Project 1 Description

*Cuneyt Akcora*

*September 26, 2018*

IBM Opportunity

IBM Pilot Program in Blockchain (forthcoming) IBM and SDSC will offer a multi-day blockchain course as part of the IBM Academic Initiative. Program Leads: Naguib Attia (IBM) and James Short (SDSC)

Date and Location: March 2019 (tentative), San Diego Supercomputer Center (SDSC) Resources: IBM Global University Program: <https://www.clds.info/education.html>

Data Description.

Our data files (shared on elearning/Ethereum datasets) contain two primary groups: token network edge files, and token price files. The **Ethereum project** is a blockchain platform, and our data comes from there. Although Ethereum started in 2015, most tokens have been created since 2016. As such, tokens have different starting dates, and their data starts from that initial date.

Token edge files have this row structure: fromNodeID\ttoNodeID\tunixTime\ttokenAmount\r\n

This row implies that fromNodeID sold tokenAmount of the token to toNodeID at time unixTime. fromNodeID and toNodeID are people who invest in the token in real life; each investor can also use multiple addresses. Two addresses can sell/buy tokens multiple times with multiple amounts. For this reason, the network is considered a weighted, directed multi(edge) graph. Each token has a maximum token count maxtmaxt; you can think of maxtmaxt as the total circulating token amount.

**Clarification for supply (10/3/2018).** There are two things; first each token has a limited supply (i.e., token count, which can be found on coinmarketcap.com as circulating amount). Then each token may have sub-units. This is similar to dollar in the US. There are around 18 trillion dollars in the economy, and each dollar is divided into 100 cents (subunits). Similarly, there is a token supply, and then there is a subunit for each token. This idea comes from Bitcoin where subunits are called Satoshis, 1 Bitcoin =10^8 satoshis. Coin market cap gives the total supply, but not sub-units, which differ from token to token. Some tokens have 10181018 sub-units. That means there can be numbers as big as totalAmount∗1018totalAmount∗1018.

Etherscan.io gives these sub-units as decimals, please see the Vechain here: It has 18 decimals, which means each Vechain token has 10181018subunits. <https://etherscan.io/token/0xd850942ef8811f2a866692a623011bde52a462c1>

Price files have no extensions, but they are text based. If you open them with a text editor (use notepad++ or similar), you will see this row structure: Date\tOpen\tHigh\tLow\tClose\tVolume\tMarketCap\r

The price data is taken from <https://coinmarketcap.com/>. Open and close are the prices of the specific token at the given date. Volume and MarketCap give total bought/sold tokens and market valuation at the date.

Primary Token Selection.

In project 1, each group will work with a single token’s data. This will be your primary token. To this end, sum the group members’ UTD Ids and take modulo 20. Suppose that your id sum is 123456 whose modulo 20 gives 16. Order the tokens by file size on disk and choose the 16th biggest token. By this scheme, we will analyze one of the top 20 tokens.

**Update 10/4/2018** If you have selected beautychain1 or beautychain2, please notify me. These tokens have failed recently, we need to change your primary token.

Preprocessing step

Find your primary token and load its data. In each token, there may be outlier amounts which are bigger than the total amount of the token. Locate these extreme outliers, if exist, and filter them out. If there are many of these (>30), investigate how many users are included in these transactions.

**Update 10/04/2018** See this news as an example of why we have these outliers: (<https://cryptoslate.com/batchoverflow-exploit-creates-trillions-of-ethereum-tokens/>)

Quality

Perhaps the most important aspect of this project is the presentation. Your report should explain each step in your solution, and provide good visuals. You may use the ggplot2library to draw plots. Data science is the art of finding and presenting actionable insights from data. This report may be a good part of your job application portfolio, so please do your best. Your output will be a doc/pdf or html file. RMD files will not be accepted, because we will have token data files access in the code. A viewer may not have these files.

Question 1 [Due 10/10/2018]

Find the distribution of how many times a user 1 - buys, 2 - sells a token. Which discrete distribution type fits these distributions best? Estimate distribution parameters.

Question 2 [Due 10/24/2018, full project report due 10/31/2018]

How can we create layers of transactions with increasing amounts? This descriptive statistic is similar to bin selection in histograms. For example, we could choose layer1layer1 as those transactions that involve 0.01×maxt0.01×maxt in amount. Find a good value for the number of layers and justify your choice.

Once you create layers, you can compute a feature in each layer. An example feature is the number of transactions, another one is the number of unique buyers. As each edge has a unix timestamp, it is easy to compute the edge time to a date. For example, 1294226315 is equivalent to 01/05/2011 @ 11:18am (UTC). See the website <https://www.unixtimestamp.com/index.php> for unix time conversion. R has functions to compute dates from unix time stamps as well. This way, for a given day you can find all layer transactions in that day.

For example, you can say on 10/12/2018 there were 25 transactions in layer 1. The price of token on that date was 3.2$. For each day in a token’s history, you can then correlate price vs feature in time.

Find an algorithm to compute the correlation of price data with each of the layers (hint: start by looking at Pearson correlation).

Question 3 [Due 10/31/2018, full project report due 11/2/2018]

**To be solved by 3 student groups only** This question is similar to the first question. You will find the most active buyers and sellers in your primary token network, and track them in other tokens. Fit a distribution for the number of unique tokens they invest in. For this question, you need to use networks of all tokens, and see if your buyers/sellers appear in them.

The second project will involve regression and hypothesis testing on token networks. It will be announced in late October.

**interesting bit** Currently, there is no global clasification for tokens. But we know that a token may be related to certain industry; storj is used to buy online space, so it is related to IT, technology, etc. We are currently working to develop a categorization for tokens based on their usage, utility etc. When you are working with tokens, keep an eye on possible categories for all tokens, and specifically for your token.