TaxVis: a Visual System for Detecting Tax Evasion Group

Hongchao Yu
School of Electronic and Information Engineering
Key Lab of Intelligent Networks and Network Security
Xi'an Jiaotong University
Xi'an, Shaanxi, China
yuhongchao@stu.xjtu.edu.cn

Huan He
School of Electronic and Information Engineering
Key Lab of Intelligent Networks and Network Security
Xi'an Jiaotong University
Xi'an, Shaanxi, China
hehuan@mail.xjtu.edu.cn

Qinghua Zheng
School of Electronic and Information Engineering
Key Lab of Intelligent Networks and Network Security
Xi'an Jiaotong University
Xi'an, Shaanxi, China
qhzheng@mail.xjtu.edu.cn

Bo Dong
School of Continuing Education
National Engineering Lab for Big Data Analytics
Xi'an Jiaotong University
Xi'an, Shaanxi, China
dong.bo@mail.xjtu.edu.cn

ABSTRACT

The demo presents TaxVis, a visual detection system for tax auditor. The system supports tax evasion group detection based on a two-phase detection approach. Different from the pattern matching based methods, this two-phase method can analyze the suspicious groups automatically without artificial extraction of tax evasion patterns. In the first phase, we use a network embedding method node2vec to learn representations that embed corporations from a Corporation Associated Network (CANet), and use LightGBM to calculate a suspicious score for each corporation. In the second phase, the system use three detection rules to analyze the transaction anomaly around the suspicious corporations. According to these transaction anomalies, we can discover potential suspicious tax evasion groups. We demonstrate TaxVis on tax data of Shaanxi province in China to verify the usefulness of the system.

CCS CONCEPTS

• **Information systems** → **Expert systems**; • **Applied computing** → *System forensics.*

KEYWORDS

TaxVis; Network embedding; Tax evasion group.

ACM Reference Format:
Hongchao Yu, Huan He, Qinghua Zheng, and Bo Dong. 2019. TaxVis: a Visual System for Detecting Tax Evasion Group. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019,San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308558.3314144

1 INTRODUCTION

Tax evasion detection is a challenging work for every country. It is reported by the Chinese government that the loss rate of tax revenue in China is above 22 percent [7]. In addition to suspicious individual detection, there is a new tendency that corporations plot together to evade tax [2, 7]. The existing solutions for detecting tax evasion groups are mainly based on graph pattern matching. In these methods, the pattern of tax evasion group must be extracted from massive cases by the tax auditors manually. As a result, there is a limitation that these patterns cannot be constantly updated to keep up with the rapid changes of tax evasion since it takes great effort for tax auditors to identify new patterns.

Given the limitations of the pattern matching based methods, we designed a two-phase approach for detecting tax evasion groups. The first phase is suspicious individual detection which use node2vec and LightGBM to detect suspicious corporations in the CANet. In [2, 6, 7], we can find that the relationships between corporations play an import role in many tax evasion patterns. As show in Figure 1, in order to utilize these relationships, we choose the network embedding method node2vec to automatically embed the corporations from the transaction and investment relationship graph. And then we use LightGBM to calculate a suspicious score for each corporation. The second phase is suspicious group analysis which can inspect whether the corporation belongs to a suspicious tax evasion group. In this phase, we use three rules to detect transaction anomalies around each corporation, and then according to these anomalies we can get some suspicious groups.

We demonstrate TaxVis based on the tax data of Shaanxi province. After selecting the detection year, our system will automatically compute the suspicious score of each corporation. Next, clicking one corporation, the system will display the details of its transactions, which include the transaction volume, the commodity category, and the transaction corporation etc. And then based on these transaction details, we use three rules to analyze the abnormal transaction of the selected corporation in a transaction flow chart. Finally, we can find whether the corporation belongs to a suspicious tax evasion group. A demonstration video is available at the following address: https://youtu.be/46M0talBclo. Our contributions are as following:

- We propose a two-phase approach to calculate the suspicious score based on the representations of corporations in CANet and detects suspicious tax evasion groups automatically.

WWW '19, May 13–17,2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3314144

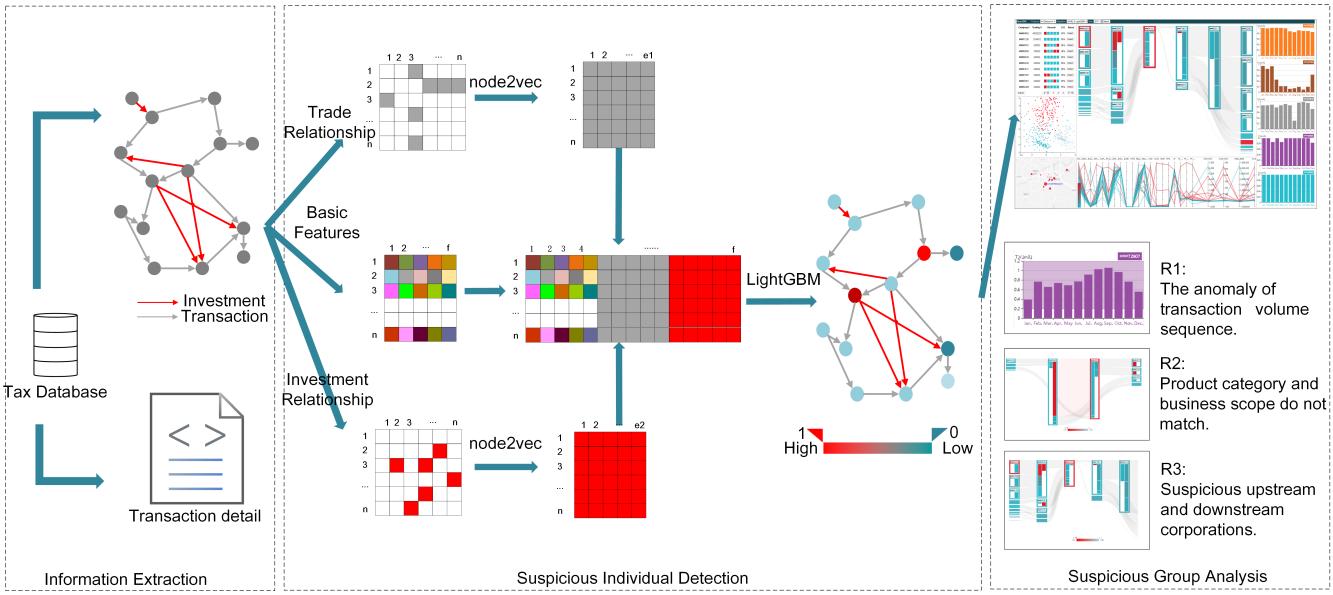

Figure 1: The structure of the TaxVis. First, we extract information from tax database. Then we detect suspicious individuals by node2vec and LightGBM. Finally, we use a visualization system to analyze the suspicious tax evasion groups.

- We design a transaction flow chart to visualize the transaction anomalies around corporations.
- We present a visual analytic system based on our proposed two-phase approach and the transaction flow chart for tax audition officer to discover suspicious tax evasion groups.

2 SYSTEM DESIGN

The structure of this system is shown in Figure 1, which consists of three components.

The first component is the information extraction, which collects all the basic features of corporations and the relationships between them, including transaction and investment relationships. The second component is the suspicious individual detection. It detects suspicious individuals by combining the node representation vectors and basic feature vectors. The last component is the suspicious group analysis. We use three detection rules to detect whether the corporation belongs to a suspicious group. The details of each component are as follows.

2.1 Information Extraction

In this component, we extract the corporationâĂŹs information from tax database. Each corporation's information consists of three parts. The first part is the corporationâĂŹs basic features, including the corporation's registration time, the registration area and the age of the legal representative, etc. The second part is the transaction and investment relationships between corporations. The third part is the transaction detail data that can be used to detect suspicious groups. The transaction detail includes the transaction volume per month, the transaction time and the commodity category, etc. Based on the above data, we build a corporation associated network (CANet) $G = \{V, E_1, E_2, A\}$, where $V = \{V_1, \ldots, V_n\}$ is a set of corporations, $E_1 = \{e^1_{ij}\}$ denotes the transaction relationships between corporations, $E_2 = \{e^2_{ij}\}$ denotes the investment relationships between corporations, and $A = \{a_i\}$ denotes the attributes of corporations.

2.2 Suspicious Individual Detection

In this component, we detect suspicious individuals in the CANet with node2vec and LightGBM. The technical details are as follows.

2.2.1 Node2vec. Node2vec is a network embedding method based on random walk that can learn continuous feature representations for nodes in a network [4]. The method is built on the assumption that the representations of adjacent nodes are approximated. Choosing node2vec is based on the following three reasons.

Transaction and investment relationships are important information in detecting tax evasion. In the previous researches about tax evasion groups, transaction and investment relationships between corporations play an important role in some suspicious tax evasion patterns [7]. Therefore, itâĂŹs necessary to combine these relationship features into the corporation features, and node2vec is suitable to learn relationship representations of corporations in the transaction and investment network.

Low time complexity. The time complexity of node2vec is O(|V|d), where V is the set of the vertices in the graph and d is the representation dimension [3]. Since in practice we have an extremely large CANet that contains millions of corporation nodes, it is necessary to choose node2vec to perform real time detection.

Both Local and global network structure can be extracted. According to the previous studies, some suspicious patterns involve not only local structure, but also global structure in the transaction and investment relationships [2, 7]. Compared with the previous methods, node2vec provides a trade-off between breadth-first (BFS) and depth-first (DFS) graph searches. Through breadth-first search,

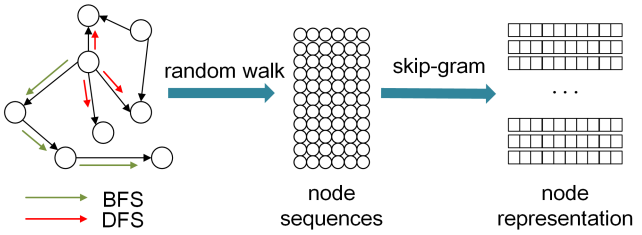

Figure 2: The procedure of the node2vec.

we can get local network structure information, and through depth-first search, we can get global network structure information. In this algorithm, we can get any combination of the two kinds of information by adjusting the parameters.

As is shown in Figure 2, node2vec consists of three steps: computing transition probability matrix, obtaining node sequences in the graph by random walk, and representing the node vectors by the skip-gram algorithm. In this system, we use node2vec to represent the transaction network and the investment network respectively, then combine the representation vectors and the basic features of the nodes, and finally combine these as inputs to LightGBM algorithm to get the suspicious score of each corporation.

2.2.2 LightGBM. LightGBM is a gradient boosting decision tree algorithm, which is effective when the feature dimension is high and data size is large [5]. In the experiment, we tested four classification algorithms (i.e., LightGBM, RandomForest, SVM and AdaBoost) for better performance. For each algorithm, we performed two experiments separately, one uses all the features of the corporation, and the other uses the basic features of the corporation. The result is show in Figure 3, in which the dotted line represents the basic feature experiments, and the solid line represents all feature experiments. As we can see, LightGBM outperforms other algorithms when we use all features. Besides, compare with using basic feature, using all features significantly improved the performances of suspicious individual detection algorithms.

2.3 Suspicious Group Analysis

In this component, we develop a visual analytic system to anticipate the suspicious groups with the following detection rules. **Rule 1: IF the sequence of transaction volume has anomaly, THEN the transaction is fraudulent**. The anomaly of the transaction volume is an important signal for fraud behaviors [1]. Tax evasion case studies found that tax evasion corporations that have been established soon will run away if they open a lot of invoices in a short time. As shown in Figure 4(c), in our system we use the $3-\theta$ principle to detect transaction anomaly. When the trading volume of a month exceeds a range calculated based on the trading volume of past 12 months, the edge in the flow chart will be marked red and the warning message will be shown when mouse moves over the edge.

Rule 2: IF the product category does not meet the business scope of the two parties, THEN the transaction is fraudulent. This rule refers to the fact that some corporations manage to increase their own input costs by purchasing invoices of other category goods. Therefore, in our approach, when a corporation buys

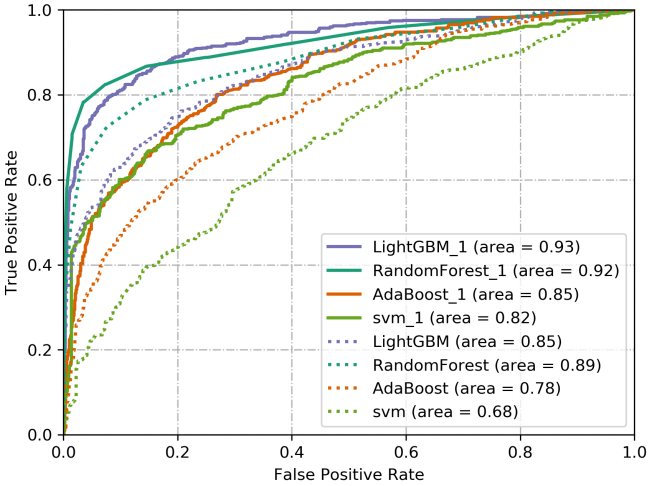

Figure 3: The ROC curves of the four algorithms.

and sells goods and the corporation's business scope do not match, we assume the transaction is fraudulent.

Rule 3: IF there exist suspicious upstream and suspicious downstream corporations, THEN the corporation is suspicious. Due to the special characteristics of the value-added tax (VAT): the amount of tax paid should be the output minus the input. In order to reduce taxes, some corporations may increase the cost of purchasing goods or reduce sales revenue. Therefore, all corporations in the VAT trading chain restrict each other and many tax evasion taxes are carried out in groups [7]. So if there exists suspicious upstream and suspicious downstream corporations, we assume that the corporation is suspicious though its suspicious score is low.

In order to analyze these anomalies, we design a transaction flow chart. As shown in Figure 4(b), the corporations are represented as rectangle blocks, and the transaction edge between corporation nodes are represented as curves. Each rectangle block consists of two components, the border and the internal influence area. The color of the border corresponds to the suspicious score (the darker red indicates higher suspicious score). The internal influence area is divided into two parts. The left area represents the influence of the upstream corporations and the right area represents the influence of the downstream corporations. Each area is divided into several areas according to the relative volume of each corporation. The color of each area represents the suspicious score of the corresponding corporation.

In the transaction flow chart, the gray edge indicates normal transaction, while the red indicates abnormal transaction (Figure 4(b-2)), which represents those transactions match the rule 1 and rule 2 aforementioned. When mouse is over the red edge, users can see the specific abnormal information. In addition, we use a histogram to show the monthly volume of transactions (Figure 4(c)). In addition to the transaction flow chart, this system also designs a parallel coordinate showing the features of all corporations, a classification result chart showing the performance of the detection algorithm and a map chart showing the geographic location of the selected corporation.

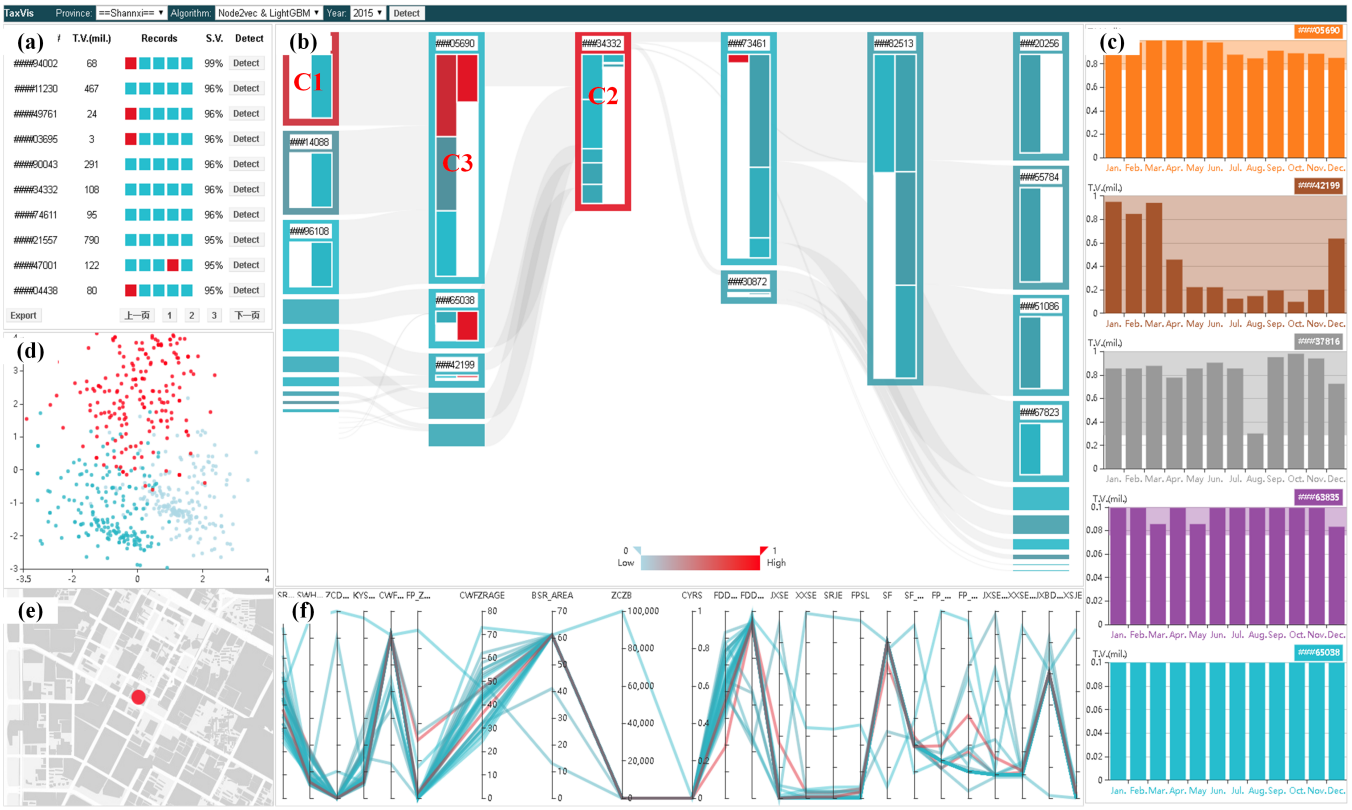

Figure 4: Screenshot of the TaxVis platform, including (a) the suspicious score table showing the suspicious individual detection result in descending order of the suspicious score; (b) the transaction flow chart depicting the trading corporations of the selected corporation and the transaction details between them; (c) the trading volume chart showing the trading volume between the corporations pre month; (d) the classification result chart showing the performance of the detection algorithm; (e) the map chart showing the geographic location of the selected corporation; (f) the parallel coordinate showing the features of all corporations.

3 DEMONSTRATION OUTLINE

We will demonstrate our system based on real tax dataset from Shaanxi province tax data in 2014-2015. There are 10,120 corporations in the dataset.

Detection of tax evasion individual. After selecting the year, our system will automatically show the detection results. The results are listed in descending order of suspicious score. In the result table (Figure 4(a)) each row corresponds to a corporation. The detail consists of the corporation id, the turnover of the past year, the tax evasion record of the past 5 years, and the suspicious score. Selecting one corporation, clicking the detect button and the rest of the page will display the corresponding information for the selected corporation.

Analysis of the tax evasion tax group. According to the previous three rules, we can check the abnormal transaction edges around the suspicious corporation to analyze the suspicious group. For example, as shown in Figure 4(b), the suspicious score of corporation C3 is very low, the suspicious scores of its upstream corporation C1 and downstream corporation C2 are very high. As a result, we can infer that the corporations C1, C2 and C3 may belong to a suspicious tax evasion group.

4 CONCLUSION AND FUTURE WORK

In this demo, we propose a two-phase approach to calculate the suspicious score based on the representations of corporations in CANet and detects suspicious tax evasion groups automatically. In future work, on the one hand, we will use some techniques to discover anomalous transactions in the second phase. On the other hand, we will apply some methods to automatically extract the tax evasion patterns to enhance the pattern matching based methods.

ACKNOWLEDGMENTS

This research was partially supported by "The Fundamental Theory and Applications of Big Data with Knowledge Engineering" under the National Key Research and Development Program of China with Grant No. 2018YFB1004500, the MOE Innovation Research Team No. IRT17R86, the National Science Foundation of China under Grant Nos. 61721002, 61532015, and Project of China Knowledge Centre for Engineering Science and Technology.

REFERENCES

[1] Bart Baesens, VÃĺronique Van Vlasselaer, and Wouter Verbeke. [n. d.]. *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to*

Data Science for Fraud Detection. John Wiley & Sons, Inc. https://doi.org/10.1002/9781119146841

[2] Walter Didimo, Luca Giamminonni, Giuseppe Liotta, Fabrizio Montecchiani, and Daniele Pagliuca. [n. d.]. A visual analytics system to support tax evasion discovery. 110 ([n. d.]), 71–83. https://doi.org/10.1016/j.dss.2018.03.008

[3] Palash Goyal and Emilio Ferrara. [n. d.]. Graph embedding techniques, applications, and performance: A survey. 151 ([n. d.]), 78–94. https://doi.org/10.1016/j.knosys.2018.03.022

[4] Aditya Grover and Jure Leskovec. [n. d.]. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (2016). ACM Press, 855–864. https://doi.org/10.1145/2939672.2939754

[5] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. [n. d.]. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3146–3154. http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf

[6] Jianfei Ruan, Zheng Yan, Bo Dong, Qinghua Zheng, and Buyue Qian. [n. d.]. Identifying suspicious groups of affiliated-transaction-based tax evasion in big data. 477 ([n. d.]), 508–532. https://doi.org/10.1016/j.ins.2018.11.008

[7] Feng Tian, Tian Lan, Kuo-Ming Chao, Nick Godwin, Qinghua Zheng, Nazaraf Shah, and Fan Zhang. [n. d.]. Mining Suspicious Tax Evasion Groups in Big Data. 28, 10 ([n. d.]), 2651–2664. https://doi.org/10.1109/TKDE.2016.2571686