Detecting Malicious Behaviors on Ethereum

CU CSPB 4502 Data Mining Fall 2024
Group 7
Dec 9, 2024

Team 7 Members

The
"DataBuffs"

BD Tinsley

4th semester in program. Expected graduation May '25.

Hallee Ray

Final semester in program. Expected graduation December '24

Questions We Sought To Answer

- Can we determine which wallets engage in fraudulent behavior?
- Can we determine a type of transaction that appears fraudulent by comparing it to general statistical expected values in conjunction with using domain knowledge?

Our Data

- Granularity – individual transactions on Ethereum Ledger
- 2.7 million transactions from August 2015 to April 2016
- 17 total features including `hash`, `from_address`, `to_address`, `value`, `nonce`, `block_address`, `gas`, `gas_price`, `input`, & `block_timestamp`

Data Preparation

01

02

03

04

Preprocess Data

Impute missing values, remove duplicates, normalize data

Preprocess Data

- `max_fee_per_gas`, `max_priority_fee_per_gas`, `transaction_type`, `max_fee_per_blob_gas`, `blob_versioned_hashes` were all features added to the Ethereum ledger after 2020
- Cleaned data contained 12 total features and 2.7 million data objects

Data Preparation

01

Preprocess Data

Impute missing values, remove duplicates, normalize data

02

EDA

Track transaction statistics, cross validate entries to ensure accuracy

03

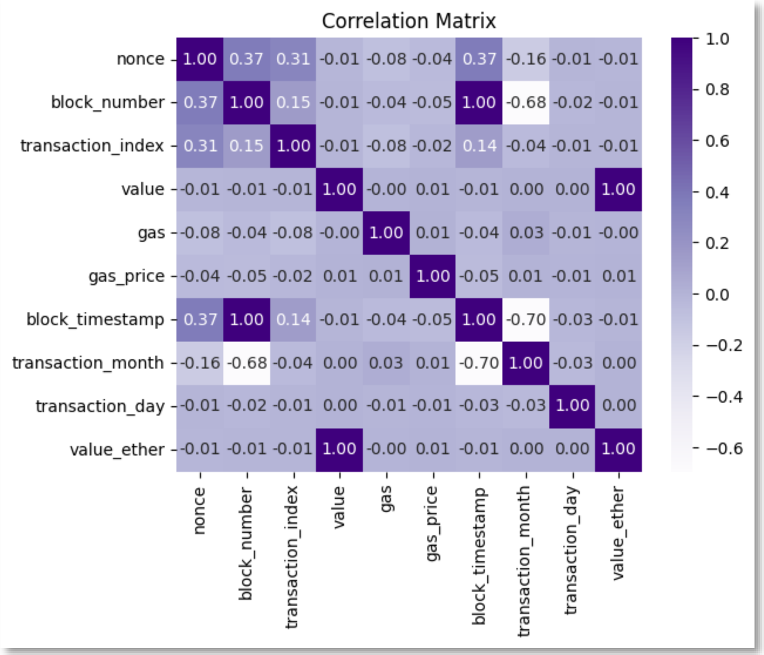
04

EDA

Correlation Matrix

- An easy way for us to discern if any of the features correlated with others in apparent or non-apparent ways

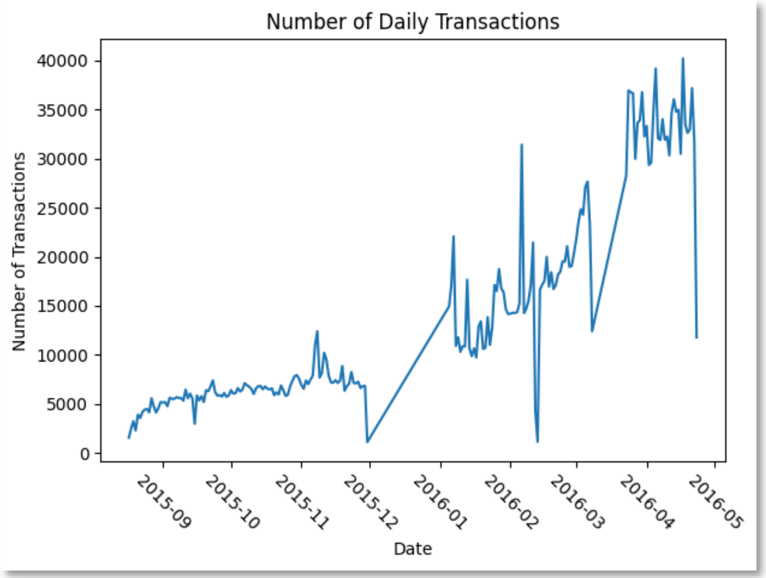
… the strong correlations all ended up being very apparent



EDA

Number of Daily Transactions

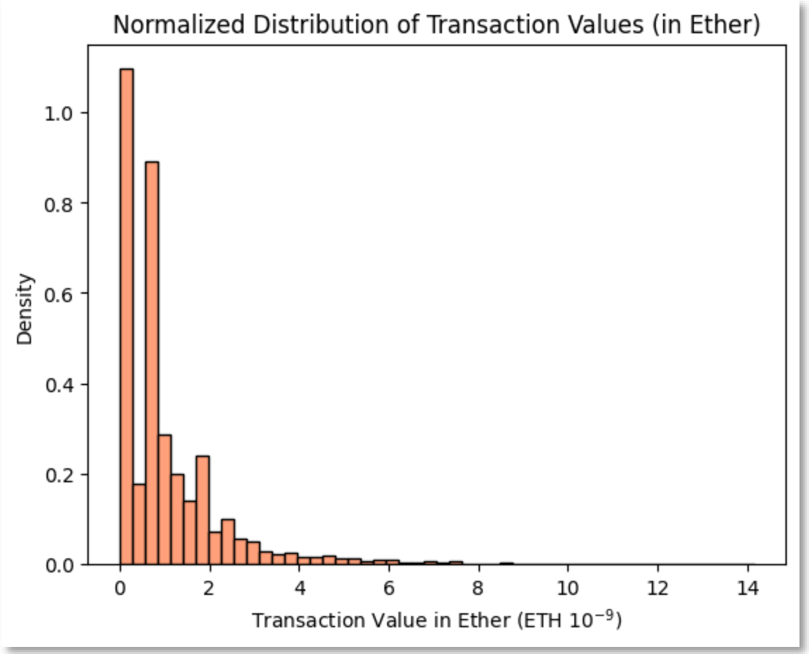
- Timestamp of the transaction as it appeared on its parent block
- The gaps in or data are denoted by the steep slopes
- Overall it revealed that the dataset captured transactions in a window where Ethereum was gaining popularity



EDA

Normalized Distribution of Transaction Values

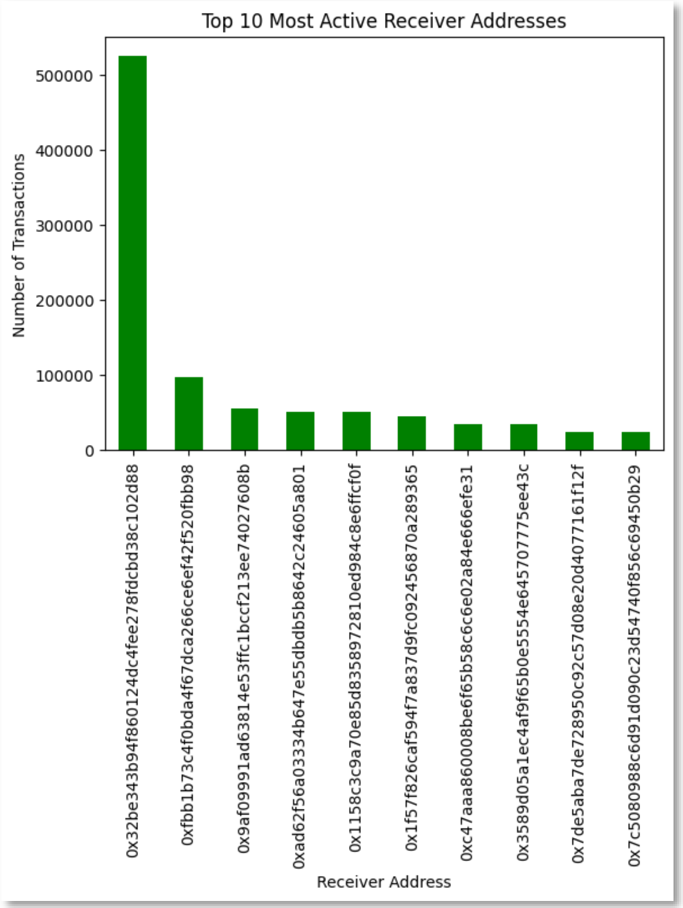
- All values in Ether (ETH 10^{-9})
- Displayed a unique multimodality with an expected right skew



EDA

Top Receiver Addresses by Number of Transactions

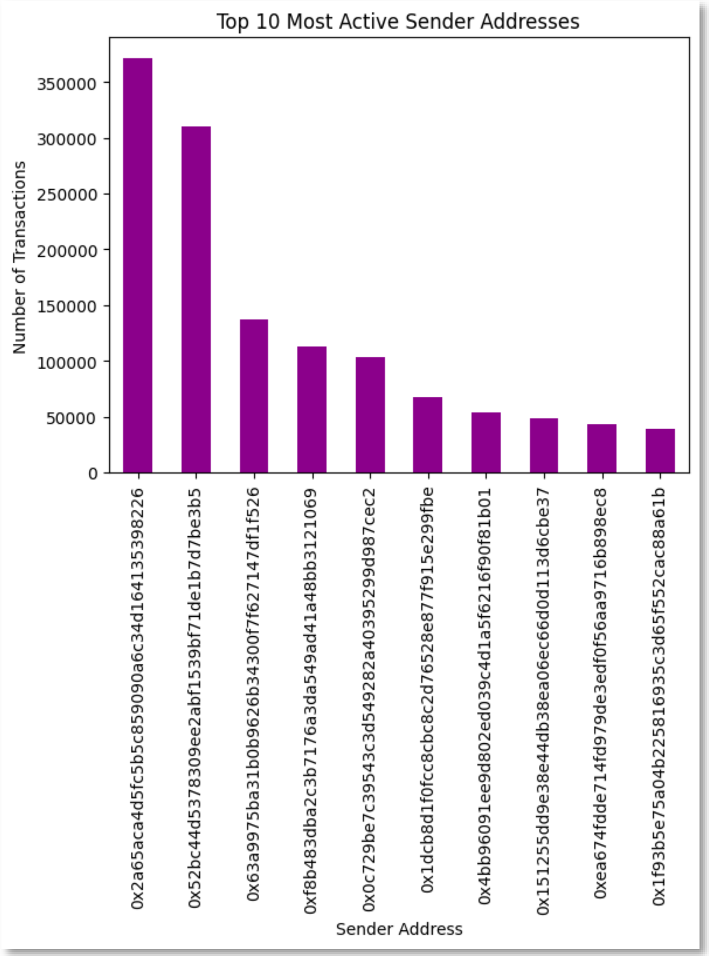
- Not by value, but raw number of transactions
- Interesting to note the dominance of the most common receiver wallet
- Note the address `0x32be34...2d88`, as we will encounter it again



EDA

Top Sender Addresses by Number of Transactions

- Again, by raw number of transactions
- A jump from 1st and 2nd, but an easier slope than receiver wallets



Data Preparation

01

Preprocess Data

Impute missing values, remove duplicates, normalize data

02

EDA

Track transaction statistics, cross validate entries to ensure accuracy

03

Feature Engineering

Generate features based inferential data and conversions

04

Feature Engineering

Additional features that give us insights into the dataset and allow us to identify potentially fraudulent activity include information regarding the frequency and values of transactions.

The goal of these features was to give us metrics about each wallet address.

- `from_frequency`: how many transactions were sent by this wallet address.
- `to_frequency`: how many transactions were received by this wallet.
- `total_frequency`: how active this wallet address is by telling us how many total transactions were processed by this wallet address.
- `from_val_total` and `to_val_total`: the total amount of ETH sent and received from this wallet address, respectively.
- `avg_value_sent` and `avg_value_received`: average value of each outgoing and incoming transaction.

Data Preparation

01

Preprocess Data

Impute missing values, remove duplicates, normalize data

02

EDA

Track transaction statistics, cross validate entries to ensure accuracy

03

Feature Engineering

Generate features based inferential data and conversions

04

Anomaly Detection

Detect outliers using K-Means clustering

Clustering Methodology

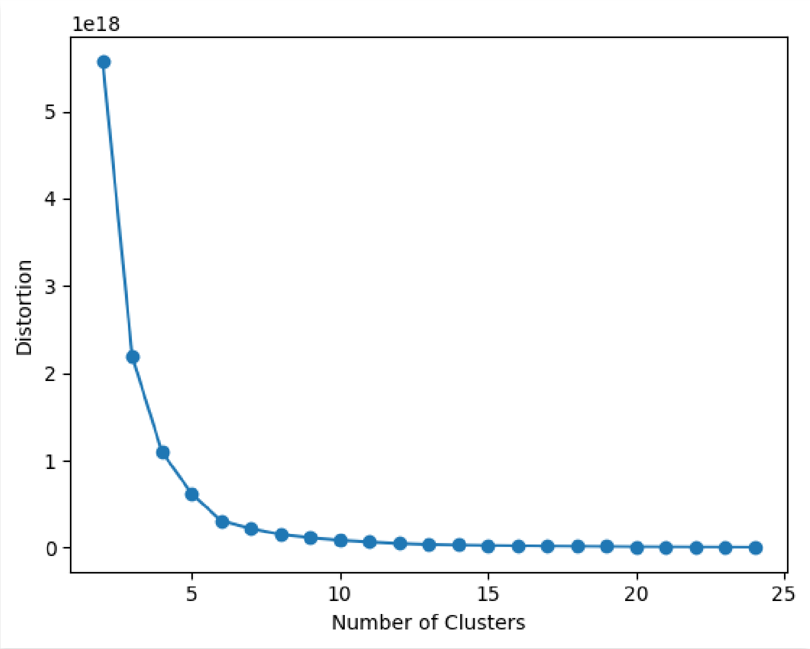
Principal component analysis

To reduce the dimensionality of the dataset from 13 to 2.

Elbow Method

Tuning the optimal k-value by finding where the "bend" in the graph.

Elbow Method



Clustering Methodology

Principal component analysis

To reduce the dimensionality of the dataset from 13 to 2.

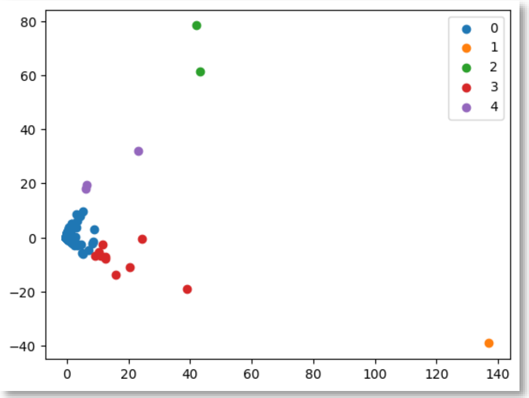
Elbow Method

Tuning the optimal k-value by finding where the "bend" in the graph.

K-Means with 5, 6 and 8 clusters

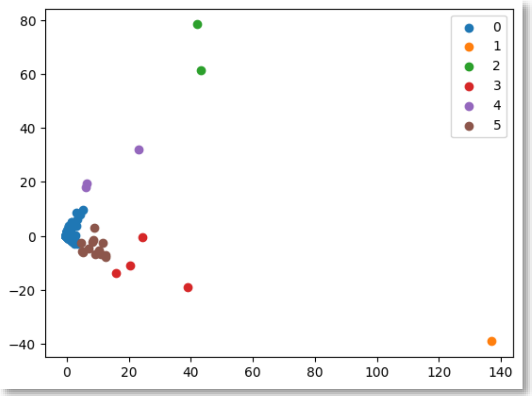
Each cluster arrangement having a silhouette score of > 0.99.

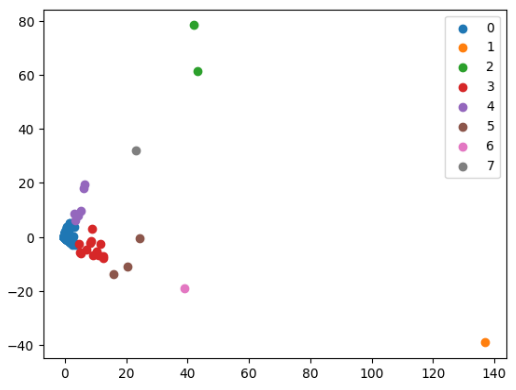
K-Means Clustering



$k = 5$

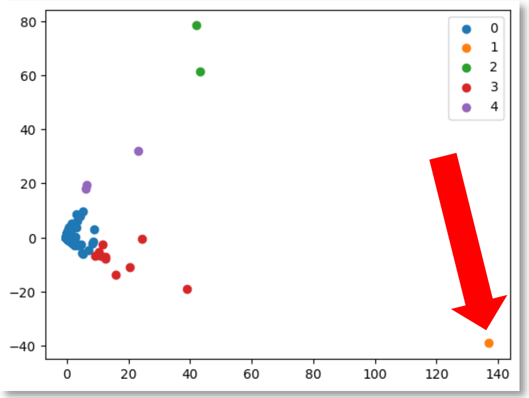
$k = 6$





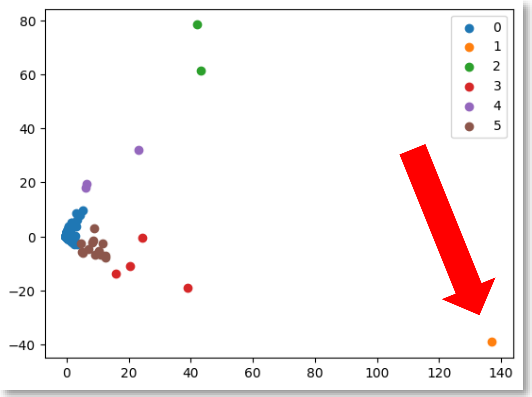
$k = 8$

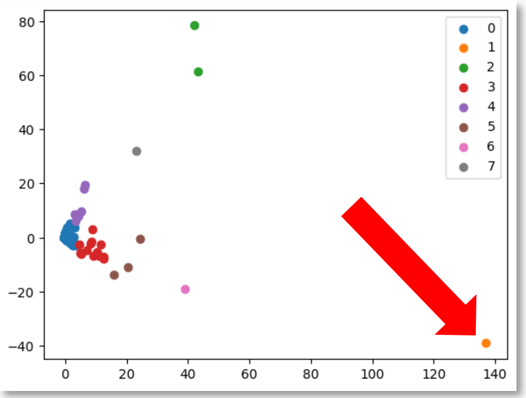
K-Means Clustering



$k = 5$

$k = 6$





$k = 8$

Knowledge Gained

Through our work, we were able to identify specific wallets that follow uncommon behaviors. With that knowledge, we created a prediction model to discern wallets that display fraudulent behaviors in different transaction data.

Knowledge Gained

We can apply our prediction model to determine if a specific wallet falls into likely fraudulent behaviors, given a range of transactions. This is useful in the case of transferring or accepting funds from unknown wallets or first-time transactional relationships.

Tools Used

Python

The backbone language for our project

NumPy

To extend Python's mathematical capabilities

pandas

For data analysis

Matplotlib/Seaborn

To visualize our results

GitHub

To house our code

CU CSPB 4502 Data Mining Group 7

Detecting Malicious Behaviors on Ethereum

December 9, 2024