Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.39490

Accounting Anomaly Detection Using Relational Graph Convolutional Networks over Heterogeneous Financial Knowledge Graphs

Xueqian Liu

Faculty of Social Science & Public Policy, King's College London, London, UK
wellington589125@gmail.com

Abstract. Accounting anomalies can occur because of relational dependencies which can be missed by independent firm-year feature vectors. In this research, the use of a heterogeneous financial knowledge graph along with R-GCN for anomaly detection is considered. Financial data is retrieved from SEC EDGAR filings and US-GAAP XBRL facts in the period 2010–2024. Enforcement releases are used to indicate the periods of misconduct. Graph nodes correspond to firm-years, accounting accounts, executives, auditors, and industries and edges of five relation types connect them. The inputs include thirty-seven financial features and entity embeddings, which are processed with the help of three-layer R-GCN with relation-specific aggregation of messages and then classified with a multi-layer perceptron classifier. Weighted binary cross-entropy is used as the loss function to handle the problem of imbalanced classes. Time-aware graph snapshots and chronological splitting are performed to avoid information leakage. GNNExplainer is utilized to produce explanatory subgraphs which undergo perturbation-based stability assessment. With ten random seeds, the obtained results are as follows: AUPRC – 0.486 ± 0.021, AUROC – 0.913 ± 0.009, F1-score – 0.571 ± 0.018. The performance is better than that of graph-feature XGBoost by 0.065 ± 0.014 in AUPRC. Ablation studies show that removal of auditor relations leads to AUPRC of 0.443 ± 0.020, while removal of temporal relations causes AUPRC of 0.451 ± 0.018. Thus, contribution of relations to detection performance can be seen from ablation results.

Keywords: Relational graph convolutional networks, Heterogeneous financial knowledge graphs, Accounting anomaly detection, GNNExplainer, Class imbalance learning

1. Introduction

The detection of accounting anomalies plays a key role in ensuring the quality of financial reporting as unusual reporting distorts the estimation of the firm's operating performance, its leverage, cash flow and risk profile. Several systematic reviews have concluded that the application of machine-learning techniques can improve the fraud detection in financial statements, while the current solutions mostly convert each firm-year observation into an independent vector of ratios or textual information [1]. The existing body of literature has identified several problems including severe

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.39490

class imbalance, non-uniform validation procedures and lack of cross-comparability of the data, especially for the classification task based on rare misconduct labels [2]. This is a crucial issue as the financial reporting process takes place within an institutional environment created by the manager, auditor-client relationship, industry characteristics, recurrent relationships between accounts, and previous year financial status.

The recent academic research has tried to introduce non-financial governance factors into the fraud prediction models. Machine-learning studies for U.S. listed firms demonstrate that the information about the firm's management and its organizational characteristics could add value for prediction beyond the traditional financial factors [3]. Dynamic ensemble learning proves that the combination of financial and non-financial features can be beneficial for prediction for the case of nonlinear decision boundary [4]. Yet, even the introduction of new tabular data leaves the relational context implicit.

In order to overcome this limitation, the present research introduces a heterogeneous financial knowledge graph, where firm-year observations are connected to accounting concepts, individuals in managerial position, the auditor-client relationship, industry and previous fiscal year.

2. Literature review

2.1. Accounting anomaly detection

Current fraud research within the finance domain has shifted from ratio-based analysis to approaches using ensemble learning, deep learning, and graphs. Several extensive surveys on applications of graph neural networks for financial fraud detection confirm the effectiveness of relational learning particularly for instances where suspicious activity is dispersed among interconnected entities rather than focused on one record [5]. Likewise, the studies on the detection of fraudulent financial statements reveal the importance of relationships between suppliers and customers to unveil signs of fraud that are not revealed based on data of a focal firm alone [6]. Nonetheless, listed-firm anomaly detection is still dominated by firm-level classification with models being trained using random splits which tend to overestimate the model performance due to the involvement of information of future years. Imbalance is an important issue since enforcement actions only make up a small part of all firm-years.

2.2. Financial knowledge graphs

Knowledge graphs allow for an explicit representation of accounting semantics and inter-entity structures. There has been recent evidence suggesting that a graph-based model of financial statements could be able to incorporate both semantic and articulation relations to form more interpretable patterns of fraud [7], and that multi-relational graph learning could incorporate multiple relationships in companies instead of using a single adjacency matrix [8]. More general literature review also points out that heterogeneous and temporal GNNs should be studied further in the case of financial fraud because there are different types of relations that have their own meaning and change through time [9, 10]. It is natural to expect a graph where different entities, including accounting accounts, governance actors, service relationships, industries, and reporting periods are separable.

DOI: 10.54254/2755-2721/2026.39490

3. Methodology

3.1. Sample and anomaly label construction

The sample consists of companies listed on the U.S. exchanges in fiscal years between 2010 and 2024. Annual 10-K filings and structured US-GAAP XBRL facts are collected from SEC EDGAR with the help of firm's CIK number, which is used as an entity identifier for the whole sample. Data points with invalid fiscal year-end date, duplicate amendments, and structurally incomparable financial institutions are filtered out. SEC Accounting and Auditing Enforcement Releases are analyzed to extract firm identifiers, dates of enforcement actions, and fiscal periods in connection with which these actions are taken because of materially false financial reporting. Only those observations where the enforcement document establishes such connection between misconduct and reporting period are labeled with an anomaly value of 1. In order to minimize contamination, observations with unlabeled anomaly from one year before to one year after a known misconduct period are not included in the negative set.

3.2. Accounting feature engineering

Facts about XBRL are linked to a limited number of common accounting concepts concerning revenues, accounts receivable, inventories, operating income, assets, liabilities, stockholder's equity, operating cash flows, capital expenditure, and some expense items. Extended tags are solved through calculation linkbases and statements context. Python 3.11, pandas, and NumPy are utilized to calculate 37 firm-years characteristics such as total accruals intensity, growth of accounts receivable, growth of inventories, gross margin variation, sales/cash ratio, income/cash flow ratio, leverage, asset structure, return on assets, and year-on-year volatility of statements. All continuous variables are trimmed at the 1% and 99% quantiles and normalized by two-digit SIC industry-year groups. The missing values are replaced with industry-year medians while keeping a corresponding missingness mask in the model.

3.3. Heterogeneous knowledge graph construction

The graph is created using Neo4j 5 and exported into PyTorch Geometric as a typed multigraph. Five node types are kept identical to the definitions used in the framework: firm-year, accounting account, executive, auditor, and industry. The graph has edges of type reports_account between firm-years and US-GAAP concepts, edges of type managed_by to named executives, edges of type audited_by to audit firms, edges of type belongs_to_industry to industry nodes, and edges of type previous_year between consecutive firm-year observations of the same issuer. Accounting amounts, normalized changes, filing lag, and missingness indicators are recorded as attributes in the graph. Edge creation is fully time aware such that for the prediction year yy, only relationships observable until the filing date are included in the graph. Relation-specific sampling limits accounting account neighbors to 32 per layer but keeps other relations intact.

3.4. R-GCN anomaly detection model

The three-layer R-GCN is coded using PyTorch 2.2 and PyTorch Geometric. The firm-year nodes start off with the 37 financial feature variables and file metadata information, while the other nodes use a combination of the trainable 64-dimensional embedding vector and any existing attributes, and combine the messages with the self-representation of the node as shown in Equation (1) - (2):

DOI: 10.54254/2755-2721/2026.39490

$$h_i^{(l+1)} = \sigma(W_0^{(l)} h_i^{(l)} + \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \frac{\alpha_{ij}^r}{c_{i,r}} W_r^{(l)} h_j^{(l)}) \quad (1)$$

Here, $h_i^{(l)}$ denotes the representation of node ii at layer l, $\mathcal{N}_i^r$ is its neighborhood under relation r, The dimensions for hidden layers are 128, 128, and 64 with ReLU activation function and layer normalization and 0.30 dropout. The resulting firm-year embeddings go to a 64-32-1 multilayer perceptron with sigmoid output. Weighted binary cross-entropy is used with the inverse of the ratio of prevalence in the training set, and AdamW optimizer uses 0.001 learning rate, 0.0005 weight decay, and gradient clipping at 2.0 (see Figure 1).

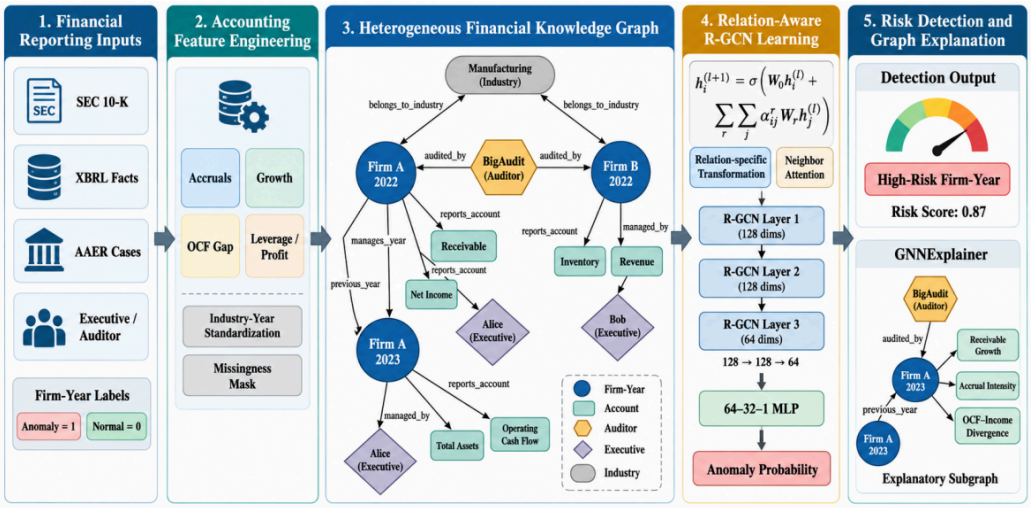


Figure 1. Heterogeneous financial knowledge graph framework for accounting anomaly detection

3.5. Graph-based anomaly explanation

Predicted values that surpass the threshold determined by the validation procedure are interpreted using GNNExplainer. The explanations are derived on the basis of two-hop ego networks, limited to keep no more than 15 edges and 10 node features. Feature and relation attributions are distinguished in order to separate abnormal accounting features from the relational exposure. Stability of the explanations is assessed by measuring the edge-set Jaccard similarity over 20 perturbation runs. If the stability value is less than 0.60, the case is considered explanation-uncertain.

4. Experimental design and validation

4.1. Temporal detection experiment

The analysis proceeds chronologically, with 2010-2019 as the training period, 2020-2021 as the validation period, and 2022-2024 as the out-of-sample period. Graph snapshots are constructed anew for each predicted year so that future executives, audits, financial values, and enforcement actions do not leak backwards. Hyperparameters are chosen by AUPRC during validation, employing early stopping if there are no improvements after 20 iterations for up to 300 iterations. Neural models are trained using 10 random seeds; therefore, the reported neural results are presented as mean ± standard deviation.

DOI: 10.54254/2755-2721/2026.39490

4.2. Benchmark and ablation tests

Comparison of the R-GCN is made against the following models: the Beneish M-Score, the Isolation Forest, the XGBoost on the engineered table features, and XGBoost on the graph features with degree, neighbor-risk prevalence, auditor concentration, executive recurrence, and previous year summary statistics. The XGBoost models are optimized over standard tree parameters and imbalance handling methods. The ablation study eliminates relations in auditor, executive, industry, accounting account, and temporal edges relation families at a time. There is an additional financial only R-GCN with removal of all relations but the accounting accounts.

4.3. Performance and robustness validation

The AUPRC serves as the first measure due to the low proportion of anomalies within the test set. AUROC, precision, recall, F1, and precision@5% are additionally provided. The 95% confidence intervals are derived from 1,000 firm-clustered bootstrapped samples that retain the same firm-years within each sample. AUPRC differences between pairs are tested using bootstrap percentile methods and Holm correction. The robustness exercise is replicated for manufacturing firms, technology firms, consumer firms, service firms, and for each tercile by market capitalization. Case explanations are carried out for anomalies found within the sample only.

5. Results

5.1. Accounting anomaly detection and benchmark comparison

The R-GCN achieved the strongest out-of-sample performance, with AUPRC of 0.486±0.021, AUROC of 0.913±0.009, and F1-score of 0.571±0.018. XGBoost produced an AUPRC of 0.391±0.017 and graph-feature The R-GCN improvement over graph-feature XGBoost was 0.065±0.014 in AUPRC, with a 95% bootstrap interval of 0.039–0.091 and adjusted $p<0.001$. Precision@5% increased to 0.638±0.026, compared with 0.521±0.024 for graph-feature XGBoost (see Table 1).

Table 1. Out-of-sample detection performance and key ablation results

Model / Variant	AUPRC	AUROC	Precision	Recall	F1-score	Precision@5%
Beneish M-score	0.214± 0.012	0.742± 0.011	0.281± 0.018	0.417± 0.021	0.336± 0.017	0.302± 0.020
XGBoost	0.391± 0.017	0.872± 0.010	0.472± 0.021	0.536± 0.023	0.502± 0.019	0.481± 0.023
Graph-feature XGBoost	0.421± 0.019	0.889± 0.010	0.501± 0.022	0.558± 0.024	0.528± 0.020	0.521± 0.024
R-GCN without auditor edges	0.443± 0.020	0.895± 0.010	0.512± 0.023	0.566± 0.023	0.537± 0.020	0.574± 0.025
R-GCN without temporal edges	0.451± 0.018	0.899± 0.009	0.519± 0.021	0.571± 0.021	0.544± 0.018	0.586± 0.024

** Adjusted $p<0.001$ versus graph-feature XGBoost for the full R-GCN.*

DOI: 10.54254/2755-2721/2026.39490

5.2. Heterogeneous relation ablation and robustness performance

The removal of auditor relationships had the largest impact on the AUPRC, which declined from 0.486±0.021 to 0.443±0.020, followed by the removal of temporal edges at 0.451±0.018. Removal of relationships between executives, accountants, and industry resulted in AUPRC scores of 0.459±0.020, 0.463±0.019, and 0.472±0.020, respectively. The pure finance-only R-GCN model had an AUPRC score of 0.414±0.018, which was a confirmation of the value provided by the governance and professional relations added information beyond graph-structured accounts.

5.3. Graph explanation of high-risk firm-year cases

The correctly identified anomaly cases resulted in the creation of explanation graphs having 8.7±2.4 edges retained and 6.1±1.8 features. Explanation graph stability obtained during perturbation trials was 0.742±0.061. Financial signals appearing frequently include growth in receivables, accrual aggressiveness, difference between operating cash flow to income, and growth in inventory. The edges appearing highly with respect to their attribution scores include audited_by and previous_year with scores of 0.284±0.047 and 0.231±0.041. This means that in the case of high confidence true positives, there is no use of extreme ratios by the model.

6. Conclusion

The paper shows that listing firms' accounting anomalies can be effectively identified by considering financial reporting as a heterogeneous graph structure and not as a separate firm-year classification task. The suggested R-GCN model, which includes XBRL features together with relations between accounts, executives, auditors, industries, and time, outperforms all ratio-based, unsupervised. Ablation analysis reveals that the auditor and temporal relations have the greatest effect on the performance of the R-GCN model.

References

[1] Ashtiani, M. N., & Raahemi, B. (2022). Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review. *IEEE Access, 10*, 72504–72525.
[2] Shahana, T., Lavanya, V., & Bhat, A. R. (2023). State of the art in financial statement fraud detection: A systematic review. *Technological Forecasting and Social Change, 192*, 122527.
[3] Schneider, M., & Brühl, R. (2023). Disentangling the black box around CEO and financial information-based accounting fraud detection: Machine learning-based evidence from publicly listed U.S. firms. *Journal of Business Economics, 93*, 1591–1628.
[4] Achakzai, M. A. K., & Peng, J. (2023). Detecting financial statement fraud using dynamic ensemble machine learning. *International Review of Financial Analysis, 89*, 102827.
[5] Motie, S., & Raahemi, B. (2024). Financial fraud detection using graph neural networks: A systematic review. *Expert Systems with Applications, 240*, 122156.
[6] Li, J., Chang, Y., Wang, Y., & Zhu, X. (2023). Tracking down financial statement fraud by analyzing the supplier-customer relationship network. *Computers & Industrial Engineering, 178*, 109118.
[7] Cai, S., & Xie, Z. (2024). Explainable fraud detection of financial statement data driven by two-layer knowledge graph. *Expert Systems with Applications, 246*, 123126.
[8] Wang, C., Wang, M., Wang, X., Zhang, L., & Long, Y. (2024). Multi-relational graph representation learning for financial statement fraud detection. *Big Data Mining and Analytics, 7*(3), 920–941.
[9] Cheng, D., Zou, Y., Xiang, S., & Jiang, C. (2025). Graph neural networks for financial fraud detection: A review. *Frontiers of Computer Science, 19*(9), 199609.
[10] Mashiko, S., Kawamata, Y., Nakayama, T., Sakurai, T., & Okada, Y. (2025). Anomaly detection in double-entry bookkeeping data by federated learning system with non-model sharing approach. *Scientific Reports, 15*, 42208.