Detecting Anomalous Business Ownership Structures with Graph Neural Networks

Final Project

Dominic Thorn

Contents

List of Figures

List of Tables

1 Introduction

2 Prior and Related Work

3 Dataset

Training a supervised classification model requires access to a dataset of labelled examples. Given the typical infrequency of fraudulent events to legitimate cases, it is common for fraud classification projects require large amounts of data in order to provide a modest number of fraudulent examples from which the model can learn.

Considering the sensitive nature of fraud investigations, it is not surprising to find that no such data is available in the public domain.

In the following section we detail the public data sources used in this study and the steps necessary for processing them. We further propose a method for simulating anomalous business ownership structures using this publically available data.

3.1 External Sources

3.1.1 People with Significant Control Register

Since 2016, it has been a requirement for all businesses in the United Kingdom to declare People of Significant Control (PSC) (*Keeping Your People with Significant Control (PSC) Register*, 2016). This includes all shareholders with ownership greater than 25%, any persons with more than 25% of voting rights, and any person with the right to appoint or remove the majority of the board of directors (*People with Significant Control (PSCs)*, 2022). The register is available as a daily snapshot, available to download from the Companies House webpage (*Companies House*, 2022).

Included in the register are the name, address, and identification number of each company for which a declaration has been received. Also listed are the name, address, country of origin, date of birth, and nature of control for each person listed as a PSC.

Rather than consume this data directly, we obtain this information from a data feed curated by the Open Ownership organisation.

3.1.2 Open Ownership Register

The Open Ownership organisation maintains a database of over 16 million beneficial ownership records, including those collated in the UK PSC register and from additional sources made available by other countries (*Open Ownership*, 2022). This data is provided in a standardised format that is conducive to machine processing and graph generation. Additional data quality processing is undertaken by Open Ownership, such as the merging of duplicated records for persons that have more than one PSC declaration (*Beneficial Ownership Data Standard*, 2022).

The Open Ownership Register is used as a canonical source of company, person, and ownership relationships for the generation of our UK business ownership graph.

3.1.3 The Free Company Data Product

The Free Company Data Product is a monthly snapshot of data for all live companies on the UK public register. Taken from the Companies House data products website, this data includes:

- basic information including company type and registered office address
- the nature of business or standard industrial classification (SIC)
- company status, such as 'live' or 'dissolved'
- date of last accounts or confirmation statement filed
- date of next accounts or confirmation statement due
- previous company names

(*Companies House Data Products*, n.d.)

This data serves as an additional source of node features for company entities in the generated dataset. It is the ability to learn from these heterogeneous sets of features, as well as the relationships between them, that distinguishes Hetereogeneous Graph Neural Networks from other implementations.

3.2 Initial Data Preparation

The Open Ownership data file consists of over 20 million records stored as nested JSON objects. This includes data for businesses and relevant persons registered outside of the UK. The Apache Spark framework (*Apache Spark*, 2022) is used for bulk data preparation due to its parallel and out-of-core processing capabilties, as well as its support for nested data structures.

As the scope of this study covers only UK registered companies and their shareholders, records for entities that do not have a shareholding interest in a UK company are discarded. Non-UK companies that are registered as a

shareholder of a UK company are also discarded, as we are unable to obtain information for these entities via Companies House. Computational resource constraints also prevent handling of a larger dataset. To further limit dataset size, and in the interests of only considering accurate and up to date information, we also filter out companies that are listed as dissolved.

While the initial nested data schema is desirable for clean representation and compact storage, we require a flat relational table structure for analytical and model training purposes. Relevant data items are extracted into a flat table for each entity type. This results in three tables of interim output: company information, person information, and statements of control that link entities to their ownership interests.

The Companies House data is joined to the company entity table via the UK company registration number.

3.3 Graph Generation

3.3.1 Nodes and Edges

The company, person, and ownership relationship tables prepared in the previous steps are used to create a graph data structure. Companies and persons are represented as nodes in the graph, and the ownership relationships represented as directed edges (from owner to owned entity) between them. The resulting graph is attributed with node features and edge weights. Since the graph consists of two types of nodes with disitinct feature sets, it can be described as an attributed heterogeneous graph (Ma et al., 2021).

3.3.2 Connected Components

Connected components are labelled in the graph to facilitate additional filtering and to allow for parallel computation of topological features.

Two nodes, u and v, are considered to be connected if a path exists between them in the graph G. If no path exists between u and v then they are said to be disconnected. Not all entities in the ownership graph are connected by any path, and so the ownership graph G can be viewed as a set of disjoint sets of nodes, or components. The graph G is described by its components thus: $G[V_1], G[V_2], ...G[V_n]$ (Bondy & Murty, 2009, p. 13).

The vast majority of businesses in the dataset have only one or two declared PSC shareholders. The initial distribution of component sizes is shown in figure 1. These small subgraphs are not of interest in this study and are excluded, along with any component that contains fewer than ten nodes or with a ratio of less than 1 in 10 natural persons to companies.

The rationale for excluding these components is threefold: first, the small number of nodes and relationships presents little information for a model to learn from; second, these small networks are less likely to be of interest in real world fraud

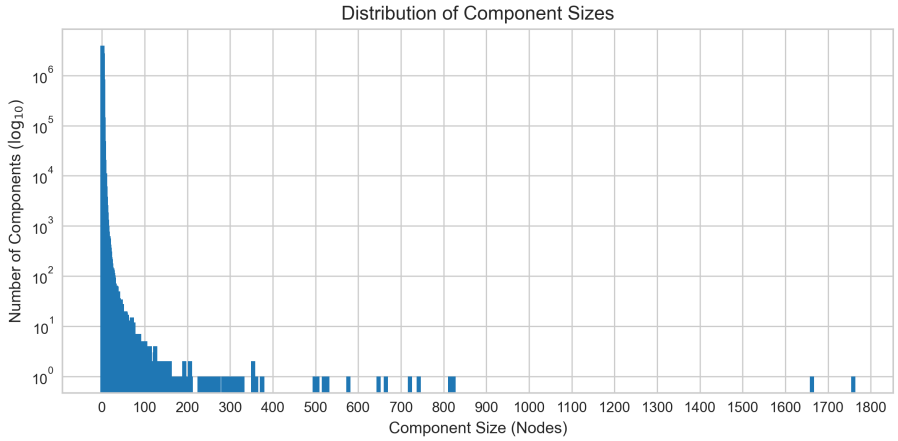

Figure 1: Initial distribution of component sizes. A single super-component consisting of 27,008 nodes is excluded from the plot.

detection, as illegitimate ownership is usually concealed behind multiple layers of ownership; and from a practical standpoint, the volume of data is brought down to a manageable size. Networks with fewer than 1 in 10 natural persons to companies are excluded, as a focus of this study is on the performance of anomaly detection methods on heterogeneous graph data, as opposed to homogeneous networks.

Finally, we address an observed data quality issue in which multiple nodes in the same component may share the same name. These nodes are merged into a single node, with the resulting node taking on all the ownership relationships of the merged nodes.

3.4 Anomaly Simulation

To simulate anomalies, 10% of the nodes are selected at random and marked as anomalous. A single outgoing edge is chosen from each anomalous node and it's source ID is replaced with that of another node marked as anomalous. This results in the anomalous nodes exchanging ownership relationships, while preserving the overall structure of the graph. The procedure is illustrated in figure 2. At conclusion, a check is performed to ensure that no anomalous nodes have been allocated their original ownership relationships.

The anomaly simulation process introduces unusual ownership relationships into the graph. The true freuqency of these occurrences is unknown, but is is expected to be far lower than 10% of nodes. A 10% anomaly rate is chosen to ensure that model training is not impossible due to extreme class imbalance, and the same effect can be achieved by oversampling the minority class (Chawla et al., 2002) or undersampling the majority class (Fernández, 2018, p. 82).

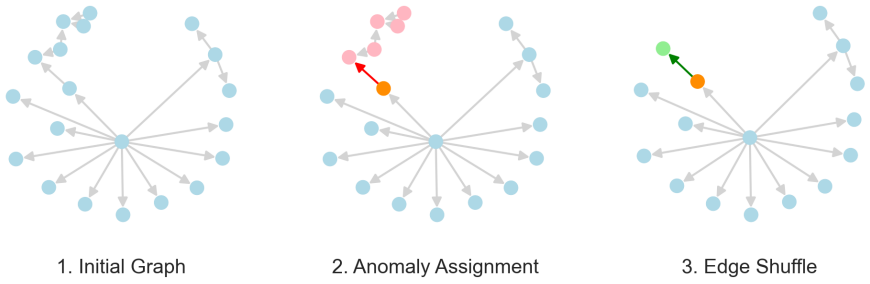

Figure 2: Process for simulating structural anomalies. The initial graph is shown in (1). A node is selected for anomalisation (orange circle) in (2). The outgoing edge (red arrow) is swapped for that of another anomalised node (green arrow), resulting in the exchange of pink nodes for green (3).

It is assumed that the data originally provided by Open Ownership is accurate and that if any anomalies are present they will have a negligible impact on experimental results. Where anomalies are present, they should have an equal effect on all models, and so should not bias the results.

3.5 Node Features

In order to train a baseline model for comparison, a set of node level features is generated to represent each node in a tabular format (Leskovec, 2021). The following topological features are extracted for each node:

- Indegree: The number of incoming edges to the node.
- Outdegree: The number of outgoing edges from the node.
- Closeness Centrality: Inverse of the sum of shortest path distances to all other nodes.
- Clustering Coefficient: Connectedness of neighbouring nodes.
- PageRank: A measure of node importance, based on the importance of neighbouring nodes (Page et al., 1998).

To capture information about the node's position in the graph, aggregate statistics are calculated for the aforementioned topological features for the node's neighbours and added as features:

- Minimum
- Maximum
- Sum
- Mean
- Standard Deviation

The count of immediate neighvours is also included as a feature.

4 Methods

4.1 Traditional Machine Learning Approaches

4.2 Graph Neural Networks

4.2.1 GraphSAGE

4.2.2 Graph Convolutional Layer

5 Results

6 Discussion

7 References

Apache Spark: A unified engine for big data processing: Communications of the ACM: Vol 59, No 11. (2022, June 13). https://dl.acm.org/doi/10.1145/2934664

Beneficial Ownership Data Standard. (2022, September 18). openownership.org. https://www.openownership.org/en/topics/beneficial-ownership-data-standard/

Bondy, A., & Murty, U. S. R. (2009). *Graph Theory.* Springer London. https://books.google.com?id=5Z71jwEACAAJ

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Companies House. (2022, June 12). http://download.companieshouse.gov.uk/en_pscdata.html

Companies House data products. (n.d.). GOV.UK. Retrieved September 18, 2022, from https://www.gov.uk/guidance/companies-house-data-products

Fernández. (2018). *Learning from Imbalanced Data Sets* (1st ed. 2018 edition). Springer.

Keeping your people with significant control (PSC) register. (2016, April 6). GOV.UK. https://www.gov.uk/government/news/keeping-your-people-with-significant-control-psc-register

Leskovec, J. (2021). *Traditional Methods for Machine Learning in Graphs.* Lecture. http://snap.stanford.edu/class/cs224w-2020/slides/02-tradition-ml.pdf

Ma, X., Wu, J., Xue, S., Yang, J., Zhou, C., Sheng, Q. Z., Xiong, H., & Akoglu, L. (2021). *A Comprehensive Survey on Graph Anomaly Detection with Deep Learning.* http://arxiv.org/abs/2106.07178

Open Ownership. (2022, May 24). openownership.org. https://www.openownership.org/en/

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank Citation Ranking: Bringing Order to the Web.*

People with significant control (PSCs). (2022, September 18). GOV.UK. https:

//www.gov.uk/guidance/people-with-significant-control-pscs

8 Appendix

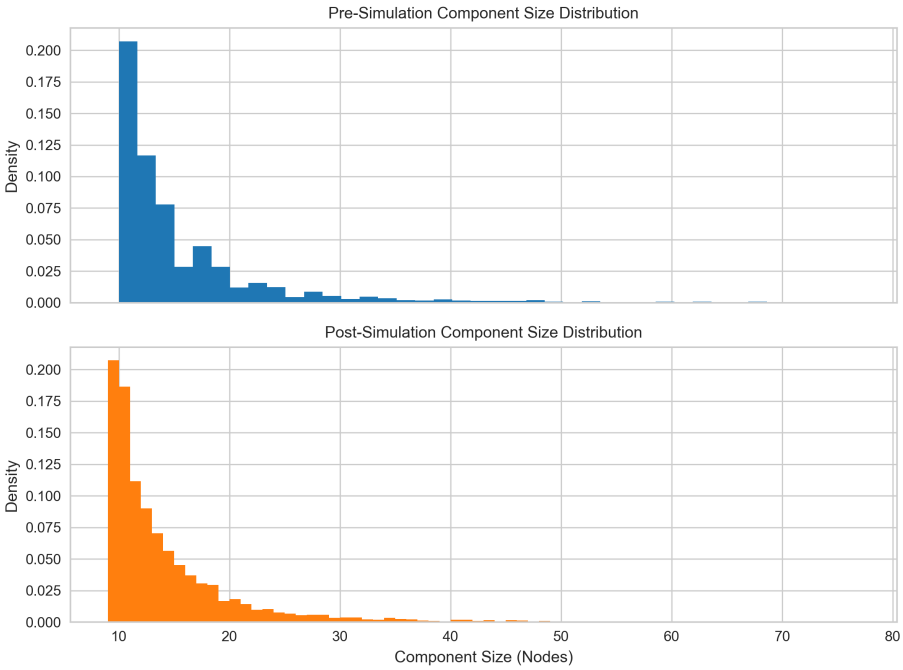

Figure 3: Distribution of component sizes before and after anomaly simulation.

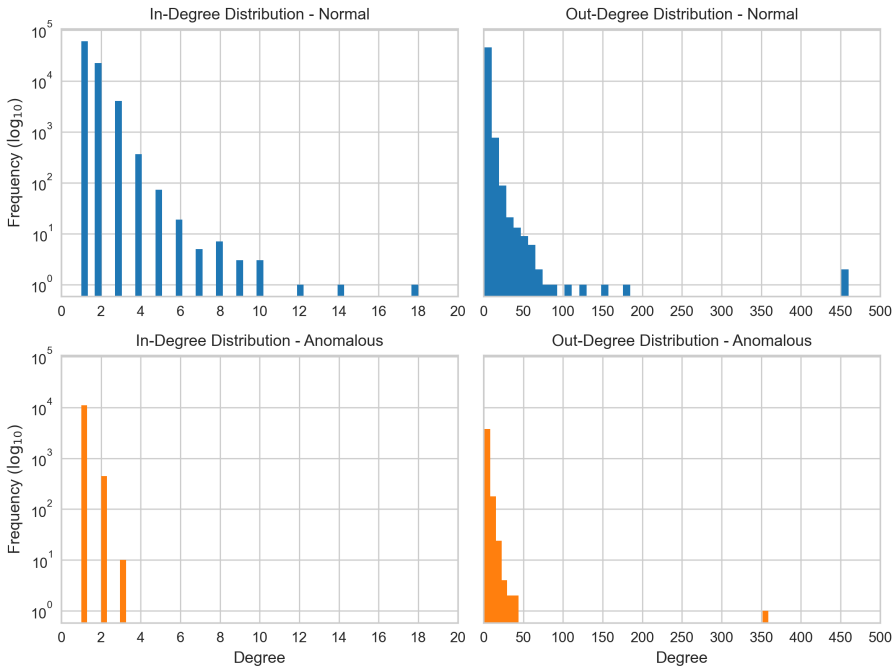

Figure 4: Distribution of node degrees before and after anomaly simulation.