After obtaining these statistics, technicians infer suspicious accounts based on their own experience. This method ignores the structural information of the fund trading network, and requires technical personnel to make manual judgment based on experience, so it is powerless in the face of large-scale data.

For network structure information mining, we often use degree centrality, eigenvector centrality, Katz centrality, PageRank, closeness centrality, betweenness centrality, transitivity [2]. PageRank is an outstanding method proposed by Google founders Larry Page and Sergey Brin in 1997 for the search
engine system. The algorithm is used to measure the importance of a specific page relative to other pages in the search engine index. With the success of Google, the algorithm has become a computing model of great concern in academia, and has been applied in a variety of network structure analysis. Our work will use transaction weight to improve PageRank algorithm to distinguish the importance of transaction network nodes, and find out the key nodes as suspicious account set.

2. Related research
There are some achievements in behavioral economics, operational research, data mining and other methods. The basic idea of these methods is to find some useful attribute information from the network to highlight the differences between network nodes, such as degree, shortest path, the amount of information contained in the path, etc. That is to fully reflect the location characteristics of nodes in the network, and to "enlarge" the significance of network nodes to define the importance of nodes [3]. Data mining technology has been widely used in the field of financial fraud, such as the application of expert system and decision support system based on rules to the detection of medical insurance fraud and credit card fraud [4]. The methods of classification, clustering and association rules are applied to the data compliance analysis at all stages of financial transactions to mine Suspicious Money Laundering behaviors [5]. With the development of machine learning technology, people try to use it in the field of economic crime. Most of these studies focus on finding abnormal transactions in a large number of transactions. But there are few studies on the identification of abnormal traders. In recent years, the research includes the cross-outlier detection model based on transaction characteristics, and the transaction outlier recognition based on the clustering algorithm also are proposed [6]. These algorithms use the transaction statistics to find individuals different from the subject. The behavior of money launderers in the transaction network is different from that of ordinary customers. The money laundering path is the flow of funds between accounts, so an effective filtering and identification of a large number of accounts has become an important part of money laundering. For example, Sharma proposed a large-scale multi-level clustering method based on the transaction characteristics of users to find out the accounts with abnormal behaviors and the associated accounts of abnormal accounts [7]. Colladon proposed a classification and mapping relationship method for customer risk, which constructed social network based on different risk factors [8].

The existing research have laid a technical foundation for the Anti-Money Laundering. Finding abnormal nodes is to find the most noteworthy entity connected with a given entity, which can be transformed into a link analysis problem. Page rank algorithm is a ranking algorithm based on the network link structure, which can be understood as the link value of the page as a vote, thus realizing the concept of link value. In this study, the page rank algorithm is combined with the characteristics of the amount of funds and transaction frequency of the fund trading network. By analyzing the voting ranking of the transaction account links, it is helpful to get the high-level nodes in the illegal trading network.

3. Account mining model in capital network

3.1. Abnormal capital transactions
Criminals usually transfer funds frequently and repeatedly through complicated transactions, so as to blur the illegal characteristics of criminal income and cover up the source and destination of criminal income, such as money laundering. In the research of social network analysis and link discovery, we usually transform the network into graph.

The three stages of money laundering process are as shown in Figure 1. The first is disposal stage, it refers to the process of putting the proceeds of crime into the cleaning system, which is the most easily detected stage. Then, the second is the cultivation stage, it refers to criminals’ transfer money frequently and repeatedly through complex transactions, so as to blur the illegal characteristics of criminal income and cover up the source and destination of criminal income. At the last is the fusion stage, it is vividly described as "drying up". Even if the illegal becomes legal, it provides a superficial
legal hiding for the financial obtained from crime. After the criminal income is put on the legal coat, the criminal income person can freely enjoy these dirty profits and collect the cleaned money for use.

For example in the money laundering financial network in Figure 2, node $V_1$ is the starting point, which is the illegal asset to be disposed in the money placement stage. The node $V_n$ is consolidation node, which is the money are collected after the legal transaction turnover. There are many scattered nodes $V_X$ in the middle, which are the turnover nodes of money transfer.

*Figure 1. Money laundering financial network. Figure 2. Key nodes in financial network.*

The money transfers out degree of starting nodes are very high, and in degree of the merging nodes are very high. And the intermediate turnover nodes are characterized by high-frequency complexity and large amount of money transfer. If the larger the money transfer weight is, the more important the bridge node is. Tracing the fund source for this node is conducive to finding out the money laundering path.

3.2. Process of abnormal account identification
Because the capital trading network is actually a graph, there are various nodes in the graph, and the nodes are connected to form an edge. In the fund transaction network, each account is a node, and people are connected with each other through transaction relationship. Account nodes have weighted value, just as people have influence. Using PageRank value to describe the importance degree of nodes in the trading network can help to find the important nodes, or the central nodes.

*Figure 3. Algorithm model framework.*

According to the requirements of PageRank algorithm, the calculation process is designed as shown in Figure 3, the process of transaction network account analysis based on improved PageRank
algorithm is mainly divided into three parts. Firstly, we need to preprocess the historical data and extract the characteristics according to the model requirements, extract the transaction behavior attributes of the account in the transaction flow, and then perform the feature processing according to the algorithm requirements. Secondly, according to the idea of the algorithm, we construct the link directed weighted graph model of the trading network, calculate the transaction weights of the nodes with the capital transaction attributes, iteratively calculate the link values of each node until convergence, and take the calculated values as the transaction heat values to visualize the distribution and list the order. Finally, according to the selected account nodes, other attribute characteristics of the account are associated, such as enterprise account information, personal account information, etc. Through the preliminary risk investigation, and then with the key account as the clue to carry on the transaction network analysis, excavates the illegal fund transfer path.

3.3. Improvement of pagerank algorithm
PageRank algorithm is a web page ranking algorithm. The two basic key points of the algorithm are: 1) the number of links. The more links a page has from other pages, the more important the page is. 2) Link quality: the higher the weight of a web page, the more important it is. Based on the above idea, the PageRank value formula of a web page is as follow:

$$S(V_i) = \sum_{j \in \text{in}(V_i)} \left( \frac{1}{\text{Out}(V_i)} \right) \times S(V_j)$$

(1)

In(V_i) is the in chain set of V_i. Similarly, Out(V_i) is the out chain set of V_i, Out(V_i) is the number of out chain. Because each page has to contribute its own score to each out chain, then \(\frac{1}{\text{Out}(V_i)}\) is the fraction that V_j contributes to V_i. Add up all the scores that V_i has contributed to its in the chain, which is the score of V_i.

Of course, the score of each web page calculated in this way is related to the score of other linked pages. The algorithm sets the score of all web pages to 1 at the beginning, and then converges the score of each page through multiple iterations. The score of convergence is the final score of the web page. If it can’t converge, the calculation can be controlled by setting the maximum number of iterations. The score that the calculation stops is the score of the web page.

PageRank is a directed and unauthorized graph, and there is a measurement of transaction relationship weight between nodes, which requires the use of weighted graph. Because in addition to considering the importance of different transaction account connections, we should also consider the degree of transaction relevance. It is defined as transaction intimacy, which is mainly determined by the amount and the number of fast in and fast out transactions. In order to unify the data dimension, the threshold reference standard for transaction volume and transaction times is standardized. Among them, the number of fast in and fast out transactions F_j is the statistical quantity which time interval between the entry time \(t_i\) and the out time \(t_o\) does not exceed the set time threshold \(t\), that is \(F_j\) counts all the quantities that meet \(|t_i-t_o|<t\).

$$W_{ij} = \sqrt{\frac{(M_j)^2}{M_t} + \frac{(F_j)^2}{F_t}}$$

(2)

When calculating the contribution of each account to the target account, it is not through the way of average distribution, but by calculating the proportion of the weight in the total weight. Here, the weight is the transaction intimacy between the two accounts. Therefore, the expression of improved PageRank is as (3).

$$S(V_i) = \sum_{j \in \text{in}(V_i)} \left( \frac{W_{ij}}{\sum_{k \in \text{out}(V_j)} W_{jk}} \right) \times WS(V_j)$$

(3)

In addition, compared with the original PageRank algorithm, the importance of nodes are not been initialized to 1 due to the large difference of transaction funds among nodes. The initial value can be classified by the whole fund transaction activity.
Table 1. Initial value of node importance.

| Level | Initial Value |
|-------|---------------|
| High (C4-C5) | 1.00 |
| Compared high (C3-C4) | 0.75 |
| Medium (C2-C3) | 0.50 |
| Low (C1-C2) | 0.25 |

As shown in Table 1, the transaction activity is calculated based on the total amount of capital inflow and outflow, and the discretization standard is set according to 4 percentile. The initialization values of node importance of different levels are given to different activity levels.

4. Experimental analysis

The experimental data comes from illegal fund evidence collection cases. Relevant fields are extracted from the transaction, and different original data are transferred to SQL after preprocessing in server2012. It is merged into a transaction flow information table. There are 180559 records in the table. The attributes are filtered and some non transaction data are removed. After the preprocessing of de duplication, empty lines and micro transactions below 100 yuan, a total of 168507 data records are sorted out, 2660 transfer out accounts and 11206 transfer in accounts. The transaction time is from April 2017 to April 2018. The data format of capital flow is shown in the Figure 4.

Figure 4. Cash flow data format.

In order to determine each threshold, the transfer out accounts are grouped, and the capital transaction times and transaction amount of each account are analyzed. The statistical results are shown in Figure 5. In the measurement of transaction relationship weight, When standardizing the transaction volume and number of transactions, the reference thresholds are set from the respective median, where are 2 and 50500. In practice, the threshold can also be set and adjusted according to the system experience value.

Figure 5. Basic statistical analysis of data.

The weights between all trading nodes can be calculated based on threshold to realize the directed weighted graph of transactions. The calculation result of transaction weights is shown in Figure 6.

Figure 6. Transaction weight calculation.
nodes, and an intuitive distinction between high-temperature link values and other nodes. According to the popularity ranking, account tracking and investigation can be carried out. As shown in Figure 8, the account information with the highest link value in the case is intercepted. Because the fund data flow is very complex, it is very difficult to find clues from the massive fund network. However, it is very time-consuming to find clues completely from the traditional method, which requires manual index by index and account by account investigation, so the improved PageRank algorithm can better filter information.

**Figure 7.** Distribution of heat value of account.  
**Figure 8.** Heat value ranking of accounts.

5. Conclusions

Aiming at the problem of identification of suspicious fund transaction, this paper combines the fund transaction network and PageRank algorithm based on the characteristics of financial transaction data flow. Through this model, transaction accounts suspetive of money laundering are screened out from a large number of suspicious transaction databases. The experimental results of the case also show that the calculation process of the model is clear and the algorithm can converge well. For account node identification, the value of link popularity not only reflects the clustering effect of high-intensity nodes, but also accurately distinguishes the high-intensity link value caused by the abnormal behavior of a large number of frequent transactions of each node. The calculation results can directly show the link heat value of each node, sorting to compare. The calculation basis is highly interpretable, being conducive to the follow-up verification and investigation.

References

[1] Jing-Jing, Sun. "Investigation Mechanism Innovation of Economic Crimes involving the Masses in the Big Data Era." Journal of Sichuan Police College, 2019(1), pp.27-33. (In Chinese).

[2] Bioglio, Livio, and R. G. Pensa. "Modeling the Impact of Privacy on Information Diffusion in Social Networks." International Workshop on Complex Networks Springer, Cham, 2017.

[3] Xiong J J, Zolotov V. Fast path traversal in a relational database-based graph structure: U.S. Patent Application 16/038,498. 2018-12-20.

[4] Kharote M, Kshirsagar V P. Data Mining Model for Money Laundering Detection in Financial Domain. International Journal of Computer Applications, 2013, 85(16), pp.61-64.

[5] Lokanan M E. Data mining for statistical analysis of money laundering transactions. Journal of Money Laundering Control, 2019, 22(4), pp.753-763.

[6] Chao, X., Kou, G., Peng, Y., & Alsaadi, F., E. Behavior monitoring methods for trade-based money laundering integrating macro and micro prudential regulation: a case from China. Technological and Economic Development of Economy, 2019, 25(6), pp. 1081-1096.

[7] Sharma A G, Kurniadi F. Applying multi-level clustering at scale to unlabeled data for anomaly detection and security: U.S. Patent 9,953,160. 2018-4-24.

[8] Colladon A F, Remondi E. Using social network analysis to prevent money laundering. Expert
Systems with Applications, 2017, 67: 49-58.