Structure and Evolution of Bitcoin Transaction Network

John Mern[1]

[1]Stanford University Department of Aeronautics and Astronautics, jmern91@stanford.edu

ABSTRACT

The Bitcoin digital crypto-currency has evolved from niche financial technology to major currency, with a market capitalization of $94 billion at the time of this writing. All Bitcoin activity is tracked in a blockchain ledger which maintains a complete history of every Bitcoin transaction. This complete history offers a unique opportunity to analyze the peer-to-peer interactions within the Bitcoin economy. This insight may allow for more powerful analysis of trends in macroscopic indicators such as price. This work analyzes the transaction network to identify trends in the macroscopic evolution of the network structure over time. A time-history of descriptive network statistics was developed and used as input into a machine learning model of Bitcoin price. In order to explore the potential of using more microscopic network data, we proposed a deep learning-based representation embedding of the network structure that compressed the entire network into a 128-word vector. The performance of a basic neural network model trained using this embedding was compared to several baselines. We found that while the embedded representation does increase the performance of the network slightly, it is not the best representation for our current application.

1 Related Work

Several works have been published in recent years in which the Bitcoin transaction network has been analyzed. In one such work by Kondor[1], the authors seek to augment traditional macroeconomic market analysis with detailed study of the underlying transaction network. Using the Bitcoin transaction ledger as the database, they study both the network structure and dynamics of money flow. A ”snapshot” of the network is analyzed by a variety of measures, including degree distribution, degree Gini coefficient, degree Pearson correlation coefficient, and clustering coefficient. Of the more significant findings, the paper reports the in- and out-degree distributions can both me modeled with power-laws. In a similar work[2], the authors extracted several of the standard descriptive statistics from the data, including degree distributions, clustering, and correlation analysis. In addition, they also calculated eigenvector centrality as a measure of how one node influences others in the network.

While these works do present a solid basis of analysis, one shortcoming is inability to adequately pre-condition the network nodes. While the data from Bitcoin is abundant, one major challenge is that many users will register multiple Bitcoin I.D.s over-time. Many I.D.s may be used for only a single transaction before becoming dormant. The reported statistics are likely polluted by these ”virtual” users.

Other works attempt to map the network structural evolution to macroeconomic trends. In a subsequent work by Kondor[3], the authors go beyond generating network statistics on the Bitcoin

database and instead seek to discover correlations between network behavior and Bitcoin market value. In doing this they first apply a heuristic approach to identify addresses belonging to the same user, which they call *contraction*. While this method is heuristic and likely sub-optimal, it is the most widely common method of contraction contraction currently used for research[4]. Improvements upon this method are an active area of research and are not pursued in this work. An unsupervised feature analysis is conducted on the graph time-history and the resulting feature vectors are shown to have strong correlation with the market price during the time period studied. In a similar work by Baumann[2], the authors map correlations between macroeconomic factors, such as market price, and the network dynamics. To this end, the work also considered the influence of external factors, such as regulatory scares, economic insolvency crisis, and others. The work successfully demonstrated the ability to identify several large entities on the network, each of which utilized several Bitcoin IDs.

2 Technical Approach

The ultimate objective of the proposed network analysis is to identify the structural features of the Bitcoin transaction network with with good explanatory power on the trading price of Bitcoin. As the prior works have suggested, there are several prevalent, intuitive features of the structure of the transaction graph that show a strong *correlation* to Bitcoin price. We seek to extend these conclusions to identify features with independent explanatory capability, implying an underlying *causal* relationship. To achieve this, we apply several techniques from the fields of unsupervised data analysis and supervised machine learning to select these fundamental features in an empirically justifiable manner. The work pipeline can be divided into four main phases as described below.

1. Contraction and data conditioning
2. Network Structural Analysis
3. Network Representational Embedding
4. Macroeconomic Machine Learning

In the first step, we gather and condition the data from the Bitcoin ledger to be suitable for network analysis. In the second step we calculate the critical descriptive statistics for the network over the time period(s) of interest, selecting those with prevalent trends as potential features for learning. In addition, we a generative model to the graph, and use the model parameters as learning inputs as well. A main contribution of this work is in the third step, in-which we propose a representational learning approach based on Convolutional Neural Networks (CNNs) to embed a representation of a sub-space of the graph spectral domain. In the final step, we apply a series of machine learning techniques to learn a model correlating the selected features with price trends.

Contraction and Data Conditioning

The data used in the experiment is the same dataset as used in[13]. This is the Bitcoin ledger as of the end of 2013. The final network in this dataset has 13M nodes and 44M edges. We are only interested in the ”steady-state” Bitcoin network, so we will only consider data from after January 2011. The first step in conditioning this dataset was to group Bitcoin addresses that likely belong to a single user. While it was originally proposed to attempt an improvement over existing methods, further study of the issue showed it was likely beyond the desired scope of this work. Instead the heuristic

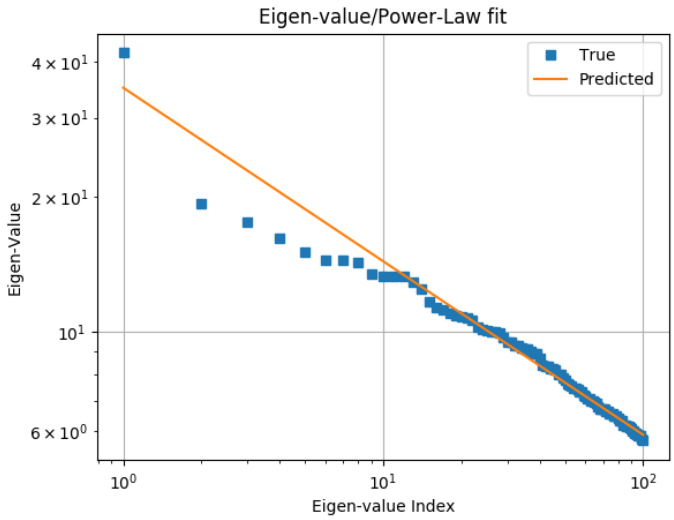
method used in prior works was employed. In any transaction involving multiple senders, all public addresses are associated, as the initiate of this transaction must have known the private-key of each of these addresses and is therefore the likely owner of all. In this way, this method provides an upper bound on the number of users on the network.

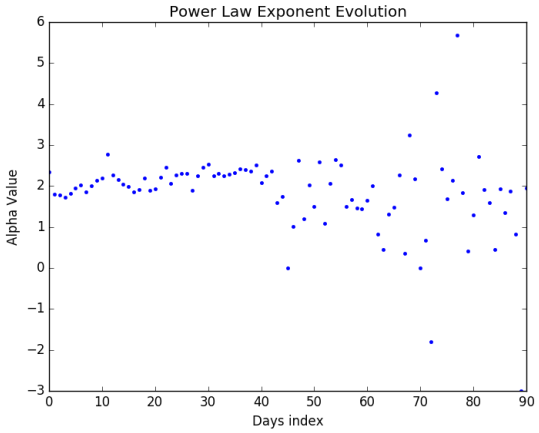
The Bitcoin ledger was downloaded and the data was parsed into an edge list of transactions in .csv format. The list was analyzed with the heuristic method, and the associated addresses were all reduced to a single user ID (in this case the lowest Bitcoin address in the transaction). This step reduced the total number of unique transactions considered from 129M to 24M over the three year period.

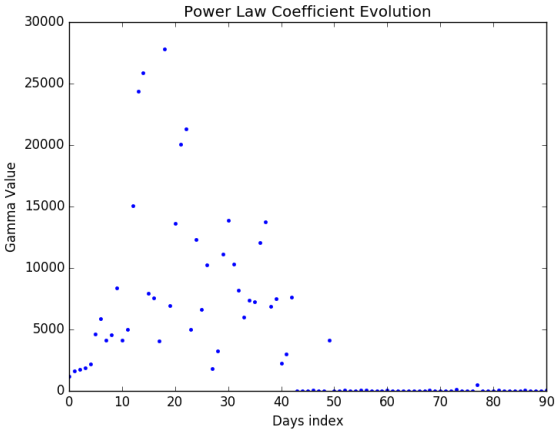
Each day, we constructed a graph of the network that included all previously seen users as nodes and prior transactions as edges, with the new users and transactions from that day added. This left us with an increasingly large and "noisy" graph as many IDs are used a single time and then become dormant. In order to address this, we pruned the network for dormant IDs each day, where a dormant ID is one that has not participated in a transaction for the past ten consecutive days. We tried this technique with multiple thresholds for dormant period, and 10 was selected as it allowed for the most aggressive pruning with little reactivation of dormant nodes. The maximum graph size then considered on any given day had no more than 100,000 nodes.

Network Structural Analysis

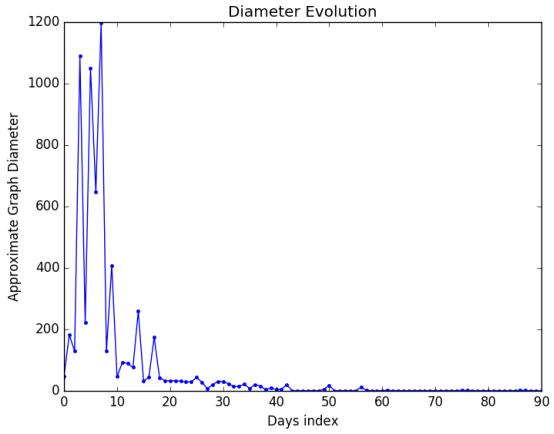
Once the user network was conditioned, we proceeded to develop a time-history of descriptive statistics. First, we confirmed findings from previous works that the degree distribution is represented by a power-law. We conducted the analysis on the final graph at the end of 2013 using a maximum likelihood estimate of the parameters of the function $f(x) = \gamma x^{-\alpha}$, and the results are shown below. As previously reported, this trend does approximately follow a power-law, though a sub-linear model is required for a better agreement. This trend suggests that a weakened preferential attachment model would be a good basis for the Bitcoin microscopic evolution. Every ten days, the parameters of this model were recalculated, and the results are shown below. As can be seen, the α value remains approximately constant over time, while the γ converges as the graph stabilizes.

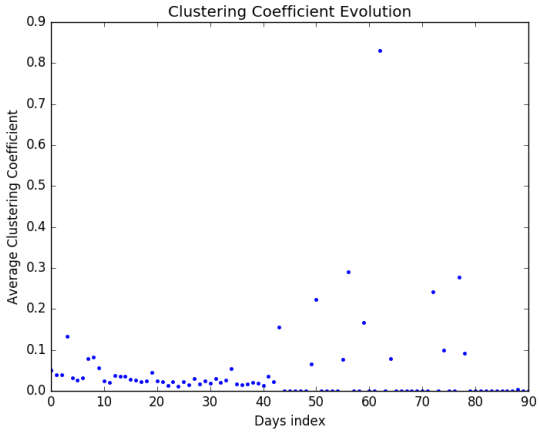


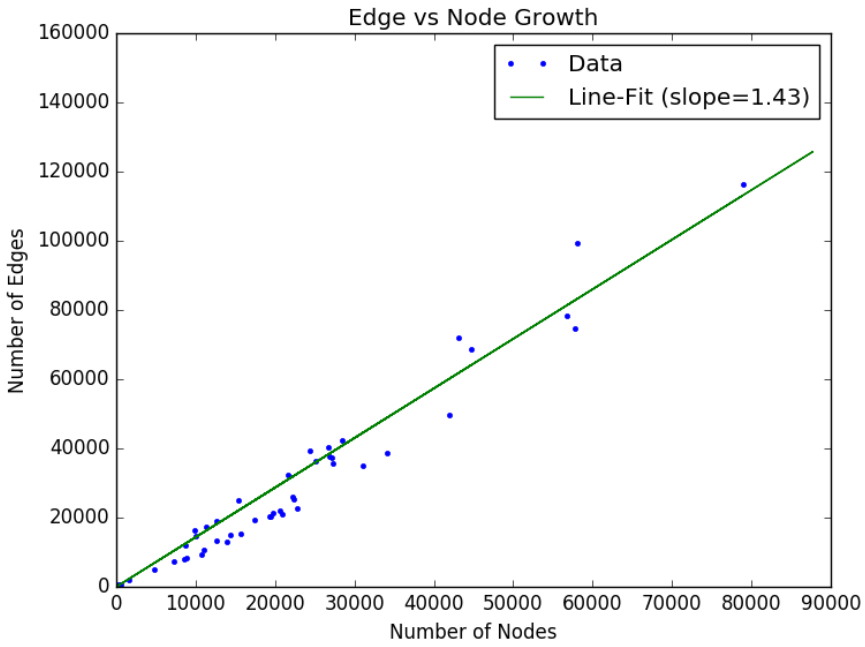




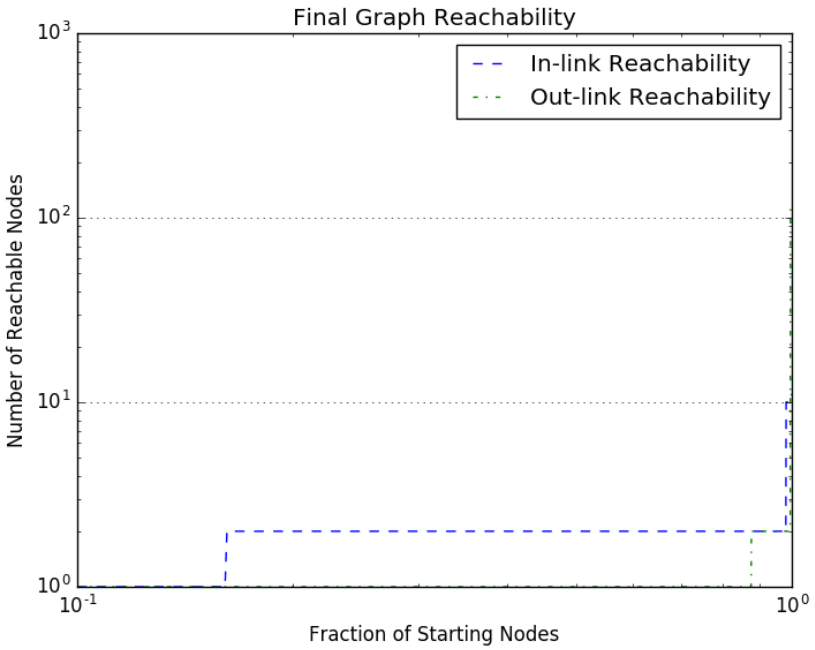
Once the power-law distribution was confirmed, we continued to develop time-series statistics of the network. Shown below are the clustering coefficient and the approximate network diameter evolution. The network diameter was calculated with a breadth-first search (BFS) originating from 1,000 randomly sampled nodes in the graph. The decrease in diameter along with the likely preferential attachment node degree sequence suggests a densification should be observed in the average degree trend over time. This was confirmed as shown by the plot of node growth vs edge growth.







Finally, in an attempt to develop an interpretable description of the network structure, we plotted node reachability at various graphs throughout the time-series. An example of the reachability curves for the final graph is shown below.



This distribution may be explained by the fact that a small number of users conduct Bitcoin mining (a process by which new Bitcoins are created) and only a single central exchange existed at the time of this analysis. The high in-link reachability suggests that a large number of users are acquiring Bitcoins, while fewer are actually spending them. This is likely driven by the growth in Bitcoin popularity attracting new users during the time of the analysis.

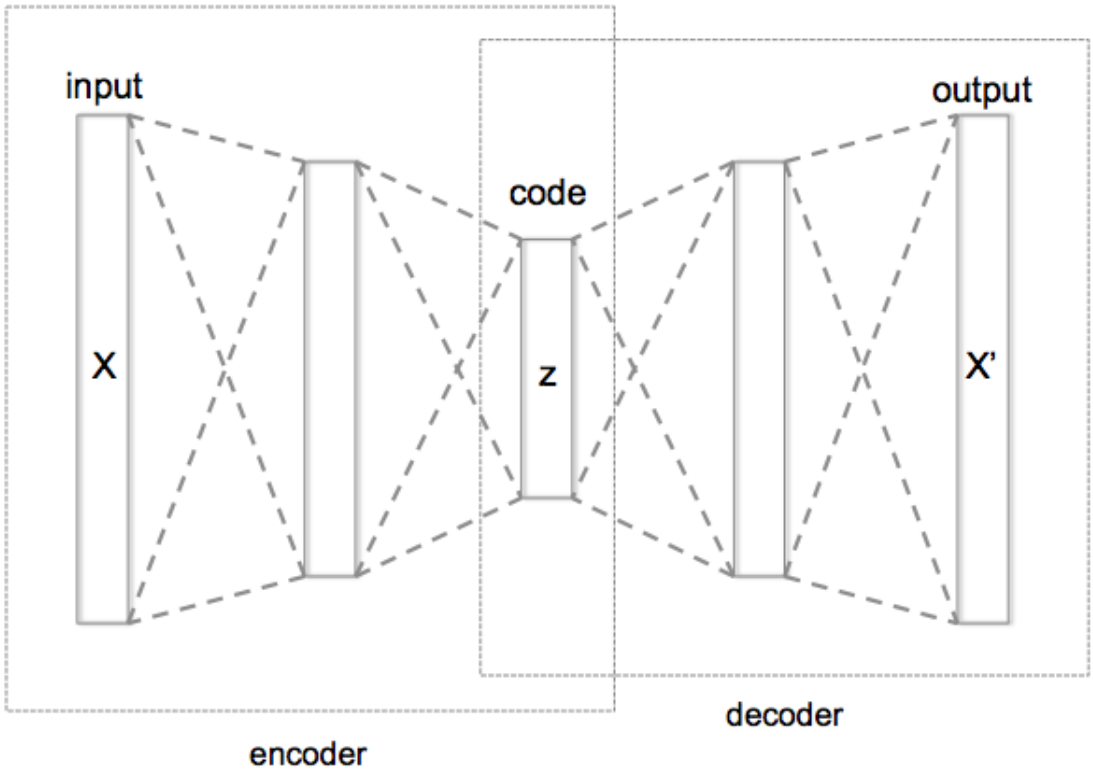
Network Representational Embedding

A hypothesis we wanted to test was that there was some explanatory power in understanding the evolution of key graph clusters over time. Instead of using conventional spectral clustering or modularity optimization, we chose to attempt a direct embedding of a sub-space of the graph spectral dimension through deep learning. The first phase of the developed method proceeds as follows.

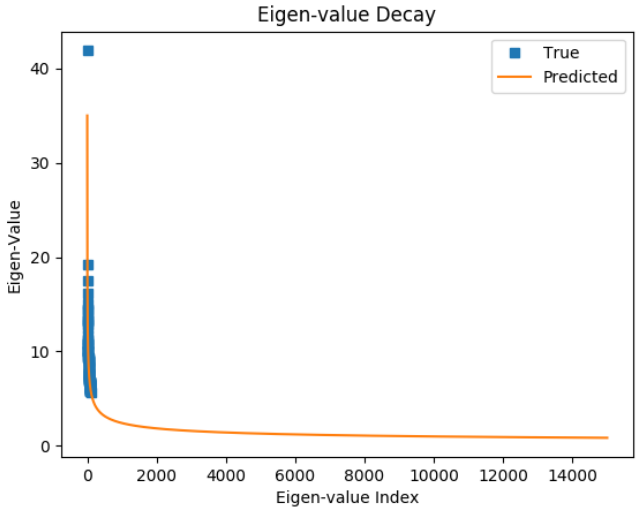
For each Graph G_i

1. Decompose the graph by eigenvalue decomposition to $G_i = U_i \Lambda_i U_i^T$
2. Truncate the eigenvalues to a targeted rank k (this is optimal via the Eckhart-Young-Minsky Theorem).
3. Create x_i an concatenation of U_i and Λ_i
4. Add X_i to the dataset X

Once this dataset is complete, a deep CNN auto-encoder was trained. A visual description of the architecture is given below. As can be seen, the network attempts to first compress the "image" (in this case x_i) to a reduced dimension at z in the encoder phase. The network then reconstructs original input through the decoder. After training is complete, the encoder can be used to generate embeddings for graph inputs.



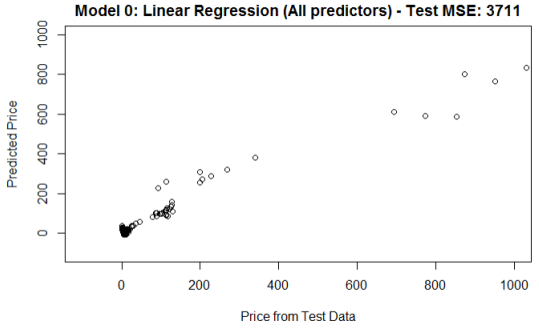
For this work, we truncated the eigenvalues to the top 100. Our theory is that this should tell us something about some form of the 100 most important clusters in the graph. To test the approximation loss incurred by this truncation, we fit a power-law to the singular value decay and compared the full graph to the truncated. As seen, a significant portion of the explanatory power is lost for a 25,000 node graph. Future work may seek to increase the retention of these features. In this work, memory limitations prevented further exploration.

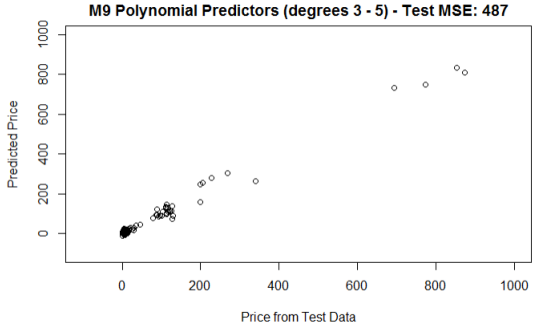


The hidden embedding z was set to a 128 length floating value vector. This was an arbitrary choice for this phase of the work. The inputs were zero-padded as required.

Machine Learning Approach

These statistics and the network embedding were used in a series of machine learning tasks to provide a correlation with Bitcoin price. As a first step, we conducted a Principle Component Analysis (PCA) on an edge-list representation of the graph over time. PCA performs a singular value decomposition on the matrix (in this case edge in dimension 0 and time in dimension 1), and selects the top singular vectors as the ideal feature basis directions. Using extrinsic market data as well as the first four singular vector values as an input features, we conducted a series of linear and non-linear regressions. The result of the first attempt and the final result of after ablation studies are shown below and the remaining ablation study graphs are in the appendix.





Additionally, we trained a simple feed forward neural network to predict change in price each day based on a set of input features. We tested this training with three different sets of input data.

Each set included the same sub-set of extrinsic market data along with a unique set of network data. The three network data sets were the first four singular vectors as described above, the embedded representation, and a collection of the previously presented descriptive statistics. An additional null model with no network data was trained. The singular vector model resulted in an mean squared error (MSE) of 262, while the null model scored approximately the same with 259. The embedded representation did slightly better, with an MSE of 215. The best performance was had from the descriptive statistics set, with an MSE of 192.

3 Conclusions

In this work, we performed an analysis of the evolution of the Bitcoin transaction network over a three year period. After adequate pre-processing of the data, we were able to identify clear structure and trends in the macroscopic growth of the network. We showed that the network structure is likely currently dominated by growth-driven trends via the addition of new users.

In order to aid in machine learning, we attempted an embedded representation of the graph spectral space via a deep CNN approach. While this embedding did capture some of the explanatory power of the graph, it did not perform as well as traditional feature selection. One likely reason for this is the loss of resolution incurred in the truncation step. Future work exploring more efficient representations allowing for larger input spaces may improve this performance. Additionally, pre-processing of this feature vector via an unsupervised technique may help to reduce the variance induced in training.

Despite the challenges, it can be concluded that there is indeed a correlation between structural dynamics of the Bitcoin network and the performance of the macroeconomic price. Further work, exploring better representation learning and more sophisticated time-series based deep learning would likely reveal further this relationship.

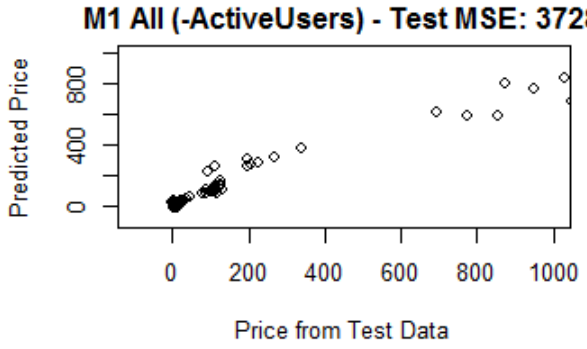
References

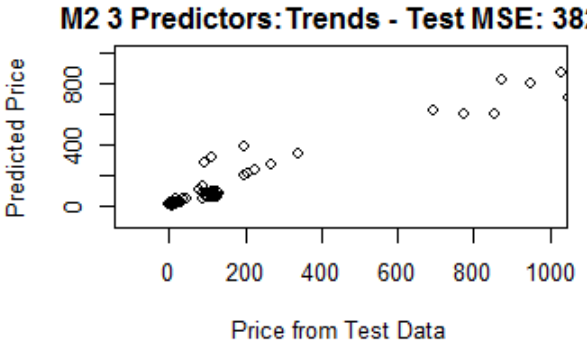
1. Kondor, D., Pósfai, M., Csabai, I. & Vattay, G. Do the rich get richer? an empirical analysis of the bitcoin transaction network. *PLOS ONE* **9**, 1–10 (2014). URL https://doi.org/10.1371/journal.pone.0086197. DOI 10.1371/journal.pone.0086197.

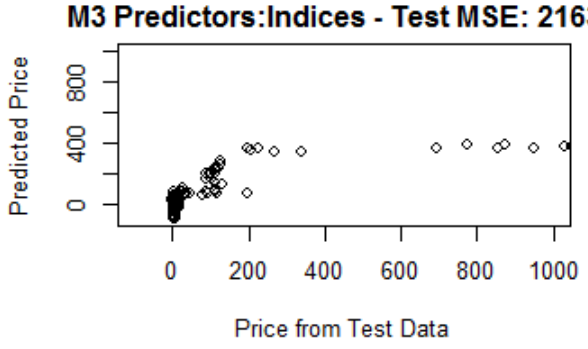
2. Baumann, A., Fabian, B. & Lischke, M. *Exploring the Bitcoin Network*, vol. 1 (2014).

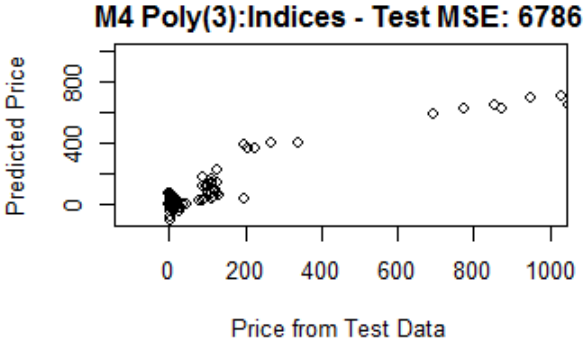
3. Kondor, D., Csabai, I., Szüle, J., Pósfai, M. & Vattay, G. Inferring the interplay between network structure and market effects in Bitcoin. *New J. Phys.* **16**, 125003 (2014). DOI 10.1088/1367-2630/16/12/125003. 1412.4042.

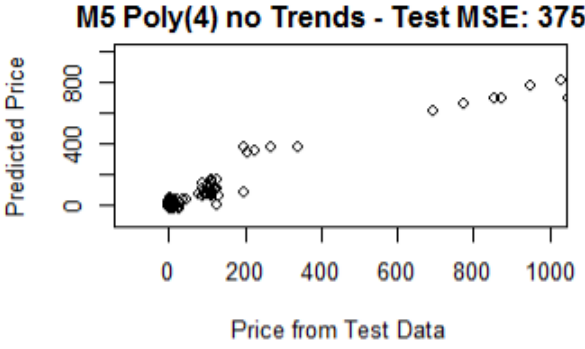
4. Spagnuolo, M., Maggi, F. & Zanero, S. *BitIodine: Extracting Intelligence from the Bitcoin Network*, 457–468 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2014). URL https://doi.org/10.1007/978-3-662-45472-5_29.

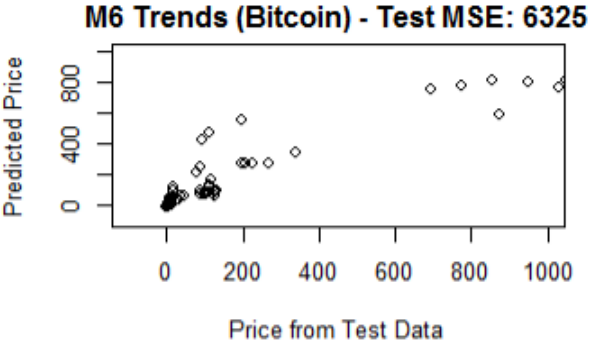

M1 All (-ActiveUsers) - Test MSE: 3728
Predicted Price
0
400
800
0
200
400
600
800
1000
Price from Test Data

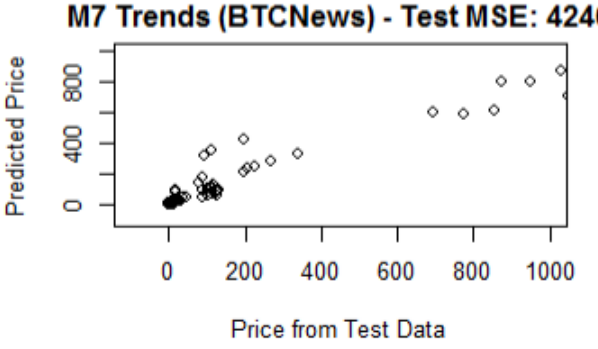

M2 3 Predictors:Trends - Test MSE: 3828
Predicted Price
0
400
800
0
200
400
600
800
1000
Price from Test Data

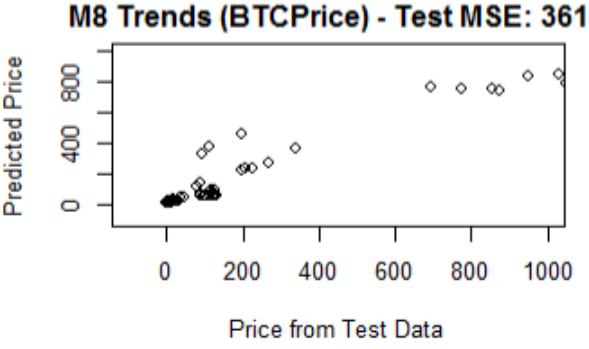

M3 Predictors:Indices - Test MSE: 21638
Predicted Price
0
400
800
0
200
400
600
800
1000
Price from Test Data


M4 Poly(3):Indices - Test MSE: 6786
Predicted Price
0
400
800
0
200
400
600
800
1000
Price from Test Data


M5 Poly(4) no Trends - Test MSE: 3758
Predicted Price
0
400
800
0
200
400
600
800
1000
Price from Test Data


M6 Trends (Bitcoin) - Test MSE: 6325
Predicted Price
0
400
800
0
200
400
600
800
1000
Price from Test Data


M7 Trends (BTCNews) - Test MSE: 4246
Predicted Price
0
400
800
0
200
400
600
800
1000
Price from Test Data


M8 Trends (BTCPrice) - Test MSE: 3613
Predicted Price
0
400
800
0
200
400
600
800
1000
Price from Test Data