Group 3 – Bitcoin Ledger Data

批注 [ST1]: Everything I have written is just a first draft, feel free to change or add anything as you see fit.

Group Members

- Hangyu Kang (hkang98)
- Shengwen Yang (syang382)
- Bodi Yang (byang89)
- Fengxia Dong (fdong6)
- Sam Tauke (sstauke)

Data Description

Our group intends to analyze data on the Bitcoin (BTC) ledger. Our data comes from a user on kaggle who scraped the BTC ledger to pull all transactions. The data is available https://www.kaggle.com/shiheyingzhe/bitcoin-transaction-data-from-2009-to-2018. The data is organized in a series of csv files that each has five columns: block height, input hash, output hash, sum of BTC transacted, and timestamp:

- **Block height:** the number of the block in which the transactions are recorded (like the page number in a ledger book).
- **Input hash:** the address(es) of the account(s) sending BTC and how much they have sent.
- **Output hash**: the address(es) of the account(s) receiving BTC and how much they have received.
- **Sum of BTC transacted:** the total amount transacted in the transaction.
- **Timestamp:** the generation time of this block.

Each row in the data corresponds to a single transaction and rewards to miners are not included in the transactions. One transaction can have multiple accounts sending and multiple accounts receiving BTC.

Analysis Questions

There are a number of possible questions that we could ask:

1. What is the average number of transactions/transaction quantity/time elapsed per block?
2. How has the number of BTC transaction / total BTC transacted changed over time? Has it been strictly increasing or have there been dips?
3. Is there a seasonal component to BTC use? (Do people transact more BTC at certain times in the year?)

Analysis Methods

For the first analysis question, we plan to use simple calculations like sum and division. For the second analysis question, we will calculate the differences between every two points and visualize the differences on the plot. To figure out whether there are extreme differences, we will use IQR to calculate proper range. Whatever the values out of the range, we will define the value as extreme differences. For the third analysis question, we will examine autocorrelation function (ACF) for a potential seasonality.

Computational Tools

We plan to use the CHTC resources to analyze this data. Because the data is nicely organized into blocks, it will be easy to use condor to separate our calculations based on the block number and then summarize the data once the initial calculations have been done. Similar to the Lyman exercise, we will write the actual calculations per block in R files and then wrap those / compile them in .sh and .sub files as necessary for distributed computation.

Download Code

The code below will download the first of 22 identically structured CSV data files. It is stored in the kaggle_downloader.sh file on our github page and assumes that you have the kaggle api setup and that a "./data/" subdirectory exists.

Data sample:

546	['1DZTzaBHUDM7T3QvUKBz4qXMRpkg8jsfB5', '1DCbY2GYVaAMCBpuBNN5GVg3a47pNK1wdi']	['1KAD5EnzzLtrSo2Da2G4zzD7uZrjk8zRAv', '1', '1DZTzaBHUDM7T3QvUKBz4qXMRpkg8jsfB5', '24']	['25']	2009-01-15 06:08

~/.local/bin/kaggle datasets download -f 0-68732.csv -d shiheyingzhe/bitcoin-transaction-data-from-2009-to-2018 --unzip

unzip *.zip -d data/

rm *68732.csv*