Krish Suchak, Rahul Rangwani, Michael Carrion, William Ansehl MACS 33002

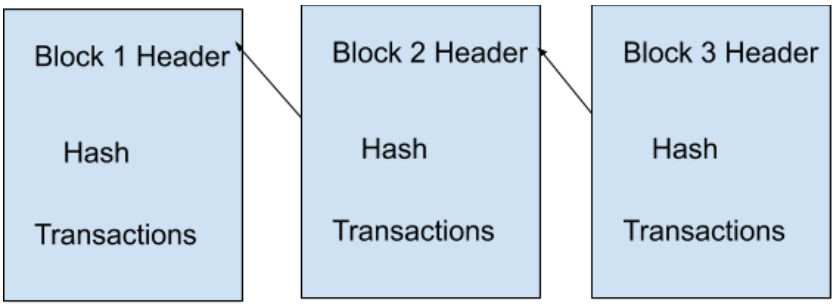
Project Proposal

Bitcoin Address Classification

I. Introduction

A. Background

Bitcoin is an example of a cryptocurrency that runs on blockchain technology; namely, it is the first example of an application of blockchain. A blockchain is an immutable ledger that records a series of blocks sequentially (and usually chronologically in time). Once a block is appended to the end of the blockchain, it is sealed cryptographically, and thus is immutable. This is because it would take half the hash power of the entire (worldwide) network to change the last block. We won't delve into the specifics of the cryptography because that is outside the scope of this project.



simplified diagram we made of linked list nature of blockchain

Each block contains a list of transactions. Each of these transactions have specific attributes including transaction hash (a unique id), transaction amount, fee amount, input addresses (who the transaction is from), output addresses (who the transaction is going to), etc.

B. Objective

First, we must understand that for any given address, we can iterate through the blockchain to get a list of transactions associated with the address. Using these transaction attributes as possible features, we aim to classify Bitcoin addresses as belonging to either a major exchange, a mining pool, or neither. Over the course of the project, this may change (like adding more classes such as darknet markets and faucets).

C. Motivation

Our motivation for creating a model that can classify addresses into entity types is to show three things:

1. Bitcoin is not as anonymous as many believe. We should be able to identify entity type for a given address with our model.
2. We can classify labeled data (addresses for which we do not have "EXCHANGE", "MINING POOL", or "OTHER" labels).
3. Another way to classify an address as belonging to a mining pool or exchange is to use a heuristic from the paper "A Fistful of Bitcoins" to identify the network of addresses belonging to a specific user / entity. This address clustering (not the machine learning technique but a heuristic) algorithm has a time complexity of $O(n^2)$ where n is the number of addresses in the blockchain. We will use this

method to label our unlabeled data. But the benefit of our model is that our model will take less time and be predictive (unlike the traditional Bitcoin address clustering algorithm).

II. Approach

A. *Process*

The step-by-step process is as follows:

1. *Data Collection and Setup:*
    a. Get the dataset using Google BigQuery and download as csv files.
    b. Load into pandas DataFrames.
    c. Set up git repo.
2. *Label Unlabeled Data:*
    a. Label x -> y pairs based on labels from results from https://chainz.cryptoid.info/btc/ (Can search name of exchange or mining pool and get main address for entity.)
    b. Implement address clustering algorithm from paper, thus labeling all addresses in a cluster with relevant entity type.
3. *Feature Construction and Selection:*
    a. Create features that may be relevant (for example time between chronologically sequential transaction associated with a given address).
    b. Use existing features in the dataset (like output value / transaction amount).
    c. Select which features should be included in the model.
4. *Classification and Preliminary Analysis*
    a. Try classification algorithms (like Random Forest, Decision Tree, KNN, etc).
    b. Analyze classification results (based on accuracy).
5. *Rework Classification based on Results*
    a. Use knowledge from class to figure why different classification algorithms differ in efficiency of classifying our dataset based on our selected features.
    b. Consult literature for alternative approaches or features.
6. *Improve Feature Selection and Classification*
    a. Construct new features if necessary.
    b. Remove old features if necessary.
    c. Try classification algorithms again.
7. *Visualizations and Writeup*
    a. Create visualizations including confusion matrix, feature importance bar chart, etc.
    b. Make technical diagrams for concepts like address clustering.
    c. Write up our results.

B. *Methodologies*

The dataset analyzed in this project is taken from Google BigQuery, a service that enables interactive analysis of large datasets. The dataset can be viewed on Kaggle (https://www.kaggle.com/bigquery/bitcoin-blockchain). The dataset is originally divided into 4 main parts:

1. Blocks: Contains 13 columns with information pertaining to the blocks that comprise the blockchain. Features consist of a Block Hash, Block Size, Timestamp and Transaction Count among others.
2. Inputs: Contains 14 columns with information pertaining to the source of transactions that comprise a block. Features include a Transaction Hash, Block Hash, Block Timestamp and addresses among others.
3. Outputs: Contains 11 columns with information pertaining to the receiving address of transactions that comprise a block. Features include a Transaction Hash, Block Hash, Block Timestamp and addresses among others.
4. Transactions: Contains 17 columns with information pertaining to each transaction as they comprise the block. Features include Transaction Hash, Transaction Size, Block Hash, Block Timestamp, input_count, output_count among others.

We will begin our analysis by selecting an initial list of features across the four originating datasets on Kaggle and joining the 4 tables along the Transaction Hash feature. Selected features are subject to change but currently consist of the following:

1. Transaction_hash: the identification of the transaction. We will use the Transaction_hash to join the 4 tables for ease of analysis.
2. Is_coinbase?: A boolean field that indicates whether or not a transaction is a block reward. This variable is key for differentiating miners from other classifications. Only mining pools receive block rewards.
3. Input_count: A count of the number of input addresses associated with that transaction. Two or more input addresses of the same transaction are likely to derive from the same entity.
4. Output_count: A count of the number of output addresses associated with that transaction
   Input_value: The total value of the inputs in the transaction.
5. Output_value: The total value of the outputs in the transaction.
6. Fee: The amount (fee) paid in this transaction. Higher fees enable transactions to be pushed to the next block faster.
7. Block_timestamp: A timestamp of the date and time that respective block was added to the blockchain
8. Size: The size of the Transaction in bytes. Transaction size increases with transaction history and transaction amount.

Note: can only aggregate (i.e. total_outvalue) according to single wallet address, not according to a specific user (that is, you can't aggregate all the wallet addresses belonging to the same user, since in our test set we'll only have one address, not all the addresses belonging to the same user)

We also intend to aggregate the data to analyze the following feature in addition (subject to change):

9. number of transactions per block involving the same address: an exchange would be more likely to have more transactions in the same block than a miner

III. Challenges

With this project, we will face a number of challenges:

1) Sourcing the Data: We have a dataset provided by Google, but will need to navigate the Google BigQuery ecosystem to get it into a manipulatable format.
2) Feature Selection: Making decisions about which feature to keep and drop will ultimately impact our model.
3) Address Clustering: To make the heuristic from the "A Fistful of Bitcoins" paper, we will need to construct an algorithm that can output a list of associated addresses given a single wallet address. We have a copy of this algorithm from the undergrad Cryptocurrencies class, but we will have to translate it to be usable for the Google data (since it was constructed to be used with data from python library Blocksci).
4) Data Volume: The sheer amount of the bitcoin ledger may make querying the data difficult. Within a single day, there can be upwards of 300,000 bitcoin transactions. If we examine too small of a timeframe, such as a day or two, we may model anomalies in the data; on the other hand, archiving too many blocks may become computationally expensive. For this reason, we aim to use a week's worth of data.
5) Single Address Classification: An entity may use several different wallet addresses, Looking at the transactions associated with a single address (as opposed to all the transactions of an address cluster) may lead to insufficient data as an input for the model. This may result in incorrect categorization.
6) Entity Stratification: By sourcing entity types ourselves using https://chainz.cryptoid.info/btc/, we may not be labeling enough data to accurately classify enough of our dataset. From previous forays into this address clustering (in the undergrad Cryptocurrencies class), we found that a given major exchange or mining pool, may be involved in as many as 5% of transactions in a week (either on sending or receiving end of a transaction). This would imply that if every entity was the size of the the top ones (like Coinbase exchange, ViaBTC mining pool), we would need to find 40 entities to label all the data. However, this is not true in reality, and as we start searching for addresses of smaller entities, we will run into diminishing returns of address clustering to label unlabeled data.

IV. Discussion / Applications

Ultimately, our project seeks to develop a novel model capable of classifying a given wallet address as belonging to either an exchange, a mining pool, or some other entity (e.g. a single user). Further, unlike the existing heuristic which takes $O(n^2)$ time, our model does not require unearthing all addresses belonging to the entity in question, and thus allows for more expedient classification.

In the future, this work can be expanded upon to more comprehensively classify different types of BTC entities; that is, in addition to classifying an address as belonging to an exchange or a mining pool, we would be able to classify it as belonging to a Faucet, HYIP, or Marketplace (see table below). Further, it would be interesting to examine features of fraudulent transactions, to see if we can build upon our model to detect fraud.

SERVICE	DESCRIPTION
Exchange	Platform where currencies can be bought and sold
Faucet	Offers free but small amounts of Bitcoin in return for completing tasks like solving CAPTCHAs, or clicking on advertisements
Marketplace	Offers various payment services (e.g. escrow)
Mining Pool	Collective group of miners who share computational resources to discover blocks

Adapted from Toyoda et al., 2018

Lastly, future work should not only seek to improve the accuracy of our current model, but should also seek to extend our model to other cryptocurrencies as well (e.g. Ethereum, Litecoin).

V. References and Relevance (for citations: http://www.citationmachine.net/mla/cite-a-book)

Google. “Bitcoin Blockchain.” *Kaggle*, 12 Feb. 2019, www.kaggle.com/bigquery/bitcoin-blockchain.

- This source provides a dataset with the type of blockchain data we will need to conduct our analysis

Day, Allen, et al. “Introducing Six New Cryptocurrencies in BigQuery Public Datasets-and How to Analyze Them | Google Cloud Blog.” *Google*, Google, 5 Feb. 2019, cloud.google.com/blog/products/data-analytics/introducing-six-new-cryptocurrencies-in-bigquery-public-datasets-and-how-to-analyze-them.

- This is an article where Google explains BigQuery which will be useful for getting the data into data frames when we conduct our analysis

Price, Will. “Bitcoin Mining Pool Classifier.” *Kaggle*, Kaggle, 31 Jan. 2019, www.kaggle.com/wprice/bitcoin-mining-pool-classifier.

- This link provides us with examples of address classification, as linked by Google from the previous source, which might be usable as a template

Lin, Yu-Jing, et al. *An Evaluation of Bitcoin Address Classificationbased on Transaction History Summarization*. Department of Computer Science, National Taiwan University & Institute of Statistical Science, Academia Sinica, 19 Mar. 2019, arxiv.org/pdf/1903.07994.pdf.

- This paper provides good information on the problem we are trying to solve with address classification, and will help in engineering features and categories

Toyoda, Kentaroh, et al. *Multi-Class Bitcoin-Enabled Service Identification Based on Transaction History Summarization*. Dept. of Information and Computer Science, Keio University & National and Kapodistrian, University of Athens, July 2018, www.researchgate.net/publication/333599819_Multi-Class_Bitcoin-Enabled_Service_Identification_Based_on_Transaction_History_Summarization.

- This paper will also help in identifying possible features to use, and help us understand some core underlying principles of Bitcoin

Meiklejohn, Sarah, et al. *A Fistful of Bitcoins: Characterizing Payments among Men with No Names*. University of California, San Diego & George Mason University, Oct. 2013, www.researchgate.net/publication/262357109_A_fistful_of_bitcoins_characterizing_payments_among_men_with_no_names.

- This paper describes the heuristic for address clustering in Bitcoin.

Harlev, Mikkel Alexander, et al. *Breaking Bad: De-Anonymising Entity Types on the Bitcoin BlockchainUsing Supervised Machine Learning*. Centre for Business Data Analytics, Copenhagen Business School & Westerdals Oslo School of Arts, Comm & Tech, 2018, core.ac.uk/download/pdf/143481278.pdf.

- Breaking Bad: De-Anonymising Entity Types on the Bitcoin Blockchain Using Supervised Machine Learning