STAT605 Project Proposal

Augustine Tang, rtang56; Jonquil Liao, zliao42;
Yudi Mu, ymu27; Ruyan Zhou, rzhou84

1 Data and Variable Description

The URL of our data is https://www.kaggle.com/shiheyingzhe/bitcoin-data-part-three-from-jan-2009-to-feb-2018here. Our data sets from Kaggle website have 19 csv files 19.82 GB in total. The data starts from 2016-09-15, and ends in 2017-07-16. Each file contains many blocks of bitcoin, and the file name shows the scope of blocks. Every file has the same structure, and each has five columns: 1. Block height(can be regarded as block names), 2. Input (address), 3. Output (address), 4. Sum (of transaction), 5.Time (of the transaction confirmation). A block includes many transactions. We also included the daily price data of the bitcoin in the time scope of our Kaggle data.

The variable Input can be seen as the name of user who sell out bitcoins, and the variable Output is the name of user who bought bitcoins from the input user, and it also contains the transaction of each user. The variable Sum is the sum of transaction that shown in the output variable. The unit of transaction is 1 bitcoin. In preprocessing, we also separated the variable Time into several columns, which are Unix time, Month, Day, Hour, and Minute.

2 Statistical Questions and Methods

2.1 Extreme Value Models detecting the transaction pattern

We may want to get better understanding of the data pattern of minute transaction amount before analyzing the price along with transactions. Most intuitively, we can plot the average transaction amount per minute/hour and see the pattern by eyes, but given the massive data size, this might be impossible. So to develop a index associated with transaction amount and scaling down the observations might be a good choice. So we apply Fisher-Tippet theorem here – extreme value of i.i.d samples of any distribution will converge to one of the generalized extreme value distribution. More specifically, we plan to do:

- Find the block maxima Q of minute transaction per hour.
- Designate the distribution of hour maxima in every month, which is one the GEV distribution: Fréchet, Gumbel, Weibull. (24$\times$30 Qs per month)
- Get the tail index $\alpha = -1/\xi$ for every GEV each month.
- Implement a time series model on αs.

In this way, we are able to see the time-varying pattern of α and do prediction on it, which indirectly tells us the fluctuation of transaction amount – small α indicating large transaction amount and vice versa.

2.2 Association Analysis

Before going further, we detect association between daily price of bitcoin and transaction amount. The main methods employed here are non-parametric Association Analysis, e.g. Spearman, Kendall's Test. If they are highly associated, we could consider multi-variate model in the following parts.

2.3 Time Series Model Application

We tried to simulate the trend of the transaction amount and daily price using time series models. The method we use is ARIMA model. Also, we found that the fluctuation of these series are huge. We may consider studying their Volatility using some time series models like GARCH. And if they are highly associated from the result of the former part, we will try multi-variate time series model as well.

3 Code Snippet

sort.R

```
arg=(commandArgs(trailingOnly=TRUE))
a = read.csv(arg[1])
a[1,]
```

Command Line

```
Height Input Output  Sum  Time
430001 3D9FUuyE4dUgYP4wywsEATcQ7nK3hyfHmb 17ACopoKchcPj1fCGiBZPvR77UkGkyDz5n,
0.00888549 2016-09-15 23:39:34
```

4 Computational tools

The platform we choose is CHTC. The programming codes include R and BASH.