This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Ruan, Jianfei; Yan, Zheng; Dong, Bo; Zheng, Qinghua; Qian, Buyue
Identifying suspicious groups of affiliated-transaction-based tax evasion in big data

Published in:
Information Sciences

DOI:
[10.1016/j.ins.2018.11.008](https://doi.org/10.1016/j.ins.2018.11.008)

Published: 01/03/2019

Document Version
Peer reviewed version

Published under the following license:
CC BY-NC-ND

Please cite the original version:
Ruan, J., Yan, Z., Dong, B., Zheng, Q., & Qian, B. (2019). Identifying suspicious groups of affiliated-transaction-based tax evasion in big data. *Information Sciences*, 477, 508-532. <https://doi.org/10.1016/j.ins.2018.11.008>

Identifying Suspicious Groups of Affiliated-Transaction-based Tax Evasion in Big Data

Jianfei Ruan^a, Zheng Yan^{b,c,*}, Bo Dong^{d,e}, Qinghua Zheng^a, Buyue Qian^a

^a*MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China*

^b*State Key Laboratory on Integrated Services Networks, School of Cyber Engineering, Xidian University, Xi'an 710071, China*

^c*Department of Communications and Networking, Aalto University, Espoo 02150, Finland*

^d*School of Continuing Education, Xi'an Jiaotong University, Xi'an 710049, China*

^e*National Engineering Lab of Big Data Analytics, Xi'an Jiaotong University, Xi'an 710049, China*

Abstract

Affiliated-transaction-based tax evasion (ATTE) is a new strategy in tax evasion that is carried out via legal-like transactions between a group of companies that have heterogeneous, complex and covert interactive relationships to evade taxes. Existing studies cannot effectively detect ATTE behaviors since (i) they perform well only for determining the abnormal financial status of individuals and ineffectively address the interactive relationships among companies, (ii) they aim at detecting ATTE from the perspective of structural characteristics, which leads to a poor false-positive rate, and (iii) few of them perform well in most sectors of companies. Effectively detecting suspicious groups according to both structural characteristics of ATTE groups and business characteristics of ATTE means (BC-ATTEM) remains an open issue. In this paper, we propose an affiliated-parties interest-related network (APIRN) for modeling affiliated parties, interest-related relationships, and their properties for identifying ATTE. Then, we identify the behavioral patterns of ATTE via topological pattern abstraction from APIRN and the-

*Corresponding author

Email addresses: `xjtu_jfruan@163.com` (Jianfei Ruan), `zheng.yan@aalto.fi` (Zheng Yan), `dong.bo@mail.xjtu.edu.cn` (Bo Dong), `qhzheng@mail.xjtu.edu.cn` (Qinghua Zheng), `qianbuyue@xjtu.edu.cn` (Buyue Qian)

oretical inference of BC-ATTEM. Based on the above, we further propose a hybrid method, namely, 3TI, for identifying ATTE suspicious groups via three steps: tax rate differential detection, topological pattern matching and tax burden abnormality identification. Experimental tests that are based on two years of real-world tax data from a province in China demonstrate that 3TI can identify ATTE suspicious groups with higher accuracy and better generality than existing works. Moreover, we identify various interesting implications and provide useful guidance for ATTE inspection based on an analysis of our experimental results.

Keywords: Graph mining, Affiliated transaction, Tax evasion, Big data

1. Introduction

Tax is one of the most important types of fiscal revenue and a means with which the government can regulate the allocation of financial resources and income distribution [18, 11, 15, 17, 38]. However, countless businesses, especially giant corporations with layered subsidiaries, attempt to lower their taxable income and shield more of their hard-earned money from the tax administration. In practice, there are two possible ways to pay less in tax: tax avoidance and tax evasion. Although they sound similar, they differ in terms of the procedures that are employed. Tax avoidance is the legitimate minimization of taxes via methods that are specified in the tax law. In contrast, tax evasion is an illegal practice of reducing taxes by not reporting income, reporting expenses that are not legally allowed, or not paying individually owed taxes, which causes a significant loss of national revenue [43]. The phenomenon of tax evasion is particularly severe in developing countries and, thus, has attracted the attention of tax authorities for a long time [11]. For example, it was reported that tax revenue loss in China accounted for 9.99% of its gross national product (\$9 trillion) in 2013 [26].

For confronting this serious tax revenue loss, the Chinese government implemented a series of tax auditing measures, including manual case selection, whistle-blowing-based selection [9] and computer-based case selection (such as financial statement cross-matching [43] and abnormal financial index screening [30]). As a result of the use of these methods, traditional tax evasion behaviors, such as writing false value-added tax invoices, fake invoices and account manipulations, have been reduced significantly [39].

In recent years, a new type of tax evasion appeared: companies work

together to conceal assets, income or information to dodge liability through affiliated transactions (ATs) [39], which is called affiliated-transaction-based tax evasion (ATTE). This type of evasion is carried out via “licit” transactions between companies that have heterogeneous, complex and covert interactive relationships, such as an interlocking shareholding relationship between companies’ shareholders or a kinship between companies’ legal representatives. These relationships are diverse and used for interest transfer to evade taxes mainly via mispricing. Consequentially, a substantial amount of tax revenue loss has been caused by ATTE. According to the statistics of the Organization for Economic Cooperation and Development (OECD), a conservative estimate of tax revenue loss due to ATTE in developing countries is approximately 4.4% of the total tax revenue in the entire developing world in 2016 [40]. At present, tax authorities are equipped with limited resources for solving the problem of ATTE. Traditional tax auditing measures are typically time-consuming with tedious processes [39]; thus, they are not feasible for detecting ATTE.

Worse, for ATTE detection in China, a huge volume of transactions and their related data must be considered. For example, there are more than 450 million taxpayers and 37,000 taxation administration offices in the country. The number of annual tax-related monthly financial reports has reached 2.5 billion, the daily peak number of transaction records has reached 2 billion, and the volume of aggregated annual data is 200 TB, which is considered big data. This volume of data further challenges traditional tax evasion detection methods.

Effectively detecting ATTE in a technical manner, especially identifying ATTE suspicious groups transactions from large-scale business transactions and related data, is a challenging issue and has attracted substantial attention in recent years. Current research on ATTE detection can be classified into two categories: the machine-learning-based auditing method (MLAM) [29, 28] and the graph-based auditing method (GAM) [39, 40]. MLAM enables the extraction and generation of knowledge from large volumes of tax data to detect and characterize ATTE behaviors and improve the efficiency of identifying problematic individuals, while GAM focuses on collecting sophisticated organizational structures of ATTE cases to identify the structural characteristics of ATTE groups. MLAM can detect the abnormal financial status of individuals with high accuracy and efficiency; however, it becomes helpless when facing heterogeneous, complex and covert interactive relationships and cannot detect ATTE groups. Meanwhile, its trained models are

typically sensitive to the training data. The results of MLAM are not intuitive and cannot provide a clear explanation of ATTE behaviors. In contrast, GAM can detect suspicious groups that exhibit the structural characteristics of ATTE. However, it neglects the business characteristics of tax evasion techniques, such as referred abnormal financial status, which is verified by tax authorities and vital for ATTE detection. Detecting ATTE based on only structural characteristics could lead to a high false-positive rate. Few methods in the literature can realize the advantages of both MLAM and GAM in ATTE detection. In addition, Existing solutions seldom perform well in identifying ATTE in most sectors of companies.

In this paper, we propose a novel approach for effectively identifying suspicious groups that exhibit both the structural and business characteristics of ATTE via a uniform detection process. First, we present an affiliated-parties interest-related network (APIRN) for characterizing essential data items for ATTE detection. Then, we identify the behavioral patterns of ATTE by extracting structural characteristics from APIRN and theoretically inferring business characteristics of ATTE means (BC-ATTEM) based on evidence. Based on the behavioral patterns, we further propose a hybrid method, namely 3TI, for identifying ATTE suspicious groups. It consists of three steps: tax rate differential detection, topological pattern matching and tax burden abnormality identification. To evaluate the effectiveness of 3TI, experiments that are based on the real-world tax data of one province in China from 2014 to 2015 were carried out. The results demonstrate that 3TI outperforms existing methods in terms of accuracy.

3TI is original and differs substantially from the previous methods of ATTE detection. Typically, those methods extract the features of ATTE from structural characteristics of tax evasion groups or the abnormal financial statuses of individual taxpayers and detect ATTE via graph-based or machine learning techniques. 3TI, which is proposed in this paper, applies topological pattern matching, along with data fusion and inference based on evidence, to process heterogeneous tax data for the detection of ATTE groups based on both structural characteristics and business characteristics of ATTE. The main contributions of this paper can be summarized as follows:

- (1) We aim at identifying ATTE suspicious groups by examining the heterogeneous, complex and covert interactive relationships and the business characteristics of tax evasion strategies that are involved in ATTE.
- (2) We find that the properties of affiliated parties (APs) and interest-related relationships (interactive relationships between APs, which are denot-

ed as IRRs) are essential for detecting ATTE. Thus, we generate an APIRN that is based on a directed attributed multi-graph-based model [4] for characterizing APs, IRRs and their properties via heterogeneous network fusion.

(3) We propose a hybrid method, namely, 3TI, for detecting ATTE suspicious groups based on the abstracted behavioral patterns of ATTE, which integrates topological patterns with BC-ATTEM.

(4) We evaluate the performance of 3TI via comparison with existing works on a real-world big dataset. The results demonstrate that 3TI substantially improves the identification accuracy, is widely generalizable, and achieves lower false-positive and false-negative rates compared with existing methods. In addition, we discuss the implications of these results and provide useful guidance for ATTE inspection via analysis of our experimental results.

The remainder of the paper is organized as follows: Section 2 presents a brief review of related works. In Section 3, we introduce key definitions and abbreviations that are used in the paper and describe the APIRN model. Then, we propose the behavioral patterns of ATTE abstraction in Section 4, followed by the 3TI method for suspicious group identification in Section 5. We present the experimental results and discuss additional implications and findings in Section 6. In Section 7, we further demonstrate the implementation of 3TI. Finally, our conclusions are presented in the last section.

2. Related Work

Many novel techniques for detecting financial fraud have been proposed in the literature. Most of them have been introduced in various comprehensive surveys [41, 7]. However, tax fraud research is still in its infancy. The techniques that are available for detecting tax fraud have various strengths and weaknesses.

The current mainstream ATTE detection methods fall into three categories: traditional tax auditing methods, machine-learning-based auditing methods and graph-based auditing methods. The traditional tax auditing methods carry out substantive audit procedures via cross-matching of financial statements or screening of abnormal financial indices. The machine-learning-based auditing methods identify ATTE patterns based on the attributes of taxpayers for detecting problematic individuals. The graph-based auditing methods focus on collecting sophisticated organizational structures of ATTE cases for finding the structural characteristics of ATTE groups.

2.1. Traditional Tax Auditing Methods

Manual case selection, whistle-blowing-based case selection and computer-based case selection are three frequently used methods of traditional tax auditing [39]. Manual and whistle-blowing-based case selection tasks are highly tedious and time-consuming and the effectiveness of these tasks heavily relies on the experience and skills of tax officers. In contrast, computer-based case selection is only used to facilitate manual case selection and focuses mainly on cross-matching financial statements [43] and screening abnormal financial indices [30]. Thus, via traditional tax auditing methods, addressing a large volume of taxation information in an efficient and effective manner is impossible. In practice, tax authorities are only able to screen a small percentage of tax reports for further auditing because of limited staff resources [24]. Thus, countering ATTE via traditional tax auditing methods is not feasible.

2.2. Machine-Learning-based Auditing Methods

To address the limitations of traditional tax auditing methods, machine learning techniques are frequently adopted to identify taxpayers who evade obligations [34]. Existing methods include association analysis [43, 33], cluster analysis [18, 30, 6, 42], classification [10, 27, 12, 21, 25, 28], simulation [5, 31, 36], and reinforcement learning [1, 19].

Association analysis. Wu et al. [43] employed association rules to a value-added tax database to uncover patterns and relationships among attributes that are useful for identifying tax evasions. Matos et al. [33] developed a method to mine frequent fraud patterns using association rules and rank taxpayers according to their potential to commit fraud.

Cluster analysis. González and Velásquez [18] adopted clustering algorithms, such as self-organizing map and neural gas, to identify taxpayer groups with similar behaviors. Assylbekov et al. [6] and Williams et al. [42] applied self-organizing maps to identify abnormal clusters with unusual behaviors that might carry out tax fraud. Liu et al. [30] used hierarchical clustering in tax inspection case selection based on seven financial indexes.

Classification. Chen and Cheng [10] proposed a hybrid model that combines a Delphi method and a rough set classifier to classify vehicle license tax payment. Junqué de Fortuny et al. [27] applied support vector machine and naïve Bayes to detect residence frauds of taxpayers.

Simulations. Antunes et al. [5] used exploratory simulations and progressively deepening models of agents to study the reasons behind tax evasion. To generate aggregated patterns of tax behaviors, Noguera et al. [36]

presented an agent-based model for the simulation of tax compliance by combining rational choices with social influence mechanisms.

Reinforcement Learning. Abe et al. [1] developed a constrained Markov decision process-based approach to the problem of optimally managing the tax, and more generally debt, collections processes at financial institutions. Goumagias et al. [19] described a Markov-based decision support model to predict behaviors of risk-neutral taxpayers and evaluate tax policies before implementation.

The above machine learning-based auditing methods partition taxpayers into “evasion” or “nonevasion” and achieve high accuracy and efficiency in individual tax fraud detection. However, none of these methods provide tax authorities with specific reasons for the classification; hence, each filing of an “evasion” taxpayer still must be manually examined by auditors. Meanwhile, they cannot identify the sophisticated organizational structures that are often created by tax evaders. To overcome this problem, it is necessary to identify an entire relationship network of tax evaders and detect their roles in it [14].

2.3. Graph-based Auditing Methods

Graph-based anomaly detection supports the investigation of sophisticated organizational structures of taxpayers. This type of method can be applied in diverse fields, such as detecting credit card fraud, calling card and telecommunications fraud, health insurance claim errors, and securities fraud [2]. Recently, graph-based anomaly detection has been applied to the detection of tax evasion [27, 14, 8, 39, 40].

Dreżewski et al. used data from bank statements and the National Court Register to construct and analyze networks in an investigation into money laundering cases [14]. However, this approach only identifies the roles of offenders and the connections between them; it cannot detect detailed fraud schemes.

Tian et al. proposed a colored-network-based model, namely, the taxpayer interest interacted network (TPIIN), and an identification method for detecting tax evasion by building a pattern tree and matching component patterns [39]. In TPIIN, APs and IRRs are characterized via tags (represented by colors). However, it fails to describe the attributes of APs and IRRs, which, however, can represent identifiers for identifying tax evasion, as we discovered during our participation in tax auditing. In addition, this identification method does not take the tax evasion strategy into account.

Tselykh et al. presented an attributed-graph-based approach to the anomaly detection problem of identifying affiliated and interdependent entities that might be at risk for ATTE [40]. However, this work places too much emphasis on clustering analysis for attributes and graph connectivity properties analysis and pays little attention to the business characteristics of tax fraud, which leads to difficulty in explaining and tracing the derived results.

In summary, graph-based auditing methods can detect the sophisticated organizational structures of tax evaders. However, they neglect the business characteristics of tax evasion, thereby leading to high false-positive rates.

3. Definitions, Abbreviations and APIRN Model

First, we introduce the key tax knowledge that is involved in the design of the APIRN model as follows:

Definition 1. Affiliated Party (AP): If one party (natural person or company; a company is also referred to as a taxpayer in the field of taxation) is involved directly or indirectly in administration, control, capital contribution or investment in any form of influence on another party, then the two parties are defined as affiliated parties (APs) to each other [20]. In particular, if an affiliated party is a natural person, he or she is defined as an affiliated person. If it is a company, it is defined as an affiliated company.

Definition 2. Interest-Related Relationship (IRR): Any form of influence that includes but is not limited to administration, control, capital contribution or investment, as well as trading relationship, between APs is defined as an interest-related relationship (IRR).

Definition 3. Affiliated Transaction (AT): A business transaction between APs is defined as an affiliated transaction (AT) and the two companies that are involved in the AT are defined as AT companies.

Definition 4. Affiliated-Transaction-based Tax Evasion (ATTE): AT companies evade taxes by trading on mispricing based on the covert interactive relationships between them. This type of tax evasion is defined as affiliated-transaction-based tax evasion (ATTE).

Definition 5. Earnings Before Tax (EBT): The money that is retained by a company before the money that is due to be paid as taxes has been deducted, which is calculated as the sales revenue minus the operating expenses and the cost of goods that have been sold [3, 32].

Definition 6. Corporation Income Tax (CIT): A direct tax that is imposed by a jurisdiction on the income or capital of a company, which is multiplied by the corporate tax rate to yield EBT [13, 22].

Definition 7. Net Profit (NP): A measure of the profitability of a company after accounting for all costs, which is calculated as CIT minus EBT [37].

Definition 8. Tax Burden Rate (TBR): The ratio of CIT to the company’s sales revenue, which is frequently used as a measure of the operating status of a company [35].

The primary strategy for identifying ATTE behaviors is to extract and model the data items that are involved in identification. Previous work focused on APs, IRRs and their categories but failed to describe the properties of APs and IRRs. By tracking and analyzing tax inspection cases and participating in tax audits, we demonstrate that both the properties of APs (e.g., location, industry sector, registration type and corporate tax rate) and the properties of IRRs (e.g., the tightness of an IRR, trading date and transaction amount) contribute significantly to ATTE identification in practice. The tightness of an IRR indicates the degree of related interests of a direct tie, which is calculated according to various economic behaviors and social relations between APs, along with other impact factors.

Therefore, based on a directed attributed multigraph model [4], we propose an affiliated-party interest-related network (APIRN) for characterizing APs, IRRs and their properties for ATTE identification. The APIRN model is described as follows:

(1) Each AP is treated as a node and the IRRs between APs are treated as edges.

(2) The category, location, industry sector, registration type and corporate tax rate of each AP are set as the properties of its corresponding node. The category and tightness of each IRR and the trading date and transaction amount are set as the properties of its corresponding edge. The categories of IRRs include kinship relationship, interlocking shareholding relationship, shareholding relationship, legal representative relationship and trading relationship. Kinship relationship and interlocking shareholding relationship are abstracted as bidirectional edges and the others are represented by unidirectional edges.

Then, we formulate an APIRN as a quadruple:

APIRN = (V, E, Property(V), Property(E)).

where

$V = \{v_p | p = 1, \dots, N_p\}$ denotes the vertex set of an APIRN, and N_p denotes the number of vertices. $V = P \cup C$, where $P = \{v_l | l = 1, \dots, N_l, N_l \leq N_p\}$ denotes the set of affiliated persons, $C = \{v_c | c = 1, \dots, N_c, N_c \leq N_p\}$ denotes the set of affiliated companies, and $N_l + N_c = N_p$ is satisfied.

$E = \{e_{pq}^{\vec{R}}\} = \{(v_p, v_q) | 0 < p, q \leq N_p\}$ denotes the edge set, where $e_{pq}^{\vec{R}} = (v_p, v_q)$ indicates that there exists an edge from the p -th vertex to the q -th vertex that represents a unidirectional relationship $\vec{R}$. In addition, a bidirectional relationship is represented by two unidirectional relationships: $e_{pq}^{\leftrightarrow \vec{R}} = e_{pq}^{\vec{R}} \cup e_{qp}^{\vec{R}}$.

$\text{Property}(V) = (VId, VCategory, Location, Industry, Registration, TaxRate)$ denotes a set of properties of vertex V , where VId is the unique tag of the vertex, e.g., a person's ID or a company's taxpayer ID. $VCategory$ denotes the category of the vertex, which is determined by the multiple roles of the vertex. The roles include affiliated company (C)/affiliated person (P), legal representative (L) and shareholder (S), which are not mutually exclusive. To represent the possible combinations of roles, we introduce a 3-bit binary encoding scheme for encoding these roles as C - 100, P - 000, L - 010 and S - 001. Then, $VCategory$ of a vertex is obtained by performing the OR operation on the encoded roles that are possessed by the vertex. For example, an affiliated person that is both a legal representative and a shareholder of a company is denoted as 011 as a result of "000 OR 010 OR 001".

When the most significant bit of a vertex's $VCategory$ is equal to 1, that is, the category holds the role of an affiliated company, then the properties that are attached to the vertex include *Location* (the location of the company), *Industry* (the sector of industry), *Registration* (the registration type of the company) and *TaxRate* (the corporate tax rate of the company), in addition to VId and $VCategory$.

$\text{Property}(E) = (ECategory, W, Date, Amount)$ denotes a set of properties that are attached to an edge. $ECategory$ is the category of the edge, W represents the tightness of an IRR, $Date$ refers to the trading date, and $Amount$ indicates the transaction amount.

$ECategory$ is represented as $ECategory \in \{\overrightarrow{KR}, \overrightarrow{IR}, \overrightarrow{SR}, \overrightarrow{LR}, \overrightarrow{TR}\}$.

- $\overrightarrow{KR} = \{e_{pq}^{\overrightarrow{KR}} | 0 < p, q \leq N_p\}$ denotes a kinship relationship between APs p and q .
- $\overrightarrow{IR} = \{e_{pq}^{\overrightarrow{IR}} | 0 < p, q \leq N_p\}$ denotes an interlocking shareholding relationship between shareholders p and q .

- $\overrightarrow{SR} = \{e_{pq}^{\overrightarrow{SR}} | 0 < p, q \leq N_p\}$ denotes a shareholding relationship between shareholder p and company q .
- $\overrightarrow{LR} = \{e_{pq}^{\overrightarrow{LR}} | 0 < p, q \leq N_p\}$ denotes a legal representative relationship between legal representative p and company q .
- $\overrightarrow{TR} = \{e_{pq}^{\overrightarrow{TR}}(i) | 0 < p, q < N_p, 0 < i \leq N_{pq}\}$ denotes a trading relationship between companies p and q , where N_{pq} denotes the number of transactions between the p -th and the q -th vertices.

Moreover, the five IRRs are divided into two types: static and dynamic IRRs.

(1) Because KR , IR , SR and LR do not change substantially over time, they are classified as static IRRs and the corresponding edges are defined as static IRR edges. For each edge with $ECategory \in \{\overrightarrow{KR}, \overrightarrow{IR}, \overrightarrow{SR}, \overrightarrow{LR}\}$, W is included in its properties, while $Date$ and $Amount$ are set to null. For $ECategory \in \{\overrightarrow{KR}, \overrightarrow{IR}, \overrightarrow{LR}\}$, W is set to 1, thereby indicating strong interest ties between APs; for $ECategory = \overrightarrow{SR}$, W is set as a shareholding rate, which quantifies the tightness of the relationship between a shareholder and its company.

(2) Because the partner, number and amount of each transaction may change over time, $\overrightarrow{TR}$ is classified as dynamic IRR and the corresponding edge is defined as a dynamic IRR edge. For each edge with $ECategory = \overrightarrow{TR}$, $Date$ and $Amount$ are included in its properties, while W is set to null.

Based on the classification of IRRs, an APIRN is divided into two sub-graphs: a static subgraph of APIRN (SAPIRN), namely, a network that is composed of all static IRR edges of an APIRN, and a dynamic subgraph of APIRN (DAPIRN), namely, a network that is composed of all dynamic IRR edges of an APIRN.

To facilitate presentation, we summarize the descriptions of the terms that are used in this paper in Table 1.

4. ATTE Behavioral Pattern Abstraction

4.1. Analysis of Typical ATTE Inspection Cases

We analyze two typical ATTE inspection cases to determine the identification criteria that are adopted by tax authorities and provide guidance for abstracting ATTE behavioral patterns.

Table 1: Term Descriptions

Abbreviations	Description
APIRN	The affiliated parties' interest-related network;
Static IRR	The kinship, interlocking shareholding, shareholding or legal representative relationship;
Dynamic IRR	The trading relationship;
SAPIRN	The static subgraph of APIRN that is constructed by all static IRR edges of an APIRN;
DAPIRN	The dynamic subgraph of APIRN that is composed of all dynamic IRR edges of an APIRN;
subSAPIRN	The graph that is constructed by one maximal weakly connected subgraph of a SAPIRN;
subDAPIRN	The graph that is constructed by all dynamic IRR edges between the company nodes of a subSAPIRN;
Static IRR trail (control chain)	The sequence of IRR edges that connect a sequence of AP vertices that are all distinct from one another, with the added restriction that the IRR edges all be in a SAPIRN and directed in the same direction;
<i>SBE-BT</i>	The topological pattern that is described as two static IRR trails with the same bidirectional edge behind a transaction;
<i>SAN-BT</i>	The topological pattern that is described as two static IRR trails with the same antecedent node behind a transaction;
BC-ATTEM	The business characteristics of the ATTE strategy;
SITL	The static IRR trail library for each subSAPIRN;
ALCS	The Average Level of the Companies with the Same location, industry sector and registration type.

[Case 1] Retaining profits through AT

As shown in Fig. 1, a chip company, namely, C3, in Shenzhen Special Economic Zone mainly produced chips and enjoyed a preferential tax rate of 15%. Eighty percent of C3's shares were held by company C1 in Shanghai City, which was an outsourcing enterprise and provided main raw materials to C3. All products of C3 were sold to a company, namely, C2, and the tax rate of C2 was 25%. Moreover, the legal representative L1 who controlled C1, who is referred to as L1, and the legal representative who controlled C2, who is referred to as L2, were brothers.

Figure 1: Retaining profits through AT Case

Tax authorities verified that (1) there was a covert relationship between the legal representatives of C3 and C2, namely, a kinship, which means they were each other's APs; (2) the transaction price between C3 and C2 was 32% higher than the local market price; and (3) TBR of C3 was 54% higher than the average level of the companies with the same location, industry sector and registration type (ALCS) and that of C2 was 41% lower than ALCS. Then, it was observed that the indirect controller of C3 (L1) and the direct controller of C2 (L2) made an agreement to trade at a price that well exceeded the local market price so that profits were retained in a low-tax-rate company, namely, C3 (which is referred to as retaining profits). Therefore, the AT between C3 and C2 led to a reduction in the overall corporation income tax (CIT) [13]. In tax inspection, this tax evasion strategy is referred to as retaining profits.

[Case 2] Transferring profits through AT

As shown in Fig. 2, a real estate development enterprise, which is denoted as C4, sold high-price properties to a real estate sale company, which is

denoted as C5, almost at cost, namely, 37% lower than the local market price. C4 had a normal tax rate of 25%, while C5 enjoyed a preferential tax rate of 13% as a startup. C4 and C5 had H1 and H2 as controlling shareholders, respectively, who held 30% and 40% of the shares. Moreover, H1 and H2 held 50% and 30%, respectively, of the shares of company C6 and acted together to control C6. This type of agreement brings H1 and H2 into a relationship that is called shareholder interlocking.

Figure 2: Transferring profits through AT Case

Tax authorities verified that (1) there existed a covert relationship between C4 and C5, namely, a shareholder interlocking relationship, which means they were each other's APs; (2) the transaction price between C4 and C5 was 37% lower than the local market price; and (3) TBR of C4 was 43% lower than ALCS and that of C5 was 48% higher than ALCS. Then, it was observed that the shareholder of C4 (H1) made an agreement with that of C5 (H2) to trade at a price that was much lower than the local market price so that profits were transferred to the low-tax-rate company, namely, C5. Therefore, the AT between C4 and C5 resulted in a reduction of CIT. In tax inspection, this tax evasion strategy is referred to as transferring profits.

Learning from the above cases, we find that two major criteria are employed by tax authorities to detect ATTE behaviors: The first criterion is the existence of a covert relationship between two trading companies, which is determined via AT detection. If this criterion is satisfied, each transaction

between the two companies is investigated to determine whether there is a trading price abnormality that leads to ATTE.

4.2. ATTE Behavioral Patterns

Inspired by the two criteria that are employed by tax authorities to detect ATTE behaviors, we abstract ATTE behavioral patterns from two aspects: structural characteristics of AT groups and business characteristics of ATTE strategies.

4.2.1. Structural Characteristics of AT Groups

According to the analysis in Section 4.1, the first criterion for detecting ATTE behaviors is to discover ATs by identifying whether a covert relationship exists between two trading companies. Moreover, the covert relationship between two trading companies, combined with the transactions between them, is described as a topology in an APIRN. In this section, we refine the structural topology characteristics of AT groups in an APIRN as topological patterns of AT, which are regarded as the topological premise of ATTE.

Based on the modeling strategy of APIRN, we abstract Case 1 as the topology in Fig. 3 (a). The three red nodes refer to the three companies: C1, C2 and C3. C1 held 80% of the shares of C3, which is represented as a purple arc, which is denoted as (C1, C3), with a weight of 0.8. The black arc, which is denoted by (C3, C2), represents a trading relationship between C3 and C2. L1 and L2 were the legal representatives of C1 and C2, which are represented as two green arcs, namely, (L1, C1) and (L2, C2), with a weight of 1. Brown bidirectional edge (L1, L2), which has a weight of 1, indicates that there existed a kinship between L1 and L2. Then, this graph indicates that the kinship between two legal representatives of two companies provides a hint as to suspicious relationships that are behind an AT. From a topological standpoint, this type of AT is mapped into a graph structure that consists of two directed static trails, namely, $(L1 \rightarrow L2 \rightarrow C2) + (L2 \rightarrow L1 \rightarrow C1 \rightarrow C3)$, and a dynamic IRR arc: $(C3 \rightarrow C2)$.

Similarly, we abstract Case 2 as the topology in Fig. 3 (b). There exist three company nodes, namely, C4, C5 and C6, and two shareholder nodes: H1 and H2. The trading relationship between C4 and C5 is represented as a black arc: (C4, C5). H1 and H2, as the shareholders of C4 and C5, both invested in C6 with shareholdings of 50% and 30%, which are represented as two purple arcs, which are denoted as (H1, C6) and (H2, C6), with weights of 0.5 and 0.3. The yellow bidirectional edge, which is denoted as (H1, H2),

represents the interlocking relationship between shareholders H1 and H2. Then, this graph indicates that the interlocking relationship between two shareholders who belong to two companies suggests suspicious relationships that are behind an AT. From a topological standpoint, this type of AT is mapped into a graph structure that consists of two directed static trails, namely, $(H1 \rightarrow H2 \rightarrow C5) + (H2 \rightarrow H1 \rightarrow C4)$, and a dynamic IRR arc: $(C4 \rightarrow C5)$.

Figure 3: Topologies abstracted from Case 1 and Case 2

As IRRs between companies with ATTE behaviors materialize in various forms, different ATTE cases are abstracted as different topologies with various IRR weights and lengths of the control chain. In other words, there exists a problem of combinatorial explosion of topologies [39] and it is impossible to list all topologies. Therefore, it is necessary to extract structural characteristics from topologies of ATTE and refine them to topological patterns.

Topological patterns serve as structural generalizations over topologies of ATTE behaviors and characterize the conditions in which there exist potential static IRRs that are behind an AT, based on which ATs can be identified. Thus, topological patterns are regarded as the topological premise of ATTE. According to the topologies of the above two cases, a topological pattern, which is described as *two static IRR trails with the same bidirectional edge that are behind a transaction* (see Fig. 4(a)), is refined (*SBE-BT* for short). Similarly, Tian et al. refined another topological pattern [39]; we do not repeat it herein. The obtained topological pattern is described as *two static*

IRR trails with the same antecedent node behind a transaction (see Fig. 4(b), which is denoted as *SAN-BT*). Then, by carefully observing many auditing cases, we verify the comprehensiveness of the proposed topological patterns, as they cover all possible ATTE instances.

Figure 4: Topological patterns

4.2.2. Business Characteristics of ATTE Means

ATs can be identified via topological pattern matching in an APIRN. However, the existence of an AT does not imply the existence of an ATTE behavior. Taking the tax audit conclusion of jurisdiction S of China in 2015 as an example, preliminary statistics reveal that the number of ATs that indeed evaded tax accounted for only 19.8% of the total ATs. Thus, in this section, we further mine the business characteristics of ATTE strategies on the basis of structural characteristic of AT to improve the behavioral patterns and detect ATTE more precisely.

Referring to Section 4.1, two ATTE strategies are commonly used: (i) retaining profits: selling products from a high-tax company to its low-tax AT company at a price that is lower than the market level and (ii) transferring profits: selling products from a low-tax company to its high-tax AT company at a price that exceeds the market level. These two ATTE strategies shift profits from a high-tax company to a low-tax one, with the aim of reducing the overall CIT. Moreover, tax authorities typically verify whether an

AT violates the ALP via a price reasonability check in an artificial manner. However, it is not feasible to identify ATTE behaviors from large-scale tax data with computer processing for the following reasons: (i) the verification operation is labor-intensive and time-consuming; (ii) it is difficult to recognize the real products from the description on an invoice due to the lack of standardization of invoice descriptions in China; and (iii) it is difficult to determine the unit price of a product when the unit of measurement is fuzzy, e.g., an invoice records that the total amount of “a batch of steel” is two million, but the unit price of the steel cannot be counted due to the fuzzy unit—“batch.”

Therefore, we carry out a theoretical study of BC-ATTEM from another point of view. Concretely, we make an inference based on evidence to quantitatively characterize BC-ATTEM from two aspects: (i) the property premise that must be satisfied to carry out ATTE behaviors and (ii) the direct consequence that results from ATTE behaviors.

The notations that are used in the inference are described in Table 2.

Table 2: Notations used in the inference

Notation	Description
S	The sales revenue;
C	The cost of goods sold;
O	The operating expense;
R	The corporate tax rate;
EBT	The earnings before tax, where $EBT = S - C - O$;
CIT	The corporate income tax, where $CIT = (S - C - O)R$;
NP	The net profit, where $NP = (1 - R)(S - C - O)$;
Y	The overall corporate income tax of both the seller and the buyer;
W	The overall net profit of both the seller and the buyer.

To distinguish symbolically between alternative trading parties of a single notation, we use a subscript k , where $k = 1$ for the seller of an AT and $k = 2$ for the buyer of the AT. For example, S_1 denotes the sales revenue of the seller; and O_2 denotes the operating expense of the buyer.

The overall CIT of two AT companies is measured by the sum of the CITs that are levied on both the seller and the buyer.

$$Y = \sum_k CIT_k = \sum_k (S_k - C_k - O_k)R_k, \text{ where } k = 1, 2 \quad (1)$$

Similarly, the overall NP of two AT companies is calculated as the sum of the NPs that are held by both the seller and the buyer.

$$W = \sum_k NP_k = \sum_k (1 - R_k)(S_k - C_k - O_k), \text{ where } k = 1, 2 \quad (2)$$

We present the following theorem and proof:

Theorem 1. *The property premise for realizing ATTE is the tax rate differential that must be satisfied by AT companies.*

Proof. For an AT, C_1 , O_1 , S_2 and O_2 are influenced by the market but keep pace with ALCS and can be viewed as constants. In addition, R_1 and R_2 are dominated by the government and can be viewed as specified values.

Suppose the amount of an AT is denoted by A_{AT} ; S_1 and C_2 are represented as follows:

$$S_1 = A_{AT} + S_{ext1}, \text{ where } S_{ext1} \text{ is the extra sales revenue of the seller.} \quad (3)$$

$$C_2 = A_{AT} + C_{ext2}, \text{ where } C_{ext2} \text{ is the extra cost of goods sold (CoGS) of the buyer.} \quad (4)$$

S_{ext1} and C_{ext2} are constants, namely, they do not depend on A_{AT} .

Based on formulas 3 and 4, the overall CIT and NP are reformulated as follows:

$$\begin{aligned} Y &= (A_{AT} + S_{ext1} - C_1 - O_1)R_1 + (S_2 - A_{AT} - C_{ext2} - O_2)R_2 \\ &= (R_1 - R_2)A_{AT} + (S_{ext1} - C_1 - O_1)R_1 + (S_2 - C_{ext2} - O_2)R_2 \end{aligned} \quad (5)$$

$$\begin{aligned} W &= (1 - R_1)(A_{AT} + S_{ext1} - C_1 - O_1) + (1 - R_2)(S_2 - A_{AT} - C_{ext2} - O_2) \\ &= (R_2 - R_1)A_{AT} + (1 - R_1)(S_{ext1} - C_1 - O_1) + (1 - R_2)(S_2 - C_{ext2} - O_2) \end{aligned} \quad (6)$$

Because C_1 , O_1 , S_2 , O_2 , S_{ext1} and C_{ext2} are constants, Y and W change only as functions of A_{AT} if R_1 and R_2 are specified. An adjusting parameter, namely, δ ($|\delta| < A_{AT}$), is added to A_{AT} as the variation. Then, the adjusted result (denoted by A'_{AT}) is represented as follows:

$$A'_{AT} = A_{AT} + \delta \quad (7)$$

Moreover, the overall CIT after adjustment (denoted by Y') is represented as follows:

$$Y' = Y(A'_{AT}) = (R_1 - R_2)(A_{AT} + \delta) + (S_{ext1} - C_1 - O_1)R_1 + (S_2 - C_{ext2} - O_2)R_2 \quad (8)$$

The fluctuation of the pre- and post-overall CIT (denoted by ΔY) is defined as:

$$\Delta Y = Y - Y' = (R_1 - R_2)\delta \quad (9)$$

Similarly, the fluctuation of the pre- and post-overall NP (denoted by ΔW) is formulated as:

$$\Delta W = (R_2 - R_1)\delta \quad (10)$$

From formulas 9 and 10, we discover that the variation trend of the overall CIT is opposite that of the overall NP over the amount of an AT. Thus, AT companies can achieve a reduction in the overall CIT and an increase in the overall NP simultaneously, which are denoted as $\Delta W > 0$ and $\Delta Y < 0$. To reach this goal, the conditions that must be satisfied for the variables that are involved in ΔY and ΔW are discussed as follows:

(1) When $R_1 < R_2$, let $\delta > 0$ (i.e., increase the amount of an AT) such that $\Delta W > 0$ and $\Delta Y < 0$. The corresponding scenario is that a low-tax company sells products to its high-tax AT company at a price that exceeds the market level (e.g., Case 1 in Section 4.1).

(2) When $R_2 < R_1$, let $\delta < 0$ (i.e., slash the amount of an AT) such that $\Delta W > 0$ and $\Delta Y < 0$. The corresponding scenario is that a high-tax company sells products to its low-tax AT company at a price that is lower than the market level (e.g., Case 2 in Section 4.1).

(3) When $R_2 = R_1$, regardless of the value of δ , $\Delta W = 0$ and $\Delta Y = 0$ always hold, i.e., the overall NP and CIT remain unchanged regardless of the amount of an AT.

Based on the derivation, we consider the tax rate differential of AT companies as the property premise of ATTE, based on which a reduction in the overall CIT and an increase in the overall NP are achieved by adjusting the amount of an AT. $\square$

Theorem 2. *When two AT companies carry out ATTE behaviors, their TBRs will be highly abnormal.*

Proof. TBR, which is defined as the ratio of CIT to the company's sales revenue, is frequently used as a measure of a company's operating status. Then, TBRs of the seller (TBR_1) and the buyer (TBR_2) are denoted as follows:

$$TBR_k = CIT_k/S_k = R_k(S_k - C_k - O_k)/S_k, \text{ where } k = 1, 2 \quad (11)$$

Combining formula 11 with formula 3, TBR_1 and TBR_2 can be further represented as:

$$TBR_1 = R_1[1 - (C_1 + O_1)/(A_{AT} + S_{ext1})] \quad (12)$$

$$TBR_2 = R_2[1 - (A_{AT} + C_{ext2} + O_2)/S_2] \quad (13)$$

When the adjusting parameter δ is introduced into A_{AT} , the fluctuations of the pre- and post-TBR of the seller and the buyer (denoted by ΔTBR_1 and ΔTBR_2) are represented as follows:

$$\Delta TBR_1 = R_1\delta(C_1 + O_1)/[(A_{AT} + S_{ext1})(A_{AT} + \delta + S_{ext1})] \quad (14)$$

$$\Delta TBR_2 = -R_2\delta/S_2 \quad (15)$$

If AT companies evade taxes by retaining profits, namely, they satisfy the condition that $R_1 < R_2$ and $\delta > 0$, it is obtained that $\Delta TBR_1 > 0$ and $\Delta TBR_2 < 0$ by combining with formulas 14 and 15, which means an increase in TBR_1 and a reduction in TBR_2 . Similarly, if AT companies evade taxes by transferring profits, namely, $R_1 > R_2$ and $\delta < 0$ ($|\delta| < A_{AT}$), it is obtained that $\Delta TBR_1 < 0$ and $\Delta TBR_2 > 0$ by combining with formulas 14 and 15, which means a reduction in TBR_1 and an increase in TBR_2 .

Combining with the average TBR (μ_{TBR_k}) by reference to ALCS, the TBR difference ratio (denoted by DR_{TBR_k}) between each of the AT companies and its corresponding ALCS is calculated as

$$DR_{TBR_k} = (TBR_k - \mu_{TBR_k})/\mu_{TBR_k}, \text{ where } k = 1, 2 \quad (16)$$

Suppose that Th ($Th > 0$) denotes the threshold of the TBR difference ratio. $DR_{TBR_1} > Th$ and $DR_{TBR_2} < -Th$ are satisfied for retaining profits, and $DR_{TBR_1} < -Th$ and $DR_{TBR_2} > Th$ are satisfied for transferring profits.

In summary, when two AT companies carry out ATTE behaviors (retaining profits or transferring profits), their TBRs exhibit abnormalities: (i) an abnormally low TBR in the seller and an abnormally high TBR in the buyer for retaining profits or (ii) an abnormally high TBR in the seller and an abnormally low TBR in the buyer for transferring profits. In general, ATTE behaviors result in an abnormally high TBR in the low-tax-rate AT company and an abnormally low TBR in the high-tax-rate company. $\square$

5. Identification of ATTE Suspicious Groups

In this section, we propose the 3TI method for detecting ATTE and identifying suspicious groups in a hybrid manner based on both structural characteristics and BC-ATTEM, which is inspired by the two premises (the topological premise and the property premise) and the consequence (abnormalities in TBR) of ATTE that are stated in Section 4. 3TI consists of three steps: tax rate differential detection, topological pattern matching and tax burden abnormality identification. First, we face a couple of challenges in topological pattern matching: (i) it is impractical to directly search for ATTE behaviors that exhibit the topological patterns in a large-scale APIRN. Thus, the low efficiency of topological pattern matching is a challenge, which is called the *LSA problem* in short, and (ii) a static IRR trail, as the main component of the topological patterns, is constructed by combining static IRR edges, which causes an explosion, which is known as the *SITCE problem* in short. To overcome these two challenges, we propose two algorithms (the *subAPIRN Partition* algorithm and the *STRL Construction* algorithm) in conjunction with Spark platform parallel processing, that make it possible to handle the large-scale topological pattern matching problem in an efficient way. Then, we introduce the main process of 3TI.

5.1. Interest Community Partition

To overcome the *LSA problem*, a straightforward scheme is designed. First, we scale down the network size of an APIRN by applying filters on its static IRRs based on a weight threshold setting because interest transfer in ATTE occurs with sufficiently influential interactions between APs. Then, we partition the scaled-down APIRN into a series of “interest communities”

by applying a divide-and-conquer strategy. The major difficulty with the partition is to achieve execution independence of the follow-up tasks in each interest community while ensuring the consistency of task results compared with those without a partition.

Thus, it is crucial to determine partition criteria for APIRN. Inspired by the observations of the topological patterns, which are described as two static IRR trails with the same bidirectional edge or antecedent node that are behind a transaction (dynamic IRR), we identify two features: (i) two static IRR trails that have the same antecedent node or bidirectional edge must be involved in a single subSAPIRN and (ii) it is impossible for a dynamic IRR that connects two subSAPIRN to be involved in the topology of an ATTE behavior. Based on these features, we identify the partition criteria, which consist of a static partition criterion and dynamic partition criterion.

- **Static partition criterion:** We define subSAPIRN as the basic unit for the partition of the SAPIRN of an APIRN, which is treated as the static part of an interest community.
- **Dynamic partition criterion:** For edges in the DAPIRN of an APIRN, we ignore the between-subSAPIRN dynamic IRR edges and group the ones for which both vertices are involved in the same subSAPIRN as a subDAPIRN, which is treated as the dynamic part of the corresponding interest community.

Based on the partition criteria, we propose an interest community partitioning algorithm, which consists of two phases: subSAPIRN partitioning and subDAPIRN partitioning. By carrying out the two phases, the SAPIRN and DAPIRN of an APIRN are segmented into series of subSAPIRNs and subDAPIRNs. Each pair of subSAPIRN and subDAPIRN groups generates an interest community, which is denoted as subAPIRN.

First phase: subSAPIRN partitioning

According to the static partition criterion, we introduce a parallel subSAPIRN partitioning algorithm for traversing all subSAPIRNs in the SAPIRN of a specified APIRN, as stated in Algorithm 1.

Suppose that each node in the SAPIRN carries a community label (*CLabel*) that denotes the community to which it belongs. First, initialize each node with a unique identification label (*IdLabel*) and initialize the *CLabel* of each node by its *IdLabel* (**initialization process**). Second, let the *CLabel* of

each node propagate to its neighbors throughout the SAPIRN (**propagation process**). Third, each node updates its *CLabel* based on its current *CLabel* and the collected *CLabel* of its neighbors (**updating process**). The detailed updating process is that each node in the SAPIRN chooses to join the community to which the minimum *CLabel* of itself and its neighbors belong. As *CLabel* propagate and update, weakly connected groups of nodes will reach a consensus on a unique *CLabel*. The propagation and updating process is performed iteratively until no node in the SAPIRN changes its *CLabel* and the nodes that have the same *CLabel* are grouped together as a subSAPIRN.

Second phase: subDAPIRN partitioning

Then, according to the dynamic partition criterion, we introduce a parallel subDAPIRN partitioning algorithm. Dynamic IRR edges for which both vertices are involved in a subSAPIRN combine with their corresponding vertices to form a subDAPIRN. Due to space limitations, no more detail regarding the process is given here.

5.2. Static IRR Trail Library Construction

Given the existence of the SITCE problem, ATTE suspicious groups may adopt various forms of static IRR trails as control chains. It is impossible to directly identify the topologies of ATTE behaviors in real time, even within the partitioned interest communities, because of the challenge that is faced in searching such complex static IRR trails for each transaction. To address this problem, inspired by the components of the topological patterns and their characteristics, we find that dynamic IRR edges are incremental and changing, whereas the edges of a static IRR trail remain unchanged for a period of time. Thus, a feasible scheme for solving the problem is to traverse all static IRR trails, construct a static IRR trail library (SITL) for each subSAPIRN offline, and reconstruct them at regular intervals. Then, for a specified transaction, the task of identifying whether it belongs to an AT can be transformed into determining whether there exist two control chains that are behind the transaction in an SITL that has been constructed in advance.

The construction of an SITL aims at recording the control chains throughout the corresponding subSAPIRN. When a control chain is long, the two parties at the ends of the chain have a distant relationship and, therefore, have little probability of being involved in the same ATTE group. Thus, we set a maximum length limit for control chains. Then, we propose a novel parallel label propagation-based path traversal algorithm, which is carried

Algorithm 1: *subSAPIRN Partition*

Input: *SAPIRN* - formed by *S_vertSet* and *S_edgeSet*
 S_vertSet - vertex set of *SAPIRN*
 S_edgeSet - edge set of *SAPIRN*
Output: *subSAPIRN* (formed by *SS_vertSet(i)* and *SS_edgeSet(i)*), $i = 1, \dots, L$, where L is the number of *subSAPIRN* in the *SAPIRN*.
 SS_vertSet(i) - vertex set of *subSAPIRN(i)*
 SS_edgeSet(i) - edge set of *subSAPIRN(i)*

```
1 begin
2   Initialize the IdLabels at all nodes in S_vertSet by unique value  $x_1, x_2, \dots, x_N$ ,  $N$  is number of nodes in S_vertSet;
3   foreach  $i$  in 1 to  $N$  do
4     Initialize CLabel of node  $x_i$  by its Idlabel,  $C_{xi}(1) = x_i$ ;
5   Set  $t=1, F=10$ ;
6   for  $t = 1, \dots, F$  do
7     foreach  $j$  in 1 to  $N$  do
8       Node  $x_j$  propagates its CLabel  $C_{xj}(t)$  to its neighbors;
9       Updates the CLabel of node  $x_j$ , let  $C_{xj}(t+1) = \min\{C_{xj}(t), C_{xj1}(t), C_{xj2}(t), \dots, C_{xjk}(t)\}$ , where  $C_{xj1}(t), C_{xj2}(t), \dots, C_{xjk}(t)$  are the CLabels of node  $x_j$ 's neighbors collected;
10      if no node changes its CLabel by comparing  $C_{xj}(t)$  and  $C_{xj}(t+1)$  then
11        break;
12  Nodes having the same CLabels are grouped together as SS_vertSet(i);
13  Edge with both nodes involved in SS_vertSet(i) are grouped together as SS_edgeSet(i);
14  SS_vertSet(i) and SS_edgeSet(i) form subSAPIRN(i);
15  return subSAPIRN(i);
```

out in each subSAPIRN to obtain its SITL, as described in Algorithm 2. The main steps of the algorithm are as follows:

Step 1: Initialization process

Suppose that each node in a subSAPIRN carries a local trail set $LTrailSet$, which stores static IRR trails (a sequence of nodes) that end with the node itself. The length of a static IRR trail is defined as the number of nodes in the sequence minus one, and its maximum length limit is defined as L . We append the unique $IdLabel$ of each node (one node as a trail) to its $LTrailSet$ for initialization, define a global trail set $GTrailSet$ for storing all static IRR trails in the subSAPIRN and initialize it to the null set.

Step 2: Propagation process

Let the $LTrailSet$ of each node propagate to its neighbors by tracing the directions of its adjacent edges. Then, we remove the nodes that have not received any $LTrailSet$ from neighbors and break the ties to their neighbors. Next, we count the number of nodes in the subSAPIRN and the current maximum trail length. If the number of nodes is equal to zero or the maximum length has reached the limit, then the algorithm is terminated and $GTrailSet$ is the SITL that records all static IRR trails throughout this subSAPIRN; otherwise, we continue.

Step 3: Updating process

Each node updates its $LTrailSet$ based on the $LTrailSet$ that was collected from its neighbors. The detailed update operation on each node consists of the following steps:

Step 3.1: pop its $LTrailSet$ to the $GTrailSet$;

Step 3.2: merge the collected $LTrailSets$ and remove the component trails that contain the $IdLabel$ of this node; and

Step 3.3: append the $IdLabel$ to the remaining component trails and update the $LTrailSet$ of this node by the set of the new trails that were obtained;

Step 2 and step 3 are performed iteratively until the termination condition is satisfied. Then, all control chains throughout each subSAPIRN are recorded in the corresponding SITL.

5.3. Identification of Suspicious Groups

Inspired by the two premises and the consequence of ATTE, which are stated in Section 4, we propose the 3TI method, which is based on the partitioned interest communities and the constructed SITLs, as shown in Fig. 5.

Algorithm 2: STRL Construction

Input: *subSAPIRN* (formed by *SS_vertSet(i)* and *SS_edgeSet(i)*), $i = 1, \dots, L$, where L is the number of *subSAPIRN* in the *SAPIRN*.

SS_vertSet(i) - vertex set of *subSAPIRN(i)*, N_i is number of nodes in *SS_vertSet(i)*

SA_edgeSet(i) - edge set of *subSAPIRN(i)*

Output: *GTrailSet(i)* (formed by static relationship trails, each of which is a sequence of *IdLabels* of nodes), $i = 1, \dots, L$.

```
1 begin
2   foreach  $i$  in 1 to  $L$  do
3     foreach  $j$  in 1 to  $N_i$  do
4       Initialize the LTrailSet of  $j$ -th node in SA_vertSet(i) by its
         IdLabel  $x_{i,j}$ ,  $LTrailSet(i, j) = \{(x_{i,j})\}$ ;
5   Set  $GTrailSet(i) = \emptyset$ ;
6   foreach  $i$  in 1 to  $L$  do
7     foreach  $j$  in 1 to  $N_i$  do
8       Node  $x_{i,j}$  propagates its LTrailSet(i, j) to its neighbors by
         tracing the directions of its adjacent edges;
9       Remove the nodes which have not received any LTrailSet from
         neighbors and break the ties to its neighbors;
10      Count the number of nodes in SS_vertSet(i) by num;
11      if  $num = 0$  or maximum of trail lengths up to the limit then
12        break;
13      Pop LTrailSet(i, j) to  $GTrailSet(i)$  ;
14      Node  $x_{i,j}$  merges the LTrailSets collected from neighbors,
          $Clt\_LTrailSet(i, j) = \cup LTrailSet(i, j_k), j_k \in K$ ,  $K$  is the set of
         node  $j_k$ 's neighbors;
15      foreach component trail  $str$  in  $Clt\_LTrailSet(i, j)$  do
16        if  $x_{i,j}$  in  $str$  then
17          Remove  $str$  from  $Clt\_LTrailSet(i, j)$  ;
18        else
19          Append  $x_{i,j}$  to  $str$ ;
20  return  $GTrailSet(i)$ ;
```

Figure 5: General view of 3TI

The main steps of the method are as follows, as described in Algorithm 3:

Step 1: Tax rate differential detection

For each dynamic IRR edge dre in $\text{subDAPIRN}(i)$ (where $i = 1, \dots, L$), compare the tax rates between dre 's source vertex src and destination vertex dst . If src and dst differ in terms of tax rate, dre is selected since it satisfies the property premise of ATTE. Otherwise, dre is ignored.

Moreover, for an additional dynamic IRR edge that corresponds to a new transaction, the process of tax rate differential detection is the same as that stated above.

Step 2: Topological pattern matching

For each dre that was selected in Step 1 for $\text{subDAPIRN}(i)$ (where $i = 1, \dots, L$), extract the $LTrailSets$ of dre 's source vertex src and destination vertex dst from the SITL of $\text{subSAPIRN}(i)$. If there exist $trails \in LTrailSet(src)$ and $trail_d \in LTrailSet(dst)$ such that the combination of $trail_s$, $trail_d$ and dre exhibits the topological pattern $SBE-BT$ or $SAN-BT$, then dre satisfies the topological premise of ATTE. That is, dre belongs to an AT and $trail_s$ and $trail_d$ serve as the control chains of the AT. Meanwhile, the companies that are involved in the dre and the two trails form a group of ATs.

If dre is an additional dynamic IRR edge, then determine whether both of its vertices are involved in a subSAPIRN before the process of topological pattern matching is carried out.

Step 3: Tax burden abnormality identification

For each *dre* that was selected in Step 2 for subDAPIRN(*i*) ($i = 1, \dots, L$), calculate the TBR of its two companies by measuring the percentages of income that they pay in CIT (see formulas 14 and 15). Then, calculate the difference ratio in TBR between each of *dre*'s companies and its corresponding ALCS (see formula 16). Herein, we denote the difference ratio of the *src* as D_s and that of *dst* as D_d . Finally, we compare the difference ratios with the preset thresholds to identify the companies that have abnormal TBRs and consider the group of ATs to which it belongs as an ATTE suspicious group. We judge an unknown company in *dre* as follows:

- If $R_s < R_d$, $D_s > Th$ and $D_d < -Th$, the company is an evader.
- If $R_s > R_d$, $D_s < -Th$ and $D_d > Th$, the company is an evader.
- If $R_s < R_d$, $D_s < Th$ or $D_d > -Th$, the company is normal.
- If $R_s > R_d$, $D_s > -Th$ or $D_d < Th$, the company is normal.

Th is the threshold set for judging ATTE based on tax burden abnormality identification.

6. Performance Evaluation

To evaluate the performance of 3TI, we first introduce our experimental design. Then, we investigate the following three research questions:

RQ 1: *What are the optimal values of the parameters in 3TI with respect to the detection accuracy?* For this purpose, we adjust the values of the parameters and evaluate their effects on the performance of 3TI.

RQ 2: *Can 3TI classify a new company into its correct category (ATTE or non-ATTE) with high accuracy?* For this purpose, we evaluate 3TI on a large real-world dataset and compare it in terms of accuracy with two baseline approaches [39, 16] across various industry sectors.

RQ 3: *How can the reliability and generality of 3TI be evaluated?* For this purpose, we performed tests in scenarios of incomplete data and various jurisdictions.

Algorithm 3: *Suspicious Group Identification*

Input: $subDAPIRN(i)$ (formed by $SD_vertSet(i)$ and $SD_edgeSet(i)$);
 $SD_vertSet(i)$ - vertex set of $subDAPIRN(i)$;
 $SD_edgeSet(i)$ - edge set of $subDAPIRN(i)$;
 $GTrailSet(i)$, STRL of $subSAPIRN(i)$;
 $i = 1, \dots, L$, L is the number of $subAPIRN$ in the $APIRN$;
 N_i is the number of edges in $SD_edgeSet(i)$;

Output: $susGroup$ table (composed of a series of components, $susGroup(i)$, each saves all the suspicious group records mined from the i -th $subSAPIRN$. Each record is composed of two static relationship trails, $srt1$ and $str2$, and one dynamic relationship edge, dre .);

```

1 begin
2   foreach  $i$  in 1 to  $L$  do
3     foreach  $j$  in 1 to  $N_i$  do
4        $dre = SD\_edgeSet(i, j)$ ;
5        $src = \text{source vertex}(dre), dst = \text{destination vertex}(dre)$ ;
6        $R_1 = \text{taxRate}(src), R_2 = \text{taxRate}(dst)$ ;
7       if  $R_1 \neq R_2$  then
8         if there exist  $trail_s \in LTrailSet(src)$  and  $trail_d \in$ 
9            $LTrailSet(dst)$  such that the combination of  $trail_s, trail_d$ 
10            and  $dre'$  meets the topological pattern  $SBE-BT$  or  $SAN-BT$ 
11            then
12              Calculate the tax burden radios of  $src$  and  $dst$ , result by
13                 $Ratio_1$  and  $Ratio_2$ ;
14              Calculate the difference in tax burden radios between
15                 $src/dst$  and the corresponding  $ALCS$ , result by  $M_1$  and
16                 $M_2$ ;
17              Select the  $dre''$  with its  $M_1$  and  $M_2$  exceeding the the
18                pre-set thresholds and save the combination of  $dre''$  and
19                its corresponding  $trail_s$  and  $trail_d$  to  $susGroup(i)$  ;
20              // see Section 4.2.2
21         Append  $susGroup(i)$  to  $susGroup$ ;
22   return  $susGroup$ ;

```

6.1. Dataset and Metrics

Dataset and Preprocessing. To evaluate the performance of 3TI, we used a set of confirmed ATTE companies and a set of known non-ATTE companies to build the samples of our dataset. To ensure the integrity of the constructed dataset, we selected companies from various industry sectors. To perform our experiments, from jurisdiction S, we collected 45,960 labeled companies that fall into various industry sectors, including retail (28.51%), wholesale (40.78%), manufacturing (12.15%), service (10.52%), construction (3.26%), insurance (2.09%), finance (0.46%), agriculture (0.65%), mining (0.42%), real estate (0.79%) and transport (0.37%). Among these companies, there are a total of 1,560 ATTE companies (positive samples) and 44,400 non-ATTE companies (negative samples). Because the actual proportion of ATTE companies in China is approximately 3.4%, we tried to make our dataset as close to reality as possible. The distribution of ATTE conditions among the sectors is shown in Fig. 6.

The information on these companies and their affiliated companies and persons that was used in our experiment was provided by the State Administration of Taxation, China. Table A.6 (in Appendix A) lists the details of the data items that were used in our experiments. Based on the collected data, an APIRN with 1,897,235 vertices (containing 845,573 taxpayers and 1,051,662 persons) and 6,274,030 edges (containing 72,104 kinship, 185,681 interlocking shareholding, 1,304,193 shareholding, 845,573 legal representative and 3,866,479 trading relations) was constructed for follow-up 3TI-based ATTE detection.

Evaluation Technique. The dataset is divided into two parts, namely, a 90% training set (containing 41,364 samples) and a 10% testing set (containing 4,596 samples), via stratified and randomly splitting. Stratified splitting aims at splitting the dataset so that each split is similar in terms of class and sector distribution. A classification setting is chosen to ensure that the training set and the testing set have approximately the same percentage of samples of the target class (either ATTE or non-ATTE) as the whole dataset. Stratified splitting is also carried out to equally distribute the sector features to ensure the same sector distribution in both the training set and the testing set.

Considering class imbalance in the dataset, a nested ten-fold cross-validation strategy is applied for parameter tuning and performance evaluation. A 10-fold inner cross-validation loop is used to tune the parameters (the weight threshold for static IRR edges, the maximum length limit for static IRR

Figure 6: The distribution of ATTE conditions for different sectors

trails and the threshold for abnormal TBR) and select the optimal detection model. Then, a 10-fold outer cross-validation loop is used to evaluate the detection model that was selected by the inner cross-validation. Specifically, we separate the entire dataset into 10 folds for outer cross-validation, each of which consists of 90% training data and 10% testing data. For each fold, the training set is again separated into 10 subfolds and inner cross-validation and grid search are used on the subfolds to determine the best parameters; second, we test the model performance on the testing set with the selected parameters. This process is repeated 10 times in the outer cross-validation loop to avoid the bias that is introduced by randomly partitioning the dataset.

Evaluation Metrics. We consider the metrics of *fraud detection accuracy* [23] to represent the ability to identify whether a taxpayer is involved in a group of ATTE or not, as listed in Table 3.

Herein, FPR is an indicator that is used to measure the probability of falsely classifying a non-ATTE company as a suspicious one; FNR is an indicator that is used to measure the probability of falsely classifying an ATTE company as a normal one. Precision is the fraction of the companies that are identified as suspicious and are truly involved in ATTE. Recall is the fraction of companies that truly carry out ATTE behaviors and are

subsequently successfully identified. F-measure is used to measure the ATTE detection accuracy using precision and recall jointly.

Table 3: Descriptions of the used metrics

<i>Term</i>	<i>Abbr</i>	<i>Definition</i>
True Positive	TP	# of companies with ATTE behaviors are correctly classified into the category <i>ATTE</i>
True Negative	TN	# of companies without ATTE behaviors are correctly classified into the category <i>non-ATTE</i>
False Negative	FN	# of companies with ATTE behaviors are incorrectly classified into the category <i>non-ATTE</i>
False Positive	FP	# of companies without ATTE behaviors are incorrectly classified into the category <i>ATTE</i>
False Positive Rate	FPR	$FP/(FP+TN)$
False Negative Rate	FNR	$FN/(FN+TP)$
Precision	p	$TP/(TP+FP)$
Recall	r	$TP/(TP+FN)$
F-measure	F_1	$2pr/(p+r)$

6.2. Sensitivity Analysis and Parameter Determination

The topological patterns of ATTE behaviors in 3TI are formed by two static IRR trails and one dynamic IRR edge. In 3TI, the construction of static IRR trails is affected by two parameters: (i) Th_1 : the weight threshold for static IRR edges and (ii) L : the maximum length limit for static IRR trails.

If the weight of a static IRR edge is high, the relationship that is denoted by the edge is tighter and, thus, more significant. We define the threshold of static IRR edge filters as Th_1 . Exceeding this threshold indicates that the filtered relationship is regarded as sufficiently influential for interest transfer in ATTE. Setting Th_1 to too small a value will cause some normal companies that have loose relationships with the ATTE companies to exhibit ATTE behavioral patterns, which could cause false positives. However, if we select a large value of Th_1 , some affiliated relationships in ATTE groups will not satisfy the detection criteria, which will yield false-negative results.

If a static IRR trail is long, the two parties at the ends of the trail have a distant relationship and, thus, low probability of being involved in the same ATTE group. We define the maximum length limit for static IRR trails as L . If a trail is within this limit, the parties that are connected by the trail are regarded as sufficiently close for interest transfer in ATTE. Similarly, setting a large value of L could lead to false positives and false negatives could be caused by setting a small value of L .

Each of the groups that exhibit the topological patterns is detected via BC-ATTEM judgement, which measures the relative difference between the TBR of the group’s trading company and the corresponding ALCS. If the relative difference is significant, the transaction is more unique and, thus, more suspicious. We define the threshold of suspicious transaction matches as Th_2 . Groups that exceed this threshold are suspected of performing ATTE. Likewise, false positives and false negatives will occur if we set an improper value of Th_2 .

In conclusion, different parameter settings of Th_1 , L and Th_2 result in accuracy variations in ATTE detection. Therefore, it is important to quantify the impact of the parameter settings on the detection accuracy and adjust the parameter settings to achieve the best detection accuracy. We refer to quantifying the impact of the parameter settings on the detection accuracy as a parameter sensitivity analysis (SA) process and define it as the process of comparing the results of multiple analyses of a dataset with various parameter settings and quantifying the differences in terms of the detection accuracy. Adjusting parameter settings is an extension of SA and we refer to it as a parameter tuning process, in which grid search with 10-fold cross-validation on a training set was conducted to identify the optimal settings of these parameters to achieve the best detection accuracy.

Based on previous descriptions, we define the ranges of Th_1 and Th_2 as $[0, 1]$ and $[0, 100]$, respectively, and the value of L ranges between 1 and 5 (see Table B.7 in Appendix B). All of these parameters are varied simultaneously over the entire parameter space. Therefore, the Sobol global sensitivity analysis approach is employed to quantify the model performance sensitivities with respect to the parameter variation, which enables the simultaneously evaluation of the relative contributions of each individual parameter and the interactions between parameters to the performance variance. To understand how the performance variance can be attributed to individual parameter variables and the interactions between the parameters, first-order sensitivity indices are calculated to accurately reflect the influences of the

individual parameters and total-order sensitivity indices are used to evaluate the overall contribution of a parameter and the interactions with other parameters. The higher the values of the sensitivity indices, the more influential the corresponding model parameters are.

Fig. 7 shows the result of the sensitivity analysis over the full range of the parameter space. Th_1 , L and Th_2 are all critical parameters to the model performance. Parameter Th_1 is the most important parameter, which contributes approximately 60% of the model performance variability, followed by parameters Th_2 and L . For L , if only the first-order sensitivity index is considered, it is not as critical as when the total-order sensitivity index is considered. This suggests that parameter L interacts strongly with other parameters.

Figure 7: Sobol global sensitivity analysis over parameters

Hence, it is concluded that the performance of 3TI is sensitive to these three parameters and they must be tuned to reach optimal performance. To identify the optimal parameter settings and prevent overfitting, we employed a standard grid-search on the training set with ten-fold cross-validation to optimize the parameters for the detection on a separate testing set.

6.3. Effects of Parameters

In this subsection, we discuss experiments for exploring the effects of three parameters (Th_1 , L and Th_2) on the ATTE detection accuracy and attempt to analyze the reasons. Based on nested ten-fold cross-validation on the whole dataset, we evaluate changes in the model performance with respect to variations in a single parameter and set other parameters according to the optimal setting portfolios so that the effect of that single parameter can be determined. The following experimental results demonstrate the influence of the diversity of the static IRR trail-construction parameters (Th_1 and L) and the variation of the threshold for BC-ATTEM judgement (Th_2) on the evaluation metrics.

Effects of the Weight Threshold. Fig. 8 shows the effects of Th_1 on the evaluation metrics of 3TI. The effects of Th_1 on 3TI performance exhibit substantial differences among various values of Th_1 . During the increase of Th_1 (from 0 to 1), the F-measure increases sharply when Th_1 changes from 0 to 0.2 and decreases smoothly when Th_1 increases from 0.2 to 1, as shown in Fig. 8 (a). The reason is that as Th_1 increases, many weak IRRs are filtered and fewer static IRR trails will be involved in the constructed SITL; hence, less evidence can be used to detect ATTE behaviors. This leads to a decrease in recall and a corresponding increase in FNR, as shown in Fig. 8 (a) and 8 (b). Meanwhile, we found that the precision increases when Th_1 changes from 0 to 0.2 and remains relatively stable when Th_1 becomes larger than 0.2, as shown in Fig. 8 (a). We analyzed the reason and found that as Th_1 increases, normal companies that are falsely regarded as suspicious were effectively filtered when we set $Th_1 = 0.2$ or even larger, as shown in Fig. 8 (c). Through the analysis, we discovered the first implication for ATTE inspection: **tax evaders tend to utilize relatively tightly controlled companies for transferring profits.**

Effects of the Maximum Length. Fig. 9 (a) shows that the F-measure increases with increasing L , especially when L changes from 1 to 2; the increase is not significant when L changes from 2 to 5. The main reason is that there are few additional static IRR trails when L becomes larger than 2. Theoretically, if the maximum length is larger, there could be more static IRR trails that are involved in an SITL. However, in practice, although there exist static IRR trails exist with large lengths that can be used as control chains for ATTE behaviors, the number of trails of this kind is small, as shown in Fig. 9 (b) and 9 (c). Worse, larger L will introduce more false positives, as shown in Fig. 9 (c). Thus, we obtain the second implication

Figure 8: Effects of Th_1 on evaluation metrics

for ATTE inspection: **tax evaders tend to utilize short control chains that are no longer than 2 to transfer profits for tax evasion.**

Figure 9: Effects of L on evaluation metrics

Effects of the Early-Warning Threshold. Fig. 10 shows the effects of Th_2 on the performance of 3TI. When Th_2 is sufficiently large, the accuracy begins to decrease. For example, when Th_2 is larger than 0.5, F-measure is decreased in Fig. 10 (a) and FNR is increased in Fig. 10 (b). We analyze the reason and find that as Th_2 increases, the judgement will be looser and fewer companies will be detected as suspicious, as shown in Fig. 10 (c). That leads to a decrease in recall and an increase in FNR. With the increase in Th_2 (from 0.01 to 0.5), the number of true positives remains relatively stable, while the number of false positives decreases sharply. This means that the companies that have a relative difference of TBR that is between 0.01 and 0.5 are typically normal. Therefore, we conclude that Th_2 cannot be set too small or too large.

Figure 10: Effects of Th_2 on evaluation metrics

6.4. Comparison with Existing Work

In this section, we compare 3TI with other ATTE detection methods in terms of performance to evaluate its effectiveness and generality.

At present, there are many ATTE detection methods that are based on machine learning and graph theory. We compare 3TI with a mainstream machine-learning-based detection method (MetaCost [28]) that derives ATTE patterns from attributes of taxpayer individuals and a graph-based detection method (TPIIN [39]) that is based on collecting sophisticated organizational structures of ATTE groups. The integrated detection accuracy rates of the three methods are listed in Table 4. 3TI outperforms the other two methods. This demonstrates its effectiveness with respect to detection accuracy.

Since the time cost is critical in tax data analysis, the computational complexity of algorithms cannot be ignored. We compare 3TI with state-of-the-art approaches [39, 28] in terms of complexity. The computational complexity of the graph-based tax evasion group detection algorithm in [39] is $O(V^3)$, where V is the number of vertices and that of the individual tax evasion detection algorithm in [28] is $O(n2^{n(n-1)})$, where n stands for the number of taxpayers. 3TI considers both the interaction between taxpayers and the abnormal financial statuses of individuals with the complexity of $O(V + E)$, where V and E represent the numbers of vertices and edges, respectively. Our method substantially outperforms those in [39, 28].

Moreover, 3TI applies to any sector of a company, in contrast to existing ATTE detection methods that are applicable only for detecting companies of specified industry sectors. Table 5 presents the comparison results of 3TI with the above two methods on detecting ATTE with various types

of companies. It shows the average accuracy that was obtained via nested cross-validation. The settings are the optimal hyperparameters, which were identified via inner cross-validation grid search. According to Table 5, compared with the other two detection methods, 3TI performs better in all industry sectors. The F-measures of MetaCost vary widely across industry sectors and TPIIN suffers from a high false-positive rate. 3TI outperforms the other two methods and always exhibits outstanding detection accuracy. The main reason is based on two aspects: First, the ATTE behavioral patterns vary across industry sectors, which cannot be well generalized by using a uniform machine-learning-based model. In 3TI, the ATTE behavioral pattern is divided into the structural characteristics and the business characteristics. The former can be abstracted as uniform descriptions and the latter are modeled separately in each industry sector. Second, the pattern of ATTE behaviors in TPIIN is just a topological pattern that is based on which of the identification results are ATs. Due to the introduction of BC-ATTEM in 3TI, the portion of AT companies that exhibit no tax evasion behaviors can be correctly identified.

In summary, 3TI is a novel and hybrid ATTE detection approach. It achieves high performance on the risk judgement of unknown companies. In addition, it is a generic and effective method that is suitable for detecting ATTE for almost all industry sectors.

Table 4: Comparison of integrated detection accuracy				
Approaches		MetaCost [28]	TPIIN [39]	3TI
Integrated detection accuracy	F-measure	60.19%	13.58%	83.31%
	FPR	2.22%	32.40%	0.33%
	FNR	29.74%	25.51%	21.92%

6.5. Assessment of Model Reliability and Generality

To evaluate the reliability and generality of 3TI, we performed tests in scenarios of incomplete data and various jurisdictions.

Detection reliability assessment. In this test, we focus on comparing the detection accuracies in the presence of incomplete declared data over various relationships. To estimate the effect of incomplete data, we intentionally remove various proportions of relations from the complete dataset. For each combination of relation types, we applied our detection model to the updated

Table 5: Comparison of F-measures of ATTE detection w.r.t. different company sectors

Industry sector	MetaCost [28]	TPIIN [39]	3TI
Retail	74.54%	16.23%	92.27%
Wholesale	66.49%	8.54%	83.88%
Manufacturing	72.60%	13.46%	83.81%
Service	61.23%	9.87%	81.02%
Construction	46.85%	17.90%	68.13%
Insurance	42.71%	27.77%	85.56%
Finance	50.67%	37.87%	59.83%
Agriculture	49.70%	67.26%	80.77%
Mining	20.74%	57.45%	73.97%
Real Estate	7.69%	22.70%	62.22%
Transport	3.70%	10.10%	63.16%

datasets with various degrees of missing-data perturbation (from 0 to 100%) to explore the effect of the annotation rate on the detection performance.

Fig. 11 shows the detection results for all the combinations of missing relation types and various degrees of missing-data perturbation. To facilitate presentation, K, I, L and S denote kinship, interlocking shareholding relationship, legal representative relationship and shareholding relationship, respectively, in Fig 11 (b), 11 (c) and 11 (d). For each combination of missing relation types, the detection accuracy decreases when the missing rate of data annotation increases. Incomplete annotation on the shareholding relation has the largest effect on the detection accuracy. However, the detection performance is relatively stable and remains satisfactory when the degree of missing-data perturbation is controlled within a range of 10%, which demonstrates the reliability of 3TI on incomplete data.

Detection generality assessment. In the previous experiments, 3TI was applied to analyze the data from a single jurisdiction (S). We achieved satisfactory detection performance. In this test, we applied 3TI to analyze the data in another jurisdiction (H). This testing dataset has the same size and distribution. We evaluated the detection performance of 3TI across various sectors in comparison with the results in jurisdiction S to evaluate the detection generality of 3TI.

The F-measures of the detection results in various jurisdictions are shown in Fig. 12. The overall F-measure of ATTE detection in jurisdiction H is 80.69%, which is similar to that in jurisdiction S (83.31%). The detection

(a) One missing relation type

(b) Two missing relation types

(c) Three missing relation types

(d) Four missing relation types

Figure 11: Comparison of the impact of different proportions of undeclared relations for detection accuracy

accuracy remains stable in various sectors across jurisdictions, which demonstrates the efficiency and generality of 3TI.

Figure 12: Detection accuracy in different jurisdictions

6.6. Further Discussion

Based on the above evaluation results, we summarize the main advantages of 3TI as follows:

Hybrid solution: 3TI realizes the advantages of both graph-based detection and machine-learning-based detection in ATTE detection by integrating structural characteristics of ATTE groups with BC-ATTEM. The performance test demonstrates that this hybrid approach outperforms existing methods in terms of accuracy and interpretability.

Generality: 3TI can be applied to detect ATTE behaviors in various types of companies because the ATTE behavioral pattern that is used is divided into the structural characteristics and the business characteristics. The former can be abstracted as uniform descriptions and the latter are separately modeled for each sector of industry. Our performance test demonstrates that

3TI can detect ATTE for various sectors of companies with higher accuracy than existing methods.

There are several disadvantages of 3TI. We need to further improve it with regard to the following aspects:

First, we must constantly gather new interactive relationships to maintain the level of detection accuracy because new ATTE behaviors that are based on covert interactive relationships continue to emerge. 3TI relies on the detection of “licit” transactions between companies that have heterogeneous, complex and covert interactive relationships. Identifying additional interactive relationships is always necessary.

Second, 3TI only analyses the transactions between companies. We did not consider cases in which companies transfer profits to personal accounts. Hence, ATTE that is implemented using personal accounts is not reflected in our graph model, which could cause false negatives.

Third, the output of 3TI indicates whether a company belongs to an ATTE group or not based on historical data. It cannot quantify the degree of suspicion and trace the evolution of behaviors of tax evaders.

7. Implementation

By adopting the method that is proposed in Section 5, we have developed a large-scale graph mining system, namely, TaxGM, for analyzing and mining suspicious tax evasion groups. TaxGM has been deployed in several provincial taxation offices in China. TaxGM offers three main functionalities: (i) a web-based interactive visualization of APIRN (see Fig. 13), (ii) discovery of the “interest communities” of a considered company from APIRN (see Fig. 14), and (iii) mining ATTE suspicious groups that are related to the company from the extracted “Interest Communities” and representing the results of evasion detection (see Fig. 15). In Fig. 15, all the normal entities and relationships are masked while all the suspicious groups and the involved transactions and affiliated relationship trails are high lighted. The highlighted graph indicates that there is an indirect affiliated relationship between two trading parties since they have a same investor. Meanwhile, one of the trading parties is an environmental protection enterprise, which enjoys a preferential tax rate and has an abnormally high TBR. This leads us to suspect that this suspicious ATTE group utilizes affiliated relationships to transfer profits to the environmental protection enterprise to lower the group’s taxable income. Using the clues that are provided by TaxGM,

163 billion RMB (approximately 24 billion US dollars) of direct tax revenue returns were restored. TaxGM offers a technical guarantee for tax collection and management.

Figure 13: Interactive Visualization of APIRN

8. Conclusions

This paper proposed a novel and hybrid approach, namely, 3TI, for detecting ATTE suspicious groups. It applies topological pattern matching and data fusion and inference based on evidence to process heterogeneous tax data to detect ATTE groups based on both structural characteristics and BC-ATTEM. 3TI outperforms other ATTE detection methods in terms of detection accuracy in various sectors of industry. In addition, it provides a satisfactory explanation of ATTE behaviors.

The results highlight the presence of a small number of covert tax evaders who behave normally in trading. This suggests that evaders in ATTE groups are not adopting regular strategies for avoiding detection during tax auditing. This demonstrates the limits of traditional individual-level tax evasion detection techniques such as abnormal financial index screening and calls for the adoption of complementary methods that help shed light where individual detection cannot reach. Given the unprecedented opportunity to adopt

Figure 14: "Interest Community" Discovery

Figure 15: Identification of ATTE Suspicious Groups

real data for our study, we focused on investigating the heterogeneous, complex and covert interactions among tax evaders. We found that tax evaders tend to utilize relatively tightly controlled and closely related companies for transferring profits to evade taxes.

Regarding future work, we plan to continue our research in the following aspects: First, we will extend the types of IRRs to make this approach applicable for ATTE behaviors that transfer profits via other IRRs, such as the IRR between a financial controller and a company. This work aims at overcoming the first and second shortcomings of 3TI. Second, we aim at quantifying the degree of suspicion of detected companies to improve the accuracy of ATTE detection to overcome the third problem.

Acknowledgments

This work is sponsored 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. 2016YFB1000903, the National Science Foundation of China under Grant Nos. 61502379, 61532015, 61672410 and 61721002, and Academy of Finland (Grant No. 308087).

References

- [1] Naoki Abe, Prem Melville, Cezar Pendus, Chandan K Reddy, David L Jensen, Vince P Thomas, James J Bennett, Gary F Anderson, Brent R Cooley, Melissa Kowalczyk, et al. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84. ACM, 2010.
- [2] Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015.
- [3] Kheder Alaghi. Operating profit, productivity indexes and net earnings. *African Journal of Business Management*, 6(11):4397, 2012.
- [4] Sandra Álvarez, Nieves R Brisaboa, Susana Ladra, and Óscar Pedreira. A compact representation of graph databases. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 18–25. ACM, 2010.

- [5] Luis Antunes, João Balsa, and Helder Coelho. Agents that collude to evade taxes. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, page 212. ACM, 2007.
- [6] Zhenisbek Assylbekov, Igor Melnykov, Rustam Bekishev, Assel Baltabayeva, Dariya Bissengaliyeva, and Eldar Mamlin. Detecting value-added tax evasion by business entities of kazakhstan. In *Intelligent Decision Technologies 2016*, pages 37–49. Springer, 2016.
- [7] Subhas Barman, Utpal Pal, Md Asif Sarfaraj, Biswajit Biswas, Animesh Mahata, and Palash Mandal. A complete literature review on financial fraud detection applying data mining techniques. *International Journal of Trust Management in Computing and Communications*, 3(4):336–359, 2016.
- [8] Alex Beutel, Leman Akoglu, and Christos Faloutsos. Fraud detection through graph-based user behavior modeling. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1696–1697. ACM, 2015.
- [9] Robert M Bowen, Andrew C Call, and Shiva Rajgopal. Whistle-blowing: Target firm characteristics and economic consequences. *The Accounting Review*, 85(4):1239–1271, 2010.
- [10] You-Shyang Chen and Ching-Hsue Cheng. A delphi-based rough sets fusion model for extracting payment rules of vehicle license tax in the government sector. *Expert Systems with Applications*, 37(3):2161–2174, 2010.
- [11] Howard R Davia, Patrick C Coggins, and John C Wideman. *Accountant’s guide to fraud detection and control*. John Wiley & Sons, 2000.
- [12] Dave DeBarr and Zach Eyler-Walker. Closing the gap: automated screening of tax returns to identify egregious tax shelters. *ACM SIGKDD Explorations Newsletter*, 8(1):11–16, 2006.
- [13] Michael P Devereux, Rachel Griffith, and Alexander Klemm. Can international tax competition explain corporate income tax reforms. *Economic Policy*, 35, 2002.

- [14] Rafał Dreżewski, Jan Sepielak, and Wojciech Filipkowski. The application of social network analysis algorithms in a system supporting money laundering detection. *Information Sciences*, 295:18–32, 2015.
- [15] Michael J Ferrantino, Xuepeng Liu, and Zhi Wang. Evasion behaviors of exporters and importers: Evidence from the us–china trade data discrepancy. *Journal of International Economics*, 86(1):141–157, 2012.
- [16] Zengan Gao. Transfer price-based money laundering: A transit trader’s perspective. In *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM’08. 4th International Conference on*, pages 1–5. IEEE, 2008.
- [17] Andrea Gebauer, Chang Woon Nam, and Rüdiger Parsche. Can reform models of value added taxation stop the vat evasion and revenue shortfalls in the eu? *Journal of Economic Policy Reform*, 10(1):1–13, 2007.
- [18] Pamela Castellón González and Juan D Velásquez. Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications*, 40(5):1427–1436, 2013.
- [19] ND Goumagias, Dimitris Hristu-Varsakelis, and A Saraidaris. A decision support model for tax revenue collection in greece. *Decision Support Systems*, 53(1):76–96, 2012.
- [20] Jane G Gravelle. Tax havens: International tax avoidance and evasion. *National Tax Journal*, pages 727–753, 2009.
- [21] Manish Gupta and Vishnuprasad Nagadevara. Audit selection strategy for improving tax compliance—application of data mining techniques. In *Foundations of Risk-Based Audits. Proceedings of the eleventh International Conference on e-Governance, Hyderabad, India, December*, pages 28–30, 2007.
- [22] Arnold C Harberger. The incidence of the corporation income tax. *Journal of Political economy*, 70(3):215–240, 1962.
- [23] Jessen L Hobson, William J Mayew, Mark Peecher, and Mohan Venkatachalam. Auditors ability to detect financial deception: The role of auditor experience and management cognitive dissonance. 2015.

- [24] Mohammad Hoseini. Misreporting in the value-added tax and the optimal enforcement. 2015.
- [25] Kuo-Wei Hsu, Nishith Pathak, Jaideep Srivastava, Greg Tschida, and Eric Bjorklund. Data mining based tax audit selection: a case study of a pilot project at the minnesota department of revenue. In *Real world data mining applications*, pages 221–245. Springer, 2015.
- [26] Shaohua Jia. Tax losses in china: Estimates, analysis and countermeasures. *Friends of Accounting*, 13:39–40, 2016.
- [27] Enric Junqué de Fortuny, Marija Stankova, Julie Moeyersoms, Bart Minnaert, Foster Provost, and David Martens. Corporate residence fraud detection. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1650–1659. ACM, 2014.
- [28] Yeonkook J Kim, Bok Baik, and Sungzoon Cho. Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications*, 62:32–43, 2016.
- [29] Wikil Kwak, Yong Shi, and Cheng-Few Lee. The fuzzy set and data mining applications in accounting and finance. In *Handbook of Quantitative Finance and Risk Management*, pages 1307–1331. Springer, 2010.
- [30] Xiaoqing Liu, Ding Pan, and Shihong Chen. Application of hierarchical clustering in tax inspection case-selecting. In *Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on*, pages 1–4. IEEE, 2010.
- [31] Toni Llacer, Francisco J Miguel, José A Noguera, and Eduardo Tapia. An agent-based model of tax compliance: an application to the spanish case. *Advances in Complex Systems*, 16(04n05):1350007, 2013.
- [32] Mandeep Kaur Mahendru. *Impact of Advertisement Expenditure on Firms Performance: A Case of Fmcg Industry in India*. PhD thesis, THAPAR UNIVERSITY PATIALA, 2014.
- [33] Tales Matos, José Antonio F de Macedo, and José Maria Monteiro. An empirical method for discovering tax fraudsters: A real case study of

- brazilian fiscal evasion. In *Proceedings of the 19th International Database Engineering & Applications Symposium*, pages 41–48. ACM, 2015.
- [34] Daniele Micci-Barreca and Satheesh Ramachandran. Improving tax administration with data mining. *White paper. Elite Analytics LLC*, 2004.
 - [35] John Neighbour. Transfer pricing: Keeping it at arms’ length. *Organisation for Economic Cooperation and Development. The OECD Observer*, (230):29, 2002.
 - [36] José A Noguera, Francisco J Miguel Quesada, Eduardo Tapia, and Toni Llàcer. Tax compliance, rational choice, and social influence: An agent-based model. *Revue française de sociologie*, 55(4):765–804, 2014.
 - [37] W Robert Reed and Cynthia L Rogers. Tax burden and the mismeasurement of state tax policy. *Public Finance Review*, 34(4):404–426, 2006.
 - [38] Prem Sikka. Smoke and mirrors: Corporate social responsibility and tax avoidance. In *Accounting Forum*, volume 34, pages 153–168. Elsevier, 2010.
 - [39] Feng Tian, Tian Lan, Kuo-Ming Chao, Nick Godwin, Qinghua Zheng, Nazaraf Shah, and Fan Zhang. Mining suspicious tax evasion groups in big data. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2651–2664, 2016.
 - [40] Alexey Tselykh, Margarita Knyazeva, Elena Popkova, Antonina Durfee, and Alexander Tselykh. An attributed graph mining approach to detect transfer pricing fraud. In *Proceedings of the 9th International Conference on Security of Information and Networks*, pages 72–75. ACM, 2016.
 - [41] Jarrod West and Maumita Bhattacharya. Intelligent financial fraud detection: a comprehensive review. *Computers & Security*, 57:47–66, 2016.
 - [42] Graham J Williams, Peter Christen, et al. Exploratory multilevel hot spot analysis: Australian taxation office case study. In *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70*, pages 77–84. Australian Computer Society, Inc., 2007.

- [43] Roun-Shiunn Wu, Chin-Shyh Ou, Hui-ying Lin, She-I Chang, and David C Yen. Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications*, 39(10):8769–8777, 2012.

Appendix A. The detailed data items of the used dataset

Table A.6: The detailed data items	
Dimensions	Items
Taxpayer	The ID of taxpayer
	The industry sector to which the taxpayer belongs
	The ID of the taxpayer’s legal representative
	The tax rate of taxpayer
	The label of ATTE (ATTE or non-ATTE)
Shareholding	The ID of shareholder
	The ID of taxpayer held
	The shareholding rate
Kinship/interlocking shareholding	The ID of one party of kinship/interlocking shareholding
	The ID of the another party of kinship/interlocking shareholding
	The details of kinship/interlocking shareholding
Transaction	The taxpayer ID of a seller
	The taxpayer ID of a buyer
	The amount of transaction
	The trading date
Tax Burden Rate	The ID of an industry sector
	The name of the sector
	The average tax burden rate of the sector

Appendix B. The algorithm hyperparameters and values

Table B.7: The algorithm hyperparameters and values	
Hyperparameters	Range of value
Th_1	$[0, 1]$
L	$\{1, 2, 3, 4, 5\}$
Th_2	$[0, 100]$

Appendix C. The source codes of the algorithms presented in this paper

For easy understanding the details of the proposed solution, we share the source codes of our algorithms in a public repository (<https://github.com/softwarefly/APIRN>).