**Group 3: Analysis of Bitcoin Blockchain First Draft**

**Group Members**

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

**Introduction:**

For the past decade, Bitcoin (BTC) has been the pre-eminent cryptocurrency. Unlike the public banking system we are used to, the recorded transactions in blocks are open to everyone and everyone shares the same transaction records once the transaction had been done between the users. It uses block-chain technology for security. Each block is a record of the transactions that have happened since the last block. And those blocks are connected in a chain, each block points to the blocks before and after it. Thus, we call the cryptographic technique the block-chain. A block is finished once a bitcoin miner has solved the mathematical equation linked to that block. The miner who solves the hash problem associated with a block is then allowed to push the verified list of transactions from the block to the blockchain and is rewarded with BTC as payment. Because the size of the block is dependent on how fast miners can solve the hash problem, the amount of time between new blocks varies. The BTC protocol aims for each block to take about 10 minutes and every 2016 blocks the protocol adjusts the problem difficulty to get closer to this 10-minute average.

Additionally, blocks are limited to at most 1 mb of transaction data. So, it is possible that the amount of transactions since the last block could exceed the amount that can fit into the current working block (in cases of large numbers of transactions or a long time to solve block). To solve this, individuals will pay fees to miners to ensure that their transactions are included in the next block, ensuring that that they don’t have to wait a long time before they can consider their transaction as valid.

Given that it is an entirely new form of currency, there is much that is unknown about BTC. Our initial report asks relatively simple questions of the data. In this initial report, we examine the velocity of transactions as well as block mining over the 370,000 blocks, which covers the period from January 2009 to August 2015. As we expected, we find that both the number of transactions per block and the amount of BTC transacted increased steadily over this period. Additionally, we find that that time elapsed per block fluctuated in the early part of this period but remained relative constant over time.

As we move forward with this project, we intend to investigate more complex questions. A defining characteristic of BTC has been its high price volatility, especially in its early period, as illustrated in **Figure A** below. This price volatility gives rise to several questions:

1. Is there a relationship between price and transaction volume? Do we see that individuals are more likely to transact BTC when prices are higher?
   1. We will investigate this relationship by estimating regressions of price on lagged transaction volume and the opposite direction.
2. Does the speed of mining increase as the price increase?
   1. This question can also be answered using simple correlations and/or regressions.

**Figure A:** Price of BTC from1/1/14–3/30/17



Source: coindesk.com

**Data Background and Preprocessing**:

Our data comes from a user on kaggle who scraped the BTC ledger to pull all blocks that have input from the genesis block in January 2009 through block 507,999 in February 2018. The data is available [here](https://www.kaggle.com/shiheyingzhe/bitcoin-transaction-data-from-2009-to-2018).

A block is considered to have no input if no transactions occurred between the mining of the previous block and the mining of that block. We expect (and observe) that this would occur more frequently earlier in the data when there were fewer people using BTC. Indeed, in the early part of our data we see periods of several hundred blocks without any transactions. 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. The data are stored in 73 CSV files that are 72.41 GB in total. Additionally, we pulled the price of bitcoin from kaggle at one-minute intervals from 1/1/12-9/14/20.

Both sets of datasets are downloaded from the kaggle API using bash scripts.

We did our initial processing of the raw data in the Linux shell. We structure our analysis though a .dag file that splits up the 22 csv files over 22 different condor jobs and the compiles the results.

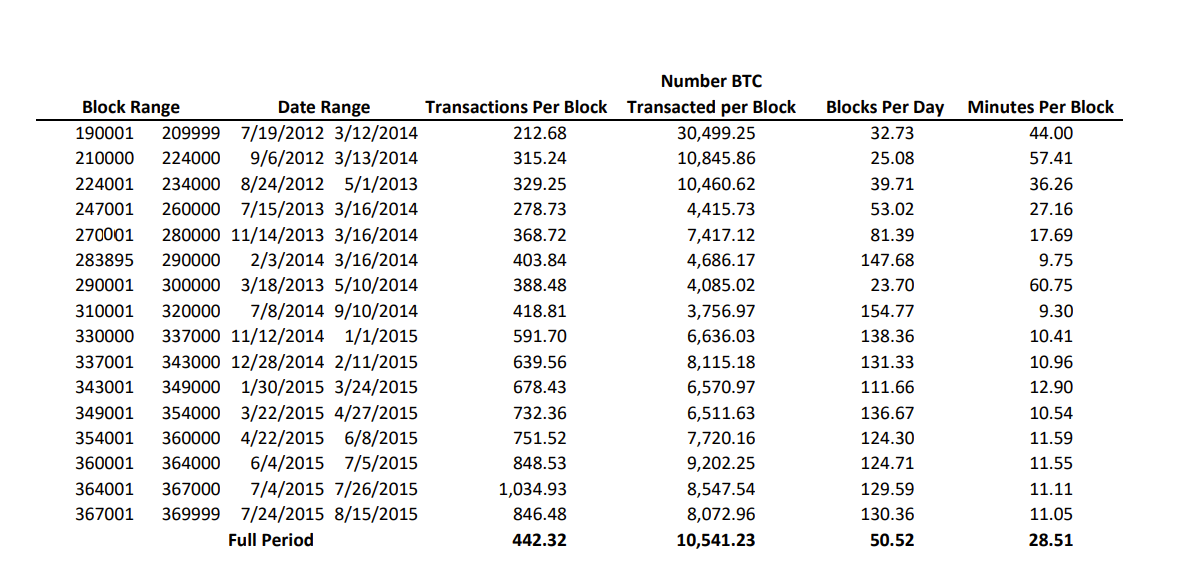
**Analysis Results**

Our first question of interest was: What is the average number of transactions/transaction quantity/time elapsed per block? We find an average of 442.32 transactions and 10,541.23 BTC transacted per block. Additionally, we find that the average minutes per block of 28.51.

As shown in **Figure B** below, the averages have changed over time. We see an increase in transactions per block from 212.68 over blocks 190,001 through 209,999 to 846.48 over blocks 367,001 through 369,999. A similar pattern emerges for the number of BTC transacted per block and the Blocks per Day. Further, we see a convergence towards the intended 10 minutes per block. All of these changes are consistent with the increased interest in BTC.

There are certain data anomalies that need to be explored further. For instance, we see a very high aver number of BTC transacted per block in the first range of blocks. Is this reflective of the true trading patterns during this period or were there outliers that skewed the numbers? Additionally, we see block ranges with nonsensical entries like dates in the future. This could be due to an error in our process or potentially indicative of an error by the individual who scraped the original ledger data.

**Figure B**: BTC Transactions 7/19/12–8/15/15



Note: Certain block ranges are excluded due to data anomalies. Blocks per day calculated as total number of blocks with at least one transaction divided by the total number of days with at least one transaction.

**Conclusion**

While far from groundbreaking, the results of this initial inquiry confirm our expectations. We observe that the volume (in terms of both number of transactions and amount transacted) has increased steadily over time. We also find that the minutes per block has tended towards 10 minutes per block.

In the future, we intend to both investigate the data anomalies that we observe here and investigate the relationship between price, transaction volumes, and mining rates.