International Journal of Applied Science and Mathematics
Volume 10, Issue 3, ISSN (Online): 2394-2894

Node Importance Measurement of Audit Relationship Network Based on Neo4j

Yifei Ye
Department of Accounting, Zhejiang University of Finance and Economics, Hang Zhou, China.
Corresponding author email id: 292576695@qq.com

Date of publication (dd/mm/yyyy): 26/06/2023

***Abstract* – The response measures for audits when facing constraints related to human and material resources are to focus on key points and the main contradictions. Therefore, how to measure the importance of each node in the audit relationship network becomes a critical issue. Based on Neo4j graph database technology and inspired by the concept of degree centrality in social network analysis, this research proposes the indicator of "The Hotness value" to evaluate the importance of nodes in the audit relationship network. Using this design, the "The Cumulative Hotness Value" is calculated to evaluate the influence of nodes, which helps audit personnel identify important nodes in the relationship map, also known as the "critical few" that the audit focuses on.**

***Keywords* – Neo4j, Node Importance, The Cumulative Hotness Value.**

I. INTRODUCTION

1.1. *Research Background*

Since the 1980s, the development of information technology has had a profound impact on various industries, including auditing. With the advancement of information technology, various non-relational databases, such as NoSQL (Not only SQL), have emerged. Among them, graph databases have attracted the attention of auditors as the best solution for processing massive data in the cloud era. Due to their different underlying technical principles, graph databases have powerful performance in processing big data, especially in locating relationships between different objects, which exceeds the capabilities of traditional relational databases. This performance matches the business needs of auditors in locating human-to-human and human-to-enterprise relationships in the audit relationship network, making them play an important role in auditing.

1.2. *Research Significance*

Currently, research on the combination of big data and auditing mainly focuses on the management level and platform construction level, with few studies on the specific application level. The application of graph database technology in auditing practice is even rarer. Based on an understanding of graph database technology and experience in participating in audit work, this study believes that the topic has the following significance: (1) Improved the traditional definition of node hotness by proposing a more practical calculation model, which provides technical support for effectively ranking nodes, thereby realizing the focus on key contradictions and critical few. (2) Helps to innovate auditing technology methods, promotes research on database-based auditing queries towards non-relational databases, and enables auditing work to adapt to the social development trend of the big data era. (3) Contributes to enriching the research content of big data + auditing and building typical cases of information technology-assisted technology-strong auditing. (4) Promotes the systematic development of "big data auditing" and encourages interdisciplinary research between auditing and computer science.

1.3. *Research Methods*

(1) The Literature Research Method.

By searching relevant literature resources in databases such as Wanfang Data, CNKI, books, and online resources, this research summarizes the development and latest progress of the combination of social network analysis technology and auditing work. Based on this, it clears up the research ideas, laying a solid foundation for subsequent research.

(2) The Simulation Research Method.

Simulation originally refers to using models (physical models, mathematical models) to replace real systems for experimentation and research. In this study, considering the sensitivity of real audit case data, the research simulates real scenarios, designs a simulation data set that includes a certain number of relationships between enterprises, legal entities, investors, and natural persons. Secondly, various real-life applications are performed on the simulation data set to improve the algorithms and compare their performance.

1.4. *Innovation*

First, this research innovatively proposes a standard "The Hotness value" for rating the importance of all nodes in the graph database based on the traditional scheme. Based on this, a model is designed to evaluate the influence of nodes, helping auditors to effectively lock important nodes in the relationship map, that is, the "critical few" that the audit focuses on.

Second, due to the use of traditional relational databases in current auditing work, which has shortcomings in discovering association relationships, and the current application of graph database technology is still in its infancy, scarce research has been conducted on its application in auditing. Therefore, this research is novel in both content and method.

II. Review of Related Theories and Literature

2.1. *Six Degrees of Separation*

The theory of social networks originates from the Six Degrees of Separation theory. In the 1960s, American psychologist Stanley Milgram conducted a chain letter experiment and proposed the famous Six Degrees of Separation theory, which states that no two strangers in the world are separated by more than six people, meaning that at most, a connection can be established between two strangers through six intermediaries. Although the Six Degrees of Separation theory has never been rigorously proven in the past decades, there are few exceptions in real life.

Figure 1 shows four nodes and three types of relationships. In this figure, we refer to the "link depth" between nodes A and D as 3.

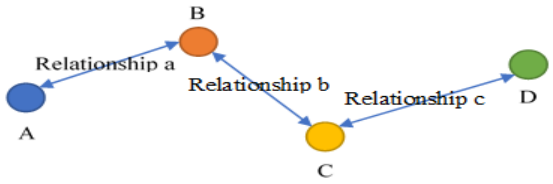

Fig. 1. Objects and the relationships between objects.

The Six Degrees of Separation points out a new direction for auditing. The discovery of relationships between people, between people and enterprises, and between enterprises is a common requirement in various audit

businesses, especially in economic responsibility audits. Business needs require technical support. In traditional relational databases, can we use SQL query statements to capture the relationships between these “objects” ? The answer is yes, but it is foreseeable that as the connection depth increases, the implementation difficulty of such query requests increases geometrically, making it increasingly impractical. Therefore, this study uses the Neo4j graph database, which has powerful performance beyond traditional relational databases, to study aspects such as locating relationships between different objects.

2.2. *Literature Review*

In recent years, there has been significant attention among scholars towards the measurement and evaluation of the importance of nodes in a relational network. Hennan et al. (2007) analyzed and summarized various methods for selecting important nodes in complex network environments, including social network analysis, system science analysis, information search domain analysis, and relative importance of nodes in a network [1]. Liao Kangli et al. (2017) introduced social network analysis into the technical method system for grassroots corruption governance, identifying and analyzing the structure of the special complex network of grassroots corruption [2]. Yang Chao et al. (2022) used social network analysis technology to construct a relational network of audit objects, introducing transactions as nodes in the social network [3]. Wei Cong (2018) extended audit data analysis to comprehensively analyze the complex relational network of audit objects [8], while Wei Cong (2022) introduces social network analysis to improve the efficiency of audit data analysis [9].

Jiang Xiaoyi (2019) suggested that the focus of social network analysis is on the interactions and connections between social individuals and their impact on social behaviour [4]. Chen Wei et al. (2019) proposed an audit method based on social network analysis and successfully implemented it using the R Programming Language [5]. Yao Xiaotao et al. (2003) elaborated on the core concepts of social network theory and the relationship between social networks and enterprise growth. Moreover, they analyzed and explored the potential intersections between social network theory and enterprise problems [6]. Jiang Xiaoyi (2019) pointed out that social network focuses on the interaction and connection between social individuals and their influence on social behavior. The activities of audit objects are always in this network. Based on this, it is proposed to use big data to automatically build a social network related to audit objects, and then use path discovery, community analysis, node role identification, visualization and other analysis methods to efficiently find audit clues [7]. Lv Tianyang et al. (2020) applied social network analysis as a new computer audit technology to extend the analysis of audit objects from isolated individuals to their relationships, thereby revealing the abuse of social relationships in the illegal behavior of audit objects. Based on the analysis of 216 publicly released audit cases, they elaborated on the mathematical foundations and basic methods of social network analysis in auditing, revealing its unique characteristics compared to similar studies in sociology and other fields in terms of the scope of analysis objects, general and specific analysis goals, unbiased and covert object behavior, and rational and hostile result validation. On this basis, they summarized the six-stage practical application path of social network analysis in auditing and the challenges faced by this method [10].

III. Simulation Case Study

3.1. *Data Preparation*

In order to conduct an in - depth study of the measurement of important nodes in the audit network and taking

into consideration the powerful performance of Neo4j graph databases in searching for relationships, this study selected Neo4j graph databases to operate. The study designed a simulated case dataset using actual audit cases as a reference, which also avoided the leakage of sensitive data. The relational tables of the dataset were distributed across four CSV files, including “The Legal Representative of the Company” with 3,000 records, “The Company investors of the company” with 1,999 records, “The Natural person investors in a company” with 11,563 records, and "Relatives" with 509 records. The above data tables are stored in csv files respectively, and their partial contents are as follows Table 1, Table 2, Table 3, Table 4. As shown:

Table 1. The Legal representative of the company.

Company ID	Corporate Representative
Com1001	1993
Com1002	1936
Com1003	2111
Com1004	2723
Com1005	1712

Table 2. The Company investors of the company.

Company ID	Company Investor ID
Com1001	Com3822
Com1006	Com3484
Com1007	Com2904
Com1008	Com3517
Com1008	Com1978

Table 3. The Natural person investors in a company.

Company ID	Natural person investor ID
Com1001	1429
Com1001	162
Com1001	1577
Com1001	9574
Com1002	7886

Table 4. Relatives.

Member 1	Member 2
7182	7992
4590	5652
4281	8262
8446	618
5381	9786

Taking Neo4j Desktop version under Windows operating system as an example, create a new project, named "Business", in the software. Under the project "Business", create a new database or graph named "test" and start the "test" database. After it is started, open the Browser tool and use the command to import the data into the graph, which are the four CSV files mentioned previously.

Execute the following import command:

```
/* load entities and relationships */
LOAD CSV WITH HEADERS FROM "file:///The Legal representative of the company.csv" AS row
MERGE (c: enterprise {enterprise number: row.company ID})
MERGE (p: personnel {personnel number: row. legal person})
CREATE (p)-[r: legal representative ]->(c);

LOAD CSV WITH HEADERS FROM "file:///The Natural Person investors of company.csv" AS row
MERGE (c: enterprise {enterprise number: row.company ID})
MERGE (p: personnel {personnel number: row. investor natural person ID})
CREATE (p)-[r: investment ]->(c);

LOAD CSV WITH HEADERS FROM "file:///The Company investors of company.csv" AS row
MERGE (c: enterprise {enterprise number: row.company ID})
MERGE (cc: Enterprise {Enterprise Number: row. Investor Company ID})
CREATE (cc)-[r: investment ]->(c);

LOAD CSV WITH HEADERS FROM "file:///Relatives.csv" AS row
MERGE (x: personnel {personnel number: row.member 1})
MERGE (y: person {person number: row.member 2})
CREATE (x)-[r: relatives ]->(y);

LOAD CSV WITH HEADERS FROM "file:///Relatives.csv" AS row
MERGE (x: personnel {personnel number: row.member 1})
MERGE (y: person {person number: row.member 2})
CREATE (y)-[r: relatives ]->(x);
```

After the command execution is completed in this example, create the necessary audit relationship network in the database, which consists of 10,370 nodes including two types - individuals and companies, and 17,580 relationships including three types - family relations, investment, and legal representative.

3.2. *Measurement of Importance of Audit Nodes*

The previous section introduced the data import, node creation, and relationship construction in Neo4j graph databases, and the use of Cypher for relationship queries can help auditors discover valuable clues. However, auditing still faces the challenge of how to effectively identify key points and major contradictions among massive amounts of data. Therefore, this study designed an index for measuring "importance" and implemented sorting by importance in Cypher to address this challenge.

The usual practice for audit in the face of human and material constraints is to focus on key points and major contradictions. Based on this approach, this study proposed the Hotness value to measure the importance of nodes in the relationship network. The Party's 20th National Congress report also mentioned "grasping the crucial minority" and "leading from the top down." It can be seen that the ideas proposed in this paper are consi-

-stent with the Party's policies and the risk-oriented audit thinking.

Many studies have shown that finding influential nodes in a relational network is a typical problem, and these few nodes in the network play a strong controlling role in its formation, evolution, and information transmission, making them valuable to auditors for discovering audit doubts and collecting audit evidence. Therefore, measuring and evaluating the importance of nodes is a key content of this research.

3.3. *The Hotness Value*

The Hotness value refers to the number of nodes directly connected to a certain node. The more nodes directly connected to it, the higher its Hotness value, indicating that the node has a higher status and greater influence. The formula for calculating the Hotness value is as follows:

$$Hotness\ (x) = x(\text{in}) + x(\text{out})$$

Therefore, in this study, the Hotness value of a node is the sum of its incoming and outgoing relationships. To obtain the Hotness value of each node, we need to follow these steps. First, we need to list the number of incoming and outgoing relationships of all personnel nodes, and sort them in descending order. The specific Cypher statement is as follows:

```
MATCH (n: personnel) --(m)
Personnel number, COUNT(m) ORDER BY COUNT(m) DESC
```

After running the above code, the Hotness value of all personnel nodes is obtained and output as a Table file. Some of the Table data is shown in Table 5:

Table 5. All personnel node heat value Hot sorting (partial).

N. Personnel Numbers	COUNT(m)
2865	12
1065	10
2831	9
4260	9
7137	8
…	…

By executing a Cypher query similar to the first section, the correctness of the results in the above table can be verified. Similarly, list the number of outgoing relationships of all company nodes and sort them in descending order. The specific Cypher statement is as follows:

```
MATCH (n: enterprise)-> (m)
Business number, COUNT(m) ORDER BY COUNT (m) DESC
```

After running the above code, the outgoing degree of all company nodes is obtained and output as a Table file. Some of the Table data is shown in Table 6:

Table 6. Sort all enterprise nodes by degree (partial).

N. Business Number	COUNT (m)
Com1158	5

N. Business Number	COUNT (m)
Com1242	5
Com1378	5
Com1605	5
Com1066	4
…	…

Similarly, list the number of incoming relationships of all company nodes and sort them in descending order. The specific Cypher statement is as follows:

```
Match (m)-> (n: enterprise)
Business number, COUNT(m) ORDER BY COUNT(m) DESC
```

After running the above code, the incoming degree of all company nodes is obtained and output as a Table file. Some of the Table data is shown in Table 7.

Table 1. Sort all enterprise nodes by click-through (partial).

N. Business Number	COUNT(m)
Com1070	12
Com1247	12
Com1621	12
Com1691	12
Com2376	12
Com2421	12
Com2567	12
…	…

Similarly, list the total number of incoming and outgoing relationships of all company nodes and sort them in descending order. The specific Cypher statement is as follows:

```
MATCH (m)--(n: corporate)
Business number, COUNT(m) ORDER BY COUNT (m) DESC
```

After running the above code, the Hotness value of all company nodes is obtained and output as a Table file. Some of the Table data is shown in Table 8:

Table 8. All Enterprise Node Heat Values Hot Sort (Partial).

N. Business Number	COUNT(m)
Com2062	15
Com1070	14
Com1621	14
Com2376	14
Com2810	14

N. Business Number	COUNT(m)
Com3719	14
Com3799	14
…	…

3.4. *The cumulative Hot value*

The Hotness value can reflect to some extent the importance and influence of a node in its relationship network, but the influence of a node is often not determined by the number of nodes it is connected to. The Hotness value of the connected nodes has an important impact on the evaluation of the node of interest. Therefore, this study proposes the cumulative Hot value based on this, to appropriately evaluate the importance and influence of a node in its relationship network.

The cumulative Hot value is defined as the sum of the Hotness value of the surrounding nodes, which reflects the comprehensive influence of the surrounding nodes.

3.5. *Analysis of Audit Hotspots*

In audit work, we can use this method to filter out key objects, so as to make targeted progress in audit work.

To grasp the main contradictions in auditing, it is necessary to understand which companies are linked to companies with relatively high Hotness value. To do this, traditional analytical thinking can be used to query all companies' in-degree and out-degree using Cypher, and sort them in descending order. The code is as follows:

```
MATCH (m)--(n: corporate)
Business number, COUNT(m) ORDER BY COUNT(m) DESC
```

See in the first 5 lines of the output Table 9:

Table 9. Ranking of hot companies.

Enterprise Number	Count (m)
Com2062	15
Com2376	14
Com2810	14
Com1070	14
Com1621	14

However, we consider that evaluating a company’s Hotness value only based on the sum of its in-degree and out-degree relationships may be less rigorous. This is because this approach does not consider the Hotness value of the in-degree companies and individuals, as well as the out-degree companies themselves. Therefore, an improved approach is to first assign a Hotness value to each individual and company, and then list the sum of all in-degree and out-degree Hotness values for all companies, sorted in descending order. This study believes that this approach to identifying hot companies is relatively more scientific.

The query code based on the above improvement approach is as follows:

```
MATCH (n: personnel) --(m)
```

```
WITH n,COUNT(m) AS count
SET n.hot = count;
MATCH (n: corporate) --(m)
WITH n, COUNT (m) AS count
SET n.hot = count;
# lists the sum of all the Hot in and out of all the chains for all the companies, in descending
order
MATCH (m)--(n: corporate)
WITH n, m.hot AS hot
Return n.Enterprise number, SUM(hot) ORDER BY SUM(hot) DESC
```

See in the first 5 lines of the output Table 10:

Table 10. Ranking of Hot Companies (Based on New Rules).

Enterprise number	COUNT (m)
Com3778	73
Com2062	72
Com3719	71
Com3731	68
Com1111	65
…	…

As can be seen, under the new rules, there are significant differences between the rankings in the new ranking list and the original Table 9. However, the accuracy of both tables can be verified item by item. This fact indicates that the new rules for identifying hot companies are more scientific and effective. Using the same approach, we can also analyze and obtain popular entrepreneurs. Whether it is hot companies or hot entrepreneurs, they can help auditors capture hot objects from the complex business relationship network, which are the objects that auditors need to focus on. This study can inspire audit practitioners to master new data analysis tools and achieve twice the result with half the effort in audit projects.

IV. Conclusion

In summary, this study draws on the concept of degree centrality in social network analysis and further proposes the Hotness value to evaluate the importance of network nodes, as well as the concept of cumulative Hot value. This helps auditors to identify important nodes in the relationship map. This study validated this approach of using cumulative Hot value through simulation case data designed in the Neo4j graph database. This provides auditors with a new analytical method for quickly identifying objects that require special attention, and provides new tools for improving audit quality and efficiency.

References

[1] Henan, LI Deyi, GAN Wenyan, ZHU Xi. Review on important node mining in complex networks. Computer Science, 2007(12): 1-5+17.
[2] LIAO Kangli, OUYANG Xiaoming. Structural characteristics of grassroots corruption networks and technical paths for collaborative governance of discipline inspection and supervision based on social network analysis. Contemporary Economics, 2017(12):24-34.
[3] Yang Chao, ZHU Jingxiang, HU Xiaozhen, GAO Qianqian. Discovery of audited object's related party relationships based on social network analysis. Finance and Accounting Monthly, 2022(22):110-115.
[4] Jiang Xiaoyi. Research on new auditing methods based on social network analysis. Market Weekly, 2019(01): 64-65.
[5] Chen Wei, Zhan Minghui, Chen Wenxia. Research on audit methods for managing risks of financial technology system users based on

social network analysis. China Certified Public Accountant, 2019(12):74-78.
[6] Yao Xiaotao, XI Youmin. Social network theory and its application in enterprise research. Journal of Xi'an Jiaotong University (Social Sciences), 2003(03):22-27.
[7] Jiang Xiaoyi. Application of social network analysis method in audit evidence collection. Nanjing Audit University, 2019.
[8] Wei Cong. Research on network analysis of audit objects and its application case. Shandong University of Finance and Economics, 2018.
[9] Wei Cong, NIU Yanfang, LU Jie. Analysis and application case of audit object network. SME Management and Technology, 2022(03):147-149.
[10] LV Tianyang, QIU Yuhui, YIN Peng. Theoretical, characteristic and practical path of social network analysis in auditing. Finance and Accounting Monthly, 2020(13):103-111.

AUTHOR'S PROFILE

Yifei Ye, from Zhejiang, China, is a second-year graduate student at Zhejiang University of Finance and Economics, majoring in auditing, and follows his mentor to the direction of big data auditing. In the future, he hopes to become a computer teacher.