Graph-Theory Based Identification of Company Groups for Tax Evasion Risk

INTRODUCTION

Modern tax administrations face the challenge of detecting increasingly sophisticated tax evasion schemes. Traditional audit approaches often fail to identify hidden networks of businesses acting together to conceal transactions or misreport revenues. This poster introduces a graph-theoretic methodology for identifying groups of companies that may pose a heightened risk of tax evasion. The core idea is to model relationships between firms as a network in which nodes represent companies and edges capture shared attributes (common shareholders, directors, addresses or frequent transactions). Using graph theory allows us to see beyond individual firms and to examine the structure of the entire network; communities or clusters within this network may indicate coordinated activity. The proposed method includes constructing the network from administrative data, applying community detection algorithms to identify dense subgraphs, and evaluating cluster risk using measures such as centrality and connectivity. The results show that suspicious groups often exhibit distinctive structural patterns—such as unusually high interconnectedness or central “hub” companies—that distinguish them from ordinary business networks. In addition, the approach can be integrated with other risk indicators, enhancing the efficiency of audit selection. By visualising the corporate network and highlighting high-risk clusters, tax authorities can allocate investigative resources more effectively and disrupt large-scale evasion schemes. Overall, this work demonstrates that graph analytics provide powerful tools for uncovering hidden relationships and offers a promising direction for risk-based tax compliance strategies.

DATA AND METHODS

The analysis used administrative tax datasets from the national revenue authority. These files contained information on corporate taxpayers (registration numbers, sector, turnover and tax declarations) and lists of directors and shareholders for each firm.

Additional data on inter-company transactions were obtained from VAT invoice registers. To create the network, the researchers extracted pairwise links between legal entities (e.g. shared board members, direct ownership links, high-value transactions) and encoded them as edges in a graph.
All data were anonymized.

The research data is stored in two matrices: the E edge matrix and the V vertex matrix. The edge matrix stores data about the edges of all layers and their characteristics:

$$E = \begin{pmatrix} L_1 & V_1^{from} & V_1^{to} & w_1 & x_{11} & \dots & x_{1p} \\ L_2 & V_2^{from} & V_2^{to} & w_2 & x_{21} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ L_K & V_K^{from} & V_K^{to} & w_K & x_{K1} & \dots & x_{Kp} \end{pmatrix}$$

where: L_i – code of the layer to which the edge belongs, V_i^{from}, V_i^{to} – codes for vertices connected by an edge, w_i – edge weight, x_{ij} –edge features; $i = \overline{1,K}$, $j = \overline{1,\ p}$.

The vertex matrix stores data about the vertices of all layers and their attributes:

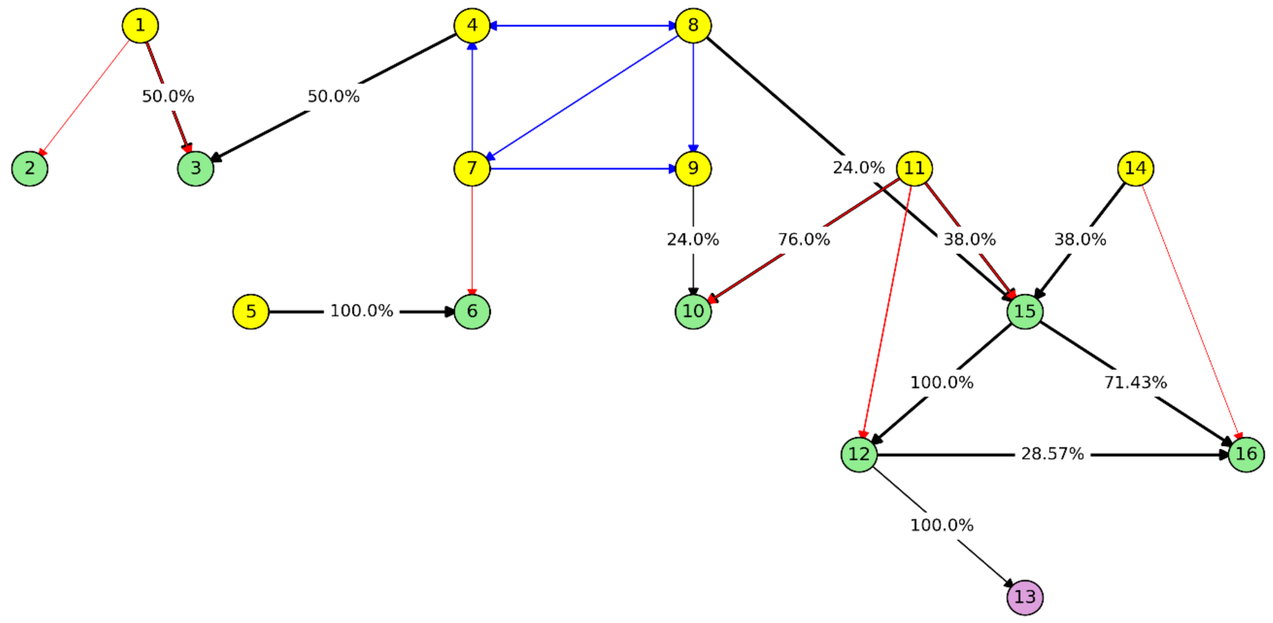
$$V = \begin{pmatrix} V_1 & T_1 & z_{11} & \dots & z_{1r} \\ V_2 & T_2 & z_{21} & \dots & z_{2r} \\ \dots & \dots & \dots & \dots & \dots \\ V_N & T_N & z_{N1} & \dots & z_{Nr} \end{pmatrix}$$

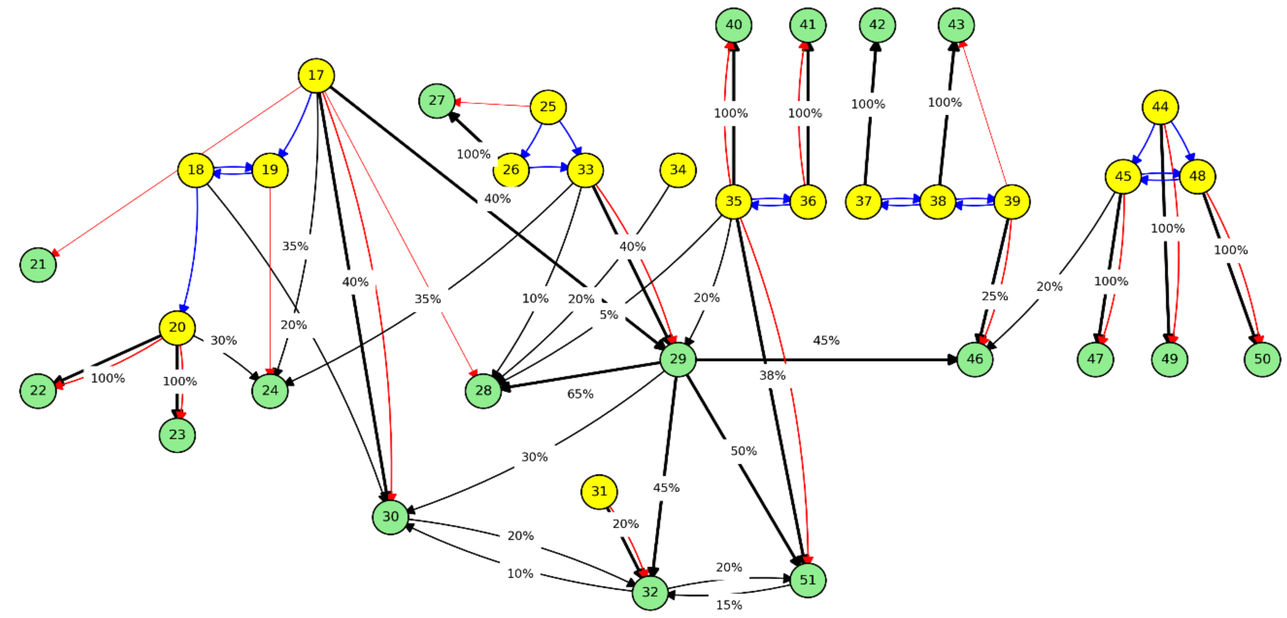
where: V_i –vertex code, T_i – vertex type (legal entity LE, natural person NP, foreign legal entity FLE, foreign natural person FNP), z_ij – vertex attributes; $i = \overline{1,\ N}$, $j = \overline{1,\ r}$.
A multi-layered graph shows the connections between vertices and pairs of layers $G_M = (V_M, E_M)$ and is defined as a set of 4 elements $M = (V_M, E_M, V, L)$. A multilayer network M can have any number d of dimensions. Each dimension corresponds to a certain set of elementary layers $L = \{L_a\}_{a=1}^{d}$. For example, one dimension could be the companies with which a person is associated (e.g., private company A, state-owned company B; a set of elementary layers of companies $L_1 = \{A, B\}$), another dimension – the type of relationship (e.g., relationship X – "owns shares", relationship Y – "holds a management position"; a set of elementary layers of relationships $L_2 = \{X, Y\}$). Here, the types of companies and connections reflect the basic layers (L_1, L_2) , and their combination (the type of relationship in each company) reflects the layers ((A, X), (A, Y), (B, X), (B, Y)).

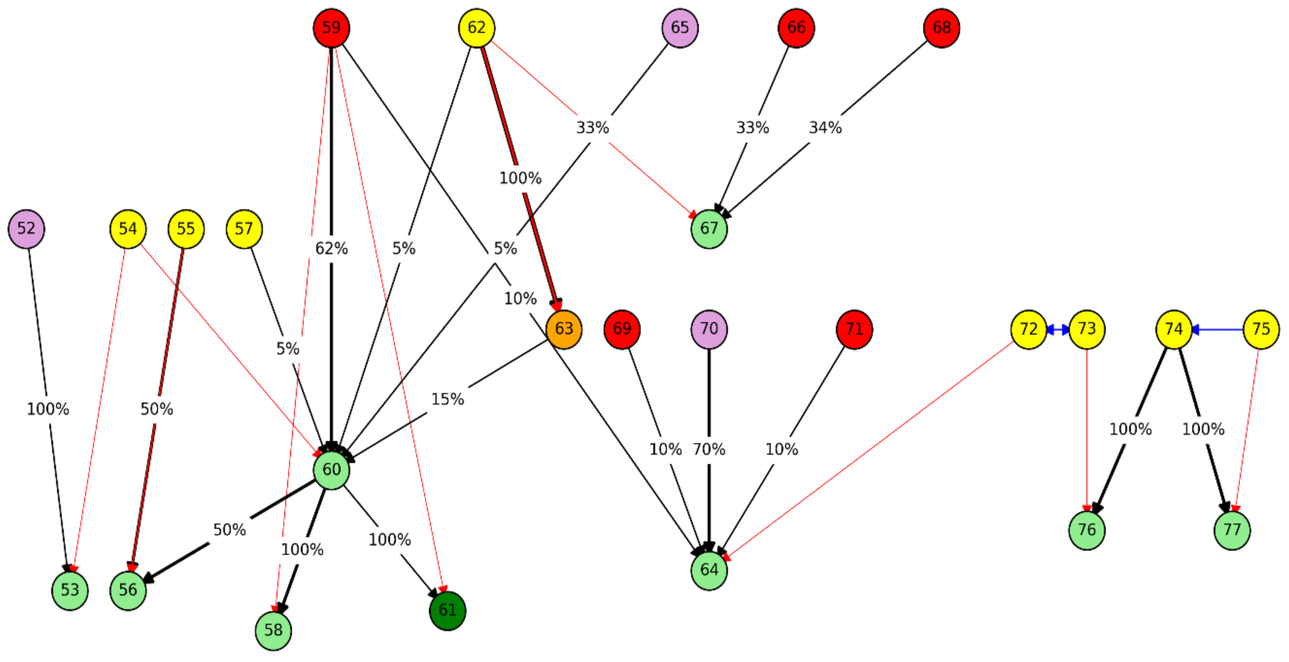
The connection (edge) and vertex matrices contain information about 113,400 legal entities (LE) and 261,600 natural persons (NP) associated with them. The edge matrix contains information about three types of connections: 156,600 (45%) shares (S), 119,800 (35%) management (M) and 67,400 (20%) family ties (F). A total of 343,800 relationships. In the top matrix, which stores information about a total of 267,700 tops, 66,800 tops are LE with financial characteristics. Financial characteristics – companies' financial performance results from profit (loss) and balance sheet reports for the full calendar year: economic activity code, number of employees, sales revenue, cost of sales, profit before tax, assets, fixed assets, intangible assets, expenses for the coming period, equity, subsidies, provisions, liabilities, accrued expenses.

RESULTS

The research constructed a company-to-company network using publicly available links (ownership, board members, address, etc.) and derived a Minimum Spanning Tree (MST). The MST revealed clusters of tightly connected companies. In total, the method identified several hundred clusters, with group sizes ranging from two to a dozen companies. Visual inspection confirmed that most clusters represented real business groups rather than random connections.







Each company and cluster were assigned a risk score based on topological metrics (degree, betweenness, closeness) and compared against historical tax-audit results. High-risk clusters corresponded well with known tax-evasion cases. Compared with baseline rule-based selection, the graph-theory approach reduced false positives and captured more companies that were later found to commit evasion.

CONCLUSION

The study demonstrates that graph-based analysis is a powerful tool for detecting potential tax-evasion schemes. By modelling corporate ownership and control networks as graphs and applying measures such as community structure and node centrality, the method detects clusters of companies and individuals whose relationships differ markedly from benign business networks.

AUTHORS

T. Ruzgas, K. Armonaitė

[1] Kaunas University of Technology, Lithuania

[2]