This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2571686, IEEE Transactions on Knowledge and Data Engineering

Mining Suspicious Tax Evasion Groups in Big Data

Feng Tian, Tian Lan, Kuo-Ming Chao, Nick Godwin,

Qinghua Zheng, Nazaraf Shah, Fan Zhang

Abstract—There is evidence that an increasing number of enterprises plot together to evade tax in an unperceived way. At the same time, the taxation information related data is a classic kind of big data. The issues challenge the effectiveness of traditional data mining-based tax evasion detection methods. To address this problem, we first investigate the classic tax evasion cases, and employ a graph-based method to characterize their property that describes two suspicious relationship trails with a same antecedent node behind an Interest Affiliated Transaction (IAT). Next, we propose a colored network-based model (CNBM) for characterizing economic behaviors, social relationships and the IATs between taxpayers, and generating a Taxpayer Interest Interacted Network (TPIIN). To accomplish the tax evasion detection task by discovering suspicious groups in a TPIIN, methods for building a patterns tree and matching component patterns are introduced and the completeness of the methods based on graph theory is presented. Then, we describe an experiment based on real data and a simulated network. The experimental results show that our proposed method greatly improves the efficiency of tax evasion detection, as well as provides a clear explanation of the tax evasion behaviors of taxpayer groups.

Index Terms—Graph mining, tax evasion, interest affiliated transaction, heterogeneous information network, big data

1 Introduction

Tax revenue collection is considered a top priority in every national and regional jurisdiction [4], [10], [16], [17], [18], [19], and China is no different. It was reported by the Chinese government that the rate of loss of tax revenue in China is above 22%. Excluding any system defects in the mechanisms of Chinese taxation collection and administration, over 12% of tax revenue in China is lost due to technical issues.

China Tax Administration Information System (CTAIS) was developed in 1996 and since 2000 it is in operation nationally boosting a revolution in the informatization supported tax administration and the data sharing between different provinces. The sharing of data has prepared the ground for deep mining and analysis of tax data. At the same time, three ways of tax inspection, manual case selection [23], computer-based case selection (data-mining-based methods [24], [7]), and whistle-blowing-based selection were adopted by China Taxation Administration in their daily operation of tax inspection. As a result of using these methods, the traditional tax evasion behaviors, such as writing false value added tax (VAT) invoices, fake invoices and the manipulation of accounts were restrained more than ever, and the number of tax evasion cases has been decreasing dramatically.

Through many years of data tracking and analysis of both domestic transactions inside China and its cross-border transactions, it has been shown that there is a new tendency for enterprises to plot together to evade tax in undetectable ways [18], [16], [4], [6], especially through legal-like-transactions. We call these kinds of transactions, Interest-affiliated transactions (IATs). In the field of accounting and management, they are called “controlled transactions”. Between the transaction parties the most important thing is that there exists a complex and covert interactive relationship. For example, if there exists: (A) a kinship between the companies’ executives or managers or between legal persons or (B) a share interlocking relationship between the shareholders. These relationships are not only heterogeneous, but also diversified and used to accomplish interest transfer between companies to evade legitimate tax.

The national tax information collection system (NTICS) deployed in China deals with a high volume of transactions and the related data involved. For example, there are more than 31,910,000 taxpayers and 48,000 taxation administration offices all over the country. The number of annual tax-related business records is up to 1 billion, the daily peak of these records is up to ten million, and the volume of annual data aggregated is 12 TB, which is self-confirmed as big data. This volume of data challenges traditional data mining-based methods of detecting tax evasion. The reasons for this are that: (A) the training data needs to be manually labeled, (B) the trained models usually are sensitive to the training data, (C) the

- *F. Tian, T. Lan and Q. Zheng are with the MoE Key Lab for INNS, Xi’an Jiaotong University, Xi’an 710049, P.R. China.*
  E-mail: fengtian@mail.xjtu.edu.cn, 1366829195@qq.com, qhzheng@mail.xjtu.edu.cn.
- *K.M. Chao, N. Godwin and N. Shah are with the Faculty of Engineering and Computing, Coventry University, Priory Street, CV1 5FB Coventry, United Kingdom.*
  E-mail: {csx240, csx014, aa0699}@coventry.ac.uk.
- *F. Zhang is with the Serveyou Group, 3738 Nan Huan Road, Bingjiang District Hangzhou 310053, P.R. China.*
  E-mail: zhf@servyou.com.cn.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2571686, IEEE Transactions on Knowledge and Data Engineering

results of the clustering-based and neural network-based methods are not explainable and they are not intuitive, (D) their efficiency is low as they need to identify the transactions (including their detail information) one by one), (E) the most important issue is that some of the covert relationships are not recorded in the NTICS, such as relationships among directors, managers and legal persons. At the same time, some companies conceal and fail to report these relationships or delay revealing changes in details of such relationships.

In agreement with the ideas presented by Wu et al. [23], we believe that the tax authorities are equipped with limited resources, and traditional tax auditing methods/strategies are time-consuming and tedious. Consequently, there is a pressing need to have further inputs to a tax avoidance database and additional information resources from big data (e.g. company director relationships, and kinships among legal persons and/or directors etc.). This will provide a more reasonable technique-enhanced taxation analysis foundation and enables the development of new methods for dealing with any new features of IATs, i. e. acquiring more covert relationships behind each IAT, improving the identification efficiency, and supporting an explainable and intuitive representation of the results mined.

After investigating classic tax evasion cases and employing the heterogeneous information network [25], [27], [28] to analyze their properties, we propose a colored network-based model (CNBM) for characterizing economic behavior, social relationship and the IATs between taxpayers. Then, we treat tax evasion detection as a two-phase process. The goal of the first phase is to discover the suspicious groups from the heterogeneous information network built based on the CNBM, in order to identify the suspicious trading relationships. We call the first phase as mining suspicious groups, *MSG*-phase for short. In the second phase, traditional methods can be used on all transactions related to the suspicious trading relationships to detect tax evasion within the set of suspicious groups. We call the second phase as identifying tax evasion, *ITE*-phase for short. The challenges in the *MSG*-phase of the proposed method are: (1) how to model a heterogeneous network that embodies all covert relationships as well as keep it simple enough to be understandable, (2) it is obvious that the diversity of types of these covert relationships not only results in the complexity of modeling but also brings challenges in detecting the suspicious groups. The direct way of representing the diverse linkages between the participants in financial dealings is to use separate colors for each type of participants and separate colors for the distinct types of linkages. In this heterogeneous network, the suspicious tax evasion groups will appear as a variety of subgraph patterns and to detect these subgraph patterns is a kind of tasks of subgraph listing. However, different forms of the suspicious tax evasion groups result in different subgraph patterns, such as triangle, quadrilateral, pentagon and hexagon, and color difference of edges in a specific subgraph pattern. So, detecting these subgraph patterns leads to a problem of combinatorial explosion and increases the computation expense. This paper attempts to address the above mentioned difficulties in the *MSG*-phase of the proposed method. The *MSG*-phase builds the CNBM, called Taxpayer Interest Interacted Network (TPIIN), based on some simplification of relationships and contraction operations on specific types of edges from data sources. After a TPIIN is built, the algorithms for constructing a pattern tree from the TPIIN database, generating a component patterns base and detecting the suspicious tax evasion groups based on a rule for finding two transaction-linked nodes with the same antecedent are proposed to overcome the problem of searching a variety of subgraph patterns (the problem can result into a combinatorial explosion). To evaluate the effectiveness of the proposed method, experiments based on real data with additional trading relationship represented by a simulated network are carried out. The results show that our proposed method can greatly improve the efficiency of detecting potential tax evasion, as well as providing clear explanations of possible tax evasion behaviors.

2 RELATED WORKS

2.1 The Concept of Business Tax Fraud

Tax evasion is illegal evasion of taxes by individuals, corporations and trusts. Generally, business tax frauds include: VAT irregularities, transfer pricing and cross-border structuring frauds; council tax exemptions and discount frauds; consumption tax fraud; sales tax and payroll tax frauds; underreporting of property rental income; intra-group transactions, interest deductions, and tax arbitrage; and the "double Irish" tax structure used by large multinational corporations to lower corporate taxes [16]. In this paper, we focus on transfer pricing [6], [20] and cross-border structuring frauds, and tax frauds in intra-group transactions or transactions between interest-affiliated entities.

2.2 The State-of-art Tax Inspection Methods Used in Many Countries and Areas

Generally, manual case selection [23], computer-based case selection (data-mining-based methods [24], [7]), and whistle-blowing-based selection are three frequently used ways of tax inspection. However, many researchers believed that manual case selection and whistle-blowing-based selection are time-consuming and tedious, while data mining techniques used by tax administrations to detect tax fraud are considered to be the most promising approaches [7]. Mechanisms, such as neural networks, decision trees [8], logistic regression, SOM (Self-organizing map), K-means, support vector machines, visualization techniques, Bayesian networks, rough set [3], K-nearest neighbor, association rules [24], fuzzy rules, Markov chains, time series, regression and simulations [2], have been used to check tax evasion [23], [12]. For example, Wu et al. [23] used a data mining technique and developed a screening framework to filter possible non-compliant value-added tax (VAT) reports that may be subject to further auditing. Chen and Cheng [3] proposed a hybrid model, which combines the Delphi method with rough sets classifier approaches, for intelligently classifying the vehicle license tax payment (called VLTP) to solve real-world problems faced by taxation agencies. Antunes et al. [2] claimed that the method of simulation with multiple agents provides a strong methodological tool to support the design of public policies. To address the tax compliance problem [11], they adopted a mean of exploring the link between micro-level motivations leading to and being influenced by macro-level outcomes, to study the complex issue of tax evasion. They believed that some relatively simple social mechanisms can explain the compliance numbers observed in real economies. Nascimento et al. [15] described a system for tax management and fiscal intelligence, called GIF, and developed a module, named the ATRe (e-TA: electronic Tax Analysis), which eases the work of the tax authorities in identifying possible tax evasions by means of data mining techniques applied on information submitted by the contractors. Fox et al. [5] developed a new way to examine tax evasion that focuses on corporate transactions, rather than corporate profits, and examined how commodity flows respond to destination sales taxes. Torrini [22] proposed a model that predicted tax evasion opportunities for self-employment in association with the law enforcement of the local authority. The number of self-employment tax evasion cases increases, if tax inspection is not enforced. Whereas when the tax inspection is carried out by the authority strictly, the number of tax evasion cases reduces. Gonz´aleza and Vel´asquez [7] adopted clustering algorithms like SOM and neural gas to identify groups of similar behaviors in the universe of taxpayers. They apply decision trees, neural networks and Bayesian networks to identify the contributory variables in order to detect behavior patterns of frauds in the groups.

In China, data mining based methods of inspecting tax violation have already been adopted. For example, Xu [24] proposed a method of mining association rules on tax inspection/audit data to find the hidden potential associated characteristics in the tax-related violation cases. Business tax frauds can occur by falsely reporting to the authorities the values lower than actual ones for their transactions in order to avoid VAT. Michael et al.'s proposed method [4] found strong statistical evidence of under-reporting exports at the Chinese border to avoid paying VAT and evidence of tariff evasion at the U.S. border, in particular concerning related-party transactions. They also found indirect evidence of transfer pricing and evasion of Chinese capital controls. Liu et al. [13] used the hierarchical clustering in the tax inspection case selection based on certain parameters. These parameters included: tax burden rate; actual tax rate; stock rate; quick ratio; asset net profit margin; cost of sales ratio; sales finance charge rate, and tax situation. The taxpayer characteristics-trigger transfer pricing audit rules are adopted in China [16] to detect possible tax frauds, with the focus on transactions above certain amounts. This method is relatively simple but may be not be effective in all cases.

Faced with the properties of big data, traditional data mining-based methods have their pros and cons. The classification-based methods need a set of sample data for training, which means the data need to be manually labeled before training takes place. Moreover, the trained model is sensitive to the sample data and will be out-of-date if behaviors in tax evasion change. In addition, the results derived from clustering-based methods and neural network-based methods are difficult to explain and trace. The worse thing is that the above mentioned data-mining-based methods need to search and evaluate each transaction in the tax-oriented big data before reliable outcomes can be derived.

In our research, the proposed method is more effective and efficient than the existing approaches, as it aims to select the suspicious relations first via other related data sources and then identify those suspicious transactions. This can overcome the above difficulty in analyzing large and dynamic data.

3 Case Study on IATs-Based Tax Evasion and Motivation

3.1 IATs-Based Tax Evasion Case Study

In this section we introduce our motivation, after giving three case studies on IATs-based tax evasion.

[Case 1] A chemistry company *C3* in Zhejiang Province mainly produced biochemical drugs and its annual net profit was negative since its establishment in 2005. All the shares of *C3* were held by a company *C1* in Shanghai City, which was an outsourcing enterprise (provides the main raw materials to *C3*). All the products produced by *C3* were sold to *C2*. The legal person, *L1*, controlling *C1* and the legal person, *L2*, controlling *C2* are brothers. The tax administration office (TAO) verified that *C3*, whose role was relatively simple, served as a producer. *C1* was responsible for investment management and supplying main raw materials to *C3*. The company *C2* was responsible for sale, delivery, marketing analysis and guidance, research and development of each product produced by *C3*. Therefore, the TAO considered that the annual net loss of *C3* was a violation of the arm's length principle (ALP) [18], [16], and applied the transaction net margin method [18], [16] to make a tax adjustment of 25.52 million RMB, to the taxable income of *C3*, by reference to the average net profit of the same products produced by the similar scale enterprises in the same industry (see Fig. 1(a)).

[Case 2] A company, *C5*, in China, sold 5000 smart meters at $20 each to *C6*, a company in Hong Kong in August 2009. The price that *C5* offered in this transaction was much cheaper than the roughly $30 that they offered to other domestic companies. After the Tax Bureau verification, it was found that *C5* and *C6* were partially owned by the same company, *C4*, which meant that *C4* held the partial share of *C5* and *C6*. Therefore, the TAO believed that this transaction between *C5* and *C6* deviated from ALP, and made a tax adjustment of $5000 to this transaction (see Fig. 2(a)).

[Case 3] A company, *C7*, in China, sold a number of

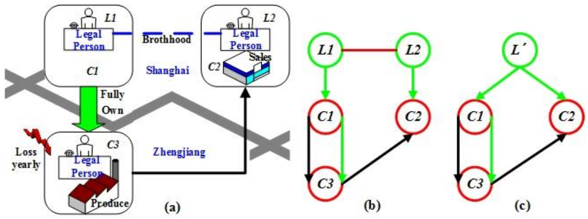


Fig. 1. Case 1.

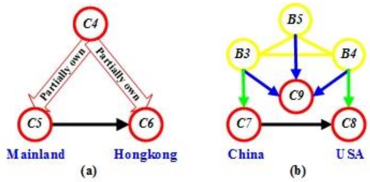


Fig. 2. Case 2 and Case 3.

BMX to *C8*, an American company, and the total volume was 90 million RMB. After TAO verified, the cost of the products produced by *C7* was 80 million RMB and the expense of selling these products was 20 million RMB in general. At the same time for *C7*, its profit rate of this kind of products was usually 9%. Meanwhile, the TAO found that both *C7* and *C8* had *B3* and *B4* (hold over 51% shares) as controlling investors, respectively; *B3*, *B4*, and *B5* together invested in a company, *C9*, and they made an agreement to act together to control this company. This kind agreement brings *B3*, *B4*, and *B5* into a situation called director interlocking. Therefore, the TAO thought that this transaction between *C7* and *C8* did not comply with ALP. After applying the cost plus method, a tax adjustment took place adding a taxable 19.89 million RMB to *C7*'s profit (see Fig. 2(b)).

After carefully observing these cases, we found that the companies and investors that plotted together for IATs-based tax evasion can be described as different kinds of graphs with colored nodes and edges. These graphs can be considered as behavior patterns between these interest-affiliated parties. For example, the Fig. 1(a) can be abstracted as Fig. 1(b). Inspired by the graph theory, three nodes refer to three companies, *C1*, *C2*, and *C3*; that *C1* has full control over *C3* is represented as a green arc, (*C1*, *C3*). The black arc, (*C3*, *C2*) denotes trading relationship between *C3* and *C2*. The similar meaning is represented by (*C1*, *C3*). *L1* and *L2* are referred to the legal persons of *C1* and *C2*, respectively, which are presented as two green arcs, (*L1*, *C1*) and (*L2*, *C2*), respectively. A brown unidirectional edge, (*L1*, *L2*), means that there exists a kinship between *L1* and *L2*. Then, this case indicates that the kinship between two legal persons of two different companies is a hint to suspicious relationships behind an IAT. This kind of hint is mapped into a graph-like structure that is consist of two directed trails, *L1* → *C1* → *C3*) + (*L1* → *L2* → *C2*) or (*L2* → *C2*) + (*L2* → *L1* → *C1* → *C3*) by the identified IAT (*C3* → *C2*). Furthermore, this kind of structure can be simplified by merging nodes *L1* and *L2* together, as shown in Fig. 1(c). So, the graph-like structure in Fig. 1(c) consists of two directed trails, (*L′* → *C1* → *C3*) + (*L′* → *C2*). This pair of trails represents the suspicious relationship of tax evasion. Trail, walk and path are defined in Appendix A.

Similarly, Fig. 2(a) and Fig. 2(b) can be abstracted as Fig. 3(a) and Fig. 3(b), respectively. In Fig. 3(a), three nodes denote three companies, *C4*, *C5* and *C6*. *C4* partially invests in *C5* and *C6*, which are represented as two red arcs, (*C4*, *C5*) and (*C4*, *C6*). Then, this case indicates that the same investor (*C4*) of two different companies, *C5* and *C6*, are a hint to suspicious relationships behind an IAT and this kind of hint is mapped into a graph-like structure (that consists of two directed trails, (*C4* → *C5*) + (*C4* → *C6*) by the identified IAT (*C5* → *C6*).

In Fig. 3(b), there exist three companies, *C7*, *C8*, and *C9*. Three director nodes, *B3*, *B4*, and *B5* are merged into a node, *B*, because they made an agreement of acting together. These result in both Fig. 3(a) and Fig. 3(b) have a similar triangle. Within node *B*, yellow unidirectional edges denote the interlocking relationship between three directors. Then, this case indicates that the directors and interlocking structure of *B3*, *B4*, and *B5* of two different companies is a hint to suspicious relationships behind an IAT. After merging the interlocking structure, this kind of hint is mapped into a graph-like structure that consists of two directed trails, (*B* → *C7*) + (*B* → *C8*) by the identified IAT (*C7* → *C8*).

Inspired by the observation of above cases, we intend to focus on better discovery of the covert (interest) relationships behind the controlled transactions, using a graph-based method that characterizes the property of these relationships as two suspicious relationship trails with a same antecedent node behind an interest affiliated transaction. These trails form a proof chain indicating the potential tax evasion between these group members. Thus, representing and discovering these covert relationships is a priority task for identifying the tax evasion behind a large number of transactions. Another benefit of doing this is to scale down the range of the suspicious groups to be searched and provide potential candidates of

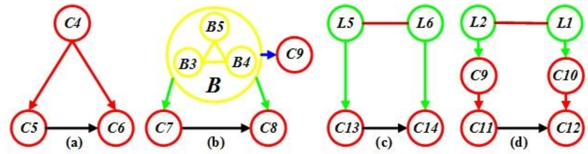


Fig. 3. Graph-based patterns.

suspicious relationship-based proof chains.

3.2 Motivation

From the case studies presented in the above section, we can summarize them as follows:

1) There exists a complex, covert interactive relationship between the transaction parties. For example, there exists: (A) a kinship between the companies' executives or managers or between legal persons; (B) a share interlocking relationship between the shareholders, or, (C) even the same controller of both parties involved in a transaction. These relationships are not only heterogeneous, but also diversified, while the transaction arcs indicate the direction of trading relationship

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2571686, IEEE Transactions on Knowledge and Data Engineering

between companies. At the same time, apart from trading relationship, five other relationships (represented by different colors: yellow, brown, blue, green and red, in figures of Section 3.1) are classified as a kind of interest affiliated relationship (or relationship of suspicious tax evasion.), which indicates the potential interaction between parties behind IATs.

2) Graph-based patterns, including triangle, quadrilateral, pentagon and hexagon (see Fig. 3) within which there only exists one trading relationship, can characterize the various behavior patterns among the parties involved with the IATs. It is obvious that Fig. 3(c) and Fig. 3(d) are the variants of Fig. 3(a) and Fig. 3(b). Note that there are many combinations of colored edges and nodes for each kind of graph-based patterns, and we do not list all of them in Fig. 3. This means that there exists a problem of combinatorial explosion of different patterns. However, a most interesting phenomenon is that there exist at least two suspicious relationship trails with a same antecedent node behind an IAT.

These observations inspire us to build a network from related information sources, and then detect tax evasion by applying a two-phase process. The goal of the *MSG*-phase is to build a well-formed heterogeneous information network and discover graph-like suspicious groups from this network. That is, if we list all potential directed trails, we can find potential suspicious groups by identifying two trails with the same antecedent node. In the *ITE*-phase, traditional tax evasion identification methods can be used to detect IATs-based tax evasion from a set of transactions in these suspicious groups. The idea is illustrated in Fig. 4.

Fig. 4 depicts not only the taxation related information sources, such as financial reports of each taxpayer and electronic receipt database (i.e. transaction database) managed by each provincial tax administration office (PTAO), but also the additional external information sources, such as the household registration database from the household registration department of public security in China (HRDPSC). Shareholding structure of each oversea company and domestic company, and recent reports from China Securities Regulatory Commission (CSRC), are mined for building various homogeneous networks. For example, an information network based on relationships between legal persons and companies/taxpayers is a homogeneous network. These mined homogeneous networks are merged into a heterogeneous information network that is modeled according to CNBM. Then, a suspicious group detection method is applied to discover all relevant patterns in the heterogeneous information network. Finally, the tax evader is identified after tax evasion judgment methods [18], [16] are applied to the transactions within these groups.

To implement this concept, the most important step is to build a heterogeneous network, and identify these suspicious groups via mining. In the reminder of this

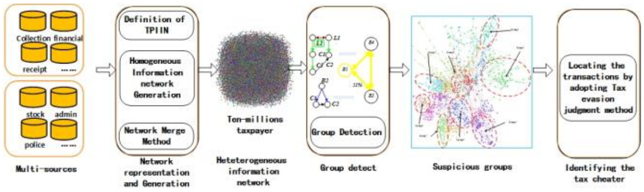


Fig. 4. Flow of tax evasion identification.

paper, we first propose a colored network-based model (CNBM) for characterizing economic behavior [21], social relationships and IATs between taxpayers. Next, we show how this type of network is generated. To evaluate the effectiveness of the proposed method for the *MSG*-phase, a proof is given based on graph theory and experiments based on real data for all the nodes, most of the edges and a trading relationship simulated network are carried out. The experimental results show that our proposed method has ability to greatly improve the efficiency of detecting possible tax evasions in the *MSG*-phase, as well as provide a clear explanation of the tax evasion behaviors of taxpayer groups.

4 Definition and Generation of Taxpayer Interest Interacted Network

4.1 Analysis of How to Model the Proposed Network

Generally speaking, there are two kinds of elements: nodes and edges in an original and un-contracted taxpayer interest interacted network. Nodes can be divided into two types, representing *Person* and *Company* respectively, while edges can be unidirectional or directed. So, the network is denoted as $Net = \{P^0 \cup C^0, E^0 \cup A^0, VC^0, AC^0\}$, where P^0 is a set of nodes representing persons, C^0 is a set of nodes representing companies, E^0 is a set of unidirectional edges, A^0 is a set of arcs (directed edges), VC^0 and AC^0 are the set of colors attached to vertices and edges respectively. To simplify the original network, we describe how to build a TPIIN as well as nodes, edges, their colors and the contraction operations on some of them as follows.

Since a person $p \in P^0$ can have a number of positions such as Chairman of the board(*CB*), Chief Executive Officer(*CEO*), Shareholder(*S*) and Director(*D*), the basic colors of *Person* nodes can be divided into four subtypes: *CB*, *CEO*, *S* and *D*. These subclasses are not mutually exclusive. In accordance with the various possible combinations of positions, there are fifteen possible disjoint subclasses of colors for *Person* nodes, defined by *CEO* and *D* and *S* and *CB*, *CEO* and *D* and *S*, *CEO* and *D* and *CB*, *CEO* and *S* and *CB*, *D* and *S* and *CB*, *CEO* and *D*, *CEO* and *CB*, *CEO* and *S*, *D* and *S*, *D* and *CB*, *S* and *CB*, *CB*, *S*, *D*, *CEO*. Considering realistic scenarios, ① in a small-scale company, there are a few investors and all of them are shareholders. A shareholder of such a company is himself a director; ② in a large-scale company, shareholders select some of them to be directors or a shareholder can be himself a director if he holds a high enough percentage of the shares. If a shareholder is at least a director, then he can be involved in the process of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2571686, IEEE Transactions on Knowledge and Data Engineering

monitoring and decision-making of a company, otherwise, he cannot. Based on this, the four subclasses (*CB*, *CEO*, *S* and *D*) of colors for *Person* nodes can be replaced by the three subclasses: *CB*, *CEO* and *D*. Therefore, fifteen possible subclasses of colors for *Person* nodes are reduced to seven possible subsets (*CEO* and *D* and *CB*, *CEO* and *D*, *CEO* and *CB*, *D* and *CB*, *CB*, *D*, *CEO*). According to Company Act of China, "a legal person (*LP*) is a unique representative of a legally and separately registered company/corporate/trust/institution". "The role of a *LP* should be assigned to a *CB* or an executive/managing Director (this is a *CEO* and *D*) or a *CEO*". Usually, a role of a *LP* in a large-scale company is assigned to a *CB* or *CEO*, while the role in a small scale company is assigned to a general manager (equals to *CEO*) or an Executive Director or a *CEO*. So, a *LP* can belong to one of these subclasses (*CEO* and *D* and *CB*, *CEO* and *D*, *CEO* and *CB*, *D* and *CB*, *CB*, *CEO*). In a well-defined network it is not necessary to have different subclasses (colors) of nodes but, when gathering persons' roles from different data sources, these subclasses (colors) will be relevant to nodes in order to distinguish them.

A *Company* node, $c \in C^0$, represents a legally and separately registered company/corporate/trust/ institution and has a unique link with a *LP* and may link with *Person* nodes with other subclasses of colors, such as a *D*.

The color for a unidirectional edge, $ue \in E^0$, is *Interdependence* that represents two kinds of relationships, kinship and interlocking, while, for arcs (directed edges), they have different colors, *Influence*, *Trading*, and *Investment,* for different relationships, influence, trading, and investment relationship, respectively. After gathering corresponding data from various information sources, different homogeneous relationship graphs are formed according to different relationships. Then, after carrying out a procedure of multi-network fusion (shown in Fig. 5) on these homogeneous relationship graphs, a taxpayer interest interacted network (TPIIN), $TPIIN = \{V, A, VC, AC\}$ is created, where V is a set of nodes, A is a set of arcs, VC is a set of colors for nodes and has two elements, *Person* and *Company*, and AC is a set of colors for nodes and has two elements, *Influence* and *Trading*. This procedure of multi-network fusion is discussed step by step as follows.

The kinship and interlocking relationships between

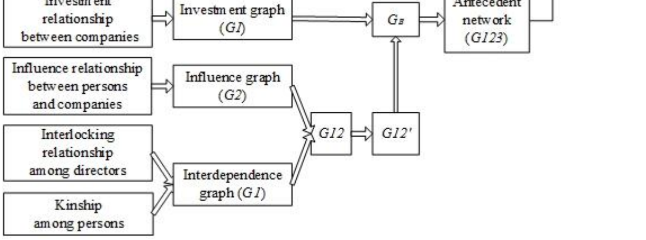


Fig. 5. Schematic procedure of multi-network fusion.

people are abstracted respectively and are represented by a single type of unidirectional edges (Note that, if there exist both a kinship and an interlocking relationship between a pair of persons, we only keep one). We call this a **homogeneous graph** *G1* **an interdependence graph**. $G1 = (V_1, E_1) = (P, E_1)$, where P contains all the persons (their roles are analyzed as above) involved in the operations of these companies' decision-making. An edge, $e \in E_1$, will be called an interdependence link (representing either kinship or interlocking relationship), and the color for all edges in E_1 is the same. The properties of *G1* are described in Appendix A.

Between a *Person* node and a *Company* node there can be a directed link which shows that the *Person* has an influence on the operations of the *Company*. There are several subclasses of this influence: (i) is-an-CEO-and-D-of; (ii) is-CEO-of, (iii) is-CB-of; (iv) is-a-D-of. After abstracting the above four subclasses individually from the relevant resources, these subclasses are represented by a single type of directed edges. We call this homogeneous graph $G2 = (V_2, A_2)$, where $V_2 = P \cup C$ and the color of each arc in A_2 is *influence*. Assume that C (each of which represents a company registered in tax offices) contains all involved companies. According to the properties of *G2* (shown in Appendix A), *G2* is thus a bipartite graph, each *Person* node, $p \in P$, must have indegree of zero and each *Company* node, $c \in C$, must have outdegree of zero. On the other hand, the outdegree of a *Person* node and the indegree of a *Company* node will be positive integers. It is possible that all *Company* nodes must have at least one *Person* node in the database connected by an influence link (so *Company* nodes must at least link with one *LP* node).

We can combine *G2* with *G1* by adding the unidirectional *Interdependence* edges between *Person* nodes, and result in a new graph, $G12 = (V_2, E_1 + A_2) = (P \cup C, E_1 + A_2)$, as $V_1 \subseteq V_2$ and E_1 *and* E_2 is disjoint ($E_1 \cap E_2 = \emptyset$). Obviously, *G12* has two kinds of edges (i.e. *Influence* and *Interdependence*). To simplify *G12* into a graph with a single kind edge, a process of edge contraction operation (shown in Appendix A) is performed on *G12*, in which each edge contraction operation will contribute to creating a syndicate, deleting an *Interdependence* edge and two nodes connected by the edge, and reattaching the corresponding arcs to the syndicate. Note that the process will work for a pair of *Person* nodes or a pair of a syndicate and a *Person* node or a pair of syndicates as these pairs are connected by the *Interdependence* edges. Repeat the above process till all *Interdependence* edges are removed and then obtain a new graph, $G12' = \{V'_{12}, A'_{12}\}$. Nodes in *G12′* are either *Person* nodes or *Company* nodes, arcs in *G12′* are all *Influence* arcs and the indegree of the *Person* nodes is zero and the outdegree of the *Company* nodes is zero (the properties of *G12′* are shown in Appendix A). Hence *G12′* is a bipartite graph, we can call any remaining *Person* nodes and syndicates of persons as *Person* nodes.

Between two *Company* instances there can be a relationship of investment, which exists if one *Company* node has a major shareholding in another. This will be a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2571686, IEEE Transactions on Knowledge and Data Engineering

directed link in a graph representation. So, after abstracting relationship of investment between *Company* instances, we get a graph *GI*, called as an investment graph, where $GI = \{V_{GI}, A_{GI}\} = \{C, A_{GI}\}$, A_{GI} is a set of *Investment* arcs. Moreover, we can combine *G12'* with *GI* by adding the directed edges between *Company* nodes. Let us denote this graph as G_B, $G_B = (V_B, E_B) = (V_B, A_3 + A'_{12})$ (A_3 and A'_{12} is disjoint, as mentioned in Appendix A) and G_B only has two kinds of arcs: *Influence* and *Investment*.

As mentioned in Appendix A, in terms of *Company* and *Investment* we may have a set of companies, in which all pairs have mutual investment arrangements (Seen in Fig. A-3 of Appendix A) and more complicated cases are presented in Fig. A-4 of Appendix A. To simplify G_B into a directed acyclic graph (DAG), firstly, we detect each strongly connected subgraph *SCS* in *GI* by applying Tarjan's algorithm [26] and save it; secondly we introduce a process of strongly connected subgraph contraction operation (shown in Appendix A) and carry it out on each *SCS* in G_B. This process will contribute to generating a DAG *G123*. This is proved and described in Appendix A. For simplicity, ① we call *Company* nodes and syndicates produced by merging *Company* nodes when applying strongly connected subgraph contraction operation as *Company* nodes. ② In *G123*, edges that are colored with either *Influence* or *Investment*. Let's consider the investment relationship as a kind of influence link. Therefore, in *G123*, nodes are colored with either *Person* or *Company*, edges are colored with *Influence*, and each of these edges is an arc. We call *G123* an influence graph (Antecedent network) that describes the influence not only between *Person* nodes and *Company* nodes,

For detecting suspicious groups behind transactions, we need to find the trading relationships between *Company* instances. A trading relationship can be represented as an arc (directed edge) between two *Company* nodes. After abstracting all the trading relationships between *Company* instances, we obtain a graph *G4*, called a trading graph, which contains only *Company* nodes and trading relationship arcs.

After combining *G123* and *G4* directly (their arcs are disjoint according to the description in Appendix A), we get a heterogeneous network, called *TPIIN*, in which, a suspicious situation is that two companies have a trading relationship and common *Person/Company* nodes that have influence on both companies. All edges in a TPIIN are arcs. In the example depicted in Fig. 6, there is a suspicious relationship between *C2* and *C3* since both nodes *C1* and *C3* are influenced directly by *P1* and there is a trading relationship from *C2* to *C3*.

4.2 Formal Definition

As mentioned in the Section 4.1, an institution/corporate/trust that pays taxes to the country legally and singly is a taxpayer. The basis of a TPIIN is a set of nodes of distinct types and a set of edges of distinct types. The two different colors of nodes are: *Company* and *Person*. The two relationships defined are: influence relationship between a *Person* node and a *Company* node or a trading relationship (from one *Company* node to another).

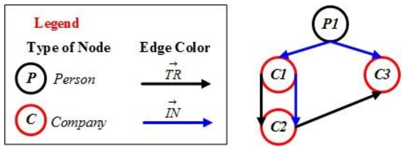


Fig. 6. Example of a TPIIN.

Definition 1. *Based on above types of nodes and arcs, a taxpayer interest interacted network is represented as a quadruple:*

$$TPIIN = \{\mathbf{V}, \mathbf{E}, \mathbf{VColor}, \mathbf{EColor}\},$$

Where

- $\mathbf{V} = \{v_p \mid p = 1, ..., N_p\}$ *denotes a set of vertices.*
- $\mathbf{E}$ *denotes a set of all existing arcs in TPIIN, and let* $E = \{e_{pq}\} = \{(v_p, v_q) \mid 0 < p, q \leqslant N_p\}$, *where* $e_{pq} = (v_p, v_q)$ *denotes that there exists an arc from the p-th vertex to the q-th vertex.*
- $\mathbf{VColor}$ = *{Person, Company}, where Company denotes the color of a vertex that represents a company or a syndicate of companies (described in section 4.1); Person denotes the color of a vertex that represents a person or a syndicate of Person nodes (such as the node B in Fig. 3(b)). According to the two colors in VColor, the vertices in a TPIIN can be represented as* $V = P \cup C$, *where* $P = \{v_l \mid l = 1, . . ., N_S, N_S < N_p\}$ *denotes all Person nodes,* $C = \{v_c \mid c = 1, . . ., N_C, N_C < N_p\}$ *denotes all Company nodes, then* $N_S + N_C = N_p$.
- $EColor = \{\vec{IN}, \vec{TR}\}$ *denotes a set of colors marked on directed arcs in a TPIIN, where* $\vec{IN}$ *denotes an influence relationship between a Person/Company node and a Company node and means that* v_p *has an influence on* v_q *directly as described in Section 4.1; and* $\vec{TR}$ *denotes a trading relationship among Company nodes and means there exists a trading relationship from* v_p *to* v_q.

From the view of influence relationship and trading relationship, there are two parts in a *TPIIN*: the antecedent network and the trading network. The antecedent network covers all relationships (investment and interdependence, etc.), which have influence on transactions between *Company* nodes, except for the trading relationship.

As known in graph theory, a directed path is represented as $\pi = \{v_1, e_{12}, v_2, ..., v_{l-1l}, e_{l-1l}, v_l\}$, $V(\pi) = \{v_1, v_2, ..., v_{l-1l}, v_l\}$ and $E(\pi) = \{e_{12}, ..., e_{l-1l}\}$. If $v_i \neq v_j$ $(i \neq j, v_i, v_j \in V(\pi))$ holds, then π is a simple directed path.

Definition 2. *Suspicious tax evasion group (Suspicious Group)*

In a TPIIN, a suspicious group consists of two simple directed trails, π_1 *and* π_2, *that have the same start and end nodes, and in* $E(\pi_1) \cup E(\pi_2)$ *there exists one and only one trading relationship incoming arc,* $e^{\vec{TR}}$ *to the end node.*

Definition 3. *Simple suspicious tax evasion group (simple suspicious group)*

In a TPIIN, a simple suspicious group is a suspicious group, whose two simple trails have no same nodes except the start and end nodes. Furthermore, in a simple suspicious group of a TPIIN, we call a simple trail from the start node to the end node as a component pattern.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2571686, IEEE Transactions on Knowledge and Data Engineering

Take Fig. 6 as an example, π_0={$P1$, $e^{\overrightarrow{IN}}_{P1C1}$, $C1$, $e^{\overrightarrow{IN}}_{C1C2}$, $C2$, $e^{\overrightarrow{TR}}_{C2C3}$, $C3$}, π_1 ={$P1$, $e^{\overrightarrow{IN}}_{P1C1}$, $C1$, $e^{\overrightarrow{TR}}_{C1C2}$, $C2$, $e^{\overrightarrow{TR}}_{C2C3}$, $C3$} and π_2 ={$P1$, $e^{\overrightarrow{IN}}_{P1C3}$, $C3$} are simple trails. π_0 and π_2 form a simple suspicious group. So, π_0 and π_2 are component patterns. π_1 and π_2 do not form a suspicious group because, in $E(\pi_1) \cup E(\pi_2)$, there exist more than one trading relationship incoming arcs, $e^{\overrightarrow{TR}}_{C1C2}$ and $e^{\overrightarrow{TR}}_{C2C3}$.

Property 1. *In any walk of an antecedent network (that is a DAG instance), there does not exist any iterated edge and any iterated node. This means that each walk in an antecedent network is a trail and as well as a path.*

Lemma 1. *If a trading relationship arc is added to a tail in an antecedent network and it forms a new walk, nw, then nw is a trail. We can call this trading-arc-added operation first-trading-arc join.*

4.3 Generation of a TPIIN

As mentioned in Section 4.1, a TPIIN is generated after a multi-network fusion method has been adopted to abstract different relationships between taxpayers from various information sources managed by CSRC, HRDPSC and PTAOS and then fuse these relationships and corresponding homogeneous networks together. Considering that the generated TPIIN is a large scale graph, our task of identifying the suspicious tax evasion groups is a three-step approach as follows:

- The first step is to segment a large scale TPIIN into small weakly connected subgraphs by applying divide and conquer strategy. This step is inspired by an intuitive idea that a trading relationship edge that connects two unconnected subgraphs (*ante*(i) and *ante*(j)) of an antecedent network is an unsuspicious trading relationship. Obviously, this means that there is definitely without one party (node) involved in two subgraphs at the same time behind the trading relationship edge. Therefore, the i-th maximal weakly connected subgraph of an antecedent network and the trading relationship links between its *Company* nodes forms the i-th weakly connected subgraph of a TPIIN, denoted as *subTPIIN*(i)
- The second step is to list all potential relationship trails in a subTPIIN (see Definition 4) in the form of *InOT-OutOSP* walk (see Definition 5) or *InOT-FTAOP* walk (see Definition 6). Inspired by the frequent pattern tree from the business transaction database [9] and considering the characteristic of a *DAG* (see Appendix A), we propose an algorithm for constructing a patterns tree and generating a potential component pattern base from each subTPIIN (a subTPIIN database is shown in Fig. 8).
- The third step executes the task of detecting the suspicious groups of potential tax evaders. The task finds any two matched component patterns both with a same antecedent element behind a trading arc in each potential component pattern base

The second and third steps are executed iteratively until every subTPIIN is processed. Based on the idea described above, a definition is introduced as follow

Definition 4. *SubTPIIN*

In a TPIIN, a subTPIIN is a graph that consists of one *maximal weakly connected subgraph (MWCS) of an antecedent network and all trading relationship arcs between the Company nodes in the MWCS.*

In this paper, a subTPIIN is in a form of edge list (a *row* * 3 array) and is a part of a TPIIN. The pseudo code of the above three-step approach is shown in Algorithm 1.

Algorithm 1. Detecting suspicious tax evasion groups

Input: Array *tpiin* (in the form of edge list: r x 3, r is a number of arcs. The *top* (m – 1) rows of a *tpiin* store all arcs in an antecedent network while other rows of the *tpiin* belong to a trading network. m indicates the index of first trading relationship arc in *tpiin*.)

Output: File *susGroup(i)*, i = 1, …, L. (a separated file, *susGroup(i)*, saves all suspicious groups that are mined from the i-th subTPIIN. L is the number of subTPIINs in the *tpiin*.)
File *susTrade(i)*, i = 1, …, L. (a separated file, *susTrade(i)*, saves all suspicious trading arcs mined from the i-th subTPIIN)

Begin

1 Abstract all *Influence* arcs from *tpiin* to form a (m - 1) * 3 matrix, *Antecendent (*an antecedent network*)*;
2 Abstract all trading arcs from *tpiin* to form a (r – m + 1) * 3 matrix, *Trade* (a trading relationship network);
3 Find each MWCS in *Antecedent* and save it into *PA_vertSet(i)* and *PA_edgeSet(i)* accordingly, where i = 1, …, L;
4 **for** i = 1, …, L **do**
5 Acquire all trading arcs between the vertices in *PA_vertSet(i)* from *Trade* and add them to *tradingEdge* (a k*3 array, k is the number of the trading arcs related to *PA_vertSet(i))*;
6 Merge *PA_edgeSet(i)* and *tradingEdge* to generate a subTPIIN, *subTPIIN(i)*, and empty *tradingEdge;*
7 Use Algorithm 2 to create the i-th patterns tree as well as generate all potential component patterns in *subTPIIN(i)* and save them into a file, *patterns(i);*
8 Carry out the pattern matching algorithm (Appendix B) to find all suspicious groups and trading arcs in *patterns(i)*, then save them in files, *susGroup(i)* and *susTrade(i)* respectively;
9 **end for**
10 **Return** all *susGroup(i)* and *susTrade(i);*

Algorithm 1 takes a TPIIN, *tpiin* (a r x 3 matrix) as input, where the first and second column of *tpiin* represent the index of start and end node of each arc, respectively, and the third column of *tpiin* indicates the color of the corresponding arc (note that, in our codes, 0 represents black while 1 represents blue; we keep the words, black and blue as shown in Fig. 8). Firstly, TPIIN is divided into two parts: *Antecedent* (a (m - 1) * 3 matrix) and *Trade* (a (r – m + 1) * 3 matrix) (Steps 1-2). Secondly, step 3 is to find all *MWCS* in *Antecedent* by employing the function *findsubgraph()* that is an improved deep-first-search strategy and described in Appendix B. Thereafter, the vertices and arcs in the corresponding antecedent network are stored in *PA_vertSet(i)* and *PA_edgeSet(i)* (i = 1, …, L), respectively, where L is the number of *MWCS*s in *Antecedent*. Thirdly, inspired by the strategy of divide and conquer, generate each subTPIIN, *subTPIIN(i)* (i = 1, …, L), and then process individually to find suspicious tax evasion groups and suspicious trading arcs (Steps 4-10). Furthermore, Step 5 is to find all trading arcs of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2571686, IEEE Transactions on Knowledge and Data Engineering

PA_edgeSet(i) from *Trade* and then Step 6 adds them into *PA_edgeSet(i)* to generate *subTPIIN(i)*. Step 7 uses Algorithm 2 to create a patterns tree as well as generate all potential component patterns for *subTPIIN(i),* where Algorithm 2 is described in the following paragraph. Step 8 finds matched component patterns and then gets all suspicious groups and suspicious trading arcs in each subTPIIN from these matched patterns, the detailed process is described in Appendix B.

Algorithm 2. Generating a patterns tree and its base

Input: Array *subTPIIN* (a subTPIIN in the form of edge list)
Output: File *patterns* (a file that stores all potential component patterns in *subTPIIN*)

```
Begin
1   Calculate the values of indegree and outdegree of each node
    in subTPIIN and save them in an array in and an array out,
    respectively, and form a 3-column matrix, Nodes ;
2   Sort the order of the elements in Nodes according to the
    increase in indegree of each node and inverted order of
    outdegree of each node, the result is saved in listD;
3   Find all nodes of degree-zero in Node and save them in an
    array indegree0;
4   for i = 1, …, length(indegree0) do
5       Put the i-th element of indegree0 into an array, str;
6       Find all children of str(1) in subTPIIN and save them
        into an array, nextnodes;
7      If nextnodes is empty then
8          Output str into the file patterns;
9      else
10       for j = 1, …, length(nextnodes) do
11           Add nextnodes(j) to the tail of str;
12             Apply deepsearchNext(nextnodes(j), subTPIIN,
               patterns, str) (shown in Appendix B) to search
               a whole trail from nextnodes(j) and end
               searching this trail until meeting criterion
               Rule1 or Rule2, and save this trail to str;
                 Rule 1: End the search of this trail if meet a
                 node that has the zero value of their out
                 degree;
               Rule 2: End the search of this trail until
               meeting a node that is the end node of a black
               arc as well as the black arc is the first trading
               relationship in this trail;
13           Output str into the file patterns;
14       end for
15     end if
16  end for
17  Return patterns;
```

Algorithm 2 shows the pseudo code of creating the patterns tree as well as generating all potential component patterns for each subTPIIN, *subTPIIN*. In which, Step 1 calculates the values of indegree and outdegree of each node in *subTPIIN* and save them in an array *in* and an array *out*, respectively, and forms a 3-column matrix, *Nodes*. The vertices in *Nodes* are sorted and the result is saved in an array, *ListD*, according to increase in indegree of each vertex and inverted order of outdegree of each vertex (Step 2). Fig. 9(a) shows an example of this kind of sorting. Next, Step 3 obtains the nodes of indegree-zero from *ListD* and put them into an array *indegree0*. Steps 4-17 describe the procedure of searching all the trails that start from each element in *ListD* and also search its children by employing an improved deep search algorithm, *deepsearchNext()* (shown in Appendix B), and its stops criterion described in *Rule1* and *Rule2* is met. (Note that, a child of a node in the patterns tree is defined as: if an arc starts from a node *A* and ends at node *B*, then we call *B* a child of a node *A*.)

Our method starts from indegree-zero nodes to search the customized walks of a subTPIIN until criterion of *Rule1* and *Rule2* is met. Based on this, we can coin two definitions as follows:

Definition 5. *Indegree-zero-start-and-outdegree-zero-stop walk (InOT-OutOSP walk)*

An Indegree-zero-start-and-outdegree-zero-stop walk is a trail belongs to a set of trails in an antecedent network and does not contain any trading arc.

Definition 6. *Indegree-zero-start-and-first-trading-arc-stop walk (InOT-FTAOP walk)*

An Indegree-zero-start-and-first-trading-arc-stop walk is a trail that adds a trading arc to the tail of a trail belongs to a set of trails in an antecedent network.

Briefly proofs of Definitions 5 and 6 are described as follows. As mentioned above, in an antecedent network, the color of all *Influence* arcs is different from the color of trading relationship arcs. Therefore, each walk generated by our proposed method starts from an indegree-zero node and ends at a *Company* node which has no children (equals to the outdegree-zero condition, *Rule1*) or is attached to a first trading arc with (*Rule2*). Obviously, for each *InOT-FTAOP* walk, it is a trail that belongs to the set of trails in an antecedent network because, during searching phase, traversing from one indegree-zero node toward an outdegree-zero node without meeting any trading arc means that the color of each arc in this walk is *Influence* and this walk still belongs to an antecedent network. Meanwhile, each *InOT-FTAOP* walk produced by a first-trading-arc join is a trail according to Lemma 1.

An un-contracted taxpayer interest interacted network is shown in Fig. 7. After executing edge contraction operation on interdependence links (interlocking link or kinship link) of the network in Fig. 7, we obtain its corresponding TPIIN as shown in Fig. 8(a). For example, *Person* nodes *L6* and *LB* in Fig. 7 are relatives, so these two are merged into a syndicate *L1* in Fig. 8. The same method is applied to *Person* nodes *B5* and *B6* in Fig. 7 and a syndicate of *Person* nodes, *B2*, is generated.

Suppose that a TPIIN is segmented according to step 3 in Algorithm 1, then we obtain only one subTPIIN, *subTPIIN*, where *PA_edgeSet* includes *Antecedent* and *Trade* accordingly (seen in Fig. 8). The *Antecedent* stores all influence relationship instances except for trading ones, while *Trade* stores all trading relationships among *Company* nodes. Then, once *subTPIIN* is input and processed by algorithm 2, the values of indegree and outdegree of each node in subTPIIN are sorted and save in *ListD* (see Fig. 9). After applying the method *deepsearchNext()* that follows each directed arc starting from the indegree-zero node until it meets criterion *Rule1* and *Rule2* (step 7 of Algorithm 2 and Algorithm B-4 in Appendix B), a patterns tree is constructed from this subTPIIN (Fig. 9) as well as potential component patterns base is easily built (as shown in Fig. 10). At the same time,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2571686, IEEE Transactions on Knowledge and Data Engineering

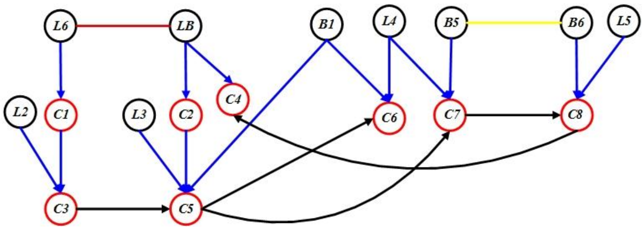
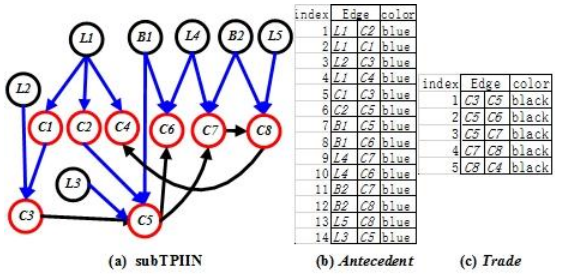


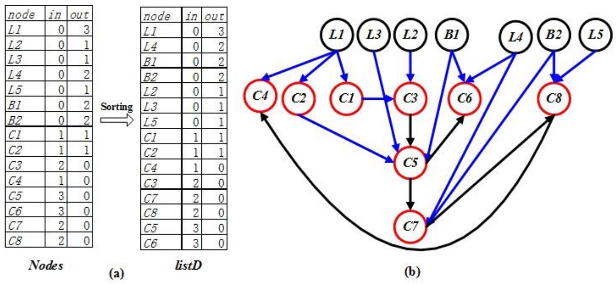
Fig. 7. Example of an un-contracted taxpayer interest interacted network.



index	Edge		color
1	L1	C2	blue
2	L1	C1	blue
3	L2	C3	blue
4	L1	C4	blue
5	C1	C3	blue
6	C2	C5	blue
7	B1	C5	blue
8	B1	C6	blue
9	L4	C7	blue
10	L4	C6	blue
11	B2	C7	blue
12	B2	C8	blue
13	L5	C8	blue
14	L3	C5	blue

index	Edge		color
1	C3	C5	black
2	C5	C6	black
3	C5	C7	black
4	C7	C8	black
5	C8	C4	black

Fig. 8. subTPIIN and its database (including *Antecedent* and *Trade*).



node	in	out
L1	0	3
L2	0	1
L3	0	1
L4	0	2
L5	0	1
B1	0	2
B2	0	2
C1	1	1
C2	1	1
C3	2	0
C4	1	0
C5	3	0
C6	3	0
C7	2	0
C8	2	0

node	in	out
L1	0	3
L4	0	2
B1	0	2
B2	0	2
L2	0	1
L3	0	1
L5	0	1
C1	1	1
C2	1	1
C4	1	0
C3	2	0
C7	2	0
C8	2	0
C5	3	0
C6	3	0

Fig. 9. Patterns tree generation.

1. *L1*, *C2*, *C5*→*C6*
2. *L1*, *C2*, *C5*→*C7*
3. *L1*, *C1*, *C3*→*C5*
4. *L1*, *C4*
5. *L3*, *C5*→ *C7*
6. *L3*, *C5*→*C6*
7. *L2*, *C3*→*C5*
8. *B1*, *C5*→*C6*
9. *B1*, *C5*→*C7*
10. *B1*, *C6*
11. *L4*, *C6*
12. *L4*, *C7*→*C8*
13. *B2*, *C7*→*C8*
14. *B2*, *C8*→*C4*
15. *L5*, *C8*→*C4*

Fig. 10. Potential component patterns base.

we call an instance in our potential component patterns base a suspicious relationship trail.

A suspicious relationship trail can be represented in two kinds of formats: for case (a) {*A1*, *A2*, ..., *Am*} (an *InOT-OutOSP* walk) and, for case (b) {*A1*, *A2*, ..., *Am*, → *Cj*} (an *InOT-FTAOP* walk), where *Ai*, $i = 1, ..., I$, denotes *i*-th node in a subTPIIN, and *Cj*, $j = 1, ..., J$, denotes *j*-th red node (company), where *j* is the number of nodes in the subTPIIN and $J < I$. An instance of the first kind of pattern is shown in the fourth line of Fig. 10, and an instance of the second kind of pattern is shown in the first line of Fig. 10.

After acquiring the potential component patterns base (as shown in Fig. 10), the task of detecting the suspicious groups of potential tax evaders is to find two matched component patterns, both with the same antecedent node, *A1*, and where one pattern is of type (b) ending in *Cj* and the other is of type (a) or (b) with one of the elements *Ai* ($1 \leq i \leq m$) ≡ *Cj*. Usually, the two matched component patterns exist in two potential suspicious relationship trails in a potential component pattern base. However, there is a special case that, if there exists a circle within an *InOT-FTAOP* walk, then this circle is a simple suspicious group and has two component patterns. For example, a *InOT-FTAOP* walk, {*A1, C4, C5,* →*C4*}, has a circle, {*C4, C5* → *C4*} and {*C4, C5*} and {*C5* → *C4*} are two component patterns in a simple suspicious tax evasion group. Using this concept of finding two matched component patterns with a same antecedent element behind a trading arc, a method of mining suspicious groups in the potential component pattern base is constructed and shown in Appendix B. Take Fig. 10 as an example for finding component patterns. There are four suspicious relationship trails related *L1*. In this set of four, the two suspicious relationship trails, {*L1, C2, C5* → *C6*} and {*L1, C1, C3* → *C5*} are such that the end node, *C5*, of the second suspicious relationship trail is included as an element of the first suspicious relationship trail. Then, we can conclude that there exists a simple suspicious group, (*L1, C1, C2, C3, C5*) and two component patterns, {*L1, C2, C5*} and {*L1, C1, C3* → *C5*}, and this suspicious group is identified by two suspicious relationship trails {*L1, C2, C5* → *C6*} and {*L1, C1, C3* → *C5*} with a same antecedent node *L1* behind an interest affiliated transaction (*C3* → *C5*). Similarly, we can find other suspicious groups: (*B1, C5, C6*) and (*B2, C7, C8*).

The proof of the completeness of our proposed method is shown in Appendix A.

Note that a node in the suspicious relationship trail can be a syndicate of *Company* nodes that definitely belong to a *SCS*, $scs_k = \{V_{scsk}, A_{scsk}\}$. Obviously, it is easy to detect the suspicious trading relationships between these *Company* nodes. For this case, if and only if there exists a trading relationship ($v_{c1} \rightarrow v_{c2}$) between two nodes, $v_{c1}, v_{c2} \in V_{scsk}$, then the trading relationship is a suspicious trading relationship. There are two basic facts for proving this. Firstly, there exists at least a trail *t* from v_{c1} to v_{c2} in scs_k according the property of a strongly connected graph. Secondly, the trail *t* and the path ($v_{c1} \rightarrow v_{c2}$) form a suspicious group. This kind of suspicious group detection lacks in the algorithm description in Section 4.3 due to the limitation of the paper length.

5 Experiments and Results Analysis

This section describes our experiments, their settings and results to verify the efficiency of the proposed method.

5.1 Experiments

Based on real data obtained from information CSRC, HRDPSC, and PTAOS in one of the Chinese provinces, we built a TPIIN with a simulated trading relationship network. Firstly, we use these information sources to abstract interlocking relationships and kinships between persons, director relationships and legal person relationships between persons and companies (taxpayers), and the investment relationships between companies. Next, corresponding graphs, *G1*, *G2*, *G3*, G123 and *G4* are produced as shown in Fig. 11, 12, 13, 14 and 15. An interdependence graph shown in Fig. 11 includes interlocking relationship and kinship between persons.

As depicted in Fig. 12, *G2* shows director relationship and legal person relationship between *Person* nodes and *Company* nodes, while *G3* in Fig. 13 includes investment relationships between companies and an antecedent network is shown in Fig. 14. (There is no strongly connected subgraph being found after applying Tarjan's algorithm [26] to *G3* and *Antecedent*. So both *G3* and *Antecedent* are a simple DAG) which is a simple and directed acyclic graph. Then, a trading network is produced according to the rules of random network implemented by Gephi [14] according to the possibility of existence of a transaction between taxpayers. During generation of the trading network, the value of trading probability of each node (company) trading with other companies in the network has a range of 0.002 to 0.1. With this range of probabilities, twenty trading networks are randomly generated. Combining an antecedent network (as shown in Fig. 14) with a trading network (an example of G4 as shown in Fig. 15) forms a TPIIN (one of which is as shown in Fig. 16). This TPIIN includes 4578 nodes that contain 776 directors, 1350 legal persons and 2452 companies. Then, we apply the algorithm proposed in Section 4.3 to discover the suspicious groups (graph-like patterns) from these TPIINs.

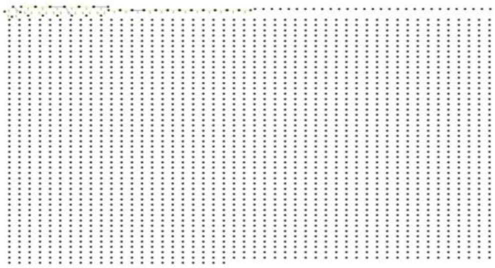


Fig. 11. Interdependence network *G1* (includes 776 directors and 1350 legal persons).

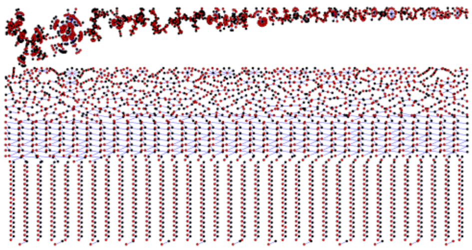


Fig. 12. *G2* (includes 776 directors, 1350 legal persons and 2452 companies).

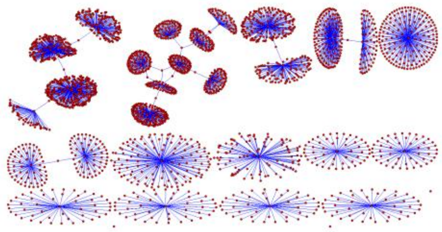


Fig. 13. *G3* (Investment relationships between companies).

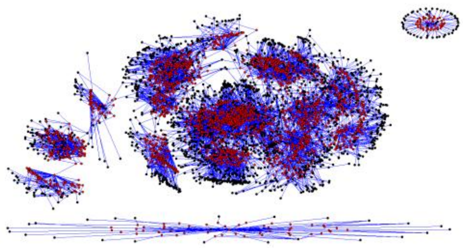


Fig. 14. Antecedent network (*G123*).

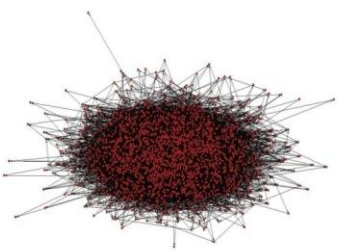


Fig. 15. Trading network *G4* (includes 2452 companies).

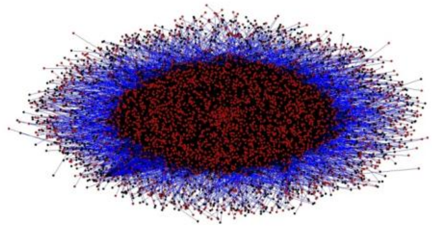


Fig. 16. TPIIN based on real and simulated data.

Note that, in all networks, each red node represents a company (taxpayer), each black node is for a person, each blue arc indicates an influence relationship between a person and a company, and each black arc shows a trading relationship between companies. In Fig. 11, each yellow edge indicates an interlocking link between two directors, while a brown edge indicates a kinship between two *Person* nodes.

Due to the high sensitivity of detailed trading information [1], the TAO did not provide us the details of each transaction, such as the volumes and types of product items traded. It is for this reason that we produced the relationship network by simulation, but it extrapolated a small and reasonable data set. In the simulation, the number of trading relationships of a specific node with other nodes increases as the value of trading probability increases. For gaining the baseline results, we implemented a global traversing algorithm that finds any component patterns behind a trading arc. The idea of this global traversing algorithm is to find all trails between any two different nodes and then check whether any two of these trails form a suspicious group.

5.2 Analysis of Results and Screening of Suspicious Tax Evasion Groups

The experimental results of detecting suspicious groups and trading relationships are shown in Table 1. As indicated in Table 1, we have identified a number of complex suspicious groups detected and simple suspicious groups detected, the number of suspicious trading relationships and the total trading relationship

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2571686, IEEE Transactions on Knowledge and Data Engineering

TABLE 1
DETECTING SUSPICIOUS GROUPS IN A TPIIN OVER VARIOUS TRADING PROBABILITY SETTINGS

Setting of trading probability	Average node degree	Counts of complex suspicious groups detected	Counts of simple suspicious groups detected	Accuracy for detecting suspicious groups	Counts of suspicious trading relationships detected	Counts of total trading relationships	Accuracy of detecting suspicious trading relationships	Percentage of suspicious trading relationships (%)
0.002	3.981	7252	1507	100%	611	11939	100%	5.1177
0.003	5.275	11506	2460	100%	881	17869	100%	4.9247
0.004	6.628	16021	3390	100%	1288	24069	100%	5.3513
0.005	7.941	19375	3977	100%	1573	30094	100%	5.2270
0.006	9.240	23071	4864	100%	1839	36036	100%	5.1032
0.008	11.847	30745	6287	100%	2445	47978	100%	5.0961
0.010	14.491	36702	7881	100%	2991	60117	100%	4.9753
0.012	17.163	44148	8989	100%	3619	72310	100%	5.0048
0.014	19.728	51023	10776	100%	4258	84064	100%	5.0652
0.016	22.424	60777	12680	100%	4895	96403	100%	5.0776
0.018	24.965	67614	13997	100%	5514	108045	100%	5.1034
0.020	27.522	75875	16103	100%	6012	119759	100%	5.0201
0.030	40.748	111885	23328	100%	9122	180401	100%	5.0565
0.040	53.793	149795	31123	100%	12126	240190	100%	5.0485
0.050	66.827	185405	38501	100%	15089	299898	100%	5.0314
0.060	79.940	226187	47361	100%	18212	359975	100%	5.0592
0.070	93.011	261367	55088	100%	21214	419914	100%	5.0520
0.080	106.276	298458	62627	100%	24150	480637	100%	5.0246
0.090	119.554	333271	69844	100%	27129	541489	100%	5.0101
0.100	132.759	372050	78252	100%	30288	602053	100%	5.0308

counts in different simulation parameter settings.

Our proposed method deals with an increase in number of transactions by focusing on the behavior patterns of the suspicious groups. The results in Table 1 indicate that the proposed method is efficient, as it scales down the range of search for the detection of the suspicious groups and suspicious trading relationship. For example, we can observe from Table 1, that total trading relationship counts in the TPIIN for each setting of trading probability increases faster than the number of suspicious groups and trading relationships that we selected as suspicious. It can be particularly noted that, in our TPIIN, a trading arc only denotes that there exists a trading relationship between its start node and its end node and it does not represent a specific transaction. The trading arc can be called as a transaction behavior. This improves the efficiency of identifying the transactions involved in tax evasion by replacing the one-by-one identification method with the proposed method of first identifying suspicious groups based on the transaction behavior pattern. Imagine that there are thousands of transactions involved in various trading relationships between companies, utilizing the patterns of tax evaders screens out the companies not involved in tax evasion from those involved in tax evasion. This contributes to the reduction of the volume of transactions that need to be considered at initial stage. It can be observed from the last column in Table 1 that our simulation results have identified 4.9247%-5.3513% suspicious trading relationships between companies considered in the simulation. Using these suspicious transaction relationships between companies to filter the transactions to be considered in next stage improves searching efficiency. Meanwhile, as shown in Table 1, the proposed algorithms can achieve 100% accuracy rate in pattern detection, if the relationship among enterprises or groups can be identified and represented as edge in the graph. At the same time, we can deduce that there is not a single complex suspicious group or simple suspicious group behind a suspicious trading relationship because the count of suspicious trading relationship detected is much less than ones of complex suspicious groups detected and simple suspicious groups detected.

6 A DEVELOPMENT SYSTEM

By adopting our proposed method in Section 4, the Servyou Group (http://www.servyou.com.cn/), a major tax management software supplier in China has developed a practical system for analyzing and monitoring taxation source. They have deployed the system in several provincial taxation offices of mainland China since 2012. The system accesses national taxation information system, manages the company list (as shown in the second line of left column in Fig. 17), collects the data from NTCIS (national tax information collection system), analyzes them to monitor the calculation of tax model (line 7 of left column in Fig. 17), and tracks the tendency of tax index (line 8 of left column in Fig. 17). The most important functions of this system are affiliated transaction analysis (The detection of suspicious trading relationships and corresponding suspicious groups of specified companies in a suspicious trading relationship was constructed by using our proposed method), application of taxation-related information, integration analysis (as shown in lines 6-11 of left column in Fig. 17). After inputting a specific company into the system, some suspicious groups and suspicious trading relationship were revealed. Fig. 17 shows a tree-like structure that describes investment relationships between companies related to a specific company. Fig. 18 shows a partial influence graph of companies monitored. Fig. 19 (its translation seen in Appendix C) shows an interface of preliminary analysis on a company and its two IATs. In which, firstly, the company's directors, its affiliated companies and their directors were analyzed and identified after carrying on affiliated transaction analysis on the company. Then, according to their suspicious trading relationship, two suspicious transactions (IATs) between the company and its affiliated companies were found and identified manually by tax audition officer.

7 CONCLUSION

The proposed method adopts a heterogeneous information network to describe economic behaviors among taxpayers and casts a new light on the tax evasion detection issue. It not only utilizes multiple homogeneous relationships in big data to form the heterogeneous information network, but also maximally utilizes the advantage of trail-based pattern recognition to select the suspicious groups. Through multi-social relationships fusion and reduction, we simplify the heterogeneous information network into a colored model with two node colors and two edge colors. Moreover, after investigating three cases, we conclude that to identify two suspicious relationship paths behind an IAT is a core problem of detecting suspicious groups. The advantage of the proposed method is that it does not lead to a combinatorial explosion of subgraphs. This improves the efficiency of detecting the IATs-based tax evasion. In contrast with other transaction-based data mining

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2571686, IEEE Transactions on Knowledge and Data Engineering

Fig. 17. Tax source monitoring and management system showing partial investment network between some companies affiliated to a specified company in Shanghai.

Fig. 18. Partial influence graph of the companies monitored.

Fig. 19. Preliminary analysis on a company and its two IATs.

techniques of tax evasion detection, the experimental results show that detecting the suspicious relationships behind a trading relationship firstly scales down the number of suspects (companies) and then the number of suspicious transactions. Furthermore, the mined results are related to many kinds of behavior patterns in illegal tax arrangement, which are intuitive for tax investigators and provide a good explanation of how tax evasion works. Our proposed method has been incorporated into a tax source monitoring and management system employed in several provinces of mainland of China.

In future work, the proposed method will be extended to deal with the situation of directed circles in investment graph, and the weight computation methods of edges during a build-in phase of TPIIN in order to help identify the tax evaders. In addition, the possibility of the introduction of more relationships into the heterogeneous information network will be investigated. Moreover, with the increasing of the size of the TPIIN, the parallel and distributed computation techniques-oriented graph processing will be employed to the proposed method to improve its efficiency and adaptability.

ACKNOWLEDGMENT

This research was partially supported by the National Science Foundation of China under Grant Nos. 91118005, 91218301, 91018011, 61472315, MoE Innovative Research Team in University under Grant No. IRT13035 and China Scholarship Council.

REFERENCES

[1] D. S. Almeling, "Seven Reasons Why Trade Secrets are Increasingly Important," *Berkeley Technology Law Journal*, vol. 27, no. 2, pp. 1092–1118, 2012.

[2] L. Antunes, J. Balsa, and H. Coelho, "Agents that Collude to Evade Taxes," in *Proc. 6th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1263–1265, May. 2007.

[3] Y. S. Chen and C.-H. Cheng, "A Delphi-Based Rough Sets Fusion Model for Extracting Payment Rules of Vehicle License Tax in the Government Sector," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2161–2174, Mar. 2010.

[4] M. J. Ferrantino, X. Liu, and Z. Wang, "Evasion Behaviors of Exporters and Importers: Evidence from the U.S.-China Trade Data Discrepancy," *Journal of International Economics*, vol. 86, no. 1, pp. 141–157, Jan. 2012.

[5] W. F. Fox, L. Lunab, and G. Schau, "Destination Taxation and Evasion: Evidence from U.S. Inter-State Commodity Flows," *Journal of Accounting and Economics*, vol. 57, no. 1, pp. 43–57, Feb. 2014.

[6] Z. Gao, "Transfer Price-Based Money Laundering: A Transit Trader's Perspective," in *Proc. 4th Int. Conf. Wireless Communications, Networking and Mobile Computing (WiCOM)*, pp. 1–5, Oct. 2008.

[7] P. C. Gonz´1´cleza and J. D. Vel´1´csquez, "Characterization and Detection of Taxpayers with False Invoices Using Data Mining Techniques," *Expert Systems with Applications*, vol. 40, no. 5, pp. 1427–1436, Apr. 2013.

[8] N. Goumagias, D. Hristu-Varsakelis, and A. Saraidaris, "A Decision Support Model for Tax Revenue Collection in Greece," *Decision Support Systems*, vol. 53, no. 1, pp. 76–96, Apr. 2012.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Burlington, Massachusetts: Morgan Kaufmann, 2011.

[10] J. Hasseldinea and G. Morris, "Corporate Social Responsibility and Tax Avoidance: A Comment and Reflection," *Accounting Forum*, vol. 37, no. 1, pp. 1–14, Mar. 2013.

[11] L. Kaplow, A. Polinsky, and S. Shavell, *Handbook of Law and Economics*, vol. 1, chap. 10, Elsevier Science Ltd, pp. 647–755, Aug. 2007.

[12] Y. Kim, Savoldi, H. Lee, S. Yun, S. Lee, and J. Lim, "Design and Implementation of a Tool to Detect Accounting Frauds," in *Proc. Int. Conf. Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP)*, pp. 547–552, Aug. 2008.

[13] X. Liu, D. Pan, and S. Chen, "Application of Hierarchical Clustering in Tax Inspection Case-Selecting," in *Proc. Int. Conf. Computational Intelligence and Software Engineering (CISE)*, pp. 1–4, Dec. 2010.

[14] B. M., H. S., and J. M, "Gephi: An Open Source Software for Exploring and Manipulating Networks," in *Proc. AAAI Int. Conf.*

Weblogs and Social Media, 2009.

[15] F. A. M. Nascimento, F. Lehnen, M. V. More, and S. A. Leizer, "Gif: A Web-Based System for Tax Management and Fiscal Intelligence in Municipal Tax Administration," in *Proc. 3rd Int. Conf. Theory and Practice of Electronic Governance (ICEGOV)*, pp. 127–133, Nov. 2009.

[16] Department of Economic Social Affairs, *Practical Manual on Transfer Pricing for Developing Countries*, United Nations, 2013.

[17] S. M. Puffer and D. J. McCarthy, "Can Russia's State-Managed, Network Capitalism be Competitive? Institutional Pull versus Institutional Push," *Journal of World Business*, vol. 42, no. 1, pp. 1–13, Jan. 2007.

[18] PwC, *Transfer Pricing and Developing Countries-Final Report*, Geneva, Swiss: European Commission, 2011.

[19] P. Sikka, "Smoke and Mirrors: Corporate Social Responsibility and Tax Avoidance," *Accounting Forum*, vol. 34, no. 3-4, pp. 153–168, Sept. 2010.

[20] P. Sikka and H. Wllmott, "The Dark Side of Transfer Pricing: Its Role in Tax Avoidance and Wealth Retentiveness," *Critical Perspectives on Accounting*, vol. 21, no. 4, pp. 342–356, Apr. 2010.

[21] C. Steinfield, J. M. DiMicco, N. B. Ellison, and C. Lampe, "Bowling Online: Social Networking and Social Capital within the Organization," in *Proc. 4th Int. Conf. Communities and Technologies*, pp. 245–254, Jun. 2009.

[22] R. Torrini, "Cross-Country Differences in Self-Employment Rates: the Role of Institutions," *Labour Economics*, vol. 12, no. 5, pp. 661–683, Oct. 2005.

[23] R.-S. Wu, C. Ou, H. ying Lin, S.-I. Chang, and D. C. Yen, "Using Data Mining Technique to Enhance Tax Evasion Detection Performance," *Expert Systems with Applications*, vol. 39, no. 10, pp. 8769–8777, Aug. 2012.

[24] S. Xu, "An Application on Association Rules Data Mining in the Tax Audit System," *Economic Supervision*, vol. 13, pp. 43–44, Nov. 2011.

[25] Yizhou Sun and Jiawei Han, "Mining Heterogeneous Information Networks: A Structural Analysis Approach," *SIGKDD Explorations*, vol. 14, no. 2, pp. 20-28, Apr. 2013.

[26] Tarjan Robert, "Depth First Search and Linear Graph Algorithms," *SIAM Journal on Computing*, vol. 1, no. 2, pp. 146–160, Jun. 1972.

[27] Yizhou Sun and Jiawei Han, "Meta-Path-Based Search and Mining in Heterogeneous Information Networks," *Tsinghua Science and Technology*, vol. 18, no. 4, pp. 329-338, Aug. 2013.

[28] Yun Xiong, Yangyong Zhu, and Philip S. Yu, "Top-k Similarity Join in Heterogeneous Information Networks," *IEEE Trans. Knowledge and Data Engineering*, vol. 27, no. 6, pp. 1710-1723, Jun. 2015.

Feng Tian is an associate Professor of Systems Engineering Institute at Xi'an Jiaotong University, China. His research interests include system modelling and analysis and cloud computing. He has over 50 academic publications. He is a member of the IEEE.

Tian Lan is currently a graduate student in Systems Engineering Institute at Xi'an Jiaotong University, China. She is interested in data mining.

Kuo-Ming Chao is a professor of computing at Coventry University, UK. His research interests include the areas of cloud computing and big data etc as well as their applications. He has over 150 refereed publications. He is a member of the IEEE.

Nick Godwin is a part time research fellow at Coventry University having previously been leading research in Computer Science at the same University. In recent years he has been supporting a wide range of research projects in Computing.

Qinghua Zheng received his B.S. and M.S. degrees in computer science and technology from Xi'an Jiaotong University in 1990 and 1993, respectively, and his Ph.D. degree in systems engineering from the same university in 1997. He was a postdoctoral researcher at Harvard University in 2002. Since 1995 he has been with the Department of Computer Science and Technology at Xi'an Jiaotong University, and was Cheung Kong Professor in 2009. His research interests include intelligent e-learning and network security.

Nazaraf Shah is a senior lecturer at Coventry University, Coventry, UK. His research interests include intelligent agents, service-oriented computing and cloud computing. He has over 50 publications in various international conferences and journals.

Fan Zhang is currently a senior engineer and a manager of department of data analytics in SERVYOU GROUP, China. He received his Bachelor's degree in Computer Science and Technology from Xidian University in 1998. He has been working in taxation information industry since 1998.