Graph Representational Learning for Internal Audit

Pai, Sumit[1,*,†], Singh, Vivek Kumar[1,†], Gupta, Sanvi[1], Chavali, Pavani[1], Siddhartha, Siddhartha[1], Bowen, Edward[2] and Tiyyagura, Sunil Reddy[1]

[1]*Deloitte & Touche Assurance & Enterprise Risk Services India Private Limited*
[2]*Deloitte & Touche LLP*

Abstract

This work aims to improve the quality of Internal Audits (IA) that are a critical part of an organization's governance structure and serves as third line of defense helping provide assurance that the controls and processes have adequate risk mitigation strategies in place. We focus on AI enabled internal audits that could improve the quality, coverage and time needed to perform them and thus improve the effectiveness and efficiency of providing assurance, to help auditors identify potential risks that may go unnoticed through traditional methods. We compare different AI methodologies that can be used in controls testing for various financial and corporate processes. We propose the use of Knowledge Graphs (KGs) and representational learning to leverage the inherent relational nature of the data and to identify potential non-compliance or fraud. The experimental results demonstrate that our proposed method exhibits a significant improvement in F1 score, outperforming standard outlier detection approaches, reducing the number of False Positives (FPs) and in turn the manual review involved.

Keywords

Internal Audit, Controls testing, Knowledge Graphs, Representation Learning

Introduction. The Internal Audit function's primary responsibility is to evaluate and advise on risk management and the related effectiveness of internal controls across an organization, including financial, operational, regulatory, IT and strategic domains. Controls are typically a set of defined processes designed to mitigate risk and some common business processes for which controls testing is done are accounts receivable/payable, employee expenses, payroll, supply chain, etc. In this paper, we specifically discuss the use-case of employee expenses.

Problem Statement and proposed solution. The key challenges with rules-based approach to controls testing are: limited scope with predefined rules that may not cater to each of the business environments, too many false positives and false negatives, scalability, manual effort especially in identifying high risk samples for testing, and heavy reliance on Subject Matter Specialists (to eliminate False Positives) to name a few. Standard outlier detection approaches, such as Isolation Forests (IF) and AutoEncoders (AE), as shown in Table 1, usually don't work well due to distribution shifts, label imbalance, and not being able to leverage the relational nature of the data (due to independent and identically distributed (i.i.d) assumption between samples). We are working towards a solution using KGs to leverage this inherent relational aspect, as well as capture the domain knowledge and thus improve upon these methods.

Knowledge Graph Design. Given a tabular representation of the time and expense dataset,

ISWC-2023: The 22nd International Semantic Web Conference
*Corresponding author.
†These authors contributed equally.
sumpai@deloitte.com (P. Sumit)

we identify relationships between the columns, and model the data as a KG. The KG Schema has five primary nodes in red, as shown in the Fig 1, each of which are described by their respective attributes in green (e.g. transaction amount for the transaction identifier node). The dataset contains continuous, discrete and textual columns, each of which are incorporated in the graph with appropriate pre-processing steps: Continuous values are binned, textual attributes are cleaned, split into keywords and semantically similar words are connected using Bidirectional Encoder Representations from Transformers (BERT)-based word embeddings. We then use a semi-supervised setup, where a small fraction of transactions ($< 1\%$) are noisy labelled as fraudulent based on a small set of controls and are assigned an edge in the KG.

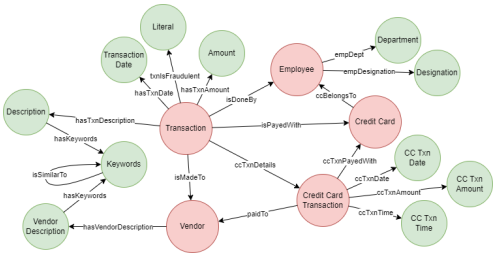

Figure 1: Representative KG Schema. Primary nodes are in red and their attributes in green

Model	Precision	Recall	F1
IF	0.25	0.45	0.32
AE	0.32	0.55	0.40
KGE	**0.59**	**0.55**	**0.57**

Table 1: Classification Results for Fraudulent Transactions.

Graph Representational Learning. We leverage the relational modeling power of graphs and learn representations of nodes and edges by propagating this relational information using Knowledge Graph Embedding (KGEs) models[1]. The trained model is calibrated on a held-out set which is made up of fraudulent and non-fraudulent transactions. The classification threshold is chosen such that it maximizes the F1 on this set and using this threshold the performance is measured on a test set. Both these sets are carved to be representative of the true data distribution where fraudulent transactions are expected to be have a very small percentage.

Results and Conclusion Compared to the other two approaches (IF and AE), as shown in Table 1, we clearly see the benefit of relational modeling with the KGs, as we were able to achieve an F1 score of 0.57 on identifying the fraudulent transactions. We provided an overview of IA and a related use-case highlighting the potential benefits of employing semantic modeling and learning based approaches to enhance controls testing. With continuous monitoring, instances of FPs would reduce, enabling greater confidence across the 3 lines of defense.

Future Work While the results of transductive models from [1] are promising, we need to retrain them from scratch as we get new batches of data due to the presence of unseen symbolic nodes. So we plan to explore inductive models, where we can approximate unseen symbolic nodes during inference thus saving huge computational costs of retraining.

References

[1] L. Costabello, S. Pai, C. L. Van, R. McGrath, N. McCarthy, P. Tabacof, AmpliGraph: a Library for Representation Learning on Knowledge Graphs, 2019.

[2] P. L. Tang, T. D. Le Pham, T. B. Dinh, Tree-Based Credit Card Fraud Detection Using Isolation Forest, Spectral Residual, And Knowledge Graph, in: MLODS, 2023.