Deep learning for Blockchain

Deep Learning DIY (ENS, 2020-2021)

Jules Baudet, Linus Bleistein

February 2, 2021

Structure

Introducing blockchain

Our Architecture

Results

Our goal

The goal of this project is to **classify ethereum addresses based on their owner**.

This can be subdivided into 2 subproblems:

- Binary classification on users: is an address owned by an **exchanges** (also called "platforms") or by another type of owner (mostly private owners)
- Cluster addresses based on ownership (i.e. which addresses are owned by the same entity).

Terms will be defined more precisely in what follows.

Introducing blockchain

What is a blockchain ?

- A Blockchain is a *chain of blocks*: decentralized transactions recorded in a public file.
- For each transaction, a *block* (log data) is added to the blockchain.
- Available data: transactions between 2017 and 2020 (addresses involved, value, timestamp).

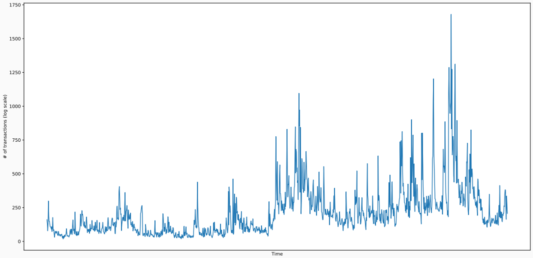

Figure 1: Volume of exchanged cryptocurrency BAT through time (2017-2020).

Our dataset

	from_address	to_address	value	unix_block_timestamp
2438955	0x88e2efac3d2ef957fcd82ec201a506871ad06204	0x67fa2c06c9c6d4332f330e14a66bdf1873ef3d2b	1.000000e+00	1496084349
2305151	0x88e2efac3d2ef957fcd82ec201a506871ad06204	0x67fa2c06c9c6d4332f330e14a66bdf1873ef3d2b	7.999999e+06	1496085435
539714	0x88e2efac3d2ef957fcd82ec201a506871ad06204	0x67fa2c06c9c6d4332f330e14a66bdf1873ef3d2b	1.200000e+07	1496085651

Figure 2: Three lines of our dataset, with features.

The dataset was provided to us by Nyctale.

- 2.7M transactions between 2017 and 2020 of BAT token.
- 95 addresses identified by Nyctale as exchanges (also called *platforms*).

Goal: classify exchanges vs 'regular' addresses.

Understanding the data

	Exchange	Non-Exchange
total # transactions sent	39 5543	230 123
total # transactions received	23 8549	245 760
Avg Amount sent	19 908	15 525
Avg Amount received	34 092.20	14 428.0
Avg # active periods	5400.9	4.4
Avg # transactions sent	4163.6	3.2
Avg # transactions received	2409.6	2.8
Avg # transactions in which involved	6573.2	6.0

In total there are about 900k addresses in the data, among which only 95 were identified as exchanges: **very imbalanced classes**.
On average, it seems that **number of active periods** and **number of transactions sent/received** are good points to discriminate in classification.

Distribution of transactions

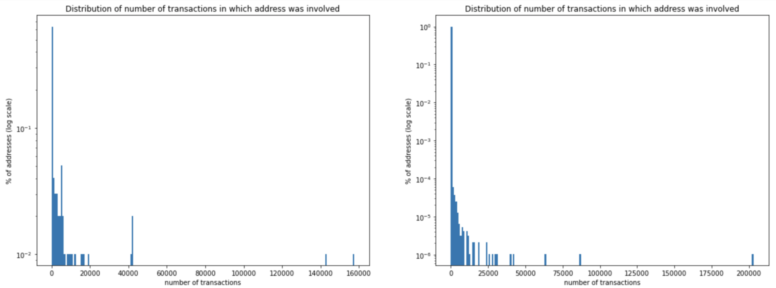

Figure 3: On the left: Exchanges, on the right: non-exchanges

Distribution of number of active periods

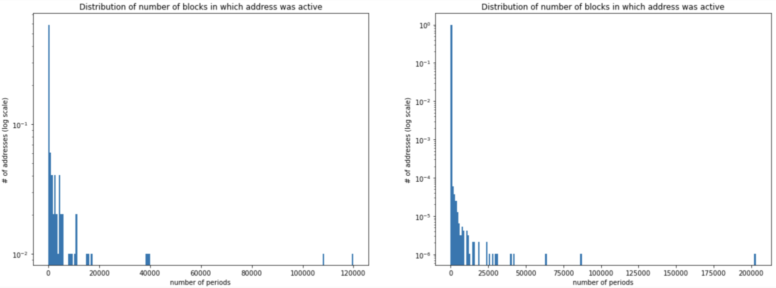

Figure 4: On the left: Exchanges, on the right: non-exchanges

Visualising the transaction graph

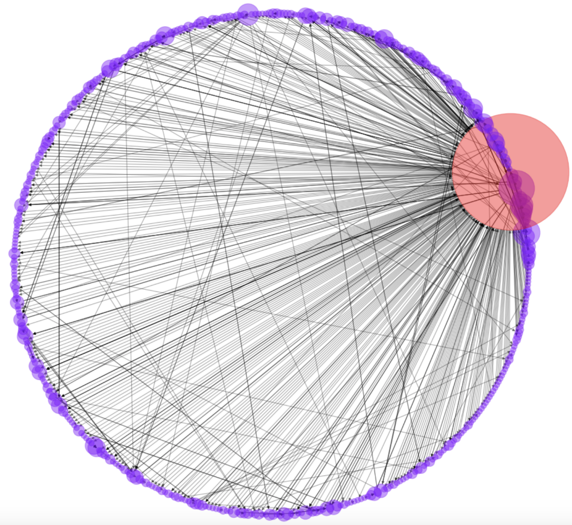

Figure 5: One day of transactions!

Visualising the transaction graph

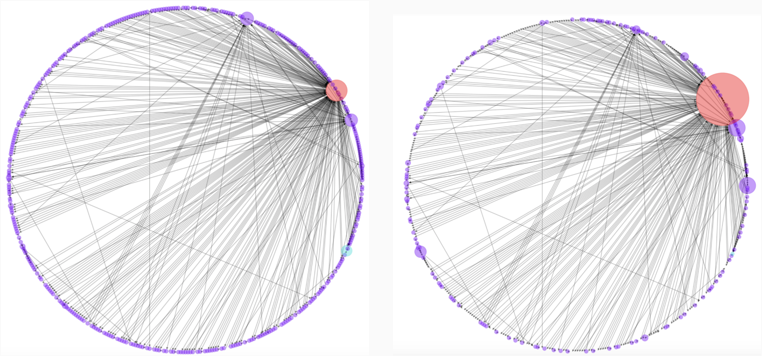

Figure 6: Left: nodes proportional to *in degree*; Right: proportional to *out-degree*

Related work

Identifying traders on blockchains is a topic of growing interest. **Why ?** Criminal activity detection, fraud detection.

- Covington, Adams, and Sargin 2016: Youtube recommandation system.
- Lin et al. 2019: manual feature extraction + classification (supervised).
- Shao et al. 2018: RNN + AM Softmax (supervised).

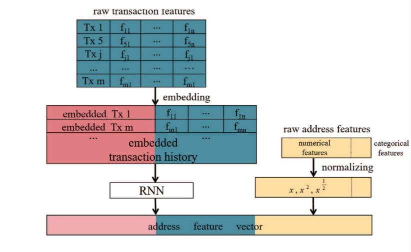

Figure 7: Shao et al. 2018's network.

- Bartoletti, Pes, and Serusi 2018: shallow learning with clustering techniques for ponzi scheme detection ($\sim$ 10 millions USD).

Embeddings ideas

Our data has important label imbalances, varies through time, lacks important information, and is composed of transactions when we are looking to classify *addresses*.

Need a nice embedding to express good discriminating features.

Our Architecture

First iterations

When starting the project we tried several (deep) architectures and embeddings.

Adjacency + MLP: the first idea we had was to represent each address by his row in the graph's adjacency matrix, and to do that for different powers of the adjacency matrix. We also appended time and address related data (average number of transactions, time between first and last transaction, standard value of transactions, average amount received and sent...). **But** very sparse data, high dimensional, and high imbalances: MLP would get good accuracy by just labeling everyone as non-exchange.

How does our network work ?

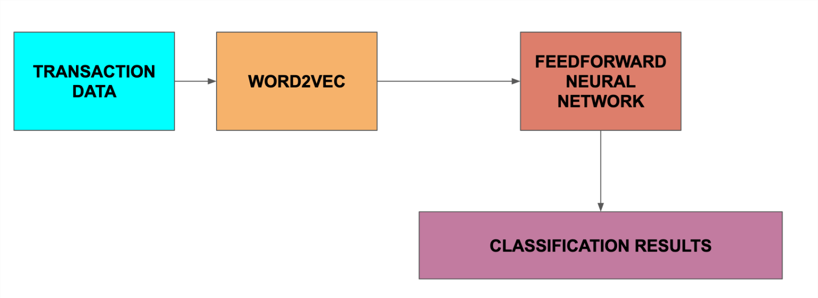

Figure 8: Our network.

Word2Vec

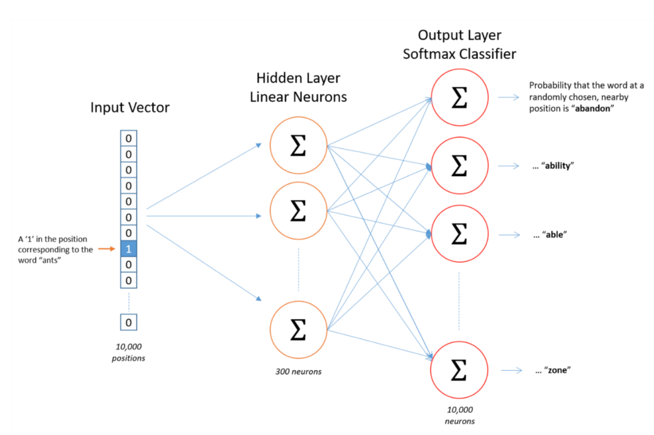

Figure 9: The Word2Vec Network. Source.

We obtain an embedding of size 150, that we can combine with statistical features. We then average them to get 1 row per address. **Maybe RNN could be used to make it more reliable**.

Results

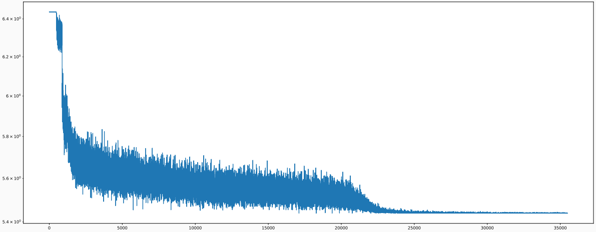

Figure 10: Training loss through time (10 epochs).

- Training results are still bad : gradient seems to vanish quickly.
- Classifier suffers from data imbalance: it learns to classify everybody as 0.

Concluding words

Untreated problems:

- Extremely unbalanced data. Undersampling, oversample ? Convex combination of features ?
- Very simple network architecture: could be hugely improved.
- Adding more statistical features: moments etc.
- Use architectures better suited for this kind of data. RNN ? Other Lose ?

References

M. Bartoletti, B. Pes, and S. Serusi. “Data Mining for Detecting Bitcoin Ponzi Schemes”. In: *2018 Crypto Valley Conference on Blockchain Technology (CVCBT)*. 2018, pp. 75–84. DOI: 10.1109/CVCBT.2018.00014.

Paul Covington, Jay Adams, and Emre Sargin. “Deep neural networks for youtube recommendations”. In: *Proceedings of the 10th ACM conference on recommender systems*. 2016, pp. 191–198.

Yu-Jing Lin et al. “An evaluation of bitcoin address classification based on transaction history summarization”. In: *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*. IEEE. 2019, pp. 302–310.

Wei Shao et al. “Identifying bitcoin users using deep neural network”. In: *International Conference on Algorithms and Architectures for Parallel Processing*. Springer. 2018, pp. 178–192.