CS 732 / Data Visualization
Project Assignment-4

Dev Patel IMT2018021
Ishmeet Saggu IMT2018030
Prateek Kamboj IMT2018057

Abstract

This technical report contains a brief overview of the methodology involved in the visualization of the provided Rabobank dataset collected from the Cooperative Rabobank, a Dutch multinational banking and financial services company to create networks and visualize them.

1 Introduction

The provided dataset contains information about 1.6 Million anonymized bank accounts. We assign two weights on each edge, which are the aggregated transferred amount and the total number of transactions between the users from the year 2010 2020. In this report we presented a detailed analysis of unweighted and both weighted networks by examining their degree, strength, and weight distributions. The analysis of bank transaction network will help in understanding the flow of money at microscopic level as well as how this contributes towards the macroscopic money transaction system.

2 Data Overview

This dataset consists of bank accounts and transactions between them. Thus, it can be be organized into a transactions network. The data was shared for 1,624,030 bank accounts and 4127043 transactions based on (from_account, to_account) pair. Considering the transactions information, we created two weighted networks (i) edge-weight is total amount of money transferred between two accounts, and (ii) edge-weight is the total number of the transactions between two accounts.

3 Tools ans Methodology

3.1 Tools

The exhaustive set of libraries used for generating the final inference visualizations involve:

- numpy
- pandas
- plotly
- gephi
- matplotlib

3.2 General Methodology

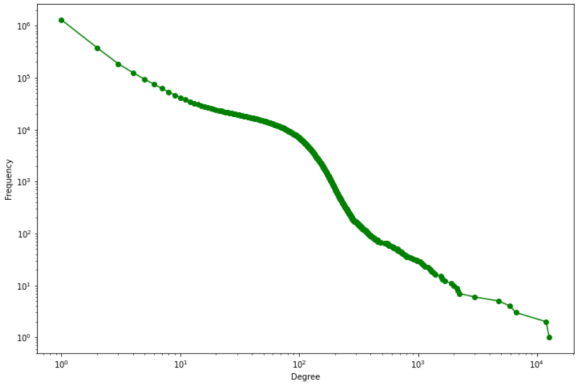
For generating the aforementioned visualizations of network dataset, we have two stage represented in two files (.ipynb and .gephi) of code to be run to get the final output. The workflow proceeds in order:

- First, the data from the csv files provided by Rabobank are pre-processed to capture only relevant information.
- Secondly, the pre-processed csv are then used to generate Node and Edge to input to Gephi.
- Lastly, we can use Gephi to generate visulizations

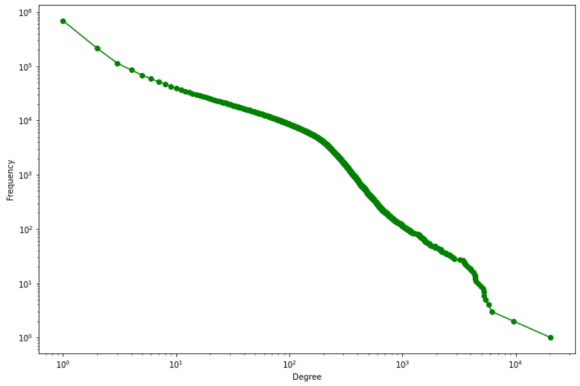
As mentioned above their are 1.6 Million nodes representing bank accounts and 4 Million edges representing transactions between bank accounts. Visualizing the network with all nodes and edges is infeasible. Hence, for visulizations done below, we have applied some sampling techniques.

4 Degree Distribution

The below Figure 1, shows cumulative in- and out- degree distribution of the whole network. The cumulative degree $P(k)$ is defined as $P(k) = \int_k^\infty p(x)dx$ where p(k) is the number of nodes with degree k. This network has a rapidly decaying degree distribution where hubs are few, and therefore this network follows the properties of traditional scale-free networks such as anomalous diffusion.



(a) Cumulative In-degree distribution



(b) Cumulative Out-degree distribution

Figure 1

5 Network diagram

5.1 Fruchterman Reingold

Methodology used for fruchterman reingold diagram

- Firstly, csv file is pre-processed and then each bank-account number is mapped to a integer number.
- Secondly, as a sampling technique we have extracted top 200 node with highest number of transaction started from(highest outdegree). And used this nodes as subgraph for visualization.
- Save the dataframe into node.csv and edge.csv

Distinct techniques used for Fruchterman Reingold diagram in Gephi

- Layout : Fruchterman Reingold
- Nodesize : outdegree Centrality (higher outdegree centrality bigger the nodesize)
- Nodecolor : Betweenness Centrality (higher betweenness leads to green color)

- Edgeweight : total amount of transaction between the accounts.

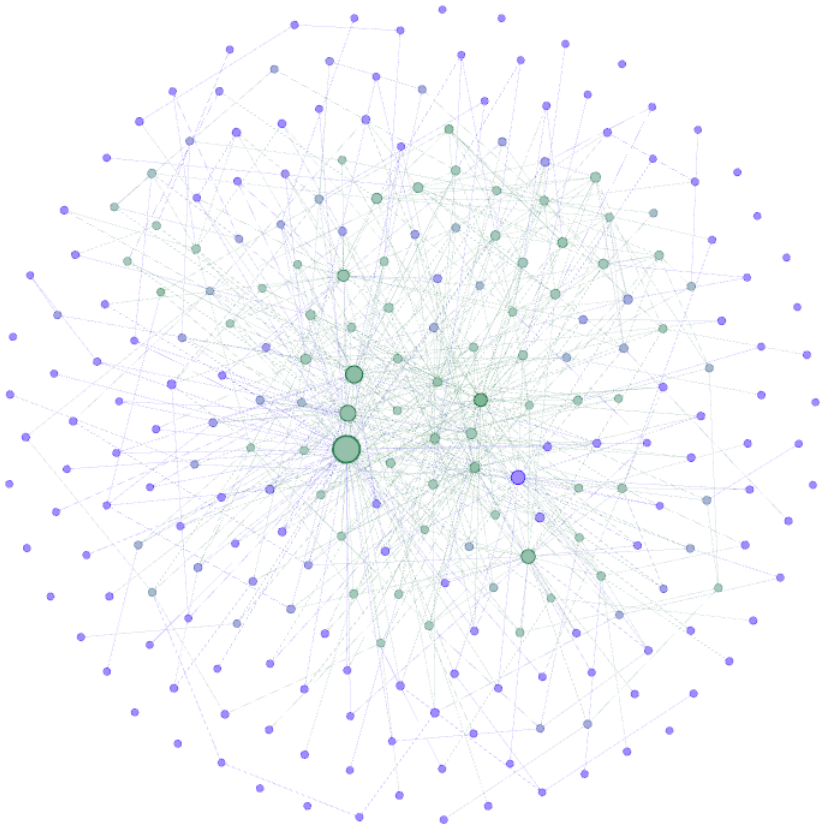


Figure 2: Reingold layout for top 200 outdegree nodes subgraph

5.2 Circular

Methodology used for Circular diagram

- Firstly, csv file is pre-processed and then each bank-account number is mapped to a integer number.
- Secondly, as a sampling technique we have extracted top 200 node with highest number of transaction ended too(highest indegree). And used this nodes as subgraph for visualization.
- Save the dataframe into node.csv and edge.csv

Distinct techniques used for Circular diagram in Gephi

- Layout : Circular
- Nodesize : indegree Centrality (higher indegree centrality bigger the nodesize)
- Nodecolor : same for each node
- Edgecolor : based on edge weight

- Edgeweight : total amount of transaction between the accounts.

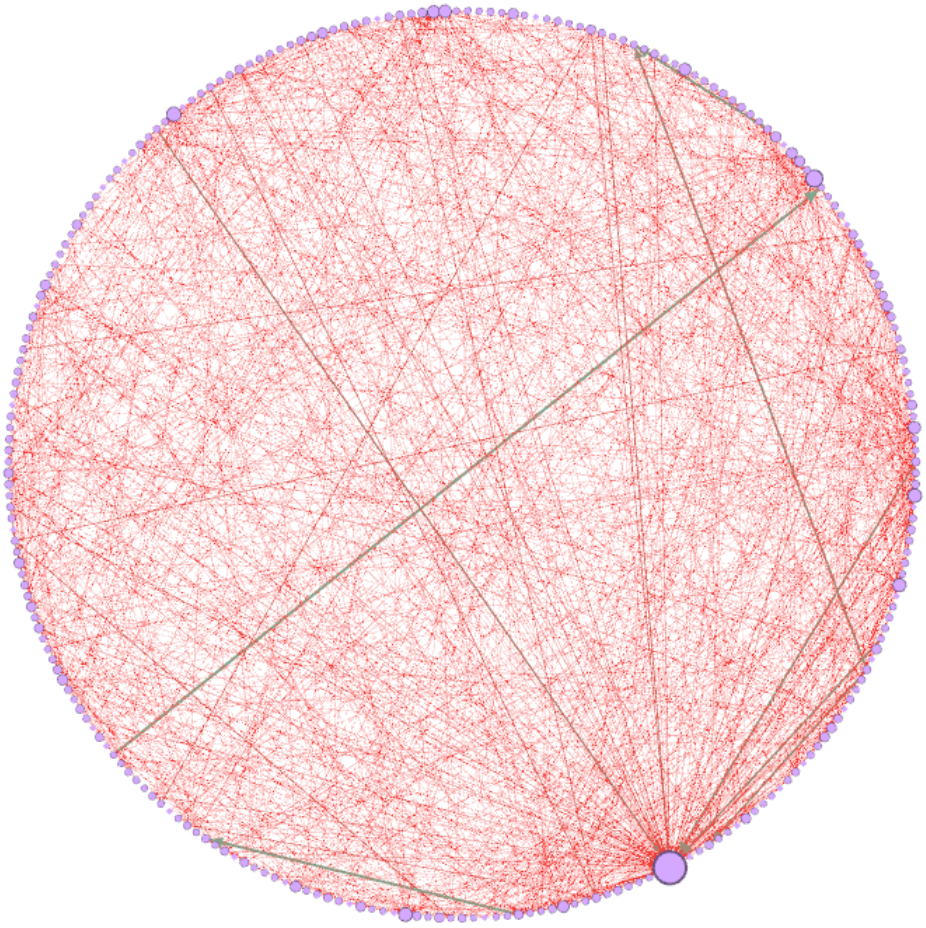


Figure 3: Circular layout for top 200 indegree nodes subgraph

5.3 ForceAtlas 2

Methodology used for ForceAtlas diagram

- Firstly, csv file is pre-processed and then each bank-account number is mapped to a integer number.
- Secondly, as a sampling technique we have used snowball sampling. In this sampling subgraph has been extracted from the network by picking a source node at random and including all nodes in the sample that are within a topological distance of $l = 3$ from the source node.
- Save the dataframe into node.csv and edge.csv

Distinct techniques used for ForceAtlas diagram in Gephi

- Layout : ForceAtlas 2

- Nodesize : Closeness centrality (higher closeness bigger the nodesize)
- Nodecolor : Degree centrality (higher the degree centrality darker the node color)

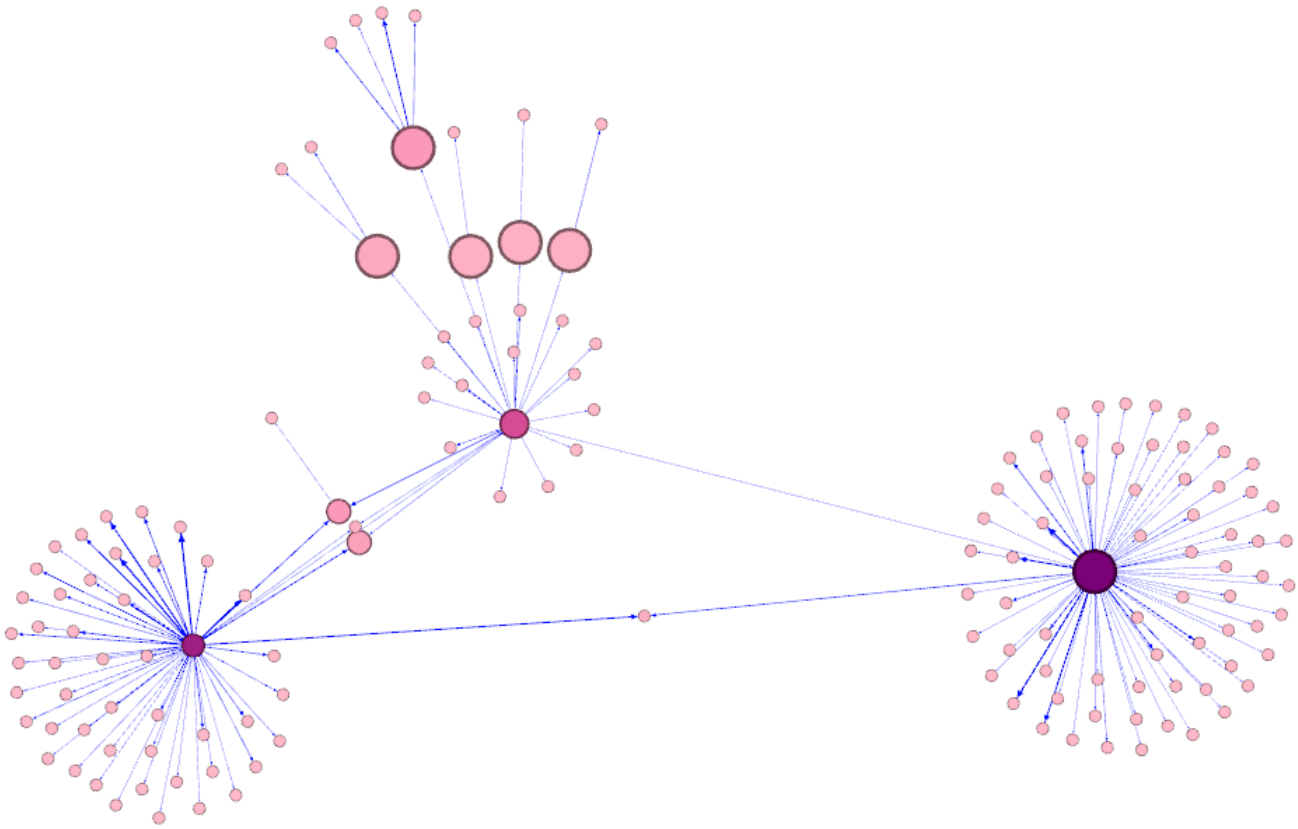


Figure 4: ForceAtlas 2 layout with snowball sampling till distance l=3

6 Total Amount Distribution

The below figure 5, show the total amount of transactions between all the account holder separated by years. With the economic growth and advance of digital finance infrastructure it can guessed that total amount of transactions will increase with the time. Which is also evident from the above image, as the total amount of transactions is continuously increasing with every subsequent year passing. With the development of UPI-based payments as well as app-based payments just pushed the boundaries and has since witnessed blossoming of a myriad of payment systems, and a gradual shift in the customer behaviour from cash to digital payments.

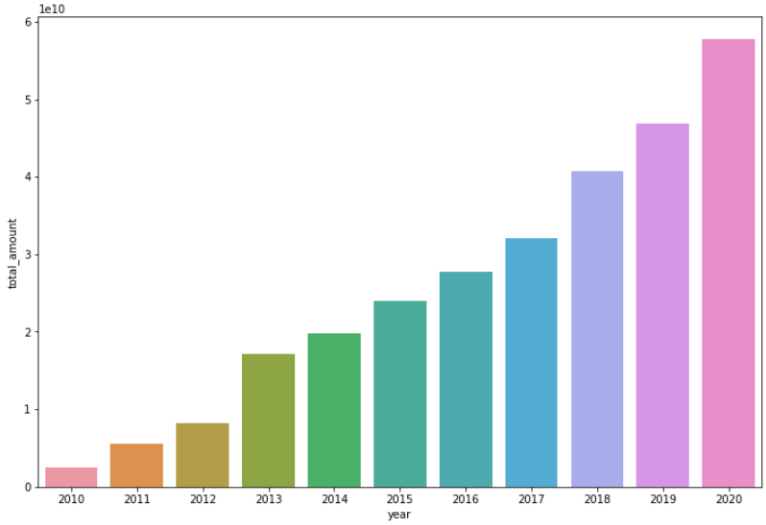
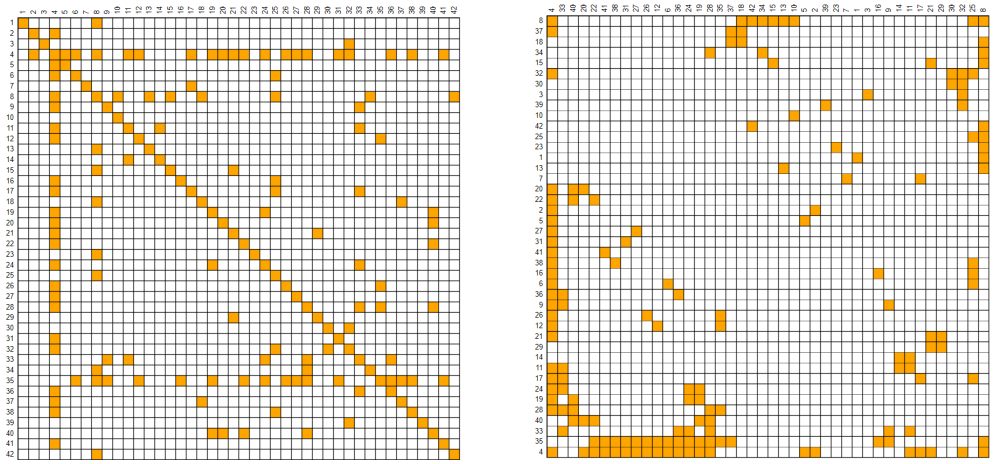


Figure 5: Total amount of transactions distribution by years

7 Adjacency Matrix Visualization



(a) Initial adjacency matrix visualization

(b) Optimal Leaf Ordering matrix visualization

Figure 6

We have first use random sub sampling technique, to sample out a graph with 50 nodes. Then we generate adjacency matrix and used standard technique to display the adjacency matrix visualization. Then we used [5] to generate Robinson Matrix Reordering (optimal leaf ordering). We can see their are some clusters forming in the reorder matrix, which could be seen as a part of community.

8 Conclusion

In conclusion, it is worth to mention that the report is useful in understanding the flow of money in society as it is derived from one-to-one money transactions. We believe that this bank transaction network and our analysis of the network's characteristics can help solve some practical tasks in the financial domain. For instance, in anomaly detection, the deviation of anomalous users' transaction behavior from structural characteristics of normal users might help in the early detection of suspicious and anomalous users.

References

[1] https://gephi.org/users/tutorial-layouts/

[2] https://towardsdatascience.com/large-graph-visualization-tools-and-approaches-2b8758a1cd59

[3] https://cambridge-intelligence.com/visualize-large-networks/

[4] https://neo4j.com/blog/5-ways-to-tackle-big-graph-data-keylines-neo4j/

[5] https://github.com/jdfekete/reorder.js

[6] https://stackoverflow.com/questions/65118158/plot-cumulative-distribution-with-networkx-and-numpy

[7] https://towardsdatascience.com/a-quick-tutorial-on-gephi-layouts-daa87fec5a20