# SEN163A Fundamentals of Data Analytics

August 10, 2020

## Introduction

The banking industry is frequently subject to fraud and thus requires strict supervision. This supervision commonly includes the monitoring of transaction logs. Since investment banks are no exception to this phenomenon a recent excerpt of transaction logs of the 'Groote Nationale Investeerbank' is investigated to detect eventual fraud. The transaction logs are extracted from the SQLite3 database "transaction_data.db".

## 1 Description of the dataset

Extracting and writing all 7734834 rows of data from the database to a pandas data frame took approximately 80 seconds. The resulting data frame consist of 10 columns, that contain information about transactions of the GNI bank. When extracted from the SQLite3 database, a row has the following format:

(1, 1, 'TRANSFER', '0.01', 'C1231006815', '170136.0', '170135.990', 'C52983754', '0.010', '0.020')

The initial dataset consist of 10 columns. The names of these columns and their corresponding interpretation are briefly explained below.

1. **ID:** Represents sequential ID of the transaction
2. **Timestamp (TS):** Timestamp corresponding to the transaction.
3. **Type:** The type of the transaction.
4. **Amount:** The amount of money that is transferred within the transaction.
5. **nameOrig (Orig):** The ID of the bank account by which a transaction is performed.
6. **oldbalanceOrig (oldbalO):** The balance of the bank account before the transaction. This is the balance of the account that will transfer the money.
7. **newbalanceOrig (newbalO):** The balance of the bank account after the transaction. This is the balance of the account that will transfer the money.
8. **nameDest (Dest):** The ID of the bank account to which money is transferred.
9. **oldbalanceDest (oldbalD):** The balance of the bank account before the transaction. This is the balance of the account to which the money is transferred.
10. **newbalanceDest (newbalD):** The balance of the bank account after the transaction. This is the balance of the account to which the money is transferred.

The extracted data set consists of 7734834 rows without any missing values. In total, there are 9073902 different bank accounts involved in all transactions. In table 1 the type, minimum value, maximum value and the amount of unique values for every column (variable) are stated.

Table 1: Columns of dataset explained

| | Columns (Variables) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | TS | Type | Amount | Orig | Dest | oldbalO | newbalO | oldbalD | newbalD |
| Type | Int | Str | Float | Str | Str | Float | Float | Float | Float |
| Min | 1 | NaN | 0.00 | NaN | NaN | -4.62·$10^6$ | -9.23·$10^7$ | -4.37·$10^6$ | -4.23·$10^6$ |
| Max | 743 | NaN | 9.24·$10^6$ | NaN | NaN | 7.74·$10^7$ | 7.74·$10^7$ | 3.57·$10^8$ | 3.57·$10^8$ |
| Unique | 743 | 1 | NR | 6353308 | 2722364 | NR | NR | NR | NR |

The values in column *Amount*, *oldbalanceO*, *newbalanceO*, *oldbalanceD* and *newbalanceO* are transformed from type string to type float in order to make calculations at a later stage. *Type*, *nameOrig* and *nameDest* remain strings (see table 1). Remarkable are the extremely low minimum values of the balance variables. In general, there is a number of (high) negative balances of bank accounts. This could be explained by the fact, that the bank allows its account holder to be in debt, for example when money is loaned to an investor. Also, the minimum value for amount is odd since it equals zero which implies that no money is transferred.

In total there are 9073902 bank accounts. From which 6353308 bank accounts transfer money to another bank account, while there are only 2722364 identified as receivers. Only 1770 bank accounts are both an original bank account as a receiver account. However, since the GNI bank is an investment bank, it is common to have a relatively small amount of receivers with respect to bank accounts that money is transferred from. This is due to the fact that there are simply more investors than companies or organisations to invest in.

The column 'type' and 'id' are deleted from the data set in a later stage. These columns did not provide any additional information. Furthermore, to prevent rounding errors that may occur in later stages, every numeric column except for timestamp is multiplied by 1000 and it's type is changed to a integer which is 64bit in Python 3.

## 2 Inconsistencies and explanations

The natural logarithm (ln) of the value that corresponds to 'amount' is taken to obtain an histogram that demonstrates the distribution from this variable (figure 1). The histogram indicates that in general high amounts are transferred. Since, the transactions are related to an investment bank this seems common practice. Also, common practice in the investment banking world is the number of high negative balances in the data set.

However, more remarkable is the inconsistency on the left of the histogram which indicates that an amount of 0.01 cents is often transferred. The right graph in figure 1 illustrates that after timestamp 400 the amount of transactions per timestamp decreases drastically.

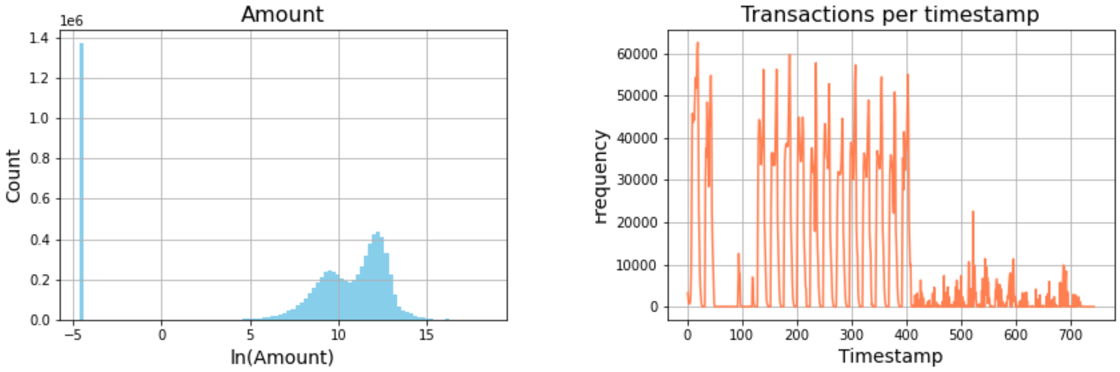

Figure 1: Left: (ln)distribution of the amount transferred per transaction Right: Frequency of transactions per timestamp

To highlight inconsistencies a new column *diff* is created. However, to easily calculate this column two other columns named *Orig_diff* and *Dest_diff* are created. These columns calculate differences in the balance before and after the transaction. Subsequently, column *diff* is created which indicates whether there is a difference and if so, how much. An inconsistency is identified if this value in the column does not equal zero. In total there are 3372180 inconsistencies identified and all of them are of the same amount: 1 cent. This leads to a total of $33721.80 that disappeared.

## 3 Fraudulent activities

The first actual step that is taken to identify fraud, is determining the bank account occurs that in most transactions. As illustrated in figure 1 there are myriad transactions where a small amount is transferred. The bank account that is involved in these transactions is referred to as C52983754 and appeared in 1372194 transactions, where one cent is transferred to C52983754 itself.

Table 2: Table that illustrates a pattern in inconsistencies

| amount | nameOrig | oldbalanceOrig | newbalanceOrig | nameDest | newbalanceDest | diff |
|---:|---|---:|---:|---|---:|---:|
| 10 | C1231006815 | 170136000 | 170135990 | C52983754 | 20 | 0 |
| 9839640 | C1231006815 | 170136000 | 160296360 | M1979787155 | 9839630 | -10 |
| 10 | C90045638 | 53860000 | 53859990 | C52983754 | 30 | 0 |
| 7817710 | C90045638 | 53860000 | 46042290 | M573487274 | 7817700 | -10 |
| 10 | C249177573 | 20771000 | 20770990 | C52983754 | 40 | 0 |
| 3099970 | C249177573 | 20771000 | 17671030 | M2096539129 | 3099960 | -10 |
| 10 | C1716932897 | 10127000 | 10126990 | C52983754 | 50 | 0 |
| 11633760 | C1716932897 | 10127000 | -1506760 | M801569151 | 11633750 | -10 |

Table 2 illustrates a pattern that occurs through the whole excerpt of transactions. Within the same timestamp an amount of 1 cent is transferred to bank account C52983754 and subsequently a higher amount is transferred to another random bank account. The random bank account than receives 1 cent less than it should which shows in the inconsistency (*diff*). The first transaction that is made to C52983754 is likely to stay unnoticed. Due to the fact that the balance of the original bank account manages to eventually go back to the amount before the 1 cent was transferred.

The bank account that occurs most as an original or sender account is referred to as C1286084959 and appears 20 times. The most remarkable thing about the transactions this account is involved in, is the fact that that this account only transfers money to only one other account, C2342523425. Which does not occur in any other logs. The total amount transferred to this account is 19999.86.

Table 3: Inconsistencies in balances from account C52983754

| timestamp | nameOrig | oldbalanceOrig | nameDest | newbalanceDest | Inconsistency |
|---:|---|---:|---|---:|---|
| 394 | C1789981744 | 50447000 | C1286084959 | 77240367116 | unknown |
| 398 | C1286084959 | 77240740716 | C2342523425 | 17973970 | 373.6 |
| 401 | C1256632176 | 22224000 | C1286084959 | 77429185321 | 188445 |
| 743 | C1286084959 | 77430968382 | C2342523425 | 19999860 | 1783.06 |

By further examining the transaction logs in which C1286084959 appears, it becomes clear that the balances corresponding to these transactions are incorrect. C1286084959 somehow has the ability to generate money, since the balance of this account seems to increase without receiving any money. This is illustrated in column *inconsistency* at table 3. This is furthermore illustrated in figure 2,3, 4 and 5. Where the figure 2 demonstrates the cumulative amount of money generated from C1286084959 and the cumulative amount that is transferred from the same bank account to C2342523425.

When plotting the cumulative amounts of the peculiar 1 cent transactions in the same graph, there is an indication that the main bank accounts that are involved in these transactions are connected. Namely, the cumulative amounts that C1286084959 creates and subsequently transfers seems to be a result of the 1 cent transactions. More specifically, the total cumulative amount of all one cent inconsistency transaction minus the total cumulative amount of one cent transactions to C52983754 equals *exactly* the amount C1286084959 transferred. This is mathematically proven, see 1.

$$Amount = 33721.80 - 13721.94 = 19999.86 \tag{1}$$

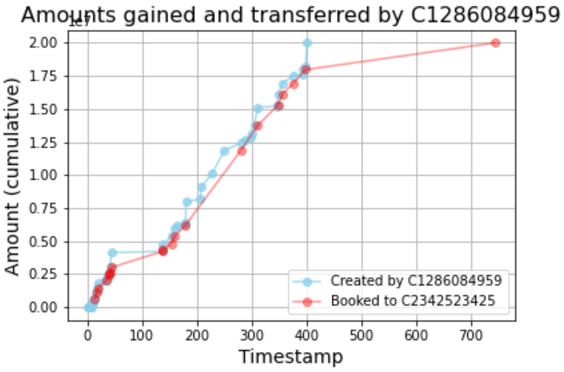

Figure 2: Money generation by C1286084959

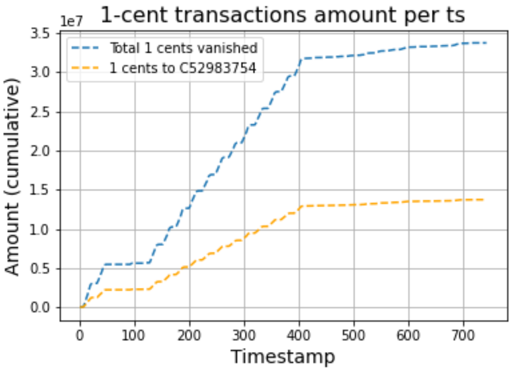

Figure 3: 1 cent transactions

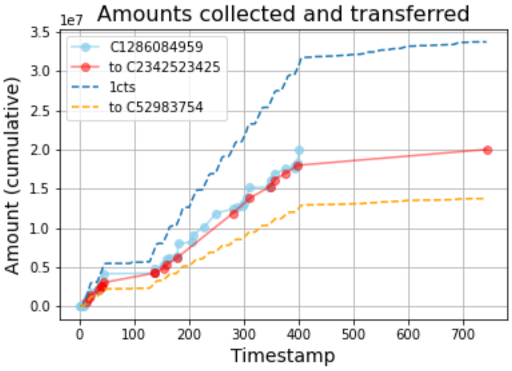

Figure 4: Cumulative fraudulent transactions

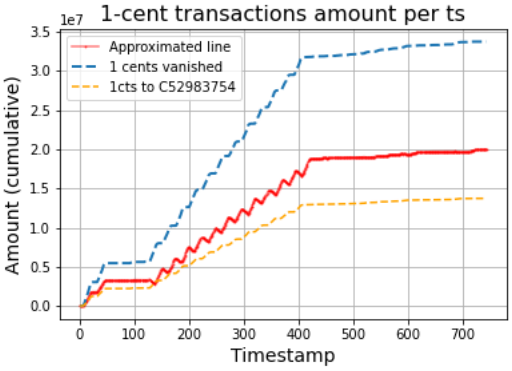

Figure 5: Approximated sum of 1 cents

## 4 Conclusions and recommendations

The cumulative amount of the 1 cents that came to light as an inconsistency, typically disappear from random bank accounts. How this amount exactly is derived and where it is stored in the meantime remains unclear. However, that bank account C1286084959 has to do with these transactions is beyond dispute. The most likely scenario is that C1286084959 has the ability to store this money on its own bank account, without making a transaction and subsequently transfers the generated money to C2342523425. This might be possible due to a leakage in the banking system, which is used to store the money unnoticed. The one cents that are transferred to C52983754 are always within the same timestamp as an inconsistent transaction. Which occurs right after the 1 cent is transferred. Note that it is not always true that the 1 cent transaction must occur before any transaction subject to an inconsistency. This one cent could function as a mask to hide the fraudulent transactions and simultaneously the total amount of 1 cents transferred to C52983754 functions as a fee. Another possibility is that this bank account has the ability to help C1286084959 with the misappropriation of money in return for a fee that consist of the cumulative amount of 1 cents transferred to C52983754. In conclusion, bank accounts C1286084959, C2342523425 and C52983754 are cooperating and the GNI bank need to impose sanctions.

## 5 Discussion

The most trivial fraudulent activities within the examined data are brought to light. However, it is possible that other fraudulent activities remain within the excerpt of transaction logs. This requires further research.