Bank Account Fraud Detection with Network Science

Cosc 421 Network Science

By Chinmay Arvind, Dmitry Kostyukov, Jerry Fan, Rylan Millar

What problem did we solve?

- Bank Account Fraud: illegal act of falsifying information for a bank account
- We are using network science to detect bank account fraud
- Problems solved:
  - Finding key fraudulent players in the network
  - Finding specific fraudulent groups within the network and their characteristics
  - Finding average profile of a fraudulent bank customer
  - Finding differences between fraudulent and non-fraudulent bank applications

What makes the problem important to solve?

- Bank Account Fraud is widespread across the globe and causes losses of funds for banks, and individuals
- Rapid growth of tech has made the risks of fraud rise, increasing the need for fraud to be detected faster
  - Phishing
  - Cryptocurrency fraud
  - Falsification of financial documents
- Fraud needs to be detected to maintain trust in financial systems

Why is it novel?

- Uses bank account applications as nodes, and similar nodes were linked to each other based on similarity of attributes
- No other papers have used this perspective to approach the problem from
- Combines network centrality and clustering metrics along with features in the data to determine fraudulent nodes
- Average fraudulent node profile and differences between fraudulent and non-fraudulent nodes not used in other papers, and can help establish a baseline of what one would consider a fraudulent node in the context of bank account fraud

What have others done to solve similar problems?

- To detect insurance fraud [5]
  - Network Centrality Measures
  - Guilt-By-Association Methods
  - Random Walks
  - Graph Neural Networks
- To detect financial fraud [6][7]
  - Social Network Analysis
  - Probabilistic Curve
  - Best Match
  - Adverse Selection
  - Density Selection

How have we solved the problem?

- Network Science
- Data Preparation
  - Kaggle dataset, PCA cleaning, stratified sampling, data validation
- Identify Key Fraudulent Players
- Profiling Fraudulent Accounts
- Detecting Fraudulent Groups
- Comparison between Fraudulent and Legitimate Accounts

Data Description & Source

Bank Account Fraud Dataset Suite:

https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022

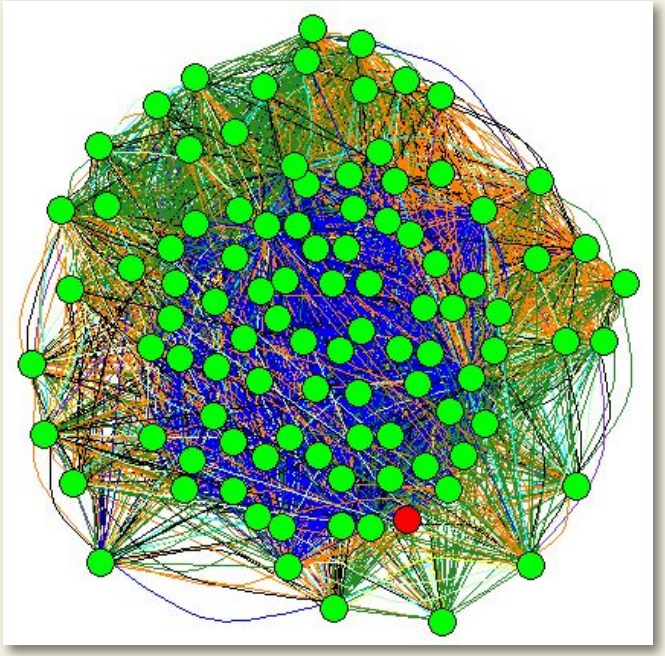
- Real World Data Set, Published at NeurIPS 2022
- 1,000,000 instances and 32 attributes for each instance.
- After Principal Component Analysis 13 attributes and 743,169 nodes remained.

Attribute	payment_type	keep_alive_session	foreign_request	email_is_free	fraud_bool	bank_branch_count_8w	zip_count_4w
Description	Credit Payment Plan Type (1 - 5)	User Option on Logout (Boolean)	Origin of Request Differs From Bank's (Boolean)	Domain of Email (Paid or Free) (Boolean)	Fraud Indicator (Boolean)	Count of Applications to Bank Branch (0 - 2521)	Count of Applications Within Same Zip (1-5767)

Attribute	name_email_similarity	bank_months_count	housing_status	velocity_6h	phone_home_valid	current_address_months_count
Description	Metric of Similarity Between Name and Email (0.0-1.0)	How Old is Previous Account (0-31 Months)	Housing Status of Applicant (1-7)	Velocity of Total Applications in Last 6 Hours (0-16.7k)	Validity of Home Phone (Boolean)	Months in Currently Registered Address (0-406)

Description of Network

- Our network is an attribute-based network.
- Nodes are individual bank applications, each application contains several attributes.
- Edges are made based on shared attributes between nodes.
- Each edge represents a certain shared attribute between two nodes.



Example Graph of a Network with n = 100 Nodes

What analyses were carried out?

Four different research topics were explored, each topic analyzed several network metrics:

1. What were the key influential players within the network?
   - Betweenness and Eigenvector Centrality
2. What is the average profile of a fraudulent account?
   - Cosine and Jaccard Similarity Coefficients
3. Can fraudulent groups within the network be identified?
   - Louvain Community Detection Algorithm
4. Are there any differences in key metrics between fraudulent and legitimate nodes?
   - Degree, Eigenvector, Closeness and Betweenness Centrality, Local Clustering Coefficient, Jaccard Similarity Coefficient.

RQ1: Key Fraudulent Players in Network

- Used eigenvector and betweenness centrality as measures of a node's importance within a network
- Nodes within the top 25 percentiles of eigenvector and betweenness centrality scores were flagged as suspicious
- Nodes that scored consistently high on both metrics can be the first to be sent as leads to domain experts in investigating bank account fraud
- High eigenvector centrality of a node = connected to other potentially fraudulent nodes

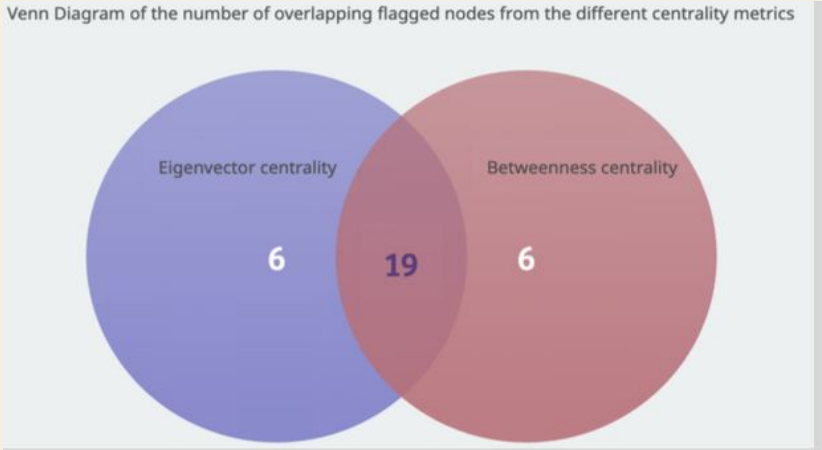
node_id	payment type	keep alive session	foreign request	email is free	fraud bool	bank branch count 8w	eigenvector centrality score
10295	2	1	0	0	0	1	1.0000000
582493	1	1	0	1	0	1	0.9781452
191649	2	1	0	1	0	1	0.970140
132333	1	1	0	1	0	1	0.9595332
638068	1	1	0	0	0	1	0.9593648
515970	2	1	0	1	0	1	0.9559428
127176	2	1	0	1	0	1	0.9533932
288461	2	1	0	0	0	1	0.9517984
661285	1	1	0	0	0	1	0.9456791
487929	2	1	0	0	0	1	0.9378288

Fig. 1. Table displaying the profile of the top 10 nodes with the highest eigenvector centrality.

RQ1: Key Fraudulent Players in Network (contd.)

- High betweenness centrality of a node = could be involved in fraud rings as middlemen as a bank account request could be similar to many other account requests, the node with high betweenness centrality could provide the link between different account requests within a fraud ring
- Venn diagram to show the number of nodes that ranked high on both centrality metrics and are the most suspicious nodes based on centrality measures

node_id	payment type	keep alive session	foreign request	email is free	fraud bool	bank branch count 8w	betweenness centrality score
512662	2	1	0	1	0	1	2.338539
10295	2	1	0	0	0	1	2.260064
127176	2	1	0	1	0	1	2.161004
677418	2	0	0	1	0	1	2.009444
132333	1	1	0	1	0	1	1.973075
7320322	2	0	0	1	0	1	1.855374
582493	1	1	0	1	0	1	1.830061
395008	2	1	0	0	0	1	1.814628
596042	2	1	0	0	0	1	1.812014
487357	2	0	0	1	0	1	1.809334



RQ2: Average Profile of Fraudulent Account

Attribute	Value
Payment_type	2
keep_alive_session	0
foreign_request	0
email_is_free	1
bank_branch_count_8w	210.1815
zip_count_4w	1648.506
name_email_similarity	0.3859064
bank_months_count	17.3637
housing_status	1
velocity_6h	5251.171
phone_home_valid	0
current_address_months_count	114.2933

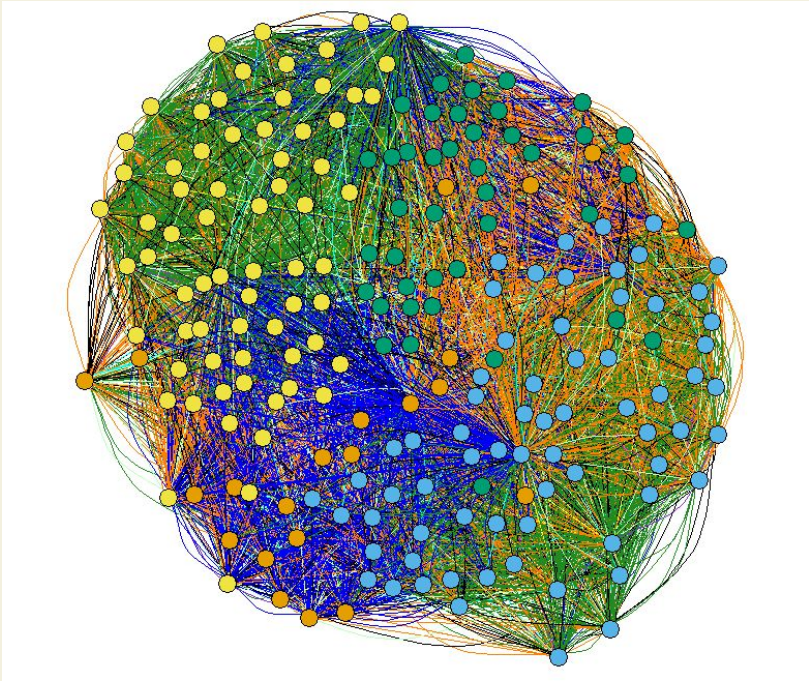
- Compute the average profile of a fraudulent account using mean and mode
- Calculate the similarity score between typical Fraud Node and any given fraudulent nodes
- Create an edge if the similarity score is above the threshold
    - Jaccard Similarity Threshold: 0.5
    - Cosine Similarity Threshold: 0.7

Metric	Value	Percentage
Number of Jaccard Edges	0	0.00%
Number of Cosine Edges	148	2.15%

Maximum possible edges: 6877

RQ3: Types of analyses carried out

- Created a baseline profile for fraudulent accounts using average attributes.
- Used the Louvain algorithm to group nodes into communities based on connections.
- Compared community profiles to the fraudulent profile to identify similarities.
- Modularity score of 0.16 indicated communities were likely random.
- Fraudulent nodes stood out as distinct from all community profiles.



- Limitations: Full dataset analysis needs more computing power.
- Future improvements: Use transaction data and machine learning for better accuracy.

RQ3: Key Mean Metric Differences Between Fraudulent and Legitimate Applications

Metrics	Betweenness (Legitimate)	Betweenness (Fraudulent)	Eigenvector (Legitimate)	Eigenvector (Fraudulent)	Degree (Legitimate)	Degree (Fraudulent)
Mean	96.518	67.956	0.649	0.556	1380.840	1182.438
Difference	42.02%		16.72%		16.78%	

Metrics	Jaccard (Legitimate)	Jaccard (Fraudulent)	Closeness (Legitimate)	Closeness (Fraudulent)	Local Clustering (Legitimate)	Local Clustering (Fraudulent)
Mean	0.701	0.658	0.844	0.819	0.851	0.862
Difference	6.50%		2.94%		1.31%	

What went wrong? And how did we fix it?

Main Challenge: Not enough computation capabilities for size of dataset (743,169 total nodes after cleaning)

Solutions:

- Took large samples of data (100-1000 nodes) which were still manageable given the available data processing capabilities.
- Utilized stratified random sampling to ensure same proportion of fraudulent (1%) to legitimate nodes (99%).
- Made use of Chi-Squared test to ensure that values within attributes of a sample are proportionate to the entire dataset.
- Aggregated results from several samples (when applicable) to manage sampling bias.

Main takeaways and what we would have done differently

- Centrality metrics flagged suspicious nodes.
- Strict similarity thresholds limited pattern detection.
- Fraudulent accounts stood out despite random network communities.
- Neural networks and transaction data could improve detection.
- Refine graph building and integrate machine learning for better results.

Citations

References

1. Dataset citation @articlejesusTurningTablesBiased2022, title=Turning the Tables: Biased, Imbalanced, Dynamic Tabular Datasets for ML Evaluation, author=Jesus, Sérgio and Pombal, José and Alves, Duarte and Cruz, André and Saleiro, Pedro and Ribeiro, Rita P. and Gama, João and Bizarro, Pedro, journal=Advances in Neural Information Processing Systems, year=2022
2. F. Grando, D. Noble and L. C. Lamb, "An Analysis of Centrality Measures for Complex and Social Networks," 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, USA, 2016, pp. 1-6, doi: 10.1109/GLOCOM.2016.7841580. keywords: Measurement;Complex networks;Correlation;Social network services;Informatics;Correlation coefficient;Analytical models,
3. Motie, Soroor, and Bijan Raahemi. "Financial fraud detection using graph neural networks: A systematic review." Expert Systems with Applications 240 (2024): 122156.
4. Girish, K.K., Bhowmik, B. (2024). Historical Analysis of Financial Fraud and Its Future. In: Thampi, S.M., Hu, J., Das, A.K., Mathew, J., Tripathi, S. (eds) Applied Soft Computing and Communication Networks. ACN 2023. Lecture Notes in Networks and Systems, vol 966. Springer, Singapore. https://doi.org/10.1007/978-981-97-2004-0_4
5. Deprez, Bruno, et al. "Network analytics for insurance fraud detection: a critical case study." European Actuarial Journal (2024): 1-26.
6. Normah Omar , Ismail bin Mohamed , Zuraidah Mohd Sanusi and Hendi Yogi Prabowo (2014). Understanding Social Network Analysis (SNA) In Fraud Detection. Recent Trends in Social and Behaviour Sciences https://www.researchgate.net/profile/Normah-Omar/publication/300376345_Understanding_Social_Network_Analysis_SNA_in_fraud_detection/links/5874560508ae8fce4924ff57/Understanding-Social-Network-Analysis-SNA-in-fraud-detection.pdf
7. Wheeler, Richard, and Stuart Aitken. "Multiple algorithms for fraud detection." Applications and Innovations in Intelligent Systems VII: Proceedings of ES99, the Nineteenth SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence, Cambridge, December 1999. Springer London, 2000.