ARTICLE

CrossMark

Analyzing outliers activity from the time-series transaction pattern of bitcoin blockchain

Rubaiyat Islam[1] · Yoshi Fujiwara[1] · Shinya Kawata[1,2] · Hiwon Yoon[2]

Abstract

In a closed economic system like blockchain, the total amount of generated cryptocurrency called bitcoin is conserved and the transaction patterns demonstrate an insight of money flow inside the blockchain. For the last 2 years, bitcoin market has grabbed an immense attention from the investors, technology entrepreneurs and currency enthusiasts. In this paper, we have come up with some findings in our investigation about the bitcoin time-series transaction patterns. We have graphically represented bitcoin's weekly patterns as a real economic currency that has been minted, stored and exchanged inside the bitcoin blockchain network. We identified outliers' activities with the help of descriptive statistical analysis. We also demonstrated transaction pattern behavioral change. The main implication of these findings is to understand some stylized facts of the time-series transaction of cryptocurrency-based fully digital financial system. Besides in our analysis, we have shown that the behavioral change of the transaction pattern is capable of explaining the system development events or major historical events that have a network impact.

✉ Rubaiyat Islam
  sb17q002@ai.u-hyogo.ac.jp

  Yoshi Fujiwara
  yoshi@sim.u-hyogo.ac.jp

  Shinya Kawata
  kawata@cmdlab.co.jp

  Hiwon Yoon
  yoon@cmdlab.co.jp

[1] Graduate School of Simulation Studies, University of Hyogo, Kobe, Japan

[2] CMD Lab, Inc., Tokyo, Japan

1 Introduction

Bitcoin is the pioneer of the increasingly popular cryptocurrencies and also the unique example of a large-scale sustainable payment system, in which all the financial transactions are publicly available (Whitepaper, Nakamoto 2008). It is issued by no central authority like government, bank or organization rather by mathematical cryptographic protocols in a distributed network system. In this system, the users pseudo-anonymously exchange bitcoins. Some special users are called miners, who only can mint bitcoins in the system by solving cryptographic puzzles by donating high computing power and in return, they earn bitcoins as their proof of work. So far, economic literature on the bitcoin issue is quite limited. Fergal and Martin (2013) and Ron and Shamir (2013) among the researchers of various background had successfully drawn the attention to the analytical aspects related to the information contained in the blockchain. Due to its still relatively low acceptance in the foreign exchange market and its poor performance as a medium of the store of value, some attention in the academic world has attracted the discussion on whether bitcoin can be considered a currency. But the trust in this currency totally comes not on the belief in central monetary authority rather computer technology and cryptography.

This paper is basically focused on three aspects of blockchain-based open source financial data. First of all, we re-structured blockchain's already extracted database (Kondor et al. 2014) for generating a hawk-eye view to observe bitcoin's daily transaction patterns to understand its economic growth with the course of time. Secondly, we had done the descriptive analysis on that database of daily transaction number and bitcoin volume to understand some of the most interesting and informative statistical distribution. Finally, we had investigated the rank distribution of some of the distinct transactions and their descriptive statistical facts to extract some network topological features.

The blockchain is one of the revolutionary databases that had been evolving since the last decade. It stores any information in a decentralized computing system and once stored, data inside it can never be altered or manipulated. It is transparently accessible to all the users logged in the database and they can view all the information published in the blockchain. Bitcoin cryptocurrency along with the sender and receivers' addresses in the form of a ledger is the financial monetary information that stored inside a blockchain and has introduced world's first successfully implemented fully digital cryptocurrency. It solved two real-world problems like "double spending" (Diego and Giovanni 2017) and "duplication problem" and made an alternative way to establish fully functional virtual currency based financial system.

2 Data set

The data used for this paper have been downloaded from the website of the Hungarian bitcoin research group (http://www.vo.elte.hu/bitcoin). Their reconstructed database comprises transaction data (sending and receiving bitcoins) with sending and

receiving addresses extracted from the blockchain network constituting the time duration from January 2009 to February 2018. The available data have been uploaded on the website in text files and some of the blockchain's extracted parameters have been mapped with randomly generated numbers in order to allocate those efficiently by the research group. Long characters of hashes have been mapped to random indicators, for example, BlockID which starts from numerical 0 value, representing the genesis block (first block of the bitcoin blockchain), and ends up to the value of 501418, which is the last block to download on the cutoff date of the month of February 2018.

For our research purpose, we have further restructured the data. The structure of the data has been shown in Fig. 1.

After reconstructing the database, we had fixed our timestamp units into each day from time duration of 1st of January 2009, when the bitcoin blockchain first initiated, to the cut off date of 8th of February 2018. We summed up the transaction count for each of those days and also summed up the input volume of bitcoin for each transaction. In each block, each bitcoin transaction follows either of the two rules as an input–output relationship in terms of bitcoin volume, i.e., input = output or input = output + transaction fees. This is the reason we have summed up the bitcoin volume of each input of transactions which represents the actual volume of bitcoin exchanged through one transaction to another. Our database excludes the transactions of the miner's bitcoin generation which are called the coinbase transactions, having no inputs, which were filtered out separately to form a separate database to merge into our analysis.

A glance at final data for our analysis has been shown in Tables 1 and 2.

In Table 3, we had the sampled price data with time duration of 9 years since bitcoin genesis block published from the beginning of January 2009. We have
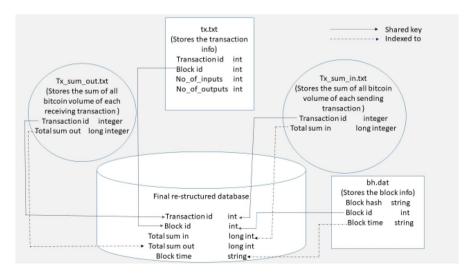

Fig. 1 The final reconstructed database that generates the summation of bitcoin volume for inputs and outputs of the transactions recorded during Jan 2009 to Feb 2018 with the block time UTC timestamp

Table 1 The sample of final data for transaction count and Bitcoin volume involved in each day transaction from a data compiled from Jan 2009 to Feb 2018

Date	Exchanged bitcoin volume/day (1 BTC $= 10^8$ satoshi)	# trans-actions/day
2009-01-12	17900000000	7
2009-01-14	6100000000	1
2009-01-15	50000000000	8
2009-01-16	20000000000	2
2009-01-18	15000000000	1
2009-01-19	77500000000	2
2009-01-20	40000000000	1

Table 2 The sample of final data for transaction count and Bitcoin volume involved in each day transaction from a data compiled from Jan 2009 to Feb 2018

Date	Mined bitcoin volume/day (1 BTC $= 10^8$ satoshi)	# mining transaction/day
2009-01-03	5000000000	1
2009-01-09	70000000000	14
2009-01-10	305000000000	61
2009-01-11	465000000000	93
2009-01-12	470000000000	94
2009-01-13	615000000000	123
2009-01-14	645000000000	129

Table 3 The sample of Market price in USD per BTC had been compiled from Jan 2009 to Feb 2018 Source: https://blockchain.info

Date	Price/BTC (in USD)
2011-01-21	0.44
2011-01-23	0.4443
2011-01-25	0.425
2011-01-27	0.4174
2011-01-29	0.446
2011-01-31	0.5
2011-02-02	0.840099
2011-02-04	0.88
2011-02-06	0.92
2011-02-08	0.9
2011-02-10	1.1

downloaded the market price data from blockchaininfo website (Market-price 2018) and used in our analysis.

3 Result and analysis

3.1 Descriptive statistics

We have taken the summary statistics from our reconstructed data and extracted some important information. Table 1 has the sample data that comprised of daily number of transactions and total volume of BTC daily transacted by regular users. Table 2 contains the sample data of the daily number of mining transactions and total mined volume daily done by the miners. We had gone through these columns of data to extract some univariate quatitaive information (Table 4).
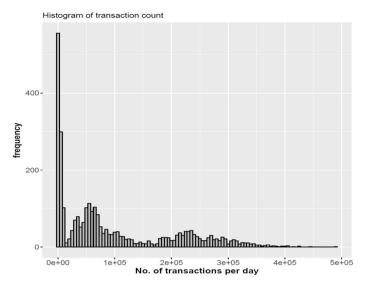
The sample size of the data is quite satisfactory. The number of transaction count in both tables is adequate to measure the statistical distribution. For the transaction count of miners per day, the mean and median values are almost same meaning that the mining transaction count is normally distributed. On an average, the mining transactions per day are 153, which are almost close to the theoretically calculated value of 144 (6 transactions per hour *24=144 transactions/day). The number of regular transaction data shows quite opposite statistics. By observing the central tendency of the data we can clearly see that the median value of both bitcoin volume and transaction count columns are smaller than that of their mean. So the data appeared to be rightly skewed. The histogram of the data shows that a majority of the data are located on the low side of the graph. Often, skewness is easiest to detect with a histogram or boxplot (Fig. 2).

Clearly, we can find some outliers on the high side that strongly affected the results of our analysis. The miner's bitcoin volume per day and mining transaction count dataset also show significant skewness and kurtosis leading to our analysis strategy of observing the data in logscale. Besides, the high mean and median values show that for the dataset it is an ideal example of a distribution that has a stronger peak, more rapid decay, and heavier tails (Fig. 3).

Table 4 Summary statistics of the complete dataset containing daily BTC volume, daily transaction count, total BTC volume daily mined and the daily count of mining transactions

| Parameter	BTC exchanged/day	Tx_count/day	BTC mined/day	Tx_mined count/day
Sample size (n)	3064	3064	3319	3319
Mean	1484291	97198.72	5131140000	153.13
Standard deviation	2316837	103734.4	2750478000	39.021
Median	1029820	58896.50	4236127000	153
Minimum	1	1	50	1
Maximum	67350000	490459	34600	692
Range	67350000	490458	34550	691
Skewness	11.35	0.97467	1.892	2.264
Kurtosis	244.71	−0.24295	11.46	32.786
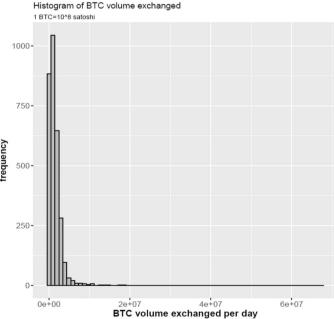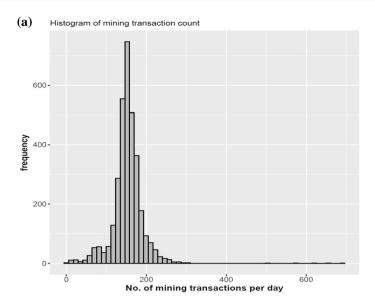
Fig. 2 Histogram of the dataset containing daily transaction count and total BTC volume transacted daily by the regular users

3.2 Auto-correlation function of BTC volume, price and number of transactions

In the previous section, we employed statistical approaches to viewing the distribution of daily BTC volume and daily number of transactions. Here, we are
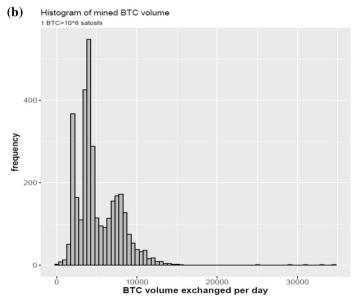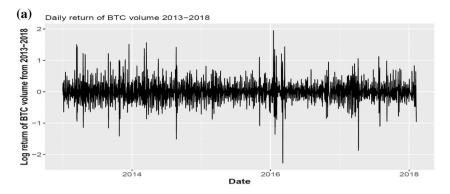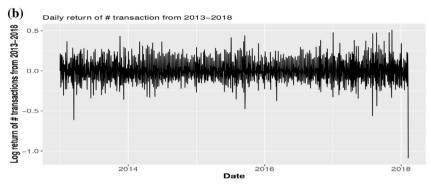
(a)

(b)

Fig. 3 Histogram of the dataset containing daily mining transaction count and daily total volume of mined bitcoin

using auto-correlation function to see if we could predict the direction of daily log returns. The log return can be defined as:

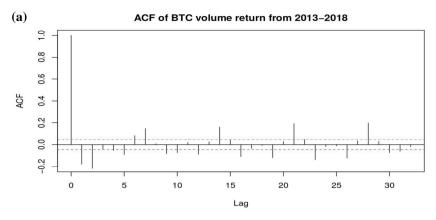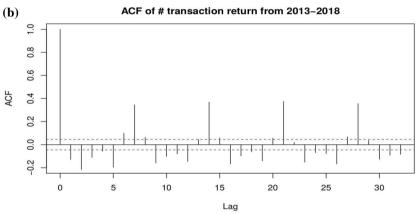$$x(t) = \log \frac{z(t)}{z(t-1)}, \tag{1}$$

(a)

(b)

(c)
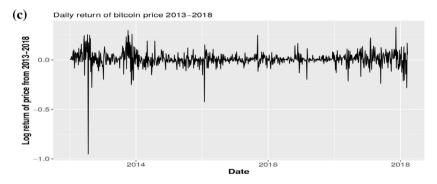
Fig. 4 The log return of **a** daily BTC volume, **b** daily number of transactions, **c** daily price

where $x(t)$ denoted log return of a variable $z(t)$ on day t. We measured the log return in order to make the series stationary for the empirical analysis. Now in our case, we calculated the log return of the BTC volume $v(t)$, and the number of transaction TX(t) and the daily price data $P(t)$, downloaded from the blockchaininfo website. We see in the Fig. 4a–c the stationary time-series log return plots of volume, the number of daily transaction and price data. The time-series data have been selected from 2013-01-01 to 2018-02-08 in order to maintain the consistency.
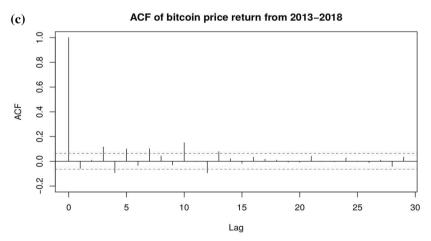
(a)

(b)

(c)

Fig. 5 Auto-correlation function of **a** BTC volume return, **b** number of transaction return, **c** price return
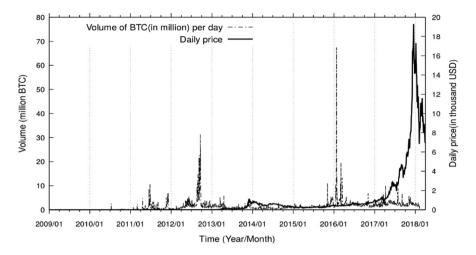
Fig. 6 The time-series data of BTC volume and number of transactions per day
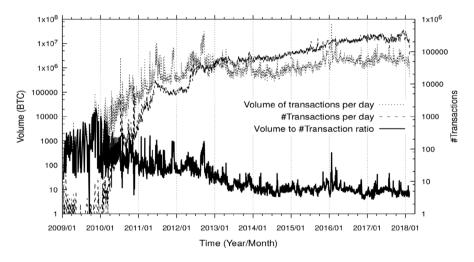

Fig. 7 Log scale plot to understand the weekly pattern of exchange of bitcoin transaction

We plotted the auto-correlation function of the three daily returns with the previous lags. The dotted line is the 95% confidence interval. For the BTC volume and number of transactions, the time-scale for relaxation was found approximately a week as shown in Fig. 5a, b. For price data, in Fig. 5c the ACF vanished at the lag of 1 day. This is reasonable as otherwise, one can do a linear prediction for up or down of tomorrow's price based on that of today.

3.3 The evolution of bitcoin transactions (a bird-eye view)

The distribution of transaction count to volume with the evolution of time has been plotted. An interesting set of observables to better understand the underlying
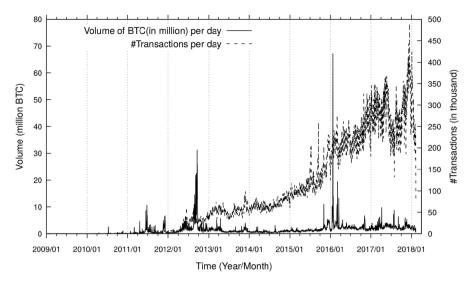
Fig. 8 Price evolution of bitcoin volume transacted per day

evolution of a unique financial system has been demonstrated. We found that there is some impulse of the volume of bitcoin transaction in the different time slots. Our research focused on this evolution of the financial system is after 2013 when bitcoin is a full-fledged matured currency used by people by trading goods and services.

Figure 6 shows that the daily exchange of bitcoin volume substantially increased after 2013 and onwards. In Fig. 7, we have plotted the same graph in log scale and showed the average volume quantity flowing through the number of daily
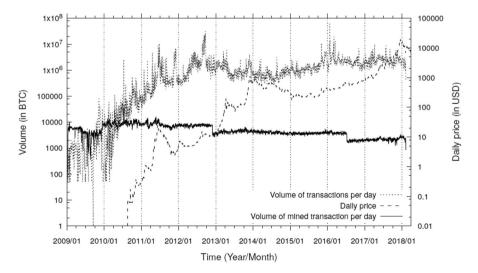

Fig. 9 Price evolution of bitcoin volume transacted and mined per day in log scale

transactions. The weekly pattern of volume flow observed in the graph proves bitcoin having a solid real economic financial system that we have statistically derived in the next section. In Fig. 8, we have portrayed the volume of BTC compared with price and observed the high price volatility.

We have plotted another graph in Fig. 9 where we showed the time evolution of price, the number of mined transactions and BTC volume. We observed that the number of supply mining transaction has quite stable throughout the time series.

3.4 The weekly pattern of bitcoin volume and transactions

We can observe that the volume per transaction became relatively stable after 2013, while it was so volatile before the year (see Fig. 6, for example). Also, it is known that bitcoin mining to generate blocks has been quite stable since the year 2013. So let us use data from January 1, 2013, in the following analysis of power spectrum.

Consider a time-series x_n with $n = 0, 1,\dots,N-1$, where N is the length of the time-series. Discrete Fourier transform of x_n is given by

$$X_k = \sum_{n=0}^{N-1} x_n e^{-2\pi nik/N}, \tag{2}$$

where $k = 0, 1,\dots,N-1$, and $i = \sqrt{-1}$, i.e., the unit of imaginary number. Obviously (1) is periodic in k with period N, so one can adopt the convention that $X_{-k} = X_{N-k}$. Because x_n is real, it follows that $X_{-k} = X_k^*$, where * denotes the complex conjugate. Frequency f_k corresponding to k is defined by

$$f_k = k/N. \tag{3}$$

The range of frequency can be regarded as $-0.5 \le f_k \le 0.5$.

Power spectrum or periodogram is defined by

$$P(f_k) = \frac{1}{N}|X_k|^2, \tag{4}$$

where $|X_k|$ denotes modulus or magnitude of X_k. Because $X_{-k} = X_k^*$, one can focus on the range $0.0 \le f_k \le 0.5$. $P(f)$ represents how much oscillator or harmonic movement with the frequency f and, equivalently, the periodicity $T = 1/f$ is contained in the original time-series x_n. Therefore, $f = 0.5$ corresponds to $T = 2$, namely most highly oscillated movement; $f \to 0$ is $T \to \infty$. One often uses smoothed periodogram by applying a filter to the raw periodogram (Bloomeld 2000).

We apply the method of smoothed periodogram for the time-series of daily volume V_n and a daily number of transactions T_n (where n denotes time in the day) in order to find periodicity in them. From Fig. 6, it is obvious that the time-series of volume and transactions have a trend of exponential growth; it would be natural to take logarithms of them and to consider the time-series, $x_n = \log V_n$ and $x_n = \log T_n$.
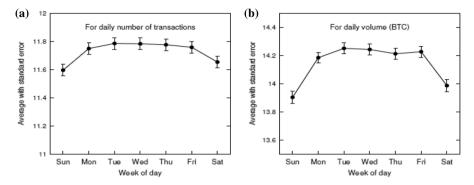
Fig. 10 Average of **a** logarithm of number of daily transaction, **b** logarithm of the daily volume of week. Error bar is the standard error

First, we segment the data into different days of the week, namely $n =$ Sun, Mon,…..Sat, and calculated averages and standard error (defined by standard deviation divided by the square root of a number of data in each collection). The result is given in Fig. 10. One can see that the level of volume or transaction is higher during weekday than weekend; in other words, there exists a weekly pattern that is not obvious in Fig. 6.

Additionally, we performed the above method of periodogram for each of the time-series. We employed a detrending by removing the mean of the series and subtracting a linear trend, a tapering with 10% at the beginning and end of the series, and a modified Daniell smoothing with successive simple moving averages of lengths 6 and 12 (Bloomeld 2000).

The result is given in Fig. 11. For both of volume and number of transactions, one can observe an obvious periodicity at $f = 1/7$ or equivalently $T = 7$ days as denoted by the dotted vertical line. Also present are higher order harmonics at $f = 2/7$ and so on. On the other hand, there is an overall increase of power spectrum towards $f \to 0$, corresponding to the trend of exponential growth which is already shown in Figs. 4 and 5.
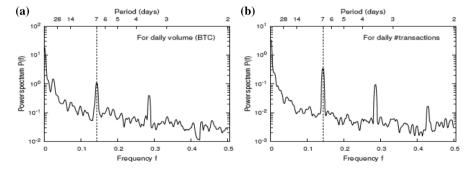

Fig. 11 Power spectrum as smoothed periodogram for the time-series of **a** logarithm of daily volume, **b** logarithm of daily number of transactions
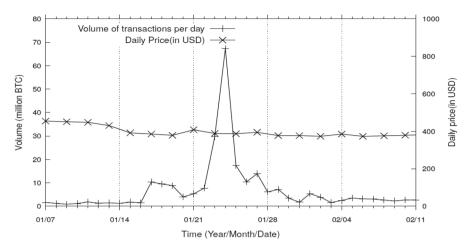
Fig. 12 A closer look at the unusually high volume of transaction happened on last week of January 2016

3.5 The outliers' transaction patterns

In our analysis, we had concentrated on two of the timeslots to find the outliers transaction pattern. Both the patterns have an unusually high spike of bitcoin volume within 1 month recorded from January 2016 to February 2016 shown in Fig. 12 and even though the there was not much variation of price there was a very big volume bitcoin circulated during the last week of January 2016.

There was an interesting finding in Fig. 13. The number of transaction appeared in January stayed in a range of 0.18–0.24 million and the BTC volume spike
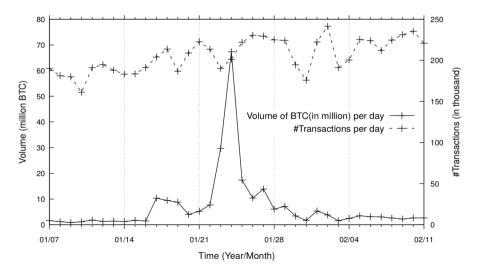

Fig. 13 A closer look of BTC volume to the number of transactions during the month of January 2016

cropped up in the last week of January. This indicates that there must be some big volume of BTC flow happened on each of some specific numbers of transactions as the total number of the daily transactions are within the regular range. This resulted in the existence of some outliers' transactions, which are responsible for the big volume of BTC transacted during that last week. By selecting this transaction pattern timeslot of 1 week (from 21st to 28th January 2016), we recalled our main reconstructed database to find out the list of individual transactions involved in that timeframe. We have created a volume to rank distribution log–log plot in Fig. 14 to understand more about the outliers activities. From the plot, it has been clearly observed that a considerable number of transactions possess the low ranks at the tail of the distribution. Besides, the steep shape of the tail suggested that there is the large rank of transactions that contained a large volume of bitcoin flow.

But which are those outliers' transactions that were distinctively traceable among the rest? The most direct method we applied is to use quantiles. The quantiles are values which divide the distribution such that there is a given proportion of observations below the quantile. Mathematically, we estimate the q quantile, the value such that a proportion q will be below it, as follows. We have n ordered observations which divide the scale into $n+1$ parts: below the lowest observation, above the highest and between each adjacent pair. We set this equal to q and get

$$i = q(n + 1). \tag{5}$$

If i is an integer, the ith observation is the required quantile estimate. If not, let j be the integer part of i, the part before the decimal point. The quantile will lie between the jth and $(j+1)$th observations. The proportion of the distribution which lies below the ith observation is estimated by $i/(n+1)$. We estimated it by $x_j + (x_{j+1} - x_j)$ times $(i - j)$.
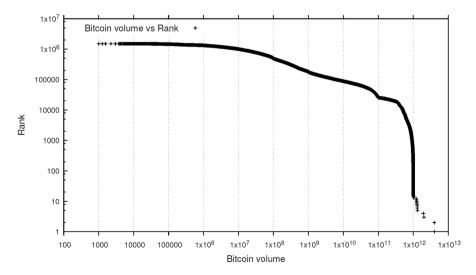

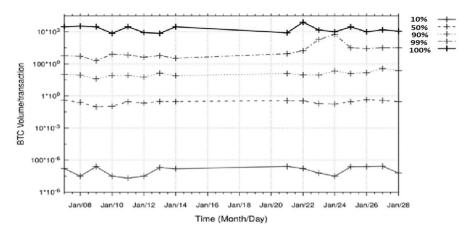**Fig. 14** Bitcoin volume rank distribution (log–log scale) from 21st Jan 2016 to 28th Jan 2016

Fig. 15 Quantile calculation for BTC volume/transaction for the January–February 2016

Now focusing on the individual transactions of the whole month of January–February 2016, we had calculated the quantile of BTC volume for each transaction which gave us the statistical insight of what percentage of daily transactions have a certain limit of BTC volume involved. As shown in Fig. 15, 99% and 100% quantiles on 21st–24th January have very interesting and have a statistical outlier pattern. For example, there is a transaction on 22nd January that has 40000 bitcoin involved in it. On 24th January, 1% of total daily transactions has 6000–10000 BTC volume at each transaction.
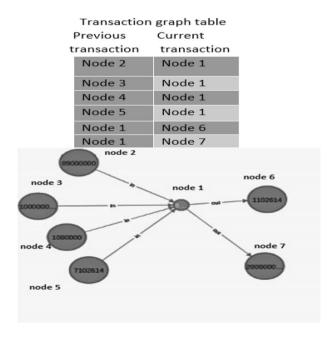

Fig. 16 Transaction graph

3.6 Directed transaction graph and degree correlation to visualize outliers' activities

The directed transaction graph represents the flow of BTC between transactions (Fergal and Martin 2013). Each node represents transactions and each directed edge between the source that is an input (previous transaction) and a target represents an output of transactions (current transactions) as shown in the Fig. 16. Each directed edge also includes a value of BTC flow (Michael et al. 2015). Thus, a transaction graph table can be constructed for each transaction as a node having a number of incoming connections called in-degree and outgoing connections called out-degree.

Mathematically, the in-degree of node i is the total number of connections onto node i and is the sum of the ith row of the adjacency matrix:

$$k_i^{\text{in}} = \sum_j a_{ij}. \tag{6}$$

On the other hand, the out-degree of node i is the total number of connections coming from node i and is the sum of the ith column of the adjacency matrix

$$k_i^{\text{out}} = \sum_j a_{ji}. \tag{7}$$

The degree correlation is the relation between k_i^{in} and k_i^{out} and sometimes can make a large difference to the effective properties of the complex network.

In our analysis, after considering the outcomes of the quantile plot, we have constructed a transaction graph. It includes all the transactions occurred in between 21st to 24th January 2016. All the transactions are represented as the nodes and links are represented as the input connections from the previous transactions to the current one. We measured the degree correlation of the graph by plotting the in-degree vs out-degree of each transaction node as shown in Fig. 17. The degree correlation graph visually described important network properties such as how many addresses involved as in-degree from the previous transactions and out-degree to the next ones. We came across an important finding that there was a transaction which has 3033 out-degrees with only 1 in-degree and also there was transaction with 633 in-degrees with 1 out-degree. This leads to the possible explanation of historical events soft fork of Bitcoin improvement protocol up-gradation called BIP-144 that was taking place during that period of time. These outlier transactions could be the result of the bitcoin developers' experiment.

4 Conclusion

We investigated on the unique transaction volume patterns and based on that we developed a methodology to extract interesting findings from a reconstructed database that has been extracted from blockchain system. We have found out that there are weekly patterns in a bitcoin volume to the price per day graph and there is a clear sign of economic financial trading of bitcoin flow among the transactions. The pattern of weekly
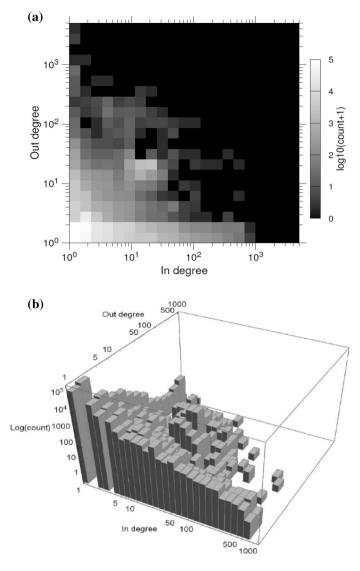
(a)

(b)

Fig. 17 The degree correlation of in-degree and out-degree **a** heat map, **b** 3D plot for transaction graph constructed in time series of 21st to 24th January 2016

trading shown in our analysis helped to investigate more on the specific impulses of transactions in a more focused timeslot. We have analyzed each transaction and bit-coin volume involved in that timeslot. The volume rank distribution helped us to iden-tify outliers transactions with the largest volume of bitcoin involved in it. The SegWit (Segregated Witness) and its effect in terms of the soft fork and hard fork debate were heating up during the beginning of January 2016 that might be one of the causes of this large amount of bitcoin flow in some outlier transactions.

The bitcoin system has historically gone through a lot of up-gradation which was termed as BIP (Bitcoin improvement proposal). In our focused tenure of January 2016, there was a big buzz in the bitcoin community that the size of block in blockchain needed to be increased. The result would make the transactions per second become faster. So BIP-144 Proposal to increase maximum possible block size starting at 8 MB was proposed in January 2016. Increasing the block size would increase the scalability but it will reduce the transaction fee which in turn would not profitable for miners since a block can hold more data and transactions that can use that new space and thus may be resulted in cheaper fees. So, there was a big debate going on between the miners and developers' community. There was a debate even for adopting either a hard fork or soft fork. Soft forks allow compatible changes. With soft forks, the old and new software can co-exist on the network. Hard forks break compatibility of all previous Bitcoin software and require every participant to upgrade to the same rules by a deadline or risk losing money. Such events can also harm network effects. After long debate and discussion, it was never merged. So, one conjecture we can make by observing these outlier transaction patterns of January 2016 that, there might be a lot of experimentation going on by the system developer community to observe network performances which might have caused this spike of the huge daily transacted bitcoin volumes.

The future research goal would be to interconnect the flow of bitcoin with address and transactions and find out more results that reveal new ways to understand the topological structure linkage with cryptocurrency evolutionary growth.

References

Bloomeld P (2000) Fourier analysis of time series: an introduction, 2nd edn. Wiley, New York

Diego R, Giovanni S (2017) Beyond bitcoin: a critical look at blockchain-based systems. Cryptography 1(2):15. https://doi.org/10.3390/cryptography1020015

Fergal R, Martin H (2013) An analysis of anonymity in the bitcoin system. Security and privacy in social networks. Springer, Berlin

Kondor D, Pósfai M, Csabai I, Vattay G (2014) Do the rich get richer? An empirical analysis of the Bit-Coin transaction network. PLoS One 9(2):e86197

Market-price (2018) https://www.blockchain.com/charts/market-price. Accessed 31 Oct 2018

Michael F, Michael SK, Sudeep P (2015) Bitcoin transaction graph analysis. http://arxiv.org/abs/1502.01657

Nakamoto S (2008) Bitcoin: a peer-to-peer electronic cash system. https://bitcoin.org/en/bitcoin-paper. Accessed 31 Oct 2018

Ron D, Shamir A (2013) Quantitative analysis of the full bitcoin transaction graph. Financial cryptography. https://eprint.iacr.org/2012/584.pdf