Sun 21st Feb, 2021

TUDelft - TPM

MSC Management Of Technology

SEN163A Fundamentals of Data Analytics

Assignment 1

Group 3

Reshma Joseph	5383595
Stijn Knoop	4608046
Amelie Müller	5432499
Boris van Overbeeke	4083164
Maxim Sachs	4236262

Contents

1 Introduction

The Groote Nationale Investeringsbank (GNI Bank) was alerted by the FIOD that there might be fraudulent transactions in their banking system. After ING received the record setting fine of 2017 for failing to spot money laundering[1], the GNI Bank is dedicated to prevent any involvement in fraudulent activity. This report presents the findings of research that was done to identify fraudulent activity in the transaction data set of the GNI bank.

2 Data set exploration

This section presents exploration of the provided data set. No information was given on currencies, therefore we assume that all amounts are in Dollars. Initial data exploration was carried out for the first 100 000 transactions.

2.1 General information

There are 7 734 834 transactions in the dataset. The mean transaction has a value of \$147 953.14, median transaction is \$34 245.66 and a total of \$1.144 trillion was transacted. The largest transaction amounted to \$92 445 520.00, the smallest to \$0.01. The distribution of balances of the sending accounts before and after payment can be seen in figure 1. Notably, some accounts are in considerable debt after their transaction. The distribution of the transaction amounts is shown in figure 2. A very high amount of small transactions is observed while the number of transactions exponentially decreases with increased transaction amounts (linear decrease on log-scale), with two outliers of extremely high amounts (\$92 445 520.00), see figure 2. The account names are constructed from a letter followed by a number. 6 923 501 addresses begin with the letter C and 2 150 401 begin with the letter M, however this likely has no relevance.

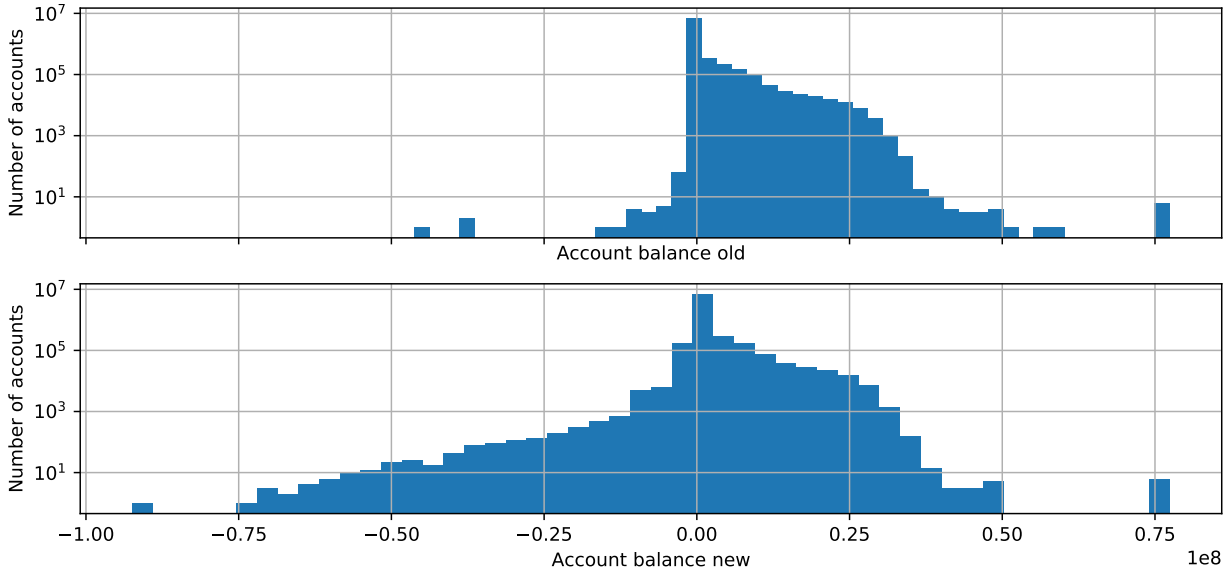

Figure 1: Histogram of account balances of sending accounts before payment and after payment

[1] https://www.fiod.nl/ing-betaalt-775-miljoen-vanwege-ernstige-nalatigheden-bij-voorkomen-witwassen/

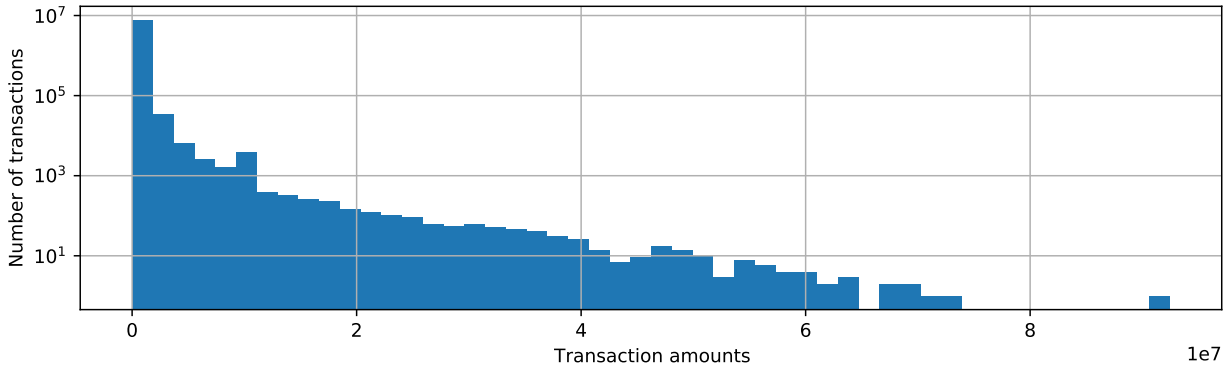

Figure 2: Histogram of transaction amounts

2.2 Transaction activity over time

As it can be seen in figure 3, the banking activities are not equally spread throughout the time stamps. The number of transactions and total transaction volume seems to fluctuate approximately every 25 time stamp, suggesting that the time stamps correspond to hours. Interestingly, transactions executed in times of little general transaction activity are especially high (see intermittent peaks of mean transaction amounts (red) and total transaction amounts (blue) in figure 3), which could be a potential indicator for fraudulent activity, but is not further investigated in this report. After tick 410, there is a significant drop in the total transaction activity and transaction volume, while the mean transaction amounts increase, meaning that fewer but higher transactions are carried out in this period. This could point to possible fraud or data loss since such a pronounced shift in transaction behaviour seems abnormal.

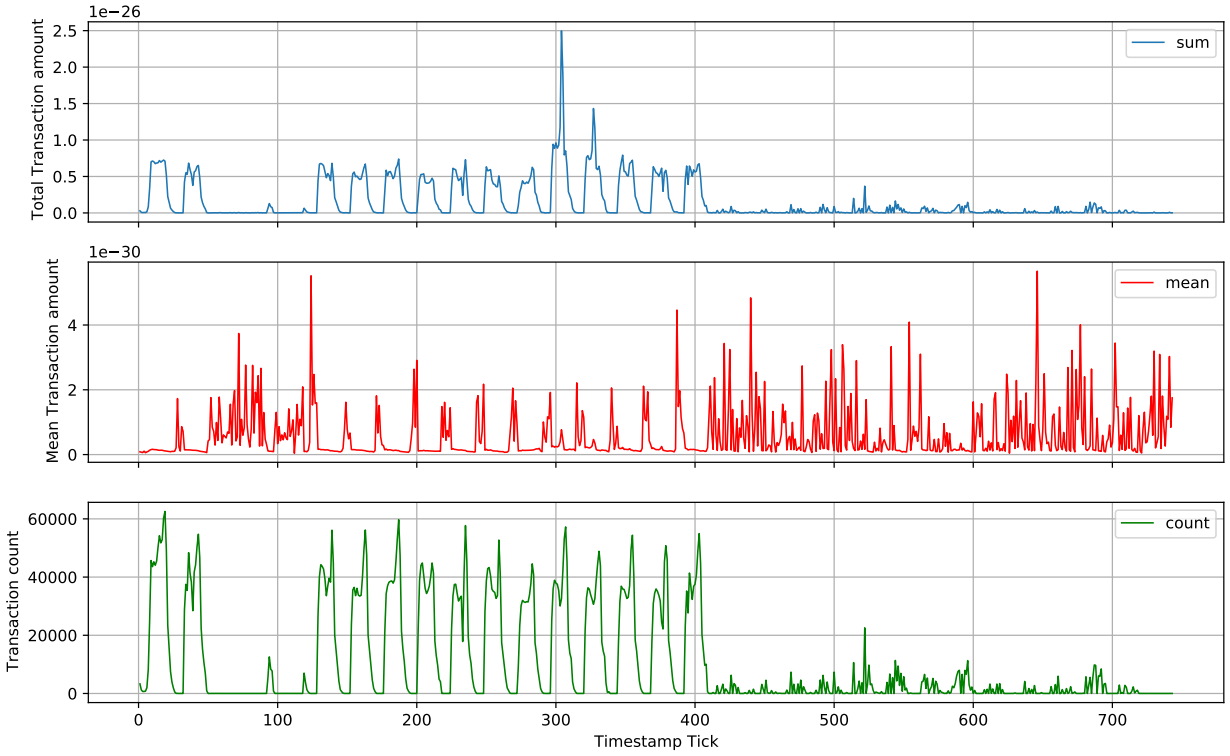

Figure 3: Transaction pattern over time

3 Dataset Consistency

We checked the data set for internal consistency. We observed a high amount of rounding errors, meaning that the transaction amounts were slightly different than the difference of old and new balance of the accounts. By rewriting the code to store the transaction with a factor of 1000, we made sure that these inconsistencies were not caused by float point errors during conversion of the database to csv-format. In total, there were 3 372 180 rounding errors in the dataset, 59.31 % amounting to \$0.01 and 40.69 % amounting to \$0.005. This shows that the data set is not internally consistent and but contains a high number of rounding errors, which could add up to rounding fraud as depicted in the movie Hacker (1995) or Office Space (1999), see section 4.2.

We observed that there are no transactions wherein money was transferred from one address to itself.

4 Identifying fraudulent activity

In order to identify fraudulent activity, our strategy was to identify any unusual pattern. Our analysis focuses on high-frequency low-volume transactions of \$0.01.

4.1 Transactions above or below certain limits

In this step, we filtered the dataset for transactions with amounts above the limit of \$1 000 000. 130 626 transactions were found. A more in-depth analysis should be carried out to investigate these low-frequency, high-volume transactions. Depending on the context of the banking accounts, sending such high amounts might be suspicious for fraud. Additionally, small amounts of \$0.1 or less were transacted 1 372 220 times. If this wasn't done for authentication purposes for example, then such accounts have to be inspected. The next section analyses the connection of these small transactions and the rounding errors observed in section 3.

4.2 Rounding error theft

Irregularities were found regarding the rounding of transactions that presumably suggest suspicious activity. After correcting for import errors, rounding errors were found on 3 372 180 transactions, see section 3. At the same time, there is one account, `C52983754`, that receives 1 372 194 transactions of \$0.01, adding up to an amount of \$13 721.94. The transaction log shows that for each transaction with a rounding error of \$0.01, there is a corresponding transaction of amount \$0.01 to the suspicious account. This transaction usually occurs right before the actually intended transaction.

This suggests that the rounding errors are in fact not errors but intended mark ups. Figure 1 shows the first 10 transactions that have a rounding error with their identified corresponding "stealing" transaction.

tx_index stolen from	timestamp	amount ($)	nameOrig	nameDest	tx_index stealing
2	1	9 839 640	C1231006815	M1979787155	1
8	1	7 817 710	C90045638	M573487274	7
15	1	3 099 970	C249177573	M2096539129	14
18	1	11 633 760	C1716932897	M801569151	17
22	1	1 563 820	C761750706	M1731217984	21
25	1	671 640	C2033524545	M473053293	24
28	1	1 373 430	C20804602	M1344519051	27
31	1	1 065 410	C1959239586	C515132998	30
34	1	311 685 890	C1984094095	C932583850	33
37	1	9 478 390	C1671590089	M58488213	36

Table 1: First 10 tx and the corresponding "stealing" transaction

However after identification of this suspicious activity, there are still 1 999 986 transactions remaining with rounding errors, of which most amount to $0.005. These could not be explained.

5 Conclusion

In this analysis the transaction sequences were investigated for suspicious activity. The main finding is the presence of rounding error theft. This is likely only achievable through access to the banks transaction processing software and would have required a modification to it to automatically transfer a rounding error to a selected account.

Additionally, further unexplained rounding errors were found and might be the result of the way the floating point implementation inside the transaction software is applied. This could be improved as well by increasing the number of significant digits for balances and transaction amounts. Moreover, potentially fraudulent patterns in transaction activities over time were identified. More information about the context of the time stamps and regular transaction histories is required to correctly interpret the observed transaction patterns.

Given the strong evidence for rounding error theft, we suggest that the GNI Bank reviews and improves their transaction processing software for malicious code as soon as possible.

6 Code Repository

The code used in the analysis above can be found at `https://github.com/maximsachs/SEN163A_Assignment1`.