Detecting Anomalous Business Ownership with Graph Convolutional Neural Networks

Project Proposal

Dominic Thorn

Contents

Introduction

Background

In October of 2021, The International Consortium of Investigative Journalists (ICIJ) revealed the findings of their Pandora Papers investigation. Through examination of nearly 12 million confidential business records, the ICIJ found evidence implicating thousands of individuals and businesses in efforts to conceal the ownership of companies and assets around the world (ICIJ 2021). The intentions behind this secrecy varied from legitimate privacy concerns to criminal activities, including money laundering, tax evasion, and fraud (European Union Agency for Law Enforcement Cooperation 2021).

To put these numbers in perspective, a 2019 study by the European Commission estimated that a total of USD 7.8 trillion was held offshore as of 2016. The share of this attributed to the European Union (EU) was USD 1.6 trillion, which corresponds to an estimated tax revenue loss to the EU of EUR 46 billion (European Commission. Directorate General for Taxation and Customs Union. 2019).

Identifying the beneficiaries of a company is challenging due to the ease with which information can be concealed or simply not declared. Further complication is introduced by the interconnected nature of businesses and individuals, as well as the ingenuity of criminals in masking illicit activity. These difficulties place significant strain on the resources of law enforcement agencies and financial institutions (Steven M. 2019).

In April 2016, the United Kingdom made it mandatory for businesses to keep a register of People with Significant Control. This includes people who own more than 25% of the company's shares ("Keeping Your People with Significant Control (PSC) Register" 2016). Ownership data is curated and processed by the Open Ownership organisation for the purposes of public scrutiny and research ("Open Ownership" n.d.). It is the data provided by Open Ownership that forms the basis of this study. Details of suspicious or illegitimate business owners are not readily available due to the sensitive nature of such records. We propose a method for simulating anomalous ownership structures as part of our experimental methods.

To model the complex network of global business ownership, it is necessary to represent companies, people, and their relationships in a graph structure. With data in this format, it is possible to consider the features of an entity's local neighbourhood when making a decision, in addition to the entity's own characteristics. Anomaly detection algorithms that operate on graph structures remain at the frontier of machine learning research.

To the best of the author's knowledge, there is indeed no published research studying the effectiveness of graph anomaly detection techniques on business ownership networks. The following proposal is a study into the application of state of the art anomaly detection techniques to business ownership graphs.

Project Title

The proposed title for this project is “Detecting Anomalous Business Ownership with Graph Convolutional Neural Networks”.

Aims, Objectives and Research Questions

Aims

The primary aim of this project is to develop an effective approach for detecting anomalous entities in a business ownership network. Second, the project will offer a comparison of Graph Convolutional Neural Network (GCN) models to traditional anomaly detection approaches on a business ownership graph. Finally, we contribute a dataset describing a real business ownership network that is suitable for graph learning.

Objectives

- Compose a business ownership graph from open data sources.
- Train and evaluate a state of the art GCN for anomaly detection.
- Perform the anomaly detection task with traditional machine learning methods.
- Compare approaches in terms of effectiveness and applicability.

Research Questions

The questions driving the research are as follows:

- What is the most effective strategy for detecting anomalous entities in business ownership networks?
- How do GCN models compare to traditional approaches in terms of classification performance?
- What are the challenges that arise in building and training a GCN model and what recommendations can be made to mitigate these?

Literature Review

There are few published studies focussed specifically on the detection of anomalous business ownership structures. We offer a review of the most relevant literature below.

Detecting Financial Fraud

Luna et al. (2018) describe a procedure for identifying suspected shell company accounts using distance and density based anomaly detection techniques. The authors were successful in detecting shell companies through observing differences in transactional behaviour. A notable caveat is that the data was simulated for the purposes of the study, which leaves questions around its applicability to real world scenarios.

Recent work by Dumitrescu, Bălțoiu, and Budulan (2022) demonstrates how local neighbourhood features and statistical scores can be used in Anti-Money Laundering (AML) models. Relevant features included unsupervised anomalous node detection techniques (Akoglu, McGlohon, and Faloutsos 2010) and local neighbourhood connectivity features (Molloy et al. 2016) calculated on *reduced egonets*. A strength of the study is that it was conducted on genuine labelled transactional data with positive results. The authors did not implement any GCN or other deep learning approaches for comparison.

Fronzetti Colladon and Remondi (2017) explore a range of social network analysis techniques for the identification of money laundering using data kept by an Italian factoring company. The authors found that constructing many networks from different projections of the graph entities improved the power of individual risk metrics. Degree centrality was determined to be a significant risk predictor in all cases, while in certain scenarios network constraint proved to be informative. It should be acknowledged that the results obtained are for the clients of a single business and that additional work is required to demonstrate wider validity.

Anomaly Detection for Graphs

Akoglu, Tong, and Koutra (2015) highlight four main reasons for the suitability of graph structures in anomaly detection:

Inter-dependent nature of the data – "Data objects are often related to each other and share dependencies." This can be observed in business ownership data through the relationships that connect individuals and companies in legal hierarchies and communities.

Powerful representation – Graphs offer a powerful way of representing inter-dependencies and long range correlations between related entities. By using different node and edge types, as well as additional attributes, it is possible to represent rich datasets. These properties are valuable in capturing the different types of entities present in a business ownership graph. A business, for example, will have attributes that are not shared by individuals, such as an industry classification code.

Relational nature of problem domains – "The nature of anomalies could exhibit themselves as relational". In context of detecting anomalous business ownership, it is evident that individuals and businesses may be anomalous predominantly through their unusual relationships to other entities.

Robust machinery – "Graphs serve as more adversarially robust tools." It is suggested that graph based systems are ideally suited for fraud detection, as bad actors will find it difficult to alter or fake their position in the global structure.

A thorough description of graph anomaly detection tasks and approaches is offered by Ma et al. (2021). Their taxonomy categorises tasks based on the graph component being targeted: nodes, edges, subgraphs, or full graphs. The authors state their belief that "because the copious types of graph anomalies cannot be directly represented in Euclidean feature space, it is not feasible to directly apply traditional anomaly detection techniques to graph anomaly detection".

Anomalous Node Detection with GCNs

Kipf and Welling (2017) propose a scalable GCN architecture for classifying nodes in a partially labelled dataset. Early attempts to apply deep learning to graph structures utilised RNN architectures which prove difficult to scale (Gori, Monfardini, and Scarselli 2005; Scarselli et al. 2009; Yujia Li et al. 2017). Kipf et al. extend prior work on spectral GCNs (Bruna et al. 2014; Defferrard, Bresson, and Vandergheynst 2017) to produce a flexible model that scales in linear time with respect to the number of graph edges.

Ding et al. (2019) combine a GCN architecture with an autoencoder in a method that identifies anomalous nodes by reconstruction error. The proposed method, DOMINANT, uses a GCN to generate node embeddings and separately reconstructs both the graph topology and the node attributes. This strategy is further developed and applied to multi-view data through the combination of multiple graph encoders (Peng et al. 2022).

An alternative method is offered by Yuening Li et al. (2019), in which a spectral convolution and deconvolution framework is used to identify anomalous nodes in conjunction with a density estimation model. The approach continues to demonstrate the importance of combining multiple perspectives of the network data, with the innovation being the use of a Gaussian Mixture Model to combine representations in a single view.

Recent Developments

Veličković et al. (2018) demonstrate the use of self-attention layers to address shortcomings in the representations captured by GCN architectures. However, a comparison of Relational Graph Attention (GAT) models to GCNs showed that relative performance was task dependent and that current GAT models could not be shown to consistently outperform GCNs on benchmark exercises (Busbridge et al. 2019).

In an application of graph attention based models to financial fraud detection, Wang et al. (2019) show that their SemiGNN model outperforms established approaches when predicting risk of default and in

attribute prediction. Baseline methods used for comparison included a XGBoost (Chen and Guestrin 2016), GCN, GAT, and LINE (Tang et al. 2015).

Methods

Data

Open Datasets

The data needed for this study is made publicly available by the UK government via Companies House. Additional processing and entity resolution is applied to UK Persons of Significant Control data by the Open Corporates organisation, and so PSC data is extracted from their own database. A summary of the data requirements is presented below.

Table 1: Data Requirements

Data	Provider	Date Retrieved
Company Data	Companies House	2022-05-24
Persons of Significant Control	Companies House	2022-05-24
Open Ownership Data	Open Ownership	2022-05-24

Simulated Data

Due to the absence of labelled shell or otherwise illegitimate company data, we are required to simulate anomalous business ownership networks in order to provide training examples for machine learning. While the full details of this procedure are the subject of ongoing experimentation, the high level steps are:

1. Identify connected graph components within the Open Ownership dataset.
2. Split these components into two groups, one to serve as the legitimate network set, the other as the anomalous network set.
3. Create anomalous networks within the anomalous network set by transferring one or more nodes from one connected component to another. The transferred nodes will be labelled as anomalous for the purposes of training.

Research Instruments and Tools

Data Preparation

In order to process the large volumes of records provided by UK companies, we will use the Apache Spark Framework (“Apache Spark: A Unified Engine for Big Data Processing: Communications of the ACM: Vol 59, No 11” n.d.) for reading and transforming the data. To construct and manipulate graphs in an efficient manner, we use the GraphFrames packages for Spark (“Overview - GraphFrames 0.8.0 Documentation” n.d.).

Graph Analysis

A number of software libraries and applications will be considered for analysing graph properties and generating graph features. These are:

- NetworkX
- igraph
- Neo4j

Deep Learning

The most widely used framework for deep learning on graphs is currently PyTorch-Geometric. A number of SOTA algorithms are made available through the library, as well as detailed documentation and tutorials.

Baseline Machine Learning Models

A selection of established machine learning algorithms will be used to benchmark performance.

- Random Forest
- Gradient Boosted Tree
- Logistic Regression

Performance Evaluation

A variety of classification metrics will be used to assess model performance. Many of these can be found in the Scikit Learn library.

Ethical Considerations

All data used in the course of this project has been made publicly available for the purposes of public scrutiny and academic research.

In the interests of personal privacy, any sensitive personal information will be transformed and / or removed from the dataset as appropriate. Specifically, we will remove address information, individual dates of birth (to be transformed to an age range), and the names of individual persons and companies.

Further to this, the study will refrain from speculating on the nature of any particular business structure or entity, and any speculation with regards to the legitimacy or legality of a particular arrangement is deemed strictly out of scope and beyond the remit of the study.

To facilitate academic openness and collaboration, the instructions and code required to produce the datasets and experimental results will be made available where it is acceptable to do so.

Anticipated Outcomes

We expect that GCN models will outperform traditional anomaly detection techniques in identifying anomalous business ownership networks. It is supposed that due to the nature of the simulated graphs, the anomalous entities will be difficult to identify from their individual attributes and local neighbourhood properties.

Project Plan

Roadmap

Task	Status
Data acquisition	Completed
Data understanding	Completed
Data preparation	Completed
Anomaly Simulation	Not Started
Feature engineering	Not Started
Preprocessing	Not Started

Task	Status
Modelling	Not Started
Evaluation	Not Started

Risks and Challenges

From reading the experiences of other researchers during the literature review, it is apparent that training a GCN may be complicated and resource intensive. It may be difficult to train a model of reasonable size on a single laptop device, and so care has been taken during the data preparation to ensure that all steps should be easily replicable on a cloud hosted virtual machine. This will allow for possible future requirements to upscale the processing power and even employ distributed training methods if feasible.

The proposed approach for simulating anomalous networks is as yet untested. It may arise that these anomalous networks are trivially easy to distinguish, leading to no differentiation in the performance of the benchmark algorithms and GCN model. On the other hand, the anomalous networks may not be sufficiently distinguishable from legitimate networks, leading to all algorithms failing to learn any mapping and little differentiation in algorithm performance. Should this be the case, additional research may need to be undertaken in order to devise a more suitable anomaly simulation process or to identify other sources of training data.

Wordcount: 2,166

References

Akoglu, Leman, Mary McGlohon, and Christos Faloutsos. 2010. “Oddball: Spotting Anomalies in Weighted Graphs.” In *Advances in Knowledge Discovery and Data Mining*, edited by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, et al., 6119:410–21. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13672-6_40.

Akoglu, Leman, Hanghang Tong, and Danai Koutra. 2015. “Graph Based Anomaly Detection and Description: A Survey.” *Data Min Knowl Disc* 29 (3): 626–88. https://doi.org/10.1007/s10618-014-0365-y.

"Apache Spark: A Unified Engine for Big Data Processing: Communications of the ACM: Vol 59, No 11." n.d. Accessed June 13, 2022. https://dl.acm.org/doi/10.1145/2934664.

Bruna, Joan, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2014. "Spectral Networks and Locally Connected Networks on Graphs." arXiv:1312.6203. arXiv. https://doi.org/10.48550/arXiv.1312.6203.

Busbridge, Dan, Dane Sherburn, Pietro Cavallo, and Nils Y. Hammerla. 2019. "Relational Graph Attention Networks." arXiv:1904.05811. arXiv. https://doi.org/10.48550/arXiv.1904.05811.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. https://doi.org/10.1145/2939672.2939785.

Defferrard, Michaël, Xavier Bresson, and Pierre Vandergheynst. 2017. "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering." arXiv:1606.09375. arXiv. https://doi.org/10.48550/arXiv.1606.09375.

Ding, Kaize, Jundong Li, Rohit Bhanushali, and Huan Liu. 2019. "Deep Anomaly Detection on Attributed Networks." In, 594–602. https://doi.org/10.1137/1.9781611975673.67.

Dumitrescu, Bogdan, Andra Băltoiu, and Ştefania Budulan. 2022. "Anomaly Detection in Graphs of Bank Transactions for Anti Money Laundering Applications." *IEEE Access* 10: 47699–714. https://doi.org/10.1109/ACCESS.2022.3170467.

European Commission. Directorate General for Taxation and Customs Union. 2019. *Estimating International Tax Evasion by Individuals.* LU: Publications Office. https://data.europa.eu/doi/10.2778/300732.

European Union Agency for Law Enforcement Cooperation. 2021. *Shadow Money: The International Networks of Illicit Finance.* https://op.europa.eu/publication/manifestation_identifier/PUB_QLAN21003ENN.

Fronzetti Colladon, Andrea, and Elisa Remondi. 2017. "Using Social Network Analysis to Prevent Money Laundering." *Expert Systems with Applications* 67 (January): 49–58. https://doi.org/10.1016/j.eswa.2016.09.029.

Gori, M., Gabriele Monfardini, and Franco Scarselli. 2005. *A New Model for Earning in Raph Domains*. Vol. 2. https://doi.org/10.1109/IJCNN.2005.1555942.

ICIJ. 2021. "Offshore Havens and Hidden Riches of World Leaders and Billionaires Exposed in Unprecedented Leak - ICIJ." https://www.icij.org/investigations/pandora-papers/global-investigation-tax-havens-offshore/.

"Keeping Your People with Significant Control (PSC) Register." 2016. *GOV.UK*. https://www.gov.uk/government/news/keeping-your-people-with-significant-control-psc-register.

Kipf, Thomas N., and Max Welling. 2017. “Semi-Supervised Classification with Graph Convolutional Networks.” arXiv:1609.02907. arXiv. https://doi.org/10.48550/arXiv.1609.02907.

Li, Yuening, Xiao Huang, Jundong Li, Mengnan Du, and Na Zou. 2019. “SpecAE: Spectral AutoEncoder for Anomaly Detection in Attributed Networks.” arXiv:1908.03849. arXiv. https://doi.org/10.48550/arXiv.1908.03849.

Li, Yujia, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2017. “Gated Graph Sequence Neural Networks.” arXiv:1511.05493. arXiv. https://doi.org/10.48550/arXiv.1511.05493.

Luna, Devendra Kumar, Girish Keshav Palshikar, Manoj Apte, and Arnab Bhattacharya. 2018. “Finding Shell Company Accounts Using Anomaly Detection.” In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, 167–74. CoDS-COMAD ’18. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3152494.3152519.

Ma, Xiaoxiao, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z. Sheng, Hui Xiong, and Leman Akoglu. 2021. “A Comprehensive Survey on Graph Anomaly Detection with Deep Learning.” *arXiv:2106.07178 [Cs]*, October. http://arxiv.org/abs/2106.07178.

Molloy, Ian, Suresh Chari, Ulrich Finkler, Mark Wiggerman, Coen Jonker, Ted Habeck, Youngja Park, Frank Jordens, and Ron Schaik. 2016. “Graph Analytics for Real-Time Scoring of Cross-Channel Transactional Fraud.” In.

“Open Ownership.” n.d. *Openownership.org*. Accessed May 24, 2022. https://www.openownership.org/en/.

“Overview - GraphFrames 0.8.0 Documentation.” n.d. Accessed June 13, 2022. https://graphframes.github.io/graphframes/docs/_site/index.html.

Peng, Zhen, Minnan Luo, Jundong Li, Luguo Xue, and Qinghua Zheng. 2022. “A Deep Multi-View Framework for Anomaly Detection on Attributed Networks.” *IEEE Transactions on Knowledge and Data Engineering* 34 (6): 2539–52. https://doi.org/10.1109/TKDE.2020.3015098.

Scarselli, Franco, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. “The Graph Neural Network Model.” *IEEE Transactions on Neural Networks* 20 (1): 61–80. https://doi.org/10.1109/TNN.2008.2005605.

Steven M., D’Antuono. 2019. “Combating Illicit Financing by Anonymous Shell Companies — FBI.” Testimony. https://www.fbi.gov/news/testimony/combating-illicit-financing-by-anonymous-shell-companies.

Tang, Jian, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. “LINE: Large-Scale Information Network Embedding.” In *Proceedings of the 24th International Conference on World Wide Web*, 1067–77. https://doi.org/10.1145/2736277.2741093.

Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. “Graph Attention Networks.” arXiv. http://arxiv.org/abs/1710.10903.

Wang, Daixin, Jianbin Lin, Peng Cui, Quanhui Jia, Zhen Wang, Yanming Fang, Quan Yu, Jun Zhou, Shuang Yang, and Yuan Qi. 2019. “A Semi-Supervised Graph Attentive Network for Financial Fraud Detection.” In *2019 IEEE International Conference on Data Mining (ICDM)*, 598–607. https://doi.org/10.1109/ICDM.2019.00070.