Thesis Proposal

Financial Fraud-detection and Network Graphs: Its Opposition, Rationale, Effectiveness, and Improvements.

Hao Yu Wang
2666479

Vrije Universiteit Amsterdam

13th February, 2024

Introduction

- **Background**

From the cryptocurrency giant FTX misusing $10 billion of customer funds to WireCard's missing $2.1 billion, Europe's largest corporate embezzlement case in recent history, financial fraud is a prevalent and adverse aspect of any society, especially in financial institutions. Ever since the dawn of human civilization, fraudulent activities have played a pervasive role in human trade. To this day, financial fraud remains a significant and difficult obstacle to mitigate and overcome as the world economy unifies and grows at a rapid pace. It is of utmost importance that financial institutions detect and prevent fraud in order to safeguard customer finances, sensitive data, trust in the economic system, and prevent the funding of illegal activities. According to Association of Certified Fraud Examiners (ACFE), in 2020, organizations were found to have lost approximately 5% of their revenue due to fraudulent activities. Moreso, according to the Federal Bureau of Investigations (FBI), the same year, losses exceeded $4.2 billion. With scientific leaps in computer technology such as artificial intelligence, fraud is becoming increasingly difficult to detect and prevent. But how can artificial intelligence be used to counter such a threat? How can institutions utilize transactional data to create AI models capable of detecting fraud at an early stage? This is where Graph Anomaly Detection (GAD) comes in as an anomaly detection methodology which utilizes graphical representation of data, network visualization, and graph algorithms. Graph Neural Networks (GNNs) being one of the most widely-used ML algorithms when it comes to fraud detection. The thesis will aim to explore the benefits of network graph representation by comparing the results to those of non-graph algorithms. The thesis will pay particular focus on the final model's Anti-Money Laundering (AML) applications.

- **Prior Work on Topic**

Several papers to consider:

1. Useful repository of all related papers
    - Related papers on the use of GNN for fraud-detection.
    - Useful to see current developments in graphical ML algorithms.
2. Catch Me If You Can: Semi-supervised Graph Learning for Spotting Money Laundering
3. VU Masters Thesis - Catch Me If You Can: Graph Neural Networks to detect fraudulent nodes to counter money laundering
    - Explores whether graph network representation is reasonable and suitable for money-laundering detection.
4. Performance of Different Machine Learning Algorithms in Detecting Financial Fraud
5. Credit card fraud detection using machine learning techniques: A comparative analysis
    - Both papers delve into non-graphical ML algorithms for fraud detection.
    - Decision Trees and Random Forest are found to outperform other classifiers.

**More sources will follow with Literature Review*

- **Expected Outcome**

According to my prior-to-research knowledge, the expected outcome is that network graphs are a suitable representation and with the use of graph neural networks (GNN) will outperform non-graphical ML models such as Random Forest (RF). Graphical representation and GNNs are

capable of representing and underlining the significance of connections between nodes, in this particular context, connections between nodes are represented as individual transactions. While traditional ML algorithms are effective, it is the hypothesis that they are only effective at representing connections and relations between columns in data due to their non-graph nature.

Thesis statement / Research Questions

> **RQ1: How does the utilization of network graph representation for customer payment information impact the accuracy and efficiency of machine learning algorithms in detecting money-laundering fraud?**

The main goal of the thesis is to investigate whether network graphs improve the effectiveness of machine learning algorithms in the context of fraud detection. Nevertheless, while this is the main goal, the thesis will also present the reasoning behind it, the extent to which it is an improvement compared to other data structures such as purely numerical data representations, and in what aspects can its effectiveness be further improved. The comparison of data representations also spills over to the debate of different machine learning algorithms and what trade offs occur between them.

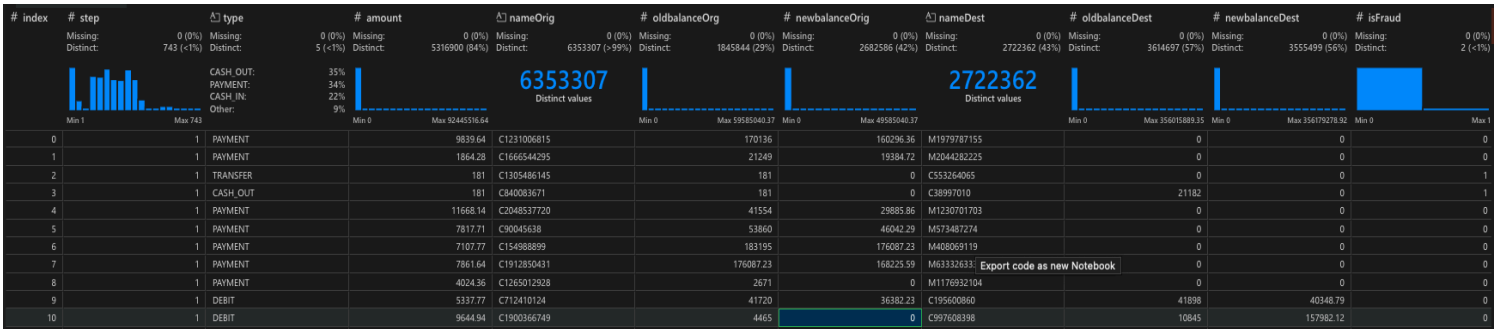
> **RQ2: What trade-offs exist between employing network graph representations and other data structures (e.g., tabular data, feature-engineered representations) concerning interpretability, computational complexity, and scalability when applied to AML detection tasks?**

This research question investigates the comparative advantages and disadvantages of network graph representations compared to alternative data structures. By examining factors such as interpretability, computational complexity, and scalability, the thesis can provide insights into the suitability of network graph representations for fraud detection tasks and identify potential areas for improvement or optimization.

Approach / Methods

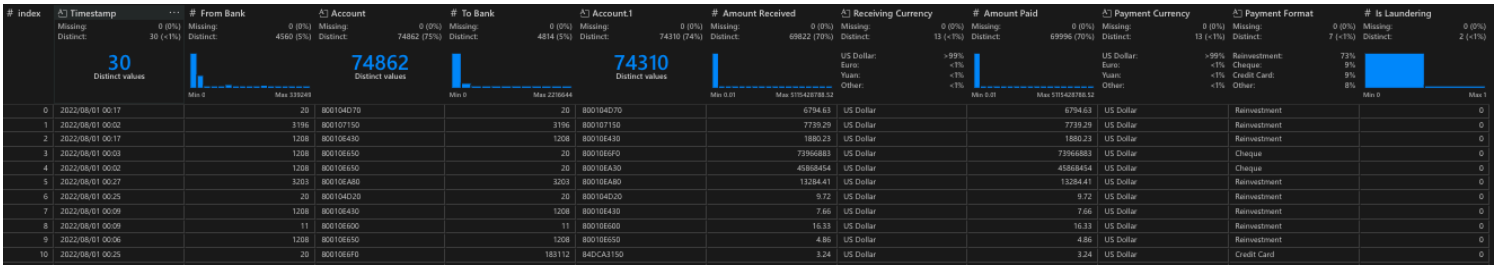
There are several datasets that can be utilized for the purposes of this thesis. All of them are labeled, synthetic, AML-oriented, and don't include sensitive data. There are 3 datasets: PaySim, AMLSim, and IBM. PaySim's dataset has 6 million sample transaction data. AMLSim has 1.3 million samples while IBM has 185 million samples.

As mentioned in the introduction, this thesis will be of a comparative and quantitative nature where two types of data representation methods will be used: graphical and non-graphical. As would follow naturally from their data representations, different machine learning algorithms will be used in order to investigate the effectiveness of graph-based algorithms. In particular, GNNs will be used for graphical representations while Random Forest will be used for non-graphical representations. Models will be created, trained, and evaluated.



index	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud
0	1	PAYMENT	9839.64	C1231006815	170136	160296.36	M1979787155	0	0	0
1	1	PAYMENT	1864.28	C1666544295	21249	19384.72	M2044282225	0	0	0
2	1	TRANSFER	181	C1305486145	181	0	C553264065	0	0	1
3	1	CASH_OUT	181	C840083671	181	0	C38997010	21182	0	1
4	1	PAYMENT	11668.14	C2048537720	41554	29885.86	M1230701703	0	0	0
5	1	PAYMENT	7817.71	C90045638	53860	46042.29	M573487274	0	0	0
6	1	PAYMENT	7107.77	C154988899	183195	176087.23	M408069119	0	0	0
7	1	PAYMENT	7861.64	C1912850431	176087.23	168225.59	M63332633	0	0	0
8	1	PAYMENT	4024.36	C1265012928	2671	0	M1176932104	0	0	0
9	1	DEBIT	5337.77	C712410124	41720	36382.23	C195600860	41898	40348.79	0
10	1	DEBIT	9644.94	C1900366749	4465	0	C997608398	10845	157982.12	0

Export code as new Notebook



index	Timestamp	From Bank	Account	To Bank	Account.1	Amount Received	Receiving Currency	Amount Paid	Payment Currency	Payment Format	Is Laundering
0	2022/08/01 00:17	20	800104D70	20	800104D70	6794.63	US Dollar	6794.63	US Dollar	Reinvestment	0
1	2022/08/01 00:02	3196	800107150	3196	800107150	7739.29	US Dollar	7739.29	US Dollar	Reinvestment	0
2	2022/08/01 00:17	1208	80010E430	1208	80010E430	1880.23	US Dollar	1880.23	US Dollar	Reinvestment	0
3	2022/08/01 00:03	1208	80010E650	20	80010E6F0	73966883	US Dollar	73966883	US Dollar	Cheque	0
4	2022/08/01 00:02	1208	80010E650	20	80010EA30	45868454	US Dollar	45868454	US Dollar	Cheque	0
5	2022/08/01 00:27	3203	80010EA80	3203	80010EA80	13284.41	US Dollar	13284.41	US Dollar	Reinvestment	0
6	2022/08/01 00:25	20	800104D20	20	800104D20	9.72	US Dollar	9.72	US Dollar	Reinvestment	0
7	2022/08/01 00:09	1208	80010E430	1208	80010E430	7.66	US Dollar	7.66	US Dollar	Reinvestment	0
8	2022/08/01 00:09	11	80010E600	11	80010E600	16.33	US Dollar	16.33	US Dollar	Reinvestment	0
9	2022/08/01 00:06	1208	80010E650	1208	80010E650	4.86	US Dollar	4.86	US Dollar	Reinvestment	0
10	2022/08/01 00:25	20	80010E6F0	183112	84DCA3150	3.24	US Dollar	3.24	US Dollar	Credit Card	0

Questions

- There is an ongoing internal debate whether I should focus on comparing graph algorithms/representation to traditional algorithms or maybe I should delve deeper into graph algorithms and compare Graph Neural Networks (GNNs) to Graph Autoencoders or to Graph Attention Networks (GATs).
    - **Further challenges that may present themselves is the bias and fairness of comparing the two methods in such a manner. Is it a fair comparison and can any conclusions be inferred from it.**
- Regarding the dataset, I am thinking of maybe adding a temporal aspect to the models I create as theoretically speaking, time plays an important role. A sequential element would be of particular interest in this case.
    - Potential RQ3: **How does the incorporation of temporal dynamics, such as transaction sequence and time dependencies, into network graph representations impact the effectiveness of ML algorithms for detecting money-laundering patterns over time?**
- How realistic is the dataset that is being used? How are fraudulent transactions labeled, under what criteria? Are these criteria relevant and realistic?
    - More research into datasets required.
- Different types of fraud can be explored, AML is one type but another prevalent type is credit card fraud. This is explored on this github.

Work plan

List the stages of your project:

1. Concept and Data Exploration -- (February)
2. Identify Potential Challenges -- (February)
3. Proposal to supervisor -- (February)
4. Formal Literature Review -- (February)
5. Solid Theory and Concept Formalization -- (March)
6. Create formal outline for thesis -- (March)
7. Programming and Model development -- (March & April)
    7.1. Start working on data representation
    7.2. Data preprocessing
    7.3. Implement models
    7.4. Model evaluation
8. Answering Research Questions -- (April)
9. Identifying Conclusion -- (April)
10. Start writing final Thesis Report -- (April & May)

Implications of Research

As fraud-related activities in the financial world become more and more sophisticated and wide-spread through the rapid proliferation of technology, it is crucial for financial institutions to continue to improve upon existing fraud-detection frameworks and to utilize new technologies to develop effective countermeasures in order to safeguard its assets, protect customer assets and data, prevent systemic risks, and maintain trust and reputation in financial institutions. By examining the effectiveness of network graph representation in fraud-detection algorithms, reasons as to why that is so can be identified and algorithms can be positively improved upon within the data representation.

List of references (*incomplete)

- https://www.dataversity.net/the-power-of-graph-databases-to-detect-fraud/
- https://arxiv.org/abs/2302.11880
- https://github.com/IBM/AML-Data?tab=readme-ov-file
- https://github.com/safe-graph/graph-fraud-detection-papers?tab=readme-ov-file
- https://ieeexplore.ieee.org/abstract/document/8123782
- https://arxiv.org/abs/1902.10191
- https://github.com/IBM/TabFormer