The Banking Transactions Dataset and its Comparative Analysis with Scale-free Networks

Akrati Saxena*, Yulong Pei*, Jan Veldsink[†], Werner van Ipenburg[†], George Fletcher*, Mykola Pechenizkiy*

*Eindhoven University of Technology, Eindhoven, The Netherlands

{a.saxena,y.pei,l.g.h.l.fletcher,m.pechenizkiy}@tue.nl

[†]Coöperatieve Rabobank U.A., Utrecht, The Netherlands

{jan.veldsink,werner.van.ipenburg}@rabobank.nl

Abstract—We construct a network of 1.6 million nodes from banking transactions of users of Rabobank. We assign two weights on each edge, which are the aggregate transferred amount and the total number of transactions between the users from the year 2010 to 2020. We present a detailed analysis of the unweighted and both weighted networks by examining their degree, strength, and weight distributions, as well as the topological assortativity and weighted assortativity, clustering, and weighted clustering, together with correlations between these quantities. We further study the meso-scale properties of the networks and compare them to a randomized reference system. This will be the first publicly shared dataset of intra-bank transactions, and this work highlights the unique characteristics of banking transaction networks with other scale-free networks.

I. INTRODUCTION

The analysis of large-scale complex networks, including social networks, technological networks, biological networks, has offered in-depth insights to uncover and explore their evolution and understand several dynamic phenomena taking place on these networks [1]. The analysis of bank transaction network will help in understanding the flow of money at the microscopic level as well as how this contributes towards the macroscopic money transaction system. It will also provide insights to improve the downstream tasks in the financial domain with guidance from the topological perspective. Some representative tasks include fraud detection and user classification in transaction networks.

In this paper, we present a detailed analysis of unweighted and weighted networks constructed from a data set of the money transactions of more than 1.6 million users of Rabobank. This paper analyzes unweighted and weighted, large-scale, one-to-one bank transaction networks. We follow the ‘cookbook approach’ for performing a systematic analysis of basic and more advanced network characteristics, which has also been followed in previous network analysis work [2]. We explain the characteristics of the dataset and networks in Section

II. We study some of the basic network characteristics in Section III. We explore meso-scale structural properties of the network, including k-cliques analysis in Section IV. We present a comparative analysis of the unique characteristics of the transaction network with other scale-free networks in Section V and, finally, discuss our findings and future directions in Section VI.

Our main contributions are mentioned below.

- 1) The anonymized banking transaction data of users. As per the best of our knowledge, this will be the first publicly available dataset of users’ banking transactions. The dataset is available at https://github.com/akratiit/RaboBank_Dataset.
- 2) We perform a detailed analysis of unweighted and weighted banking transaction networks and highlight the similarities and differences of these networks with other types of networks. As per the best of our knowledge, this is the first work to analyze intra-bank transaction network. We hope the shared data and this work will help further understand the evolution of money flow in society.

II. DATASET AND NETWORKS

The dataset is collected from the Coöperatieve Rabobank U.A.¹, a Dutch multinational banking and financial services company. This dataset consists of bank accounts and transactions between them. For any pairs of accounts with one or more transactions, we also collected more information, including the numbers of transactions between two accounts and the total amount of money transferred from one account to another over a period of 11 years from 2010 to 2020. Thus, it can be organized into a transaction network, either unweighted or weighted. The data was shared for 1,624,030 bank accounts and 4,127,043 transactions based on (from_account, to_account) pair.

We create the network from this data, having 1,624,030 nodes which are accounts, and 3,823,167 edges which represent that the respective users performed one or more transactions. Next, we identify the weakly connected components in the network, and it contains 723 connected components, where the largest weakly connected component contains 1,622,173 nodes and

This work is licensed under a Creative Commons Attribution International 4.0 License.

ASONAM '21, November 8–11, 2021, Virtual Event, Netherlands

© 2021 Copyright is held by the owner/author(s).

ACM ISBN 978-1-4503-9128-3/21/11

<https://doi.org/10.1145/3487351.3488339>

¹<https://www.rabobank.nl/>

Fig. 1: A sample subgraph extracted from the black node (chosen uniformly at random) till distance $l = 3$.

3,821,514 edges. The information of connected components is: 1 component has 1,622,173 nodes, 3 components have 27, 15, and 13 nodes, and the rest of the components have less than ten nodes. The weakly largest connected component will be referred to as $G(V, E)$ where V is the set of node and E is the set of edges. In our analysis, we mainly consider the weakly largest connected component (WLCC) if not mentioned otherwise.

Considering the transaction information, we create two weighted networks (i) edge-weight is the total amount of money transferred between two accounts, and (ii) edge-weight is the total number of transactions between two accounts. The first network will be referred to as G^T , and the second will be referred to as G^N .

III. BASIC NETWORK CHARACTERISTICS

We show a small subgraph of the network in Fig. 1 that is sampled using snowball sampling [3]. The sample has been extracted from the network by picking a source node at random (shown with black color) and including all nodes in the sample that are within a topological distance of $l = 3$ from the source node. The node size and color are scaled based on the total degree (in-degree + out-degree) of the node. It appears from this figure that the network consists of a few hubs that are well connected with each other and small clusters of nodes as previously observed in other scale-free networks [2], [4].

To further understand network structure, we show in Fig. 2 the average number of nodes $N_s(l)$ as a function of distance l , where $N_s(l)$ represents the number of nodes within a distance l from a given source node. The plots are shown for several choices of the source node using solid colored lines, and the average value of $N_s(l)$ for uniformly sampled 0.1% source nodes is shown by the black dashed line. The $N_s(l)$ is computed for a small sample due to the high computational complexity. To a good approximation, we analyze that the curve follows the Boltzman equation $B + (A - B)/(1 + (x/x_0)^p)$, where A , B , x_0 , and p are the parameters.

Fig. 2: Average Number of nodes ($N_s(l)$) at distance l as a function of distance (l) for uniformly sampled 0.1% source nodes (dashed black lines) and the number of nodes at distance l for some random nodes (solid thin lines).

Fig. 3: Cumulative In-degree and out-degree distribution.

We further analyze the macro-scale properties of the network. In Fig. 3, we show the cumulative in- and out-degree distribution of the weakly largest connected component (WLCC). The cumulative degree distribution $P_>(k)$ is defined as $P_>(k) = \int_k^\infty p(x)dx$, where $p(x)$ is the degree probability density function. The cumulative degree distribution $P_>(k)$ for the WLCC is approximated well by the power-law equation of the form $P(k) = c*(k)^{(-\gamma)}$, having the exponent $\gamma \approx 1.43$ and 1.13 for the cumulative in-degree and out-degree distribution, respectively.

For a better understanding of the network, we plot the out-degree versus in-degree in Fig. 4; the plot shows that the out-degree is not correlated with in-degree and has the Spearman correlation coefficient -0.15 . It indicates that there is no correlation between in-degree and out-degree. That is reasonable because most accounts that belong to regular customers have more transfer-out operations, e.g., purchasing products, than transfer-in operations, e.g., receiving salaries.

By taking into consideration the transaction information, i.e., times of transactions and amount of money, we construct two weighted networks, G^T , and G^N . The cumulative edge-weight distributions are shown in Fig. 5.

The weights on edges contain richer information of transactions, so we compute the strength where the in-strength $s^{in}(i)$ and out-strength $s^{out}(i)$ of a node i is the sum of weights of incoming edges and out-going edges, respectively. We investigate the cumulative in- and out-strength distributions

Fig. 4: Out-degree versus In-degree for WLCC.

 Fig. 5: Cumulative edge-weight distribution for G^T and G^N .

of both weighted networks, and the plots are shown in Fig. 6. As expected, both distributions follow the power law. The distributions of in-strength and out-strength are similar because in the process of data collection, we filter out special accounts such as gas stations that have much higher in-degrees than out-degrees.

A. Neighborhood Analysis

For further understanding, we analyze the characteristics of the local neighborhood of nodes. We first compute the assortativity of degree and strength of the network that measures the similarity of connections in the network with respect to the node degree or strength, respectively [5]. The assortativity values are mentioned in Table I, and both, degree and strength, do not follow assortativity or dis-assortativity. These results are not in correlation with the assortativity observed in other scale-free networks [5], [2].

Fig. 6: Cumulative In-strength and Out-strength distribution.

Assortativity	Value
In-degree Assortativity	-0.018
Out-degree Assortativity	-0.011
In-strength Assortativity G^T	-0.004
Out-strength Assortativity G^T	-0.005
In-strength Assortativity G^N	-0.005
Out-strength Assortativity G^N	-0.003

TABLE I: In and Out- Degree and Strength Assortativity values for different networks.

Fig. 7: Average neighbor In-degree versus In-degree and average neighbor out-degree versus out-degree.

We plot the average neighbor In-degree versus In-degree $\langle k_{nn}^{in} | k^{in} \rangle$ and average neighbor out-degree versus out-degree $\langle k_{nn}^{out} | k^{out} \rangle$ in Fig. 7, and observe that the most of the users have a high average neighbor degree irrespective of in/out-degree of a user. The average neighbor In-strength versus In-strength $\langle s_{nn}^{in} | s^{in} \rangle$ and average neighbor out-strength versus out-strength $\langle s_{nn}^{out} | s^{out} \rangle$ for G^T network are shown in Fig. 8 and no correlation is observed due to the basic network characteristics. The similar results were observed for G^N [6].

Next, we analyze the clustering coefficient of nodes for unweighted and weighted networks [7]. The average clustering coefficient versus degree is plotted in Fig. 9. The clustering coefficient decreases with the total degree of nodes. It shows that when a user makes more transactions with different accounts, it is less probable that the neighboring accounts will make a transaction with each other. The clustering coefficient follows the same pattern as observed in other large-scale scale-free networks [5], [2].

 Fig. 8: Average Neighbor In-strength versus In-strength and average neighbor out-strength versus out-strength for G^T .

Fig. 9: Average clustering coefficient versus total degree.

Fig. 10: No. of Communities versus community size.

free networks [2], [8].

In Table II, we summarize the characteristics of the considered networks that include the minimum, maximum, mean, standard deviation, skewness², and Kurtosis³ values of various basic network properties.

IV. MESO-SCALE NETWORK CHARACTERISTICS

In this section, we analyze the meso-scale characteristics of the network. First to analyze the local structure of the network, we compare the presence of cliques of different orders as these are considered the basic building blocks and moderate different functionalities of the network. We compare the cliques of order $k = 1, 2, \dots, 8$ with random networks in which the probability to have a connection is the same for all pairs of node. The random networks are created using Erdős-Rényi (ER) network generative model [9]. The link formation probability in the ER graph is computed as, $p = 2m/(n(n-1))$, and the expected number $E[X]$ of cliques with k nodes is computed as $E[X] = \binom{n}{k} (k!/a) p^l$, where $l = k(k-1)/2$ and $a = k!$ is the number of automorphic graphs, those are isomorphic to node-edge order preserved permutation of the graph [10]. The cliques are counted by considering the undirected version of the network, and the results are shown in Table III. The presence of high-order cliques (cliques beyond order three) in the Rabobank network makes it starkly different from a random network that has a very low probability of high-order cliques being present.

To further understand how the users are connected in a group and how these groups are further connected with each other, we analyze the community structure of the network. For the Rabobank data, the information of ground truth communities is not known. However, the unavailability of the ground truth data is a well-known issue in analyzing network data, and the researchers have proposed several methods to identify communities using the structural properties of the network. We use the Leiden community detection method [11]. The distribution of the number of communities versus community size is shown in Fig. 10. We observe that there are several

small-size communities and a few big-size communities as observed in several scale-free networks [12]. The communities for network G are linearly binned with bucket size 100. The detailed results for community analysis are available at [6]. The results show that the bank transaction networks contain community structure as observed in other networks, such as social networks, co-authorship networks, financial networks, and so on [12].

We also analyze the core-periphery structure of the network using the k-shell decomposition, that is another well-known meso-scale network structure [13]. In a banking transaction network, the core-periphery network will provide insights into the flow of money from peripheral to more central nodes. We observe that the inner core layers have a very small number of nodes and has the core-periphery structure as observed in other networks [13].

V. TRANSACTION NETWORK VS. SCALE-FREE NETWORKS

The banking transaction network follows the characteristics of scale-free networks; still, it has some clear similarities and differences with other scale-free real-world networks. The degree distribution, strength distribution, edge-weight follows power law as observed in other social, information, biological or technological networks [4]; however, there is no correlation between in-degree vs. out-degree, and in-strength vs. out-strength as observed in other Information networks, such as WWW [14]. In social and Information (transportation) networks, the strength of nodes increases with degree [2], [15]; however, the banking transaction network has different pattern due to the nature of the network evolution (please refer to Section III). These basic characteristics show that neighborhood connectivity is different from other kinds of networks. The network is neither assortative nor disassortative, as we observe in most of the other scale-free networks including trade and finance networks [5], [8]; therefore, no correlation with the neighborhood nodes' degree is observed.

The evolution of the network is not random and regulated by an underlying evolving mechanism. The network has community, and hierarchical structure [12], [13] as observed in other scale-free networks as shown in Section IV. One another main difference that we observed is the correlation of weak ties with edge-weights. In social networks, the weak ties, or also referred to as bridges, are the connections between communities

²The skewness measures the asymmetry of the given probability distribution about its mean.

³Kurtosis computes how heavily the tail of the distribution differs from the tail of a normal distribution.

		Min	Max	Mean	Std	Skewness	Kurtosis
G	In-degree	0	2.04×10^4	2.36	32.66	2.17×10^2	1.05×10^5
	Out-degree	0	1.27×10^4	2.36	20.26	3.35×10^2	1.84×10^5
	Clustering Coefficient	0	1	0.02	0.12	7.29	53.95
G^T	In-strength	0	1.57×10^{10}	1.54×10^5	1.34×10^7	1.01×10^3	1.18×10^6
	Out-strength	0	4.04×10^9	1.54×10^5	6.53×10^6	4.71×10^2	2.61×10^5
	Edge-weight	1	3.35×10^9	6.56×10^4	2.75×10^6	8.99×10^2	9.43×10^5
	Weighted Clustering Coefficient	0	2.72×10^{-3}	2.31×10^{-7}	7.30×10^{-6}	1.62×10^2	4.22×10^4
G^N	In-strength	0	5.04×10^6	68.16	4.08×10^3	1.17×10^3	1.44×10^6
	Out-strength	0	3.11×10^6	68.16	2.94×10^3	8.19×10^2	8.01×10^5
	Edge-weight	1	2.58×10^4	28.93	1.19×10^2	34.02	3.24×10^3
	Weighted Clustering Coefficient	0	5.47×10^{-2}	1.41×10^{-5}	1.83×10^{-4}	71.15	1.24×10^4

TABLE II: Summary of descriptive network statistics.

Order	Empirical Count	ER Expectation
1	1622173	1622173
2	3318903	3318903
3	417461	11.42
4	34638	7.43×10^{-11}
5	3815	9.76×10^{-28}
6	417	2.70×10^{-50}
7	26	1.61×10^{-78}
8	0	1.12×10^{-112}

TABLE III: Number of cliques of order $k = 1, 2, \dots, 8$ in the undirected Rabobank network (empirical count) and their expectation values in a corresponding ER network.

and therefore have a lower edge-weight as the social interaction between people belonging to two different communities is not very strong [2]. However, in the banking transaction network, no such correlation is observed. The transactions that happened between users belonging to different communities might have lower or higher strength based on the transferred amount and the total number of transactions. The evolving model for banking transaction networks is still an open question, and the above statistical observations mentioning the similarities and differences will help pin down the underlying evolving mechanism.

VI. CONCLUSION

In this work, we constructed an unweighted network from bank transaction records of Rabobank and two weighted networks where edge-weights are assigned using the aggregated amount of transactions and the total number of transactions from 2010 to 2020. To our knowledge, it is the first intra-bank transaction network studied so far and will be the first bank transaction dataset that will be publicly available. The micro and meso scale network analysis shows that the network evolution is not random. To further analyze, we study the network's community structure that shows that the users are organized in the smaller groups making more transactions between them and fewer transactions outside the group.

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] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.
- [2] Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, Gábor Szabó, M Argollo De Menezes, Kimmo Kaski, Albert-László Barabási, and János Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New journal of physics*, 9(6):179, 2007.
- [3] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Physical review E*, 73(1):016102, 2006.
- [4] Mark Ed Newman, Albert-László Ed Barabási, and Duncan J Watts. *The structure and dynamics of networks*. Princeton university press, 2006.
- [5] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.
- [6] Akraati Saxena, Yulong Pei, Jan Veldsink, Werner van Ipenburg, George Fletcher, and Mykola Pechenizkiy. The banking transactions dataset and its comparative analysis with scale-free networks. *arXiv preprint arXiv:2109.10703*, 2021.
- [7] Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105, 2007.
- [8] Stefano Schiavo, Javier Reyes, and Giorgio Fagiolo. International trade and financial integration: a weighted network analysis. *Quantitative Finance*, 10(4):389–399, 2010.
- [9] Paul Erdos, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.
- [10] Béla Bollobás and Bollobás Béla. *Random graphs*. Number 73. Cambridge university press, 2001.
- [11] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.
- [12] Alex Arenas, Leon Danon, Albert Diaz-Guilera, Pablo M Gleiser, and Roger Guimera. Community analysis in social networks. *The European Physical Journal B*, 38(2):373–380, 2004.
- [13] Akraati Saxena and SRS Iyengar. Evolving models for meso-scale structures. In *2016 8th International Conference on Communication Systems and Networks (COMSNETS)*, pages 1–8. IEEE, 2016.
- [14] Yihong Hu and Daoli Zhu. Empirical analysis of the worldwide maritime transportation network. *Physica A: Statistical Mechanics and its Applications*, 388(10):2061–2071, 2009.
- [15] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. Modeling the evolution of weighted networks. *Physical review E*, 70(6):066149, 2004.