

Unsupervised Identification of Corporate Tax-Risk Profiles Using Multilayer Network and Financial Characteristics

Tomas Ruzgas

tomas.ruzgas@ktu.lt

Kaunas University of Technology

Alina Stundziene

Kaunas University of Technology

Rozita Susniene

Kaunas University of Technology

Karolina Armonaite

Kaunas University of Technology

Klaudija Petrauskaite

Kaunas University of Technology

Jurgita Arnastauskaite Zenceviciene

Kaunas University of Technology

Research Article

Keywords: tax-risk profiling, unsupervised learning, reduced clustering method, multilayer networks, knowledge discovery, corporate groups

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11084724/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

This study proposes an unsupervised knowledge-discovery framework for identifying heterogeneous tax-risk profiles among corporate groups by integrating multilayer network characteristics with group-level financial deviations. Corporate relationships are represented through ownership, management and control, and kinship layers, from which interpretable structural characteristics are derived and combined with indicators reflecting deviations in assets, revenue, profit, and liabilities. Clustering is performed using the Reduced Clustering Method (RCM), a recently developed unsupervised method based on modified inversion-formula density estimation and dimensionality reduction. Unlike rule-based or supervised approaches, the framework does not require predefined risk criteria or labelled outcomes. It is evaluated using anonymised administrative data from Lithuania, with the final analytical sample comprising 722 corporate groups and 7,050 legal entities. Separate clustering of financial and network characteristics identifies distinct financial and structural profiles, while their joint analysis reveals more nuanced combinations of financial anomalies and organisational complexity. The identified profiles include groups characterised by extreme financial deviations, complex ownership and control structures, and substantial foreign participation. The results demonstrate how density-based clustering of graph-derived and financial characteristics can support interpretable, group-level tax-risk profiling and provide a data-driven basis for further expert investigation.

1. Introduction

Tax avoidance and aggressive tax planning remain persistent challenges for tax administrations, particularly when economic activities are organised through groups of legally distinct but financially and structurally interconnected companies. Empirical studies indicate substantial revenue losses associated with corporate tax avoidance (Crivelli et al., 2016; Cobham & Janský, 2018). Such structures motivate analytical approaches that move beyond isolated taxpayers and consider both financial behaviour and relationships among economic entities.

Traditional tax-risk assessment commonly relies on expert judgement, financial statement analysis, and rule-based detection procedures (Liu et al., 2010; West & Bhattacharya, 2016). Although such approaches remain useful, their scalability is limited when administrative systems contain large volumes of heterogeneous financial and relational data. This has encouraged the application of data-mining and machine-learning methods to tax compliance and financial fraud problems, including clustering, classification, anomaly detection, and related analytical techniques (Assylbekov et al., 2016; Castellón González & Velásquez, 2013; Hajek & Henriques, 2017). Unsupervised methods are particularly relevant when reliable labels of confirmed risky behaviour are unavailable, as they enable latent structures and atypical profiles to be identified directly from observed data.

Graph-based methods provide an additional analytical perspective by representing dependencies among companies and individuals that cannot be captured by financial attributes alone. Previous studies have modelled ownership links, management ties, transactions, and other relationships for tax and fraud analysis (Junqué de Fortuny et al., 2014; Didimo et al., 2018; Ruan et al., 2019; Tian et al., 2016). In corporate groups, however, different types of relationships may coexist and convey different structural information. Ownership, management and control, and kinship relationships can therefore be represented as separate layers rather than collapsed into a single network. Such a multilayer representation makes it possible to derive interpretable characteristics of organisational complexity, including indirect control, circular structures, connectivity, and foreign participation.

Despite these developments, an important methodological gap remains between relational analysis and unsupervised financial profiling. Financial indicators and network structures are often analysed separately, whereas corporate-group risk assessment may require both sources of information to be considered jointly. Moreover, multilayer corporate networks produce heterogeneous structural characteristics that must be transformed into a representation suitable for systematic comparison across groups. This creates a knowledge-discovery problem in which interpretable graph-derived characteristics and financial deviations need to be integrated within an unsupervised analytical framework capable of revealing both dominant and atypical group profiles.

To address this problem, the present study develops an unsupervised knowledge-discovery framework that combines multilayer network-derived characteristics with group-level financial deviation measures and applies the Reduced Clustering Method (RCM) introduced by Lukauskas and Ruzgas (2023). RCM is a density-based clustering approach built on modified inversion-formula density estimation and dimensionality reduction. Its methodological properties and comparative performance were examined in the original methodological study; the present research instead investigates its use in a substantially different setting, where heterogeneous relational and financial information must be integrated to identify interpretable corporate-group profiles. The resulting framework therefore combines an established clustering methodology with a new domain-specific representation of corporate groups rather than proposing a new graph-clustering algorithm.

The framework is evaluated using anonymised administrative data on Lithuanian corporate entities and associated individuals connected through ownership, management and control, and kinship relationships. The underlying relational data comprise approximately 113,400 legal entities and 261,600 natural persons, while data-completeness and group-selection criteria yield a final analytical sample of 722 corporate groups comprising 7,050 legal entities. The empirical analysis examines financial characteristics, network characteristics, and their joint representation to determine whether the proposed framework can reveal distinct and interpretable corporate-group profiles relevant to tax-risk assessment.

Main Contributions

This study makes three main contributions:

1. It develops an interpretable group-level representation that integrates structural characteristics derived from heterogeneous corporate-network layers with financial deviation measures, enabling relational and financial information to be analysed within a common feature space.
2. It applies the Reduced Clustering Method (RCM), previously developed and benchmarked as a density-based unsupervised clustering approach, to a new knowledge-discovery setting involving heterogeneous corporate-group data and the identification of both dominant and atypical profiles.
3. It demonstrates, using real administrative data, how separate and joint clustering of financial and network characteristics can reveal distinct corporate-group profiles that may support tax-risk prioritisation when labelled outcomes and predefined risk rules are unavailable.

2. Literature Review

Research relevant to corporate tax-risk profiling spans tax compliance analysis, machine learning, financial fraud detection, and network-based knowledge discovery. This section reviews four areas directly related to the proposed framework: traditional approaches to tax-risk assessment, unsupervised and supervised machine-learning methods, graph and multilayer-network approaches, and the methodological gap between relational representation and financial profiling. Particular attention is given to methods applicable when labelled outcomes are unavailable, as this is a central characteristic of the empirical problem considered in this study.

2.1. Traditional Approaches to Tax Avoidance Risk Assessment

Tax administrations have traditionally relied on expert-driven procedures, financial statement analysis, ratio-based indicators, document inspection, and rule-based systems to identify taxpayers requiring further investigation (Liu et al., 2010; West & Bhattacharya, 2016). These approaches provide interpretable decision rules and remain important in practice, but their effectiveness depends on predefined indicators and expert knowledge. Their scalability also becomes increasingly challenging when administrative data include large numbers of interconnected entities and heterogeneous information sources.

A further limitation of firm-level assessment is that potentially relevant financial behaviour may involve several legally distinct but economically or structurally connected entities. Profit allocation, intra-group financing, ownership arrangements, and other forms of functional specialisation can only be partially represented by indicators calculated for individual companies (Loretz et al., 2017; TPA, 2009). Corporate-group analysis therefore requires methods capable of combining entity-level financial information with the relationships linking companies and individuals.

2.2. Machine Learning Approaches in Tax Evasion and Financial Fraud Detection

Machine-learning and data-mining methods provide an alternative to exclusively rule-based assessment by allowing patterns to be learned from multidimensional administrative and financial data. Both supervised and unsupervised approaches have been investigated in tax and financial-fraud applications. Supervised models can exploit known outcomes: Hsu et al. (2015), for example, applied data-mining techniques to tax audit selection, while Kim et al. (2016) and Hajek and Henriques (2017) examined classification methods for fraudulent or misleading financial reporting. Their applicability, however, depends on the availability and reliability of labelled observations.

Unsupervised learning addresses a different problem by identifying latent structures or atypical observations without predefined outcome classes. Assylbekov et al. (2016) combined Self-Organising Maps with hierarchical clustering to examine VAT-evasion patterns, while Castellón González and Velásquez (2013) combined clustering and association-rule mining to characterise taxpayers associated with false invoices. Such methods are particularly relevant to exploratory tax-risk analysis when confirmed outcome labels are incomplete, delayed, or unavailable. However, conventional clustering applications generally operate on tabular attributes and do not inherently represent relationships among companies and individuals. This motivates the integration of unsupervised learning with relational representations of corporate groups.

2.3. Graph-Based Approaches to Fraud and Tax Evasion Detection

Graph-based analysis complements attribute-based methods by explicitly representing relationships among companies, individuals, and economic activities. In tax and fraud applications, relational structures have been used to analyse corporate ownership, affiliated transactions, suspicious substructures, and connections among potentially related entities. Junqué de Fortuny et al. (2014) examined corporate residence fraud using relational information, Didimo et al. (2018) developed the TaxNet visual analytics system integrating network information with administrative data, Ruan et al. (2019) analysed suspicious groups associated with affiliated-transaction-based tax evasion, and Tian et al. (2016) applied graph-based techniques to identify suspicious tax-evasion groups in large-scale data.

Recent graph-based fraud research increasingly treats relational structure itself as an important source of information. For example, Haghighi et al. (2025) proposed a hypergraph-learning framework for fraud detection that moves beyond pairwise relationships and represents higher-order dependencies among entities. Although such supervised graph-learning approaches differ substantially from the unsupervised setting considered here, they illustrate the growing importance of richer relational representations in fraud-related knowledge discovery.

Nevertheless, graph-based and attribute-based analyses address different aspects of the problem. Network topology can reveal connectivity, indirect control, cycles, or unusual participation patterns, whereas financial variables describe the economic roles of entities within a group. For corporate-group

risk profiling, the analytical challenge is therefore not simply to detect suspicious graph structures but to construct an interpretable representation in which structural and financial characteristics can be examined jointly.

2.4. Multi-layer Graph Models and Their Relevance

Multilayer or multi-relational networks provide a natural representation when the same entities are connected through different types of relationships. In corporate data, ownership, management and control, and kinship links convey different information and can therefore be retained as separate relational layers. Collapsing these relationships into a single undifferentiated network may obscure their distinct roles in the organisation and control of corporate groups.

The broader network-analysis literature increasingly addresses the methodological challenges created by multilayer structures. Recent work has developed, for example, community-extraction methods specifically for multilayer networks (Nakano et al., 2025). Such approaches demonstrate the value of preserving layer-specific information, although their objective differs from that of the present study: rather than extracting communities directly from a multilayer network, we derive interpretable structural characteristics from corporate-group networks and integrate them with financial characteristics before clustering.

This distinction is important because the unit of analysis in the present study is the corporate group rather than an individual network vertex or community. The multilayer network is therefore used as a source of structural information, while unsupervised clustering operates on group-level feature representations. This design enables heterogeneous relational information to be combined with financial deviations within a common analytical space.

2.5. Research Gap and Positioning of This Study

The reviewed literature indicates that three methodological components relevant to corporate tax-risk profiling have largely developed along separate lines: financial and attribute-based risk analysis, relational or graph-based analysis, and unsupervised clustering. Financial approaches provide information on the economic behaviour of individual entities but may overlook dependencies within corporate groups. Graph-based methods represent such dependencies explicitly, yet network structure alone does not describe how financial roles are distributed among connected companies. Unsupervised clustering, in turn, provides a mechanism for discovering latent profiles without labelled outcomes, but its results depend critically on how heterogeneous information is represented.

The methodological problem addressed in this study is therefore the integration of these components at the corporate-group level. We construct an interpretable feature representation from multilayer corporate relationships and group-level financial deviations and analyse this representation using the Reduced Clustering Method developed by Lukauskas and Ruzgas (2023). RCM is based on modified inversion-formula density estimation combined with dimensionality reduction and was previously evaluated against established clustering approaches across more than 20 datasets. The earlier study demonstrated the methodological feasibility and comparative performance of RCM; the present study focuses instead on its application to heterogeneous graph-derived and financial characteristics in a real administrative knowledge-discovery problem.

Accordingly, the novelty of the present framework lies not in proposing another graph-clustering algorithm, but in combining a recently developed density-based clustering methodology with a multilayer, group-level representation of corporate relational and financial characteristics. This enables the identification and interpretation of both dominant and atypical corporate-group profiles without predefined risk rules or labelled tax outcomes.

3. Methodology

This section presents the methodological framework for unsupervised identification of corporate-group profiles relevant to tax-risk assessment. The framework combines two complementary sources of information: structural characteristics derived from multilayer corporate networks and financial deviation measures calculated at the group level. These characteristics are integrated into numerical feature representations and analysed using the Reduced Clustering Method (RCM) developed by Lukauskas and Ruzgas (2023).

The methodological principle is to separate relational representation from clustering. Corporate networks are not clustered directly; instead, their multilayer structure is used to derive interpretable group-level characteristics describing organisational complexity, ownership and control patterns, and foreign participation. These characteristics can then be analysed separately or combined with financial deviations within a common feature space. RCM provides the unsupervised mechanism for identifying latent group profiles in this representation without requiring predefined risk classes or labelled outcomes.

The analytical procedure consists of four main components:

1. Construction of a multilayer corporate network,
2. Extraction of group-level structural characteristics from the network,
3. Computation of group-level financial deviation indicators,
4. Unsupervised clustering of the resulting feature representations using RCM.

Figure 1. Analytical pipeline for corporate anomaly detection

3.1. Multi-layer Graph Representation of Corporate Groups

Corporate relationships are represented as a multilayer network in which legal entities and natural persons may be connected through different types of relations. Preserving these relations as separate layers allows their distinct structural roles to be retained during subsequent feature extraction. Formally, the corporate system is represented as:

$$M = \{G^{[1]}, G^{[2]}, \dots, G^{[L]}\}$$

where each layer $G^{[l]} = (V, E^{[l]})$ corresponds to a specific relationship type. In this study, three layers are considered:

- ownership/shareholding relations,
- management and control relations,
- kinship (family) relations.

The vertex set V includes domestic and foreign legal entities and natural persons. Edges are defined according to the corresponding relationship layer and may contain layer-specific information, such as ownership shares. Maintaining the distinction between relationship types prevents premature aggregation of heterogeneous relational information and provides the basis for constructing interpretable structural characteristics at the corporate-group level.

3.2. Graph-Based Structural Feature Extraction

Rather than clustering the corporate networks directly, the multilayer representation is used to derive group-level structural characteristics. Each corporate group is represented by a numerical vector describing selected properties of its induced network structure. This transformation preserves interpretable information about organisational complexity while producing a representation suitable for subsequent unsupervised clustering.

The following features are computed:

- $X_1(C_k) = |V(C_k)|$ representing the total number of entities (legal and natural persons) in the group.
- $X_2(C_k) = \frac{2|E(C_k)|}{|V(C_k)|(|V(C_k)|-1)}$ measuring the overall connectivity of the group.
- The number of simple cycles within the group network $X_3(C_k) = \text{cycles}(G|_{C_k})$, which may indicate circular ownership or control structures often associated with opacity.
- $X_4(C_k) = \frac{\text{number of foreign entities in } C_k}{|V(C_k)|}$, capturing cross-border complexity.

$$X_5(C_k) = \text{number of entities controlled through ownership chains of length } \geq 2$$

• .

- Counts of predefined shareholder structure types (T1–T11), describing ownership concentration and fragmentation patterns.

Together, these characteristics provide an interpretable group-level representation of network connectivity, ownership and control complexity, cross-border participation, and shareholder composition. They are treated as descriptive structural characteristics rather than direct indicators of tax evasion.

3.3. Financial Deviation Indicators

Structural characteristics describe how a corporate group is organised but do not capture the distribution of economic roles among its members. They are therefore complemented with financial deviation indicators calculated at the corporate-group level.

For each financial variable $F \in \{\text{revenue, assets, liabilities, profit, cost of goods sold}\}$, the relative contribution of company i within group C_k is defined as:

$$s_{i,F} = \frac{F_i}{\sum_{j \in C_k} F_j}$$

The deviation from the group's typical structure is measured as:

$$d_{i,F} = |s_{i,F} - \text{median}(s_{i,F})|$$

The group-level financial deviation indicator is then defined as:

$$D_{i,F}(C_k) = \max_{i \in C_k} d_{i,F}$$

For each corporate group, the resulting financial representation consists of five maximum relative-deviation characteristics derived from assets, revenue, profit, liabilities, and other operating results. Large deviations indicate that particular entities perform disproportionately large financial roles relative to

other members of the same group. These deviations are treated as characteristics for exploratory risk profiling rather than as direct evidence of specific tax-avoidance practices.

3.4. Feature Vector Construction and Normalisation

Each corporate group C_k is represented by a numerical feature vector:

$$\mathbf{X}(C_k) = (X_1, X_2, X_3, X_4, X_5, \dots, D_{\text{profit}}, D_{\text{assets}}, D_{\text{liabilities}}, \dots)$$

All features are standardised using Z-score normalisation to ensure comparability across different measurement scales. Three feature representations are subsequently considered: financial characteristics only, network characteristics only, and their joint representation. This design makes it possible to examine the information revealed by each source separately and in combination.

3.5. Density-Based Clustering Using the Reduced Clustering Method

Clustering is performed using the Reduced Clustering Method (RCM), based on the methodology developed by Lukauskas and Ruzgas (2023). The method extends clustering based on modified inversion-formula density estimation by incorporating dimensionality reduction, thereby allowing the density-based procedure to be applied to higher-dimensional feature spaces. In the original methodological study, this extended procedure is referred to as reduced clustering based on the modified inversion formula density estimation (RCBMIDE); the shorter term RCM is used throughout the present study.

The underlying approach estimates cluster-specific densities using the modified inversion-formula framework and assigns observations according to the resulting cluster structure. Dimensionality reduction is incorporated before density-based clustering to obtain a lower-dimensional representation suitable for the estimation procedure. The complete mathematical derivation, parameter-estimation procedure, and clustering algorithm are provided by Lukauskas and Ruzgas (2023) and are therefore not reproduced here.

The methodological study of RCM evaluated the approach on more than 20 datasets and compared its clustering performance with established alternatives, including centroid-, mixture-, hierarchical-, and density-based methods. The results showed that the reduced extension enabled application to higher-dimensional data and produced competitive clustering performance across the evaluated datasets. The present study does not repeat this generic algorithm benchmark; instead, it examines the application of RCM to corporate-group representations derived from financial characteristics, multilayer network characteristics, and their combination.

For each of these feature representations, RCM produces clusters of corporate groups exhibiting similar multidimensional profiles. The resulting clusters are subsequently interpreted using their financial and structural characteristics rather than being treated as predefined tax-risk classes.

3.6. Interpretation of Corporate-Group Profiles

The clusters are interpreted as data-driven corporate-group profiles rather than predefined tax-risk categories. Interpretation is based on cluster size and the distributions of the financial and structural characteristics used in clustering, with particular attention to pronounced financial deviations, network cycles and indirect control, foreign participation, and shareholder structure. Small or atypical clusters exhibiting unusual combinations of these characteristics are considered potentially relevant for further expert assessment. This interpretation does not establish tax evasion or assign a confirmed risk status to individual groups; instead, it identifies profiles that may support subsequent risk prioritisation and investigation.

3.7. Methodological Summary

The proposed framework combines multilayer relational representation, interpretable group-level feature extraction, financial deviation analysis, and density-based unsupervised clustering. Its methodological contribution lies in integrating heterogeneous corporate-network and financial information within a common group-level representation and applying RCM to discover latent profiles without labelled outcomes. The framework retains the interpretability of the original financial and structural characteristics, allowing the resulting clusters to be examined in terms relevant to tax-risk analysis and expert decision support.

4. Data and Experimental Setup

This section describes the administrative data, corporate-group selection procedure, feature construction, and experimental design used to evaluate the proposed unsupervised framework. The analysis is organised around three feature representations—financial characteristics, network characteristics, and their combination—allowing the resulting corporate-group profiles to be examined separately and jointly.

4.1. Data Description

The empirical study uses anonymised administrative data provided by the State Tax Inspectorate under the Ministry of Finance of the Republic of Lithuania. The data combine relational information on legal entities and associated natural persons with company-level financial information, providing the basis for constructing multilayer corporate networks and group-level financial characteristics.

The relational dataset contains approximately 113,400 legal entities (LE) and 261,600 associated natural persons (NP), connected by approximately 343,800 relationships. Three relationship types are considered: ownership/shareholding, management/control, and kinship/family relations. These relationships constitute the three layers of the corporate network used for structural feature extraction.

Each entity is classified as one of the following types:

- domestic legal entity (LE),
- domestic natural person (NP),
- foreign legal entity (FLE),
- foreign natural person (FNP).

This distinction enables domestic and cross-border participation to be represented separately when constructing group-level network characteristics.

Financial information is derived from annual profit-and-loss statements and balance sheets available to the tax authority. The available variables include sales revenue, cost of goods sold, profit before tax, total assets, total liabilities, and supplementary accounting indicators such as equity, fixed assets, provisions, and other operating results. Five core accounting indicators—sales revenue, cost of goods sold, profit before tax, total assets, and total liabilities—are used to assess financial-data completeness during the group-selection procedure. The financial characteristics subsequently used for clustering are group-level deviation measures derived from assets, revenue, profit, liabilities, and other operating results, as described in Section 3.3.

Complete information for all five core indicators is available for approximately 22,500 companies, corresponding to about 20% of the legal entities in the relational dataset. The corporate-group selection procedure described below accounts explicitly for this incomplete financial coverage.

4.2. Identification and Selection of Corporate Groups

Corporate groups are identified as connected components of the multilayer relational network, in which entities may be linked directly or indirectly through ownership, management/control, or kinship relations. The initial network decomposition yields approximately 58,400 groups of varying size. To obtain groups with sufficient financial information for calculating the group-level deviation measures while retaining non-trivial organisational structures, the following selection criteria are applied:

1. The group contains at least three legal entities with complete information for the five core financial indicators.
2. At least 25% of the legal entities in the group have complete information for these indicators.
3. The group contains between 5 and 50 legal entities.

After applying these criteria, the final analytical sample consists of 722 corporate groups comprising 7,050 legal entities.

The resulting sample therefore represents the subset of corporate groups satisfying the predefined size and financial-data completeness requirements and should not be interpreted as representative of all groups in the underlying network.

4.3. Feature Construction

For each selected corporate group, structural and financial characteristics are constructed according to the procedures described in Sections 3.2–3.4. Structural characteristics capture group connectivity, cycles, indirect ownership or control, foreign participation, and shareholder composition. Financial characteristics quantify the largest within-group deviations in the relative financial roles of individual companies. All characteristics are standardised before clustering.

Three feature representations are analysed: (1) financial characteristics only, (2) network characteristics only, and (3) the combined set of financial and network characteristics. This design allows the profiles revealed by each information source to be examined separately before assessing their joint representation.

4.4. Experimental Design

The empirical analysis follows a fully unsupervised design. Each of the three standardised feature representations is analysed separately using RCM: financial characteristics, network characteristics, and their combined representation. For each analysis, the resulting cluster solution is characterised by the number and size of clusters and by the distributions of the variables defining their profiles.

The clustering solutions are subsequently examined to identify dominant and atypical group profiles and to determine which characteristics primarily distinguish them. Particular attention is given to small clusters and to combinations of pronounced financial deviations or unusual structural characteristics, as these profiles are potentially informative for exploratory tax-risk assessment.

4.5. Evaluation Strategy

Confirmed outcome labels indicating whether individual corporate groups were involved in tax evasion or other forms of non-compliance are not available. Consequently, the empirical analysis does not evaluate predictive accuracy or treat cluster membership as a validated risk classification. Instead, evaluation focuses on the interpretability and differentiation of the profiles produced by the unsupervised procedure.

The cluster solutions are examined with respect to cluster size, financial deviation patterns, structural characteristics, and the co-occurrence of multiple distinctive characteristics within the same profiles. The financial-only, network-only, and combined representations are also compared descriptively to assess how the integration of both information sources changes the segmentation of corporate groups. These analyses provide an exploratory basis for identifying profiles that may warrant subsequent expert assessment.

5. Results and Analysis

This section presents the empirical results obtained by applying the proposed clustering framework to Lithuanian corporate groups. The analysis focuses on how financial deviations, network structure characteristics, and their combination differentiate groups of companies and allow the identification of profiles relevant for tax risk assessment.

Each corporate group is characterised by two sets of quantitative indicators:

5.1. Descriptive Statistics of the Corporate Network

The underlying relational network contains 113,424 legal entities and approximately 261,600 natural persons connected through ownership/shareholding, management/control, and kinship relationships. Table 1 summarises the shareholder structure types observed among the legal entities in the relational dataset. Because the subsequent clustering analysis is conducted at the corporate-group level, these figures describe the broader relational system from which the analytical sample was selected rather than the final clustering sample itself.

Table 1
Types of shareholder structures in Lithuanian companies

Type	Number of shareholders and their shareholding	Number of companies	Percentage
T1	1 person: 100	82 755	73%
T2	2 people: 99 / ~1	1 302	1%
T3	2–3 people: 90 / ~10	1 609	1%
T4	2 people: 80 / 20	2 197	2%
T5	2 people: 70 / 30; 66 / 33	2 517	2%
T6	2–3 people: 51 / ~30 / ~20	4 619	4%
T7	2 people: 50 / 50	11 823	10%
T8	3 people: 33 / 33 / 33	2 787	2%
T9	3–4 people: 33 / 33 / 25 / <20	645	1%
T10	3–4 people: ~20–30	2 297	2%
T11	5 people: 20 / 20 / 20 / 20 / ~20	873	1%
Total		113 424	100%

The descriptive statistics illustrate the substantial reduction from the complete relational network to the final analytical sample, primarily because financial information is not available for all legal entities. The selection criteria described in Section 4.2 retain 722 corporate groups comprising 7,050 legal entities. All subsequent clustering results therefore refer to these selected groups unless stated otherwise.

5.2. Clustering Based on Financial Characteristics

Applying RCM to the standardised financial characteristics identifies eight clusters of corporate groups. The solution accounts for 70.1% of the variation in the financial feature representation. Cluster sizes differ substantially, indicating that the financial profiles are not distributed uniformly across the analytical sample. The largest clusters represent comparatively common financial configurations, whereas several small clusters capture less frequent patterns characterised by pronounced deviations in particular financial dimensions.

The largest cluster (Cluster 6; 262 groups, 36%) represents comparatively typical financial structures, with generally moderate deviations across the analysed characteristics. Several smaller clusters exhibit more distinctive profiles. Cluster 3 (63 groups, 9%) is characterised by comparatively large asset deviations, Cluster 4 (39 groups, 5%) by distinct revenue-generation centres, and Cluster 5 (58 groups, 8%) by pronounced liability deviations. The two smallest clusters are particularly atypical: Cluster 7 contains seven groups and is characterised by exceptionally large profit deviations, while Cluster 8 contains only three groups and exhibits extreme deviations in other operating results. These patterns are interpreted as empirical financial profiles rather than evidence of specific tax-avoidance mechanisms.

An illustrative example of an atypical financial profile is Group ID 3871, which belongs to Cluster 8. Its multilayer network structure is shown in Fig. 2, while the financial characteristics of individual group members are presented in Table 2. The group exhibits pronounced asymmetries in the distribution of financial roles among its members. In particular, company ID12 accounts for substantial shares of several financial quantities, whereas company ID15 contributes only a small share of group revenue but a considerably larger share of profit. These differences illustrate the type of within-group financial heterogeneity captured by the financial clustering.

Table 2
Group ID 3871 characteristics*

Vertex ID	Company type	Shareholder structure type	Owns X5	Owned by X5	Asset str	Revenue str	Profit str	Str of other operating results	Liabilities str	comp_median
12	LE	T1	0	3	56%	85%	66%	1080%	64%	66%
6	LE	T1	0	0	14%	6%			14%	14%
15	LE	T9	1	0	10%	1%	32%	-984%	1%	1%
16	LE	T4	0	3	10%	0%	2%	4%	0%	2%
10	LE	T4	0	0	6%	2%			12%	6%
3	LE	T7	0	0	3%	5%			9%	5%
13	FLE	T1	0	4						

* Here and further in the results tables, str denotes the relative share, comp_median denotes the median of the company's relative share.

5.3. Clustering Based on Network Characteristics

Applying RCM to the standardised network characteristics identifies 12 clusters, accounting for 43.5% of the variation in the network feature representation. Compared with the financial analysis, the network-based solution contains a larger number of profiles and a more dispersed distribution of groups across clusters. This reflects heterogeneity in the structural organisation of corporate groups, including differences in group size, connectivity, ownership and control patterns, foreign participation, and shareholder composition.

The network-based clusters demonstrate that corporate groups with broadly similar sizes may nevertheless differ substantially in their internal organisation. Particularly distinctive profiles are associated with combinations of cycles, indirect ownership or control, foreign legal or natural persons, and less common shareholder structures. These characteristics describe organisational complexity and cross-border participation; individually, they should not be interpreted as evidence of non-compliance. Their relevance for tax-risk profiling arises from the possibility of examining such structural characteristics jointly and, subsequently, in combination with financial deviations.

Table 3 summarises the profiles of the 12 clusters obtained from the network characteristics. Several relatively small clusters exhibit particularly distinctive structural configurations. Cluster 1 is characterised by the presence of cycles and comparatively high indirect control, Cluster 9 stands out for the number of foreign natural persons, while Cluster 12 exhibits exceptionally high values of indirect control and foreign legal-entity participation. These profiles illustrate that structural differentiation among corporate groups arises from different combinations of network characteristics rather than from a single indicator.

Group ID 5774 provides an illustrative example of a structurally complex group from Cluster 1. Its multilayer network is shown in Fig. 3 and the corresponding company-level characteristics are reported in Table 4. The example illustrates how a complex relational structure may coexist with heterogeneous financial roles among individual companies.

Table 3
Network group clusters

Clusters	1	2	3	4	5	6	7	8	9	10	11	12
Number of groups	10	30	89	214	52	31	83	43	13	93	58	6
Number of groups, %	1%	4%	12%	30%	7%	4%	11%	6%	2%	13%	8%	1%
X2	0.07	0.02	0.05	0.05	0.05	0.03	0.05	0.05	0.04	0.11	0.05	0.02
X3	1.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17
<i>Owns X5</i>	11.50	7.63	1.37	0.91	3.48	3.06	1.52	6.65	5.54	1.83	3.71	46.17
X4	0.61	0.52	0.50	0.51	0.51	0.44	0.45	0.73	0.42	0.66	0.53	0.63
<i>FLE X6</i>	0.30	0.80	0.45	0.18	0.62	0.35	0.30	2.72	0.62	0.25	0.60	7.67
<i>FNP X6</i>	0.30	1.10	0.40	0.21	0.46	0.87	0.41	0.35	6.69	0.19	0.36	1.67
T1	6.20	13.80	4.73	4.91	5.65	6.19	4.29	9.72	5.23	4.13	5.72	22.33
T2	0.20	0.27	0.03	0.00	0.12	0.06	0.01	0.07	0.15	0.04	1.28	0.33
T3	0.60	1.27	0.20	0.13	0.19	0.23	0.18	0.30	0.08	0.10	0.41	0.67
T4	0.90	0.33	1.36	0.00	0.48	0.32	0.05	0.35	0.15	0.11	0.28	1.17
T5	0.70	1.03	0.38	0.26	0.38	0.35	0.34	0.28	0.69	0.05	0.47	1.67
T6	0.90	1.50	0.44	0.64	0.60	0.55	0.49	0.40	0.77	0.19	0.57	2.33
T7	1.10	4.57	1.47	1.36	1.27	1.81	1.36	0.98	1.69	0.96	1.21	3.00
T8	0.40	0.80	0.21	0.24	0.19	2.65	0.18	0.21	0.92	0.13	0.19	0.50
T9	0.50	0.23	0.01	0.00	1.10	0.06	0.01	0.00	0.38	0.01	0.03	0.00
T10	0.30	1.13	0.19	0.01	0.38	0.61	1.43	0.09	1.00	0.14	0.28	0.67
T11	0.30	0.93	0.07	0.10	0.13	0.13	0.17	0.05	0.69	0.04	0.03	0.83

Table 4
Group ID 5774 characteristics

Apex ID	Type of company	Type of shareholder structure	Owns X5	Owned by X5	Asset str	Revenue str	Profit str	Str of other operating results	Liabilities str	comp_median
29	LE	T8	0	0	25%	35%	58%	49%	12%	35%
47	LE	T1	0	0	30%	27%	7%	33%	54%	30%
30	LE	T9	2	5	11%	10%	12%	11%	7%	11%
51	LE	T10	2	6	13%	10%	11%		8%	10%
32	LE	T11	1	5	8%	9%	13%		5%	9%
41	LE	T1	0	0	6%	3%	1%	4%	6%	4%
24	LE	T9	0	0	2%	1%			2%	2%
49	LE	T1	0	0	4%	1%	-2%	1%	5%	1%
43	LE	T1	0	0	1%	1%	2%	2%	0%	1%
46	LE	T10	0	3	1%	2%			1%	1%
28	LE	T5	0	1	0%	0%			0%	0%
27	LE	T1	0	0	0%	2%	0%	0%	0%	0%
40	LE	T1	0	0	0%	0%	0%		0%	0%

5.4. Combined Financial and Network Clustering

The final analysis combines the standardised financial and network characteristics within a single group-level representation. Applying RCM to this combined feature space identifies 17 clusters accounting for 45.5% of the variation. The larger number of clusters relative to the separate financial and network analyses indicates a more differentiated segmentation when both sources of information are considered jointly. The combined solution reveals profiles that cannot be described solely by the magnitude of financial deviations or by network structure, because groups may exhibit different combinations of these characteristics. The characteristics of the 17 combined clusters are summarised in Table 5.

The percentages reported for the three clustering analyses refer to different feature representations and should therefore not be interpreted as directly comparable measures of relative clustering quality. The purpose of the combined analysis is instead to examine whether integrating financial and structural information produces additional, interpretable corporate-group profiles.

Table 5
Results of clustering based on mixed characteristics

Clusters	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Number of groups	42	3	32	180	24	7	51	88	50	58	10	37	2	62	25	42	6
Number of groups, %	6%	0%	4%	25%	3%	1%	7%	12%	7%	8%	1%	5%	0%	9%	3%	6%	1%
X2	0.05	0.06	0.04	0.05	0.02	0.04	0.05	0.06	0.05	0.12	0.07	0.05	0.02	0.06	0.04	0.05	0.02
X3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.10	0.00	0.00	0.00	0.00	0.00	0.17
Owens X5	1.71	4.67	2.06	0.95	8.29	3.14	1.71	1.22	4.08	1.83	11.50	9.05	2.00	1.27	2.48	3.76	46.17
X4	0.42	0.57	0.45	0.50	0.52	0.53	0.50	0.53	0.54	0.69	0.61	0.72	0.42	0.55	0.44	0.52	0.63
FLE X6	0.36	0.67	0.34	0.31	0.79	0.29	0.31	0.30	0.76	0.22	0.30	3.08	0.00	0.21	0.08	0.64	7.67
FNP X6	0.86	1.00	1.13	0.21	1.42	0.57	0.24	0.42	0.52	0.09	0.30	0.46	14.00	0.23	0.80	0.48	1.67
T1	4.86	4.67	4.97	4.74	14.21	8.14	4.92	4.36	5.46	4.17	6.20	9.35	10.00	5.77	5.76	6.36	22.33
T2	0.05	0.00	0.03	0.00	0.33	0.43	0.02	0.07	1.32	0.05	0.20	0.05	0.00	0.03	0.08	0.14	0.33
T3	0.02	0.00	0.16	0.01	1.29	0.29	1.29	0.09	0.42	0.02	0.60	0.19	0.00	0.13	0.12	0.19	0.67
T4	0.14	0.67	0.22	0.26	0.38	0.00	0.65	0.25	0.40	0.17	0.90	0.43	0.00	0.10	0.48	0.55	1.17
T5	0.26	0.33	0.22	0.35	1.00	0.29	0.27	0.20	0.44	0.07	0.70	0.41	2.00	0.18	0.44	0.50	1.67
T6	0.64	0.33	0.59	0.59	1.88	0.71	0.45	0.47	0.60	0.26	0.90	0.38	0.50	0.29	0.52	0.67	2.33
T7	1.45	1.00	1.31	1.49	4.71	1.14	1.04	1.19	1.20	0.93	1.10	1.27	5.50	1.40	1.68	1.26	3.00
T8	0.38	0.33	0.28	0.26	0.83	0.57	0.08	0.25	0.22	0.10	0.40	0.30	2.00	0.13	2.80	0.17	0.50
T9	0.12	0.67	0.09	0.00	0.21	0.00	0.02	0.00	0.04	0.02	0.50	0.03	0.50	0.05	0.08	1.12	0.00
T10	0.31	0.00	2.22	0.20	1.25	0.14	0.25	0.25	0.28	0.16	0.30	0.08	4.00	0.15	0.60	0.26	0.67
T11	1.14	0.00	0.09	0.00	0.88	0.43	0.02	0.02	0.02	0.03	0.30	0.05	1.00	0.03	0.08	0.10	0.83
Maximum relative deviation of assets	0.10	0.11	0.10	0.09	0.16	0.22	0.09	0.28	0.16	0.11	0.14	0.20	0.12	0.12	0.14	0.13	0.08
Highest relative deviation in revenue	0.11	0.26	0.10	0.09	0.17	0.13	0.12	0.14	0.13	0.12	0.14	0.16	0.25	0.44	0.11	0.13	0.12
The largest relative deviation in other operating results	0.25	15.75	0.15	0.17	0.22	0.24	0.20	0.24	0.19	0.19	0.19	0.23	0.20	0.40	0.17	0.19	0.20
Maximum relative deviation of liabilities	0.13	0.16	0.20	0.10	0.18	0.19	0.14	0.39	0.17	0.12	0.19	0.24	0.16	0.12	0.19	0.18	0.08
Largest relative deviation in profit	0.95	0.23	1.61	0.83	1.62	42.09	0.59	0.79	0.63	0.45	1.17	0.86	0.81	0.97	0.77	0.89	0.33

The distributions of selected financial and network characteristics across the combined clusters are presented in Fig. 4. The combined profiles show that pronounced financial deviations and structural complexity do not necessarily occur in the same groups. Some clusters are differentiated primarily by financial characteristics, others by network structure, while several exhibit distinctive combinations of both. In particular, profiles involving substantial

financial deviations together with cycles, indirect control, foreign participation, or uncommon shareholder structures illustrate the additional descriptive information obtained by integrating the two feature sets. These combinations are of particular interest for exploratory tax-risk assessment because they allow financially atypical behaviour to be interpreted together with the organisational structure in which it occurs.

6. Discussion

This study examines whether heterogeneous relational and financial information can be integrated within an unsupervised framework to identify interpretable corporate-group profiles relevant to tax-risk assessment. The framework combines structural characteristics derived from multilayer corporate networks with group-level financial deviations and analyses these representations using the Reduced Clustering Method (RCM) developed by Lukauskas and Ruzgas (2023). The empirical results are discussed below in terms of the information revealed by financial, network, and combined representations, the methodological role of RCM in this setting, and the potential use and limitations of the resulting profiles for tax-risk analysis.

6.1. Interpretation of Identified Clusters

The three clustering analyses show that financial and network characteristics provide complementary descriptions of corporate-group heterogeneity. Financial characteristics produce comparatively clear profiles associated with different patterns of asset, revenue, profit, liability, and other operating-result deviations. Network characteristics reveal a different form of differentiation, associated with organisational structure, including cycles, indirect ownership or control, foreign participation, and shareholder composition. The combined representation provides a more differentiated segmentation in which financial and structural characteristics can be interpreted jointly.

The financial-only analysis indicates that most corporate groups belong to relatively common profiles with moderate deviations, whereas several smaller clusters are characterised by pronounced deviations in specific financial dimensions. Such clusters are analytically important because they identify uncommon distributions of financial roles within corporate groups. However, these deviations do not by themselves establish the presence of tax avoidance. Rather, they provide empirical signals that can be examined together with other information when selecting groups for further assessment.

The network-based analysis similarly distinguishes several structurally unusual profiles. Cycles, extensive indirect control, foreign legal or natural persons, and uncommon shareholder structures occur with different intensities across the identified clusters. These characteristics should not be interpreted individually as indicators of non-compliance; their analytical relevance lies in describing how corporate groups are organised and in allowing structurally atypical configurations to be identified systematically.

The combined analysis is particularly informative because it shows that financial atypicality and structural complexity do not always coincide. Some profiles are distinguished mainly by financial deviations, others by network structure, while several combine characteristics from both domains. This result supports the use of an integrated representation for exploratory tax-risk profiling: it allows financial behaviour to be interpreted within the organisational structure of the corporate group rather than treating either information source in isolation.

6.2. Added Value of the Integrated Multilayer Representation

The methodological value of the proposed framework lies in the representation of corporate groups rather than in treating the underlying networks as objects to be clustered directly. Multilayer relationships are first transformed into interpretable structural characteristics, which can then be analysed separately or combined with financial deviations within a common feature space. This design preserves information about different relationship types while allowing the clustering procedure to operate on comparable group-level representations.

RCM provides the unsupervised mechanism for analysing these representations. The method was originally developed and benchmarked by Lukauskas and Ruzgas (2023) as a density-based clustering approach using modified inversion-formula density estimation and dimensionality reduction. The present study extends the application domain of RCM rather than introducing a new clustering algorithm: it demonstrates how the method can be used with heterogeneous characteristics derived from a real administrative multilayer network and financial data.

The empirical results also clarify the value of integrating the two information sources. The combined analysis does not simply replace the separate financial and network analyses; instead, the three representations serve different interpretative purposes. Financial clustering highlights how economic roles are distributed within groups, network clustering characterises organisational structure, and the combined representation identifies profiles in which these dimensions intersect. This layered interpretation is particularly useful in settings where labelled outcomes are unavailable and the objective is exploratory knowledge discovery rather than predictive classification.

6.3. Implications for Tax-Risk Analysis

From an applied perspective, the identified profiles can support tax administrations by providing an additional exploratory layer between large administrative datasets and case-level expert assessment. Rather than applying a single risk threshold to all corporate groups, analysts can examine groups within profiles characterised by particular financial deviations, structural configurations, or combinations of both.

The framework is therefore best viewed as a decision-support component rather than an automated classification system. Cluster membership does not establish that a corporate group is non-compliant, nor does it provide a probability of tax evasion. Its practical value lies in reducing a heterogeneous population of corporate groups to a set of interpretable profiles and making the characteristics underlying those profiles explicit. Analysts can subsequently combine this information with audit history, tax returns, transaction data, or other administrative evidence not included in the present study.

This distinction is important for practical implementation. Unsupervised profiling is particularly useful when confirmed labels are unavailable or delayed, but expert judgement remains necessary when profiles are translated into operational decisions. The proposed framework can therefore complement existing risk-assessment procedures by identifying structurally or financially atypical groups for further examination rather than replacing established compliance-management systems.

6.4. Limitations and Directions for Future Research

Several limitations should be considered when interpreting the results. First, complete financial information is available only for a subset of legal entities, and the analytical sample is therefore restricted to corporate groups satisfying predefined data-completeness and size criteria. The resulting profiles should not be interpreted as representative of all corporate groups in the underlying administrative network. Second, confirmed labels of tax evasion or other forms of non-compliance are unavailable. Consequently, the identified clusters cannot be validated as risk classes, and the analysis evaluates their interpretability and distinctiveness rather than predictive accuracy.

Third, the empirical application is based on data from a single national tax administration and captures corporate structures and financial characteristics for the analysed period rather than their evolution over time. The transferability of the identified profiles to other institutional environments therefore remains to be examined. Finally, the study focuses on selected network and financial characteristics; additional information such as transactions, tax payments, audit outcomes, or temporal changes in ownership and control could provide further discriminatory information.

Future research should therefore focus on external and temporal validation of the identified profiles. Longitudinal network data could be used to examine whether changes in ownership, management, or financial roles precede changes in tax-risk assessments. Where reliable audit outcomes become available, the unsupervised profiles could also be evaluated against confirmed cases or incorporated into hybrid frameworks combining unsupervised clustering, expert assessment, and supervised prediction. Cross-country applications would provide an additional test of the generalisability of the proposed representation and clustering framework.

7. Conclusions

This study developed and empirically examined an unsupervised framework for corporate tax-risk profiling based on the integration of multilayer network characteristics and group-level financial deviations. The framework applies the Reduced Clustering Method (RCM) developed by Lukauskas and Ruzgas (2023) to interpretable feature representations derived from corporate relationships and financial data. The contribution of the present study lies not in proposing a new clustering algorithm, but in demonstrating how a recently developed density-based method can be applied to heterogeneous corporate-group information for exploratory knowledge discovery.

The empirical application to anonymised Lithuanian administrative data shows that financial and network characteristics identify different forms of heterogeneity among corporate groups. Financial clustering distinguishes profiles associated with the distribution of assets, revenue, profit, liabilities, and other operating results, while network clustering reveals differences in organisational structure, including indirect control, cycles, foreign participation, and shareholder composition. The combined representation allows these dimensions to be interpreted jointly and produces a more differentiated set of corporate-group profiles without requiring predefined classes or labelled outcomes.

The combined analysis partly reproduces the dominant patterns observed in the separate financial and network representations, but it also enables additional interpretation of profiles in which financial and structural characteristics occur together. This is important because pronounced financial deviations and organisational complexity do not necessarily coincide in the same corporate groups. The integrated representation therefore adds descriptive value by placing financial behaviour within the relational structure in which it occurs, rather than by replacing the information provided by the separate analyses.

From a practical perspective, the framework can support tax administrations by structuring heterogeneous corporate-group data into interpretable profiles that may guide further expert examination. It should be viewed as a decision-support component rather than an automated tax-evasion detection system. Cluster membership does not establish non-compliance or provide a validated probability of tax risk; instead, it helps identify groups with distinctive combinations of financial and structural characteristics that may warrant additional investigation.

Several limitations should be considered when interpreting the findings. The analysis is restricted to corporate groups satisfying predefined financial-data completeness and size criteria, and the resulting sample is therefore not representative of all groups in the underlying administrative network. In addition, confirmed labels of tax evasion or other forms of non-compliance are unavailable, preventing direct validation of the identified clusters as risk classes. The empirical application is also limited to one national setting and does not incorporate temporal changes in corporate structures or financial behaviour.

Future research should focus on external and temporal validation of the proposed framework. Longitudinal data could be used to examine changes in ownership, management, and financial roles over time, while confirmed audit outcomes would allow the relationship between unsupervised profiles and observed non-compliance to be assessed directly. Additional data sources, such as transactions, tax payments, or cross-border flows, could further enrich the group-level representation. Cross-country applications would provide an additional test of the transferability of the identified profiles and the robustness of the framework across different institutional settings.

Declarations

Author Contribution

T.R. contributed to the study conception, methodology, data analysis, and manuscript preparation. A.S., R.S., K.A., K.P., and J.A.Z. contributed to data preparation, analysis, interpretation of the results, and manuscript revision. All authors reviewed and approved the final manuscript.

Data Availability

The data that support the findings of this study are not publicly available due to confidentiality restrictions concerning taxpayer information. The data were provided by the Lithuanian State Tax Inspectorate and cannot be shared by the authors.

References

1. Assylbekov Z, Melnykov I, Bekishev R, Baltabayeva A, Bissengaliyeva D, Mamlin E (2016) Detecting Value-Added Tax Evasion by Business Entities of Kazakhstan. *Intelligent Decision Technologies*. pp. 37–49. 10.1007/978-3-319-39630-9_4
2. Castellón González P, Velásquez JD (2013) Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Syst Appl* 40(5):1427–1436. 10.1016/J.ESWA.2012.08.051
3. Cobham A, Janský P (2018) Global distribution of revenue loss from corporate tax avoidance: re-estimation and country results. *J Int Dev* 30(2):206–232. 10.1002/jid.3348
4. Crivelli E, de Mooij R, Keen M (2018) Base Erosion, Profit Shifting and Developing Countries. *FinanzArchiv*. Vol. 72, no. 3, p. 268. 10.1628/001522116X14646834385460
5. Didimo W, Giamminonni L, Liotta G, Montecchiani F, Pagliuca D (2018) A visual analytics system to support tax evasion discovery. *Decis Support Syst* 110:71–83. 10.1016/J.DSS.2018.03.008
6. Haghighi V, Soltani B, Shabani N et al (2025) Beyond pairwise relationships: a transformer-based hypergraph learning approach for fraud detection. *Knowl Inf Syst* 67:7987–8022. <https://doi.org/10.1007/s10115-025-02476-5>
7. Hajek P, Henriques R (2017) Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems*. Vol. 128, pp. 139–152. 10.1016/J.KNOSYS.2017.05.001
8. Hsu K-W, Pathak N, Srivastava J, Tschida G, Bjorklund E (2015) Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue. *Real World Data Mining Applications*. pp. 221–245. ISBN 978-3-319-07812-0
9. Junqué de Fortuny E, Stankova M, Moeyersoms J, Minnaert B, Provost F, Martens D (2014) Corporate residence fraud detection. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*. pp. 1650–1659. 10.1145/2623330.2623333
10. Kim YJ, Baik B, Cho S (2016) Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Syst Appl* 62:32–43. 10.1016/J.ESWA.2016.06.016
11. Liu X, Pan D, Chen S (2010) Application of Hierarchical Clustering in Tax Inspection Case-Selecting. *2010 International Conference on Computational Intelligence and Software Engineering*. pp. 1–4. 10.1109/CISE.2010.5676711
12. Loretz S, Sellner R, Bucci V (2017) Aggressive tax planning indicators: Final Report. p. 189. 10.2778/822243
13. Lukauskas M, Ruzgas T (2023) Reduced Clustering Method Based on the Inversion Formula Density Estimation. *Mathematics*, Vol. 11, no. 3, 661. 10.3390/math11030661
14. Nakano S, Yamashita T, Shin A et al (2025) Weighted APPR with restart for community extraction from multilayer networks. *Knowl Inf Syst* 67:8919–8939. <https://doi.org/10.1007/s10115-025-02488-1>
15. Ruan J, Yan Z, Dong B, Zheng Q, Qian B (2019) Identifying suspicious groups of affiliated-transaction-based tax evasion in big data. *Inf Sci* 477:508–532. 10.1016/J.INS.2018.11.008
16. Tian F, Lan T, Chao K, Godwin N, Zheng Q, Shah N, Zhang F (2016) Mining Suspicious Tax Evasion Groups in Big Data. *IEEE Trans Knowl Data Eng* 28(10):2651–2664. 10.1109/TKDE.2016.2571686
17. TPA Centre For Tax Policy and Administration (2009) Forum on Tax Administration: Compliance Management of Large Business Task Group. Experiences and Practices of Eight OECD Countries
18. West J, Bhattacharya M (2016) Intelligent financial fraud detection: A comprehensive review. *Computers Secur* 57:47–66. 10.1016/J.COSE.2015.09.005

Figures



Figure 1

Analytical pipeline for corporate anomaly detection

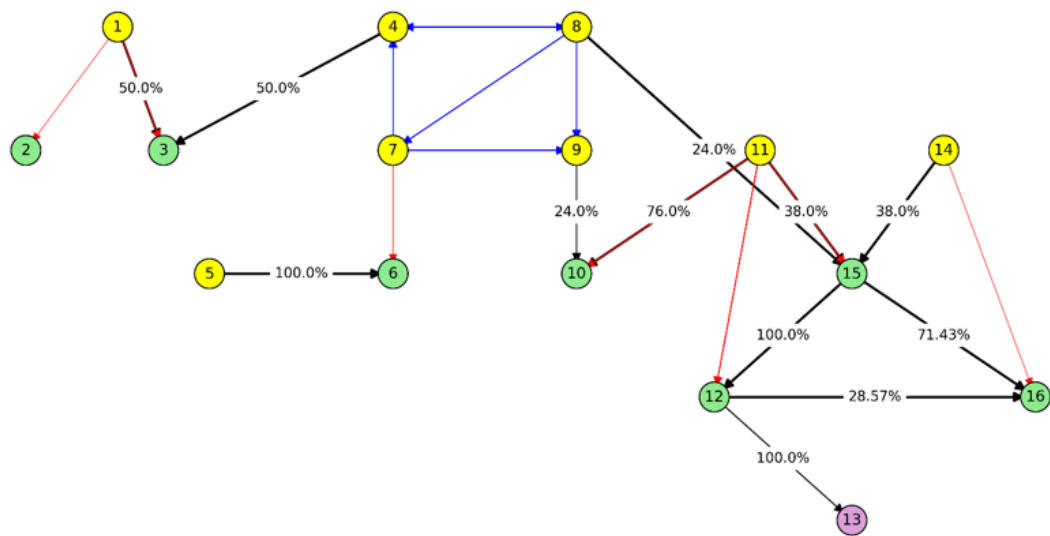


Figure 2

Group ID 3871 multi-layer graph

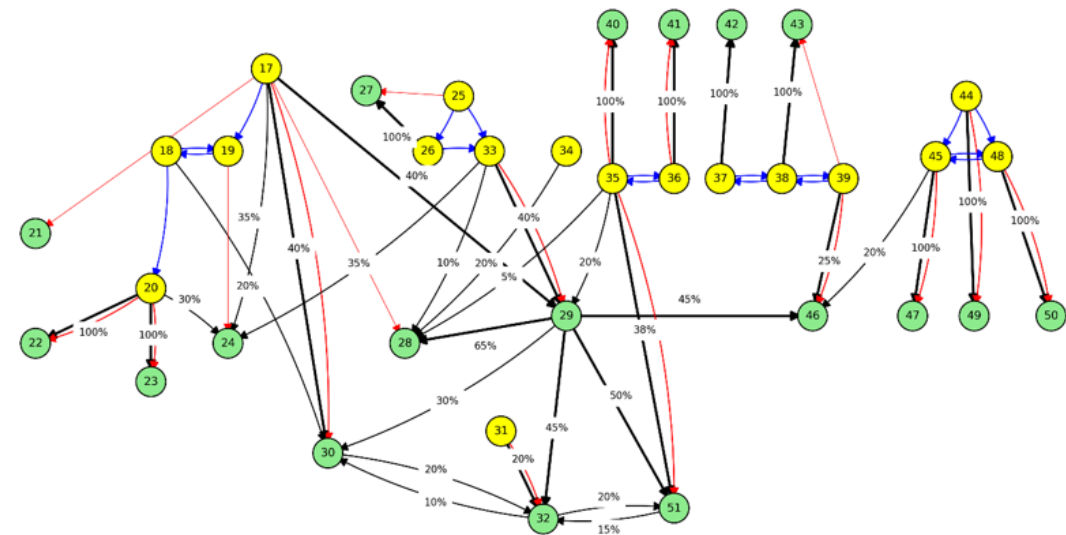


Figure 3

Group ID 5774 multi-layer graph

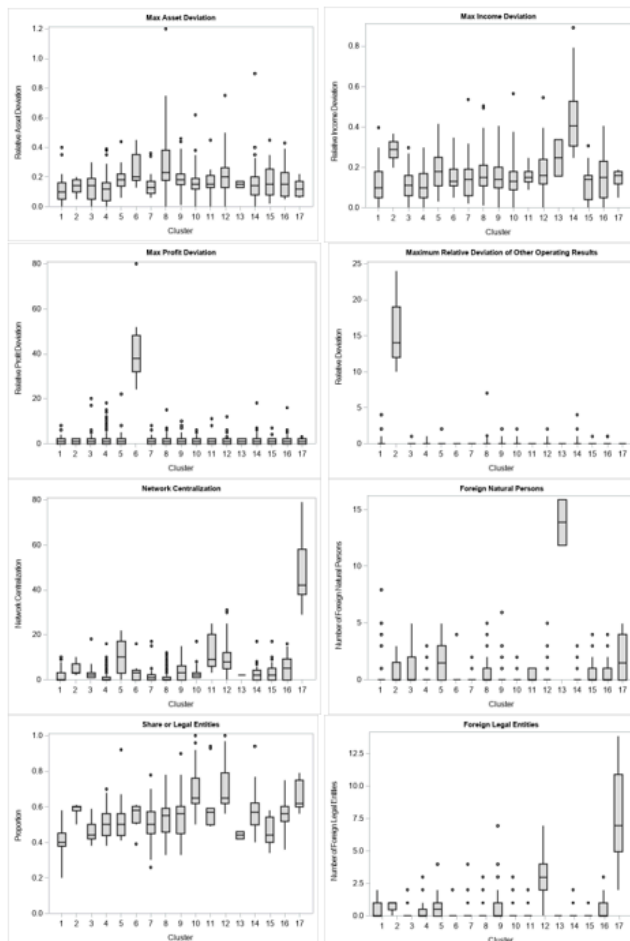


Figure 4

Comparison of cluster characteristics