Report on Application of Network Analysis in Money Laundering Detection

Virendrasinh Chavda

School of Mathematics, Statistics and Actuarial Sciences,
University of Essex, UK

Abstract

Detection of money laundering in banking systems has become more complex as criminals develop technologies themselves to cheat the system. Nonetheless, banks and other financial institutes are deep into developing better and better solutions for the purpose of detecting suspicious trails of money. These systems are known as anti-money laundering systems or AML systems [1].

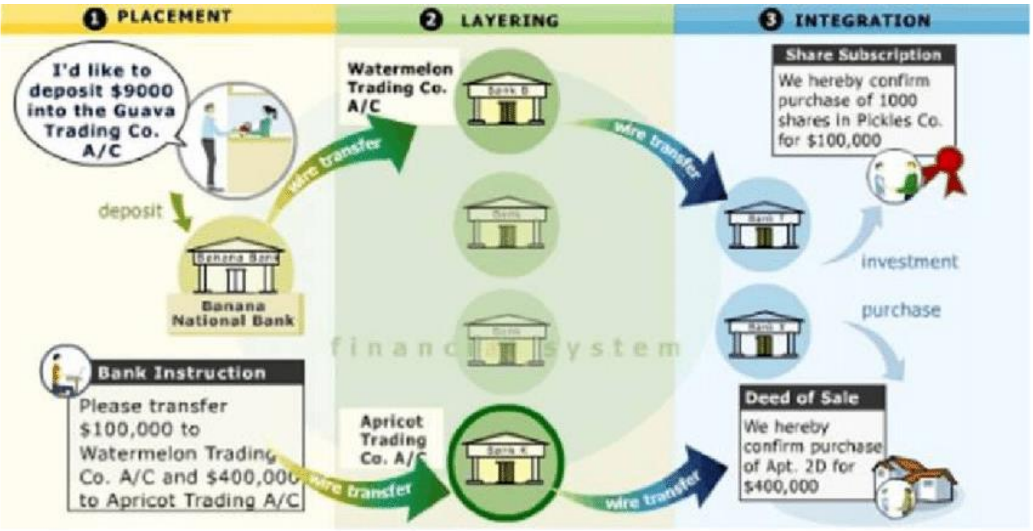
As the rules for doing businesses and starting companies are eased in many European countries and tax havens, it is getting easier to register companies or entities which can be used for illicit activities [2]. These companies are known as shell companies and generally of limited liability. Shell companies can be used as mediators and can create a large and complex network in combination with other shell companies, which can then be used for laundering money [2].

Due to this networking nature of money laundering, it makes sense to use network analysis techniques which can identify important accounts in any system and then apply machine learning models to identify whether those accounts can be considered as fraudulent or not [3]. Literature review of a few research papers on different methods of using network analysis for AML is included in this paper and important methods are discussed. Since combining network analysis and machine learning methods will be a huge project, this report concentrates only on network analysis side of money laundering detection. This report also includes various techniques that can be used for detection of important accounts in a network and results of these techniques. For this purpose, synthetic banking data, which imitates the real-life money laundering patterns, is used [4].

Keywords: Money Laundering, Shell Companies, Detection, Network Analysis, Machine Learning

1. Introduction

Money laundering is defined in the Proceeds of Crime Act 2002 [5] as “the process by which the proceeds of crime are converted into assets which appear to have a legitimate origin, so that they can be retained permanently or recycled into further criminal enterprises”. The key elements of money laundering involve the acquisition of funds through criminal activities and the subsequent manipulation and integration of these illegally obtained assets into the legitimate financial system, often achieved through a complex process of placement, layering, and integration [6]. The three-layer model of money laundering systems is very popular and is shown in Fig. 1.1 [7].



1 PLACEMENT
2 LAYERING
3 INTEGRATION
I'd like to deposit $9000 into the Guava Trading Co. A/C
deposit
Banana National Bank
Bank Instruction
Please transfer $100,000 to Watermelon Trading Co. A/C and $400,000 to Apricot Trading A/C
wire transfer
Watermelon Trading Co. A/C
Apricot Trading Co. A/C
financial system
Share Subscription
We hereby confirm purchase of 1000 shares in Pickles Co. for $100,000
investment
purchase
Deed of Sale
We hereby confirm purchase of Apt. 2D for $400,000

Source: © AUSTRAC on behalf of the Commonwealth of Australia

Fig. 1.1 Three-layered model for money laundering [7]

The first step is placement. In this step, proceeds of the crime need to either enter a financial system or be used to buy an asset. The second step is layering, by which a launderer, through some financial transactions, tries to conceal and disguise the source of the money. This step is done by breaking down the money into small amounts and transferring it to different financial institutions. In the final stage, integration, the money is assimilated along with all other assets in the system to make the money appear as if it were obtained legally [7]. This sophisticated network used for money laundering makes it difficult to identify accounts associated with illicit transactions.

Traditionally, banks, governments and financial institutions have relied upon rule-based systems to flag suspicious transactions or accounts, but due to evolving scenarios, rulebooks are becoming complex and false positive rates are going high [8]. This led to institutions searching for more reliable methods to detect financial crimes, and due to the networking nature of money laundering, network analysis is used in modern AML systems.

2. Literature Review

Anomaly detection was the rule of thumb for detection of suspicious financial transactions a few years back. As explained by Gómez, Agudelo, and Patiño [9], anomaly detection is done using past data and machine learning models. Defining anomalies can be a tedious task and can change in different scenarios. It has been established that machine learning algorithms like decision trees, support vector machines, logistic regression, k-means clustering, k-nearest neighbors, etc. can be effectively used in anomaly detection in case of simple fraud detection tasks like credit card fraud, as discussed by Hilal, Gadsden, and Yawney [10]. In the past few years, with advancements in deep learning models and availability of cloud computing, deep learning models such as convolutional neural networks (CNN), long short-term memory networks (LSTM), autoencoders and GANs are becoming more popular. Generally, deep learning models have shown remarkable results over traditional machine learning algorithms for fraud detection as discussed by Kute, Pradhan, Shukla, and Alamri [11]. But where transparency and explainability is priority, deep learning models take a backseat. Due to the difficulty of tracing back the influence of features and its black box nature, deep learning models, as explained by Dobson [12], are less attractive for implementation in financial crime detection, as establishing proof of crime becomes difficult.

Another popular approach for detecting financial crime is rule based detection as shown in work by Oztas and colleagues [13]. In this method, certain rules are defined such as looking for unusual patterns in transactions, or searching for higher cash-based transactions, etc. like machine learning based anomaly detection, rule-based detection also relies heavily on past data. For this reason, it becomes difficult to detect fraud transactions if they are channeled through complex systems. Also, with emerging cyber technologies, criminals are investing in more complex and smarter networks for laundering money. This issue is addressed by Salazar and Vargas [14]. With constantly changing technologies, rulebooks for rule-based detection models are also becoming complex and result in higher false positive rates, sometimes equal to 90%, as highlighted by Jensen and Iosifidis [15].

A more modern approach is to use network science in combination with machine learning algorithms to detect suspicious networks and accounts. Network analysis can identify large networks very efficiently with community detection models as shown by Alshantti and Rasheed [16]. Measurement of centrality can be used to identify important nodes in the network, i.e. accounts with higher degree of transactions, and properties like closeness centrality and betweenness centrality can be used to determine mediator accounts as discussed by Dreżewskia and colleagues [17]. The benefit of this graph-based approach is that it is not limited to any rules, nor does it rely too heavily on past data. This network-based approach is discussed in this report and is implemented on a synthetic bank transactions dataset to study how network analysis works for money laundering detection.

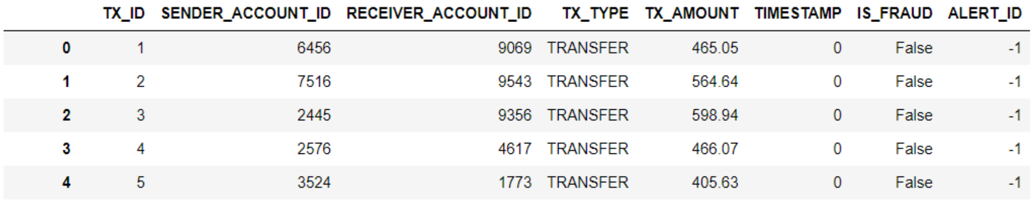
3. Overview of Dataset

3.1. Source of Dataset

Authors have used open-source data, IBM AMLSim Example Dataset [4], to demonstrate the network analysis techniques. The dataset is available on Kaggle. This is synthetic data, created to mimic banking transactions with known fraudulent patterns. Since it reflects properties of real banking transactions, this dataset can be used to test network analysis and machine learning algorithms to predict and identify fraudulent accounts and transactions.

3.2. Dataset description

The dataset consists of 1323234 rows and 8 columns. The rows are entries of transactions from one account to another and columns show attributes of each transaction. TX_ID refers to transaction ID, SENDER_ACCOUNT_ID refers to account number from which the transaction is done to receiver account, RECIEVER_ACCOUNT_ID. TX_TYPE shows the type of transaction such as transfer or received. This data is created to show transactions from one account to another, so there is only one value in TX_TYPE, i.e. ‘TRANSFER’. TX_AMOUNT is the transaction amount of any particular transaction, while TIMESTAMP refers to timestamp of that transaction. Column IS_FRAUD shows binary values TRUE or FALSE, depending on whether the transaction is fraudulent or not, i.e. TRUE for fraudulent and otherwise FALSE. ALERT_ID columns show categories of alert that need to be generated when a potentially fraudulent activity is detected. Values for columns IS_FRAUD and ALERT_ID are empirically known from experience. Fig. 5.1 shows data of the first 5 transactions from the dataset.



	TX_ID	SENDER_ACCOUNT_ID	RECEIVER_ACCOUNT_ID	TX_TYPE	TX_AMOUNT	TIMESTAMP	IS_FRAUD	ALERT_ID
0	1	6456	9069	TRANSFER	465.05	0	False	-1
1	2	7516	9543	TRANSFER	564.64	0	False	-1
2	3	2445	9356	TRANSFER	598.94	0	False	-1
3	4	2576	4617	TRANSFER	466.07	0	False	-1
4	5	3524	1773	TRANSFER	405.63	0	False	-1

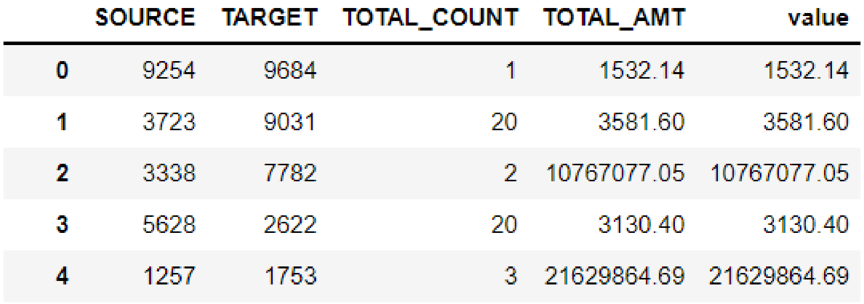
Fig. 5.1: Overview of few transaction entries from dataset.

3.3. Sampling method

This dataset has 1323234 rows and 8 columns and will require huge time complexity to load the data. For this reason, authors decided to use random sampling technique and made a subset of 2000 entries to balance out lack of computational power.

During the sampling process, authors have used group by function to accumulate all transactions from same account to one entry. SENDER_ACCOUNT_ID column is renamed as SOURCE and RECEIVER_ACCOUNT_ID as TARGET for ease of use. Moreover, TX_AMOUNT was sliced to two columns, namely TOTAL_COUNT and TOTAL_AMT. TOTAL_AMT column represent the total number of times transactions have been done between same pair of accounts, and TOTAL_AMT column represents total amount that has been transacted between those pairs. Authors have not considered columns TX_TYPE, TIMESTAMP, IS_FRAUD, and ALERT_ID, as developing machine learning model for prediction is not in the scope of this report.

A random sample of 2000 rows have been generated from the original dataset, and random state is set to 2 to get repetitive results for ease of comparison. Fig. 5.2 shows the first five entries in sampled dataset.



	SOURCE	TARGET	TOTAL_COUNT	TOTAL_AMT	value
0	9254	9684	1	1532.14	1532.14
1	3723	9031	20	3581.60	3581.60
2	3338	7782	2	10767077.05	10767077.05
3	5628	2622	20	3130.40	3130.40
4	1257	1753	3	21629864.69	21629864.69

Fig. 5.2: Overview of few transaction entries from sampled dataset.

4. Results of Experiments

4.1. Methods used for analysis.

A transaction network is assumed to be directional as transactions have dimensions, i.e. from one specific account to another specific account.

Fig. 6.1 shows the plotting of this sampled network.

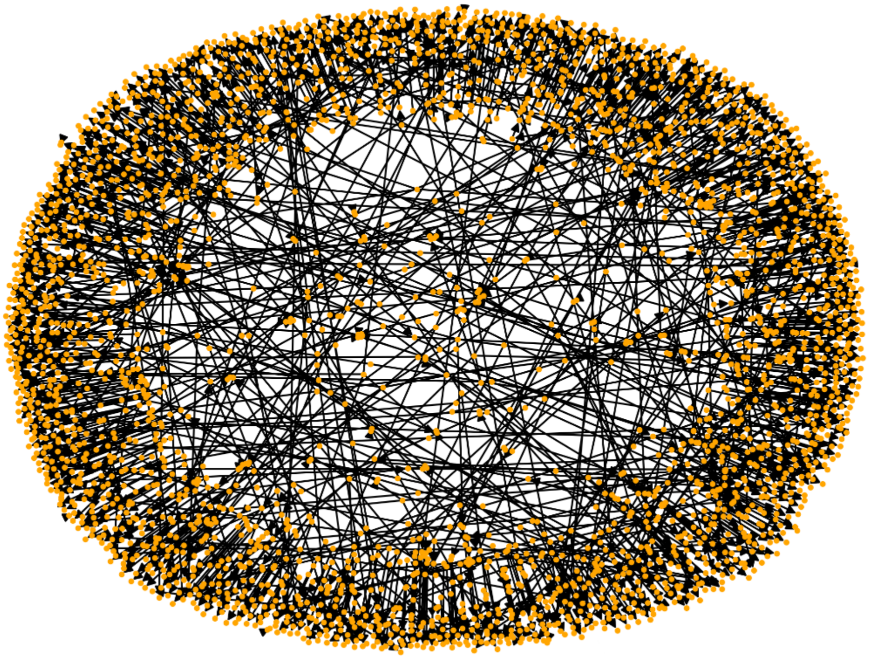


Fig. 6.1: Overview of few transaction entries from dataset.

The number of nodes for the network shown in graph in Fig. 6.1 is 2997 and number of edges for same network is 2000. To check the presence of important or central nodes, degree distribution is plotted as shown in Fig. 6.2. From the plot it is visible that most of the nodes has just 1 to 2 degrees, but a few nodes have degrees higher than 5 degrees. Accounts associated with these nodes can be of importance as these nodes have a high flow of incoming or outgoing transactions and thus can be used for illicit transactions. Table 6.1 shows the count of nodes per degree, which confirms that there is indeed one node which has 11 degrees, while 3 nodes have 7 degrees. These nodes can be further investigated to see how they are connected within one group.

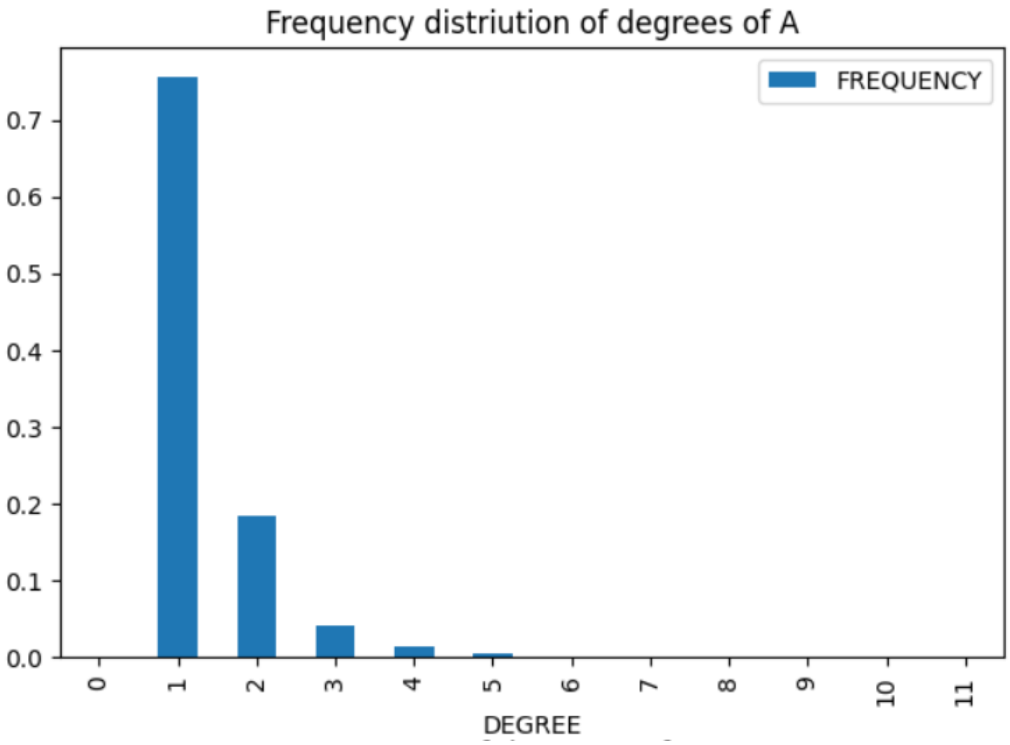
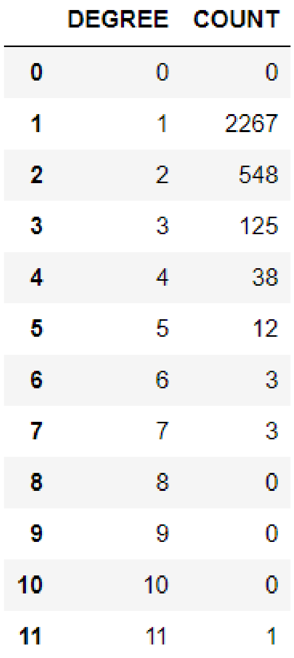


Fig. 6.2: Degree distribution of the network.



	DEGREE	COUNT
0	0	0
1	1	2267
2	2	548
3	3	125
4	4	38
5	5	12
6	6	3
7	7	3
8	8	0
9	9	0
10	10	0
11	11	1

Table 6.1: Count of nodes per degree

For further analysis, degrees, closeness centrality and betweenness centrality for each node of the network are calculated. Node 9993 has the highest degree of 11 and maximum closeness centrality of 0.0033 within the network, and node 9990 has highest betweenness centrality of 4.01e-06. These nodes should be considered for further investigation as they are nodes of importance within the network. No result can still be reached without checking the roles of these nodes in their local groups.

Methods like k-cliques and k-components cannot be implemented on directed networks, so to analyze the networks, communities are generated using Girvan-Newman algorithm. Execution of Girvan-Newman algorithm is slow and so to adjust for the computational power and time complexity, first 2 components from all communities are considered for further analysis.

A few communities are highlighted with different colors among all nodes as shown in Fig. 6.3.

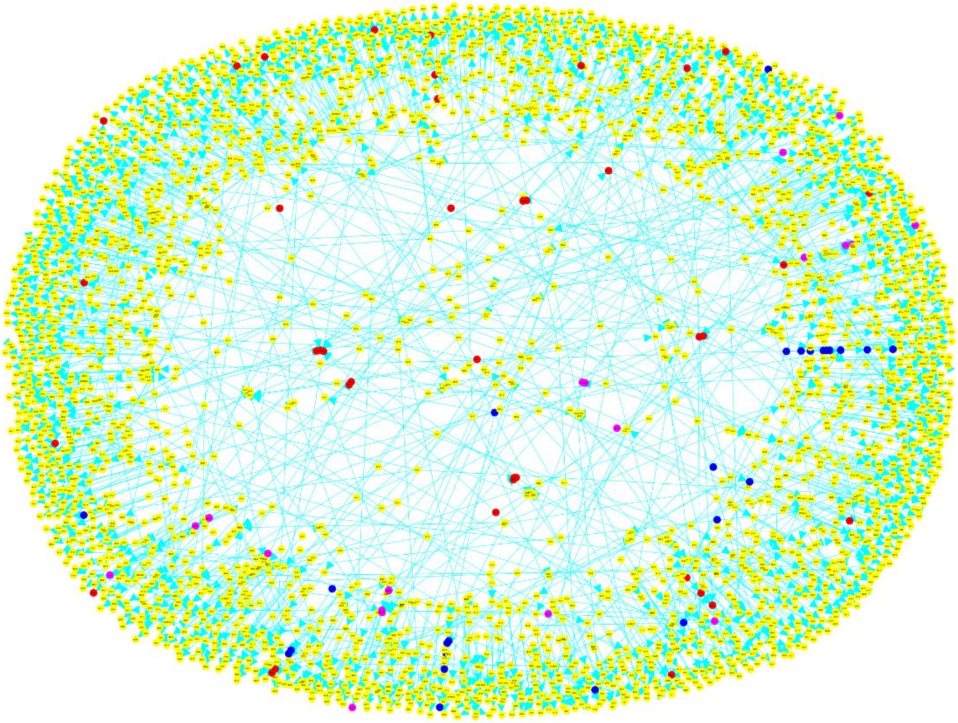


Fig. 6.3: Nodes from same communities are highlighted with similar color.

4.2. Results of the experiment.

For the first analysis, authors have considered the community with node 9993 within it, as this node has highest degree and closeness centrality. For second analysis, community with node 9990 is considered as it has maximum betweenness centrality, and for third analysis, largest community is chosen to check if any suspicious node is within it.

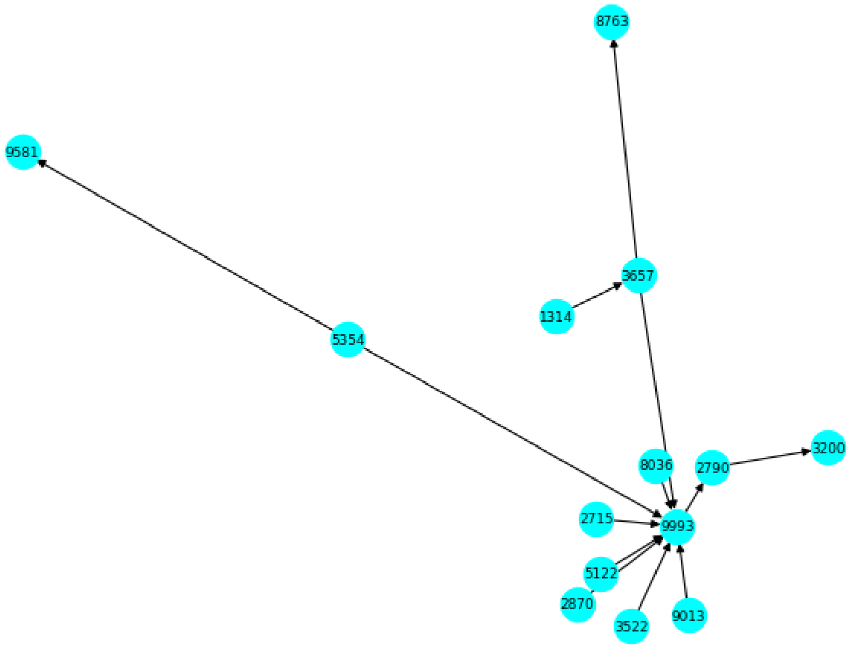
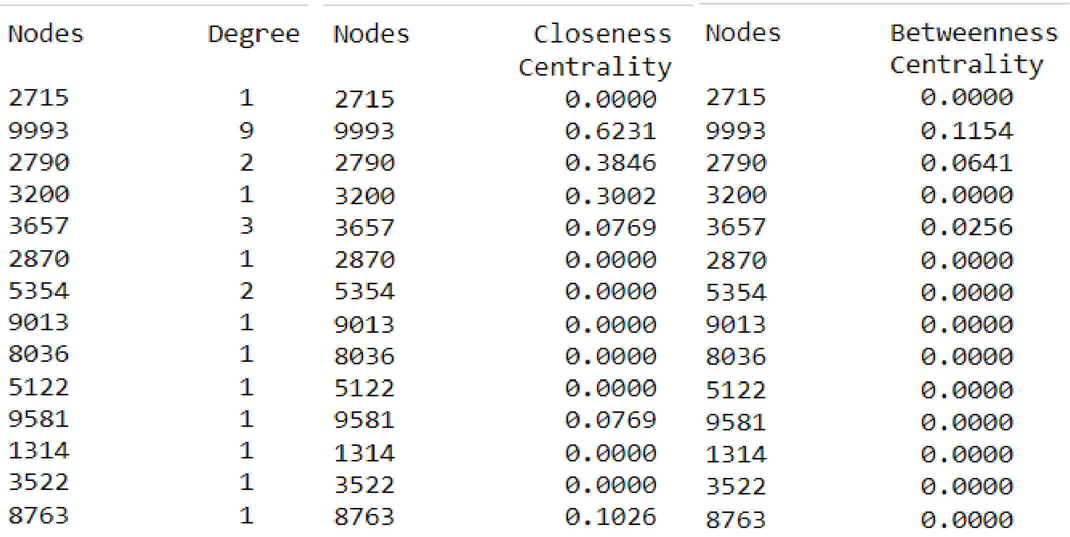


Fig. 6.4: Community with node 9993

Fig. 6.4 shows the community with node 9993. This community has 14 nodes and 13 edges. Table 6.2 shows the degree, closeness centrality and betweenness centrality for all nodes in this network. Node 9993 has the highest degree of 9, highest closeness centrality of 0.6231 and highest betweenness centrality 0.1154 within the community. This node is surely a node of importance within the group as its higher degree shows that is well connected with the community, higher closeness centrality shows that it has shortest paths to other nodes of this community and higher betweenness shows that it can act as mediator between two nodes. All these properties of this node show that it can be used for faster distribution of money within the community. But 9993 has an in degree of 8 and out degree of 1 which specifies that this node can act as a collector within the community. But a legit account such as a business account can be a collector too. Hence, without more information like location of the account, transaction amounts, type of account, trustees of account, it is difficult to determine that this node is involved in any fraudulent transaction. If this account is based in a tax haven country, then it can be flagged as a risky account [18].



Nodes	Degree	Nodes	Closeness Centrality	Nodes	Betweenness Centrality
2715	1	2715	0.0000	2715	0.0000
9993	9	9993	0.6231	9993	0.1154
2790	2	2790	0.3846	2790	0.0641
3200	1	3200	0.3002	3200	0.0000
3657	3	3657	0.0769	3657	0.0256
2870	1	2870	0.0000	2870	0.0000
5354	2	5354	0.0000	5354	0.0000
9013	1	9013	0.0000	9013	0.0000
8036	1	8036	0.0000	8036	0.0000
5122	1	5122	0.0000	5122	0.0000
9581	1	9581	0.0769	9581	0.0000
1314	1	1314	0.0000	1314	0.0000
3522	1	3522	0.0000	3522	0.0000
8763	1	8763	0.1026	8763	0.0000

Table 6.2: Measurement of centrality for first community

For the second analysis, the community chosen with node 9990 is shown in Fig. 6.5. This community is made up of 18 nodes and 17 edges. Node 9990 has the highest degree of 7, maximum closeness centrality of 0.3971, and maximum betweenness centrality of 0.1324 within this group. Table 6.3 shows the degree, closeness centrality and betweenness centrality for all nodes in this network. It has in degree of 6 and out degree of 1. As seen with node 9993 in first community, node 9990 in second community is a node of importance and likewise, further information related to the account is needed to determine whether this account is used as a collection account in money laundering network.

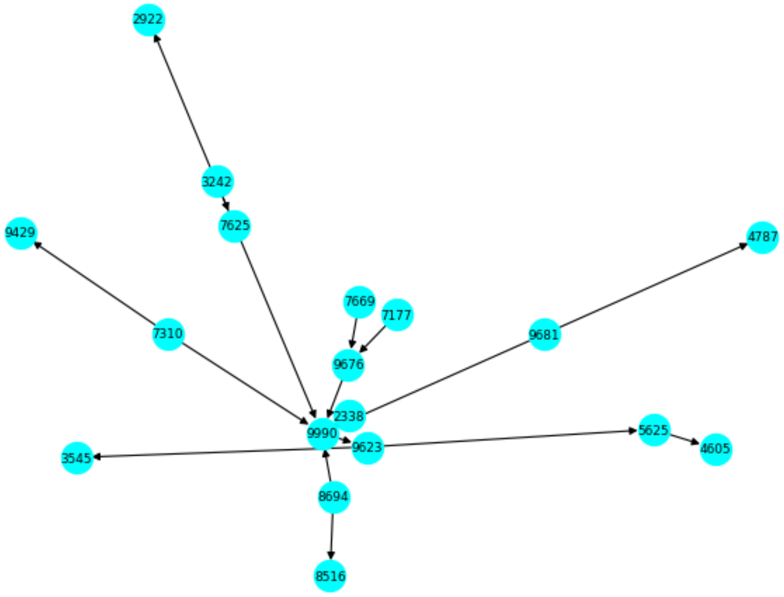
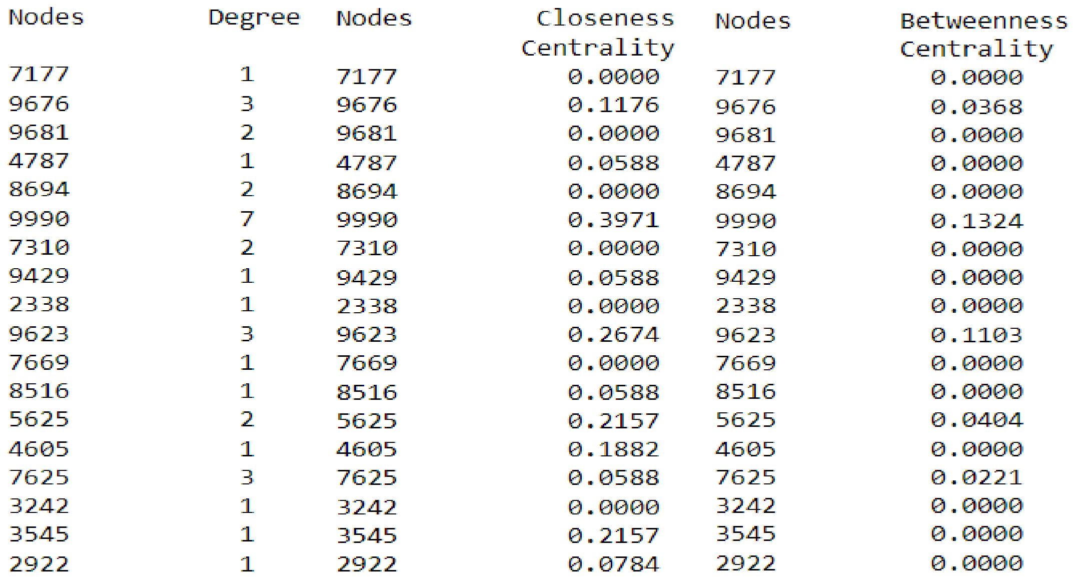


Fig. 6.5: Community with node 9990



Nodes	Degree	Nodes	Closeness Centrality	Nodes	Betweenness Centrality
7177	1	7177	0.0000	7177	0.0000
9676	3	9676	0.1176	9676	0.0368
9681	2	9681	0.0000	9681	0.0000
4787	1	4787	0.0588	4787	0.0000
8694	2	8694	0.0000	8694	0.0000
9990	7	9990	0.3971	9990	0.1324
7310	2	7310	0.0000	7310	0.0000
9429	1	9429	0.0588	9429	0.0000
2338	1	2338	0.0000	2338	0.0000
9623	3	9623	0.2674	9623	0.1103
7669	1	7669	0.0000	7669	0.0000
8516	1	8516	0.0588	8516	0.0000
5625	2	5625	0.2157	5625	0.0404
4605	1	4605	0.1882	4605	0.0000
7625	3	7625	0.0588	7625	0.0221
3242	1	3242	0.0000	3242	0.0000
3545	1	3545	0.2157	3545	0.0000
2922	1	2922	0.0784	2922	0.0000

Table 6.3: Measurement of centrality for second community

For the third analysis, the largest community is chosen from the list of all communities. The largest community has 21 nodes and 20 edges and the graph for this community is shown in Fig. 6.6.

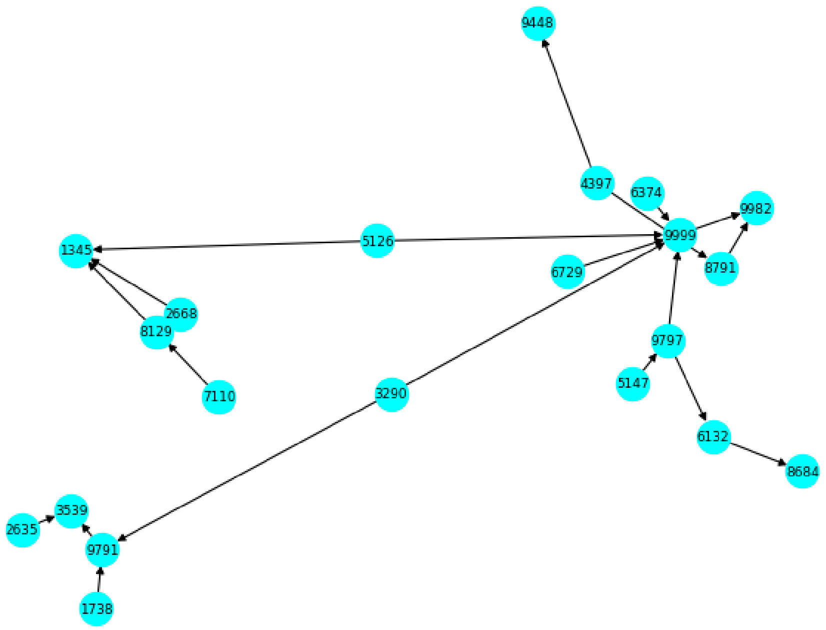
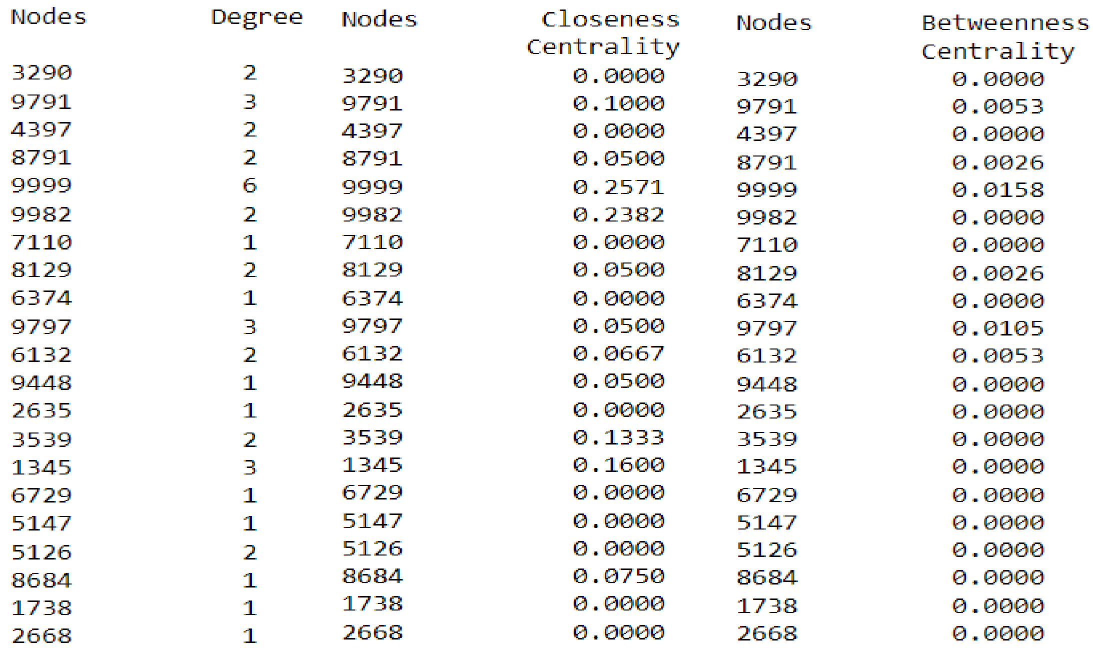


Fig. 6.6: Graph representing largest community.

Table 6.4 shows measurement of centrality for third community. Node 9999 has the highest degree of 6 as shown in Fig. 6.4. High degree of any node shows how densely is connected to the community. This can be the property of an account indulged in fraudulent activities as this type of account has frequent and well distributed transactions [19].

Checking in and out degrees for this community gives highest in-degree for node 9999, i.e. 5, and highest out-degree for node 9999 is only 2, which is comparable to other nodes in the community. This can mean that money is flowing more in this account compared to money going out. This can mean that this node can be a collection agent in the community [18].

When conducting closeness centrality too, the same node (9999) is found to have highest closeness centrality of 0.25 as show in Fig. 6.5., which signifies that account associated with node 9999 has shortest paths in the network compared to other nodes. Thus, this account can be used to distribute money faster and hence it can be used as distribution account, which makes this account riskier compared to other accounts [18].



Nodes	Degree	Nodes	Closeness Centrality	Nodes	Betweenness Centrality
3290	2	3290	0.0000	3290	0.0000
9791	3	9791	0.1000	9791	0.0053
4397	2	4397	0.0000	4397	0.0000
8791	2	8791	0.0500	8791	0.0026
9999	6	9999	0.2571	9999	0.0158
9982	2	9982	0.2382	9982	0.0000
7110	1	7110	0.0000	7110	0.0000
8129	2	8129	0.0500	8129	0.0026
6374	1	6374	0.0000	6374	0.0000
9797	3	9797	0.0500	9797	0.0105
6132	2	6132	0.0667	6132	0.0053
9448	1	9448	0.0500	9448	0.0000
2635	1	2635	0.0000	2635	0.0000
3539	2	3539	0.1333	3539	0.0000
1345	3	1345	0.1600	1345	0.0000
6729	1	6729	0.0000	6729	0.0000
5147	1	5147	0.0000	5147	0.0000
5126	2	5126	0.0000	5126	0.0000
8684	1	8684	0.0750	8684	0.0000
1738	1	1738	0.0000	1738	0.0000
2668	1	2668	0.0000	2668	0.0000

Table 6.3: Measurement of centrality for largest community

Finally, betweenness centrality for all nodes is calculated and again the node with highest betweenness centrality of 0.15 is the node 9999 as shown in Fig. 6.6. Higher betweenness centrality means this node can be a mediator in many transactions throughout the network. This confirms that node 9999 is a node of interest and it can be flagged risky. If frequent transactions are done through the account associated with node 9999, then it can be considered suspicious [18].

5. Conclusions and Future Work

In this report we explored an approach applying network analysis techniques to detect money laundering. We found that the Girvan-Newman algorithm can effectively separate clusters or communities within a large network and these clusters are further utilized for in depth analysis.

Techniques for finding centrality of nodes enable us to find the role and engagement of each node within the community and to analysis how cash is flowing in and out through any node. Closeness centrality and betweenness centrality shows us how easy it is for any node to get involved in transactions with other nodes and how it can act as an intermediary to conceal the source of transactions.

These techniques combined with visualization allow identification of risky and suspicious accounts. Though it is difficult to be 100% determinant, these techniques can be used in addition to machine learning algorithms for better results. For example, columns can be generated in original dataset where degrees of nodes and their centrality values can be stored. Moreover, columns in the original dataset, which have not been taken into sampled datasets due to computational and time constraints, can also be added in machine learning model, which may or may not give better results in fraudulent transactions detection.

Moreover, when network analysis identifies an account as a risky account or important account, all the entries of transactions related to that account can be associated with a value that indicates the importance of that account. This data can then be used by machine learning models as a feature to predict if an account is involved in any illicit transactions. Additionally, data like age of the account and ownership of the account can also help in determining the nature of use of this account. If an account is registered to any company, attributes associated with that company, such as type of company, number of employees, trustees, parent or child company, profit/loss sheet can add valuable features to the dataset. This additional information can be used to assign roles to the accounts of importance and other connected accounts can also be investigated [20].

Finally, the analysis done in this report [section 5.2] proves that network analysis methods are effective in isolating communities of complex network and accounts with higher measurement of centrality can be identified and can be examined for money laundering with additional information. Due to the effectiveness of network analysis in combination with machine learning, these techniques are used in anti-money laundering software [21]

References

[1] https://en.wikipedia.org/wiki/Anti%E2%80%93money_laundering_software.

[2] Department of the Treasury Financial Crimes Enforcement Network, The Role of Domestic Shell Companies in Financial Crime and Money Laundering: Limited Liability Companies, November 2006.

[3] Maryam Mahootiha, s. Alireza Hashemi G., Designing a New Method for Detecting Money Laundering based on Social Network Analysis. Conference: 2021 26th International Computer Conference, Computer Society of Iran (CSICC), March 2021.

[4] https://www.kaggle.com/datasets/anshankul/ibm-amlsim-example-dataset

[5] https://www.legislation.gov.uk/ukpga/2002/29/part/7

[6] https://financialcrimeacademy.org/key-elements-of-money-laundering/

[7] Hamed Tofangsaz, A New Approach To The Criminalization of Terrorist Financing And Its Compatibility With Sharia Law. The University of Waikato, Journal of Money Laundering Control, October 2012.

[8] https://www.cylynx.io/blog/network-analytics-for-fraud-detection-in-banking-and-finance/

[9] Daniel Otero Gómez, Santiago Cartagena Agudelo, Andrés Ospina Patiño, Anomaly Detection applied to Money Laundering Detection using Ensemble Learning. Mathematical Engineering Department of Mathematical Sciences, School of Sciences, Universidad EAFIT, December 2021.

[10] Waleed Hilal (a), S. Andrew Gadsden (a), John Yawney (b), Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. (a) McMaster University, Canada, (b) Adastra Corporation, Canada - January 2022.

[11] Dattatray Kute - UNSW Sydney, Biswajeet Pradhan-University of Technology Sydney, Nagesh Shukla - Griffith University, Abdullah Alamri, Deep Learning and Explainable Artificial Intelligence Techniques Applied for Detecting Money Laundering: A Critical Review, DOI:10.1109/ACCESS.2021.3086230, June 2021.

[12] J.E. Dobson, on reading and interpreting black box deep neural networks.

Int J Digit Humanities 5, 431–449 (2023). https://doi.org/10.1007/s42803-023-00075-w, november 2023.

[13] Berkan Oztas, Deniz Cetinkaya, Festus Adedoyin, Marcin Budka, Enhancing Transaction Monitoring Controls to Detect Money Laundering Using Machine Learning. Department of Computing and Informatics, Bournemouth University, United Kingdom.

[14] José-de-Jesús Rocha-Salazar and María-Jesús Segovia-Vargas, Money Laundering in the Age of Cybercrime and Emerging Technologies. DOI: 10.5772/intechopen.1004006, January 2024.

[15] Rasmus Ingemann Tuffveson Jensen(a) (b), Alexandros Iosifidis (a), Qualifying and raising anti-money laundering alarms with deep learning. (a) Department of Electrical and Computer Engineering, Aarhus University, Finlandsgade 22, 8200 Aarhus, Denmark, (b) Spar Nord Bank, Skelagervej 9, 9000, Aalborg, Denmark, November 2022.

[16] Abdallah Alshantti, and Adil Rasheed, Self-Organising Map Based Framework for Investigating Accounts Suspected of Money Laundering. DOI: 10.3389/frai.2021.761925.

[17] Rafał Dreżewski(a), Jan Sepielak(a), Wojciech Filipkowski(b), The Application of Social Network Analysis Algorithms in a System Supporting Money Laundering Detection. (a) AGH University of Science and Technology, Department of Computer Science, Kraków, Poland, (b) University of Białystok, Faculty of Law, Białystok, Poland.

[18] David Jancsics, Shell Companies and Government Corruption. School of Public Affairs, San Diego State University Imperial Valley Campus, San Diego, CA, USA.

[19] Brigitte Unger, Utrecht University - Joras Ferwerda, Utrecht University-Hans Nelen, Maastricht University, Money Laundering in the Real Estate Sector: Suspicious Properties. DOI:10.4337/9781781000915, January 2011.

[20] Ian Goodrich, Mapping the Laundromat: A Network Analysis of Money Laundering in the United Kingdom. Central European university, Budapest, Hungary – 2019.

[21] https://financialcrimeacademy.org/network-analysis-in-anti-money-laundering/