AML Cyclic Pattern

Introduction

In the realm of digital financial systems, the battle against money laundering stands as a paramount challenge, especially the face of sophisticated networks of anomalous accounts and mule operators. As the IMPS landscape evolves so do the tactics employed by illicit actors seeking to exploit vulnerabilities within established systems. The sheer volume of transactions in digital financial systems like IMPS exacerbates the challenge, making it increasingly difficult to uncover patterns indicative of illicit activities.

With the IMPS ecosystem being established at a ubiquitous scale, it becomes complicated to represent the data in a traditional relational based database system and deploy machine learning algorithms at such a huge scale to identify trends and uncover hidden patterns in the data. Overcoming the drawbacks of a traditional relational database system, graph machine learning plays a pivotal role by leveraging interconnected data points to identify suspicious patterns and activities. By representing accounts and transactions as nodes and edges in a graph, Graph-based machine learning algorithms can analyse complex networks by visualizing the cash flow at different levels of the underlying network.

This project employs the Louvain method, an unsupervised machine learning algorithm, to partition the graph into smaller segments, facilitating individual analysis. Utilizing this segmentation, the project applies various indices—such as the Gini index, Small World index, and Reciprocal Ratio—to identify segments exhibiting high inequality, small-world properties, and concentration of bidirectional links. These characteristics are indicative of suspicious and laundering behaviours within financial networks. Furthermore, the project conducts cyclic pattern detection on these segments, identifying elementary circuits in directed graphs that signify circular transaction patterns commonly associated with money laundering activities. This comprehensive approach enables the targeted identification and analysis of potentially illicit behaviours within complex financial transaction networks.

Methodology

A graph network is represented by its nodes and edges. The accounts represent the nodes and the transactions between these accounts are represented by its edges. This allows us to model relationships using graph-based algorithms which increases the efficacy of modelling large scale data.

The IMPS data for the 60-days' time period contains 74 million nodes and 74 million edges and a sample of the data is shown below.

Node id	Account number	Anomalous flag
1	SBI123	0
2	ICI345	0
3	AXI789	1

Table1: Node data

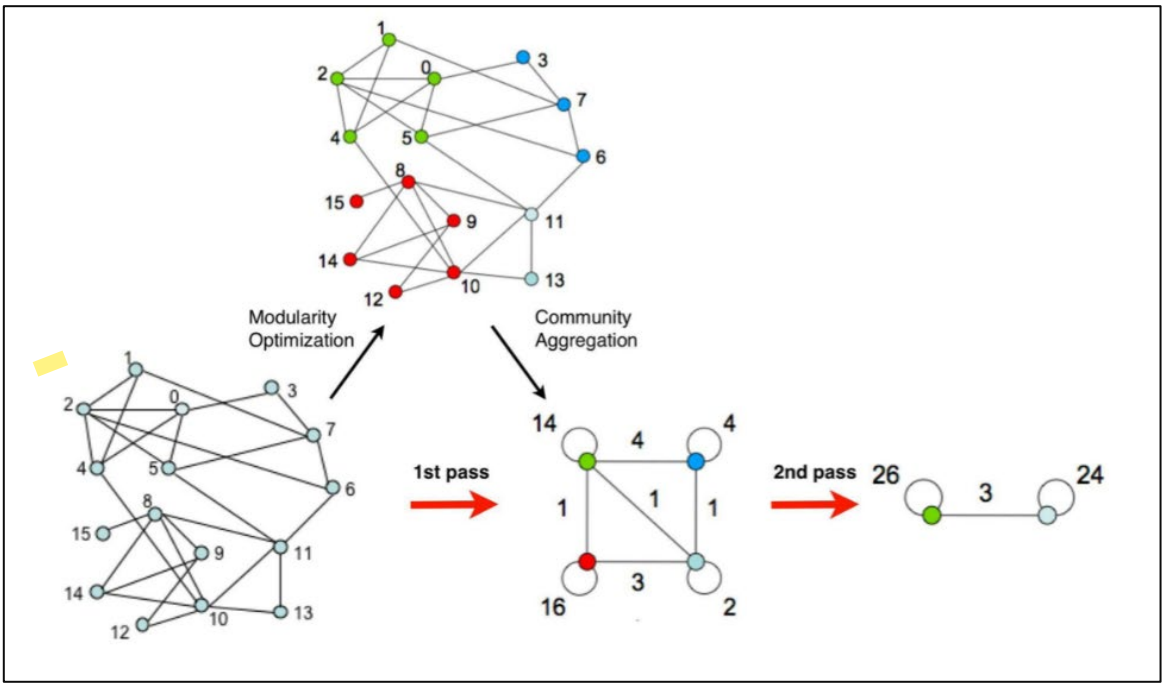
Modularity is a measure of the structure of networks, identifying the strength of the division of a network into communities. Higher modularity indicates dense connections between the nodes within communities but sparse connections between nodes in different communities. Modularity optimization aims to maximize this measure to find the best community structure.

From Node id	To Node id	Total Amount	Total count
1	3	6000	1
2	1	100	2
1	2	3000	2

Table2: Edge features

The IMPS aggregated data for a 60-days' time period is used for the creation of the graph. The node file and the edge file are pre-processed and loaded into a cu data frame. The anomalous data file is pre-processed using suitable pre-processing techniques for the analysis. The anomalous flag for the node property for the 60-days data is then extracted and updated accordingly with the mapped node IDs. A graph is created using cuGraph with accounts as nodes and transactions as the relationship between these nodes.

To identify suspicious accounts pertaining to money laundering activities, it is essential to partition the graph for a scalable analysis. This is done using the cuGraph Louvain algorithm which is an unsupervised machine learning algorithm based on the idea of optimising a measure called modularity. Modularity quantifies the quality of a community structure by comparing the number of edges withing communities to the expected number of edges if the network were randomly connected. The algorithm aims to maximise the modularity score, indicating a strong community structure.



Visualisation of Louvain Algorithm

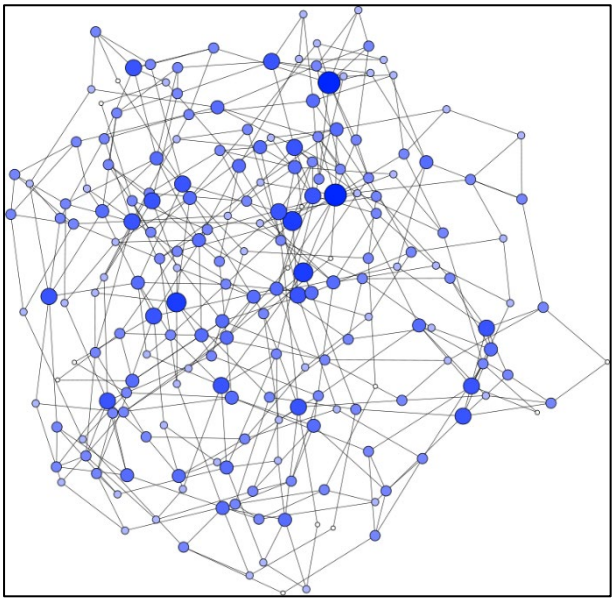
Following modularity optimisation, the algorithm performs hierarchical clustering to further refine the partitions. It merges communities that yield the greatest increase in modularity, resulting in hierarchical structure of nested communities. The Louvain method returns a cu data frame with partition IDs and the respective node IDs belonging to those partitions. Once, the communities are formed, the anomalous accounts reported by the banks are mapped within the communities and only such communities containing the bank reported anomalous accounts are

included for further analysis. This gives us the filtered cu data frame. This allows for the identification of highly suspicious communities for a more focussed analysis.

Indices like the Gini Index, Small World Index and Reciprocal Ratio are further applied to these partitions to detect money laundering communities using network analysis. For the application of the indices, the partitions are processed individually, by taking each partition and the subset of edges pertaining to that partition from the edge data frame.

Reciprocal Ratio in a network is calculated by taking the ratio of number of links that are bidirectional (i.e., node A to B and B to A) to the number by the total number of links in the community. A higher value indicates more mutual transactions, which could be suspicious in the context of money laundering as launderers might create circular patterns of transactions to obfuscate the money's origin.

The Gini Index is calculated by taking the absolute differences between all pairs of nodes' transaction amounts, summing these differences, and then dividing by twice the square of the number of nodes multiplied by the average value of the transactions. A high Gini Index indicates a high level of inequality in transaction connections among the nodes, which could be a sign of centralized control in money laundering operations. In money laundering, a few nodes (individuals/entities) might handle most transactions, suggesting control or coordination indicative of laundering activities.



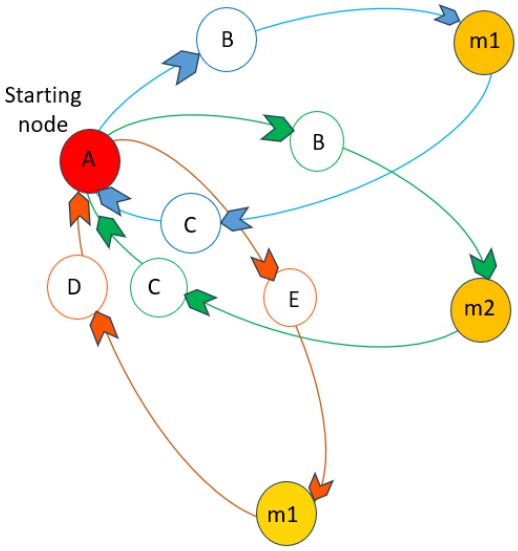
Small World Index

To determine if a network exhibits small world properties, both its clustering coefficient and average path length are calculated. The clustering coefficient measures how interconnected the nodes are, and the average path length assesses the average number of steps needed to connect any two nodes. The clustering coefficient is calculated using NetworkX library's average clustering algorithm and the average path length is measured using NetworkX library's average shortest path length algorithm. A network is considered a small world if it has a higher clustering coefficient and approximately the same average path length as a random network. In money laundering detection, a high small world index might indicate a closely knit group, potentially engaging in coordinated laundering activities.

A total of 12 communities including anomalous mapped accounts and 12 communities not including any anomalous mapped accounts are sampled and the indices are applied to them. The threshold for each index is determined by where most anomalous activities lie within the resulting data frame. Subsequently, these thresholds are implemented on partitions containing anomalous accounts, giving us partitions indicative of potentially suspicious activity characterized by high inequality, small world properties, and a concentration of bidirectional links—signalling potential money laundering behaviours.

Cyclic pattern detection is crucial to our use case since identifying circular flows exposes attempts to conceal the origin or destination of illicit funds. It also uncovers complex layering techniques employed in money laundering.

For the identification of cyclic patterns, individual partitions are scrutinized. Within each partition, anomalous accounts are singled out, forming directed subgraphs inclusive of these accounts, their neighbouring nodes, and corresponding edges sourced from the edge data frame. These subgraphs are then subjected to NetworkX library's simple cycles method, furnishing a comprehensive list of cycles originating from nodes within the subgraph.



Cycles for a node

Features including transactional characteristics such as in bound and out bound amounts and counts, in and out degrees of accounts, along with the features based on cyclic patterns originating from specific nodes are extracted. A meticulous threshold analysis is conducted, and thresholds are refined based on the relative presence of anomalous accounts across different bins, subsequently identifying suspicious accounts, culminating in the final selection of targeted accounts.

Analysis

We focus on determining optimal threshold values for Gini index, Small World index, and Reciprocal ratio. This involves a systematic approach where we sample sets of partitions containing anomalous accounts and sets devoid of such accounts based on their density.

Specifically, we identify the top k partitions comprising anomalous accounts and the top k partitions without any anomalous accounts, sorted by the descending order of the number of nodes within each partition. Subsequently, we calculate the Gini index, Small World index, and Reciprocal ratio for these sampled partitions.

By observing the distribution of these indices separately, we ascertain threshold values where the majority of partitions containing anomalous accounts lie. This approach enables us to establish robust criteria for distinguishing between partitions indicative of potentially anomalous activity and those reflecting normal transaction behaviour.

Once partitions surpass predefined threshold values, we employ the simple cycles method to identify potential cyclic patterns within the data.

Subsequently, we focus our analysis on the accounts from which these cycles originate. The relevant account data, along with the detected cycles, are stored in a cu data frame, with an "anomalous" flag mapped based on node IDs. Various key metrics such as total transaction amounts, transaction counts, maximum and minimum values, as well as average values of counts and amounts, in degrees and out degrees, are aggregated from the edge data frame. Additionally, features such as cycle count per account and cycle length are extracted for further insights.

To capture the nuanced behaviour of suspicious accounts, we only consider cycles of lengths ranging from 3 to 12. This is based on the observation that anomalous accounts often operate within smaller, tightly interconnected communities. We also compute ratios between various transaction metrics to identify potential mule behaviour. An account is flagged as suspicious if it exceeds thresholds in at least five of these metrics.

The resultant accounts are divided into two categories, accounts with more than 50 cycles originating from them and accounts with 50 or less than 50 cycles originating from them. On the accounts with cycle length 50 or less, further validation is performed by analysing the ratio of total in and out amounts at the cycle level, excluding transactions occurring outside the identified cycles. Suspicious nodes are those whose ratio of total transaction amounts at the cycle level closely approaches 1, indicating potentially anomalous behaviour.

Conclusion

Identifying unseen anomalous trends in the data helps us to capture the characteristics of suspicious activities happening in the network which is critical for a financial institution in the payment sector. Rapid advancements in the technological landscape have made it possible to employ the latest techniques. Graph-based algorithms provide a platform to understand, analyze and provide comprehensive insights on the flow of the transactions.

This paper demonstrates the implementation of the Louvain algorithm for graph partitioning to subsequently perform network analysis using Gini index, Small World index, reciprocal ratio for identification of suspicious networks. Furthermore, cycles are generated in highly suspicious partitions in the subgraphs made by anomalous accounts and their neighbors. Features based on transaction features like amounts and counts are derived along with features based on cycles. The number of accounts falling in any five features are identified.

By employing this comprehensive approach, we aim to provide financial institutions with a robust framework for detecting and mitigating suspicious activity, ultimately safeguarding against financial fraud, and ensuring regulatory compliance.