Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Entrepreneurial Risks

Bitcoin Bubbles: Epidemic-Diffusion Analyses and Models

Term Paper

Pedro Daniel Partida Güitrón

daniepar@ethz.ch

Chair of Entrepreneurial Risks
ETH Zürich

Supervisors:
Prof. Dr. Didier Sornette
Sumit Kumar Ram
Dr. Vasile Gradinaru (study advisor)

March 1, 2019

Acknowledgements

This semester project is part of the curriculum of the M.Sc. program in Computational Science and Engineering at ETH Zürich. The research was carried out at the Chair of Entrepreneurial Risks.

I want to thank my supervisor, Prof. Dr. Sornette, who have supported continuously and helped me while performing this work. I would especially also thank my scientific adviser, Sumit Kumar Ram for his valuable guidance, advice, and supports during this project. I am deeply indebted to them, for all their invaluable advice and correction in preparing this work. I would also like to thank Dr. Gradinaru, my study advisor; without him, this work would never have been done.

Last but not least I own the deepest thanks to my family, especially my parents, for their love, encouragement, and support. Without them, I would not have been able to pursue this study.

Zürich, March 2019

Pedro Daniel Partida Güitrón

Abstract

Since the creation of Bitcoin, the popularity of cryptocurrencies has increased over the last years. Large and rapid price movements together with frequent bubble periods have characterized the cryptocurrency market. As a consequence, studying the behavior of Bitcoin prices aiming to predict the next bubble has taken on prime importance.

In this semester project, exogenous factors are taken into account to analyze the dynamics of Bitcoin price. We use the time series of Bitcoin price, YouTube views and Google trends related to Bitcoin news to study the relaxation response of a social system after exogenous bursts of activity using the time series of daily views for nearly 250000 YouTube videos. This analysis is based on the epidemic model presented in [1].

Furthermore, we validated the necessary stationary conditions for the first order difference of the time series in order to carry on with the rest of the analysis. Thus, we focus on the returns of the time series to observe the behavior of their volatility over different rolling windows and calculate the lag length among all the time series. Based on the reflexivity theory we verify the one-directional causality that prices affect news in social media. Additionally, in this project, we go one step further and explore the possible existence of bidirectional causality. Meaning that not only price has an effect on the news, but also that public interest could influence prices.

Keywords: Bubbles, Financial Time Series, Bitcoin, YouTube, Google trends

Contents

Acknowledgements **i**

Abstract **ii**

1 Introduction **1**

1.1 Background: Bitcoin Bubbles 1

1.1.1 Reflexivity 2

1.2 Objectives 3

1.3 Outline 3

2 Data **4**

2.1 Types of Data 4

2.1.1 Bitcoin Price 4

2.1.2 YouTube Data 4

2.1.3 Google Trends Data 6

2.2 Preprocessing of Data 6

3 Methodology **7**

3.1 Power-law 7

3.2 Time Series Analysis 8

3.2.1 Stationarity 9

3.2.2 Volatility Analysis of Returns 10

3.2.3 VAR Analysis of Returns 11

3.2.4 Granger Causality 11

4 Results **13**

4.1 Behaviour of Data 13

4.2 Fitting the Power-law 15

4.3 Stationarity 19

4.3.1 Correlogram 19
4.3.2 Dickey–Fuller test 20
4.3.3 Johansen Test 22
4.4 Rolling Window Volatility 23
4.5 Rolling Window VAR Analysis 25
4.6 Overall VAR 29
4.6.1 Response functions 33
4.7 Granger Causality 36

5 Discussion **41**

Bibliography **42**

A Appendix I **A-1**
A.1 VECM A-1
A.2 Empirical Mode Decomposition A-3

B Declaration of Authorship **B-1**

Chapter 1

Introduction

1.1 Background: Bitcoin Bubbles

Bitcoin has been the subject of discussion over the last years. One core of these discussions has been the large price movements and the frequent appearances of bubbles. A bubble is defined as a period of unsustainable growth, when the price of an asset increases ever more quickly, in a series of accelerating phases of corrections and rebounds [2]. Furthermore, during a bubble phase, the price follows a faster-than-exponential power-law growth process, often accompanied by log-periodic oscillations. This dynamic ends abruptly in a change of regime that may be a crash or a substantial correction.

Since Bitcoin was introduced in the market, the cryptocurrency has experienced frequent financial bubbles periods followed by a crash. Notorious examples are the financial bubble starting mid-2013 finalizing with a significant correction at the end of November 2013, the bubble in mid-2017 with the crash in December before Christmas. Those processes have a hyperbolic shape that reflects the steep rise in the price at the final stage of the bubble [2]. The frequent repetition of cryptocurrency bubbles in only a decade since they were introduced makes them very interesting to analyze. Furthermore, cryptocurrencies and especially Bitcoin have been the scope of discussion in news, videos, and trends, especially during the bubble periods.

In these discussions, the collective interest of the commons plays a significant role in determining the value of the financial assets. Hence, the time evolution of interest of collectives can be beneficial in the prediction of future interest and prediction of the future price of the assets. Further, the reaction of the interested people to the incoming information tells us about the impact/importance of the information. Considering the above facts, we are trying to understand the dynamics of collective interest growth about the crypto-currency topics on YouTube and Google over time.

Interested people react to the incoming information, and their reaction conveys the importance of the information. An example of an external force acting as incoming information affecting price markets happened on December 13th, 2003 [3]. That morning Bloomberg's headlines were. "US Treasuries Rise; Hussein Capture May Not Curb Terrorism". Half an hour later, Bloomberg's alert headlines changed: "US Treasuries Fall; Hussein Capture Boosts Allure of Risky Assets". The explanation that prices changed dramatically was that Saddam's capture (exogenous force) had caused the price to rise or fall. In this project, We consider the crypto-currency related videos on YouTube and Google trends as the incoming sources of information and herding of viewers to view the video as the collective interest/reaction to the information.

From the previous findings [1], we can consider these videos and trends as exogenous shocks to the complex network consisting of YouTube views and Google trends as the response to the shocks. These videos and trends contain information, which has importance and relevance only up to a certain time period from the publication of the information and becomes almost irrelevant after that. Hence, the total number of views and total watch time per day follow a power-law decay curve [1].

1.1.1 Reflexivity

George Soros [4] has first proposed reflexivity. His theory differs from the general equilibrium theory that markets reflect the economic fundamentals after reaching equilibrium. Reflexivity states that prices influence the fundamentals and change the expectation. Thus, markets move towards disequilibrium till reaching a point where the effect is reversed going to the opposite direction. In other words, financial markets can create inaccurate expectations and then change reality to accord with them. This is the opposite, which always assume that financial expectations adapt to reality, not the other way round.

An example of reflexivity is the pro-cyclical method of lending. Meaning that banks have a willingness to ease lending standards for real estate loans when prices are rising, then raising standards when real estate prices are falling, reinforcing the boom and bust cycle.The interest in reflexivity has increased following the crash of 2008, with academic journals, economists, and investors like Larry Summers, Joe Stiglitz, and Paul Volker discussing the theory [5].

From previous studies [6], we know that during the build-up phase of a bubble, there is a growing interest in the public for the commodity in question. Thus, bubbles and crashes are times where the consensus is too strong. This discovery acted as motivation to explore the dynamics of Bitcoin prices taking into account external factors such as YouTube views and Google trends.

1.2 Objectives

Find whether there is reflexivity in social data (YouTube views + Google trends) and the financial data (Bitcoin price). Moreover, our initial hypothesis is that Bitcoin price Granger causes YouTube views and Google trends. This intuition is obvious at specific time windows. During the last quarter in 2017 Bitcoin price was increasing every day. Since those prices were hitting an all-time high, the news spread quickly, and it became a common conversation among peers. After the crash in December 2017, the price dropped significantly and maintained stably for the upcoming months. The same effect occurred to the YouTube views related to Bitcoin and Google trends. We aim to verify this one directional Granger causality and further analyze the possibility of a bi-directional causality.

1.3 Outline

The rest of the project is structured as follows. The data used for the analysis is described in Chapter 2. Different types of data are briefly presented. The necessary preprocessing steps to analyze the data are described in Section 2.2.

In Section 3, we will discuss the methodology that guided us to do this research. We briefly explain the theory that acted as our basis for our analysis. First, we present the epidemic model to analyze the dynamics of viewing. Secondly, we present different techniques to check for stationarity. Thirdly, we define the rolling window volatility analysis. Finally, a VAR model is presented.

In Chapter 4, different results are presented. Useful information about the predictability of the financial time series trends might be hidden inside the amplitude and the exponents of these decay curves, as the view counts are generated by the collective human dynamics and signify the evolution of collective interest over time. The returns of the different data sets are further analyzed. We are especially interested to see the behavior of the volatility and optimal lag values over different rolling windows.

In Chapter 5, some significant conclusions drawn from this work are discussed.

Finally, some other interesting results are incorporated in the appendix of this report.

Chapter 2

Data

In this chapter, it is described the data used in this project. These data are clustered into three different categories. Section 2.1 outlines these different types of data. Furthermore, we describe the acquisition method used for each category.

2.1 Types of Data

Throughout this project, we gathered information about Bitcoin's price, YouTube statistics and Google trend searches.

2.1.1 Bitcoin Price

CoinMarketCap [7] has the Bitcoin price information starting from mid-2013 till today. This information includes dates, open, high, low and close price, as well as volume and market capitalization.

Acquisition

We exported this data to a .csv file in order to later import it to a python data-frame. Another useful source was CoinCheckup [8] for the financial time series analysis.

2.1.2 YouTube Data

YouTube videos give a useful hint to know about the popularity of Bitcoin at specific time frames. With this information, we aim to analyze the price of Bitcoin with the total daily amount of views of an extensive database of videos time windows. We gathered approximately 250000 videos related to Bitcoin.

Acquisition

The data acquisition of the YouTube statistics was done by the Chair of Entrepreneurial Risks previously. In the start of this project I received this data. The following points describe the necessary steps to acquire the YouTube statistics:

1. Crawling YouTube to get the links of the YouTube channels who publish contents related to crypto-currencies
2. Making a list of videos which are relevant for our study
3. Crawling individual videos for the information about the daily view statistics

Crawling YouTube for relevant channels A bot was written which navigates through YouTube to find out the relevant channels for our studies. The bot acts like a typical YouTube user and follows YouTube's recommendation system for finding similar channels and navigates through YouTube and collects the links for the channels which are related to crypto-currencies. This procedure acts like a Quasi-Monte Carlo sampling method. The bot is written in Python3 and uses selenium automated web-drivers for the automated browsing of Firefox browser. The links of the channels are saved in a MongoDB database.

Filtering relevant videos The auto-generated subtitle files were downloaded for all the videos from the list of channels, obtained from the previous step. We analyze the contents by processing the subtitle files using "webvtt" python library (which helps in making the subtitle files to a human-readable text transcript file) and keyword matching library "difflib" (which helps in the fuzzy matching of words). As the words in the transcript (generated by the text to speech engines) might have a different spelling than usual, we make a fuzzy match of all the words in the transcript to a crypto-vocabulary corpus that is created by us, to find out if the video/content is related to crypto-currencies. Through this process, we filter the videos and make a list of videos relevant videos from the channels. We store the links of the videos the content along with the extracted keywords, publication date, total views, total likes, total shares of the videos in the MongoDB database.

Crawling for daily view statistics YouTube provides the daily view, daily watch time, daily share and daily like statistics for each of the videos. It is a private information for the content creator and usually not shown to the public. However, this particular information is always there and hidden inside the meta data of the video web-pages. We scrape the video pages to find out the view statistics for the specific videos and store them in our database for further analysis.

2.1.3 Google Trends Data

Another interesting source to analyze is the total daily amount of Google searches related to Bitcoin together with the Bitcoin price.

Acquisition

Google trends data is available at GitHub [9]. This source uses a python script that traces back the Google searches for Bitcoin starting from mid-2010 till end-2018. Similar to the acquisition process of Bitcoin price 2.1.1, we imported the statistics to a .csv file to later incorporate it into our python data-frame.

2.2 Preprocessing of Data

In order to properly analyze the data in our model, it is necessary to prepossess the data and apply some modifications. The following points illustrate the important steps within the data preprocessing:

1. After wrapping all data statistics from the YouTube videos in a .json format, it is necessary to build a function that converts it to a python data-frame.
2. The daily amount of views on YouTube videos is hugely noisy. To decrease this effect, we apply a smoothing function. We use the Savitzky-Golay filter to smooth the data with a window length of 37 and a polynomial of order one to fit the samples.
3. Since we are interested in comparing two time series, it is crucial to evaluate each series on the same date. After receiving the different data sets described in Sections 2.1.1, 2.1.2 and 2.1.3, we realize that the dates are not identical in the whole series. To solve this issue, we sorted the arrays and appointed every value to the closest date compared to the other time series.
4. After sorting the time series, we rescaled the values of each time series so that the values are between $[0, 1]$.
5. Finally, we take the *log* values of the data that we are interested in order to have a meaningful financial analysis. Since our initial guess was, that the *log* values of price, views, and trends would not be stationary, we also take the first difference of the *log* values to achieve stationarity. It is also more interesting to evaluate the returns of the data.

Chapter 3

Methodology

Several factors lead to viewing a video on YouTube or conduction a Google search related to Bitcoin. Those causes include triggering from email, linking from external websites, discussion on blogs, newspapers, and television, from social influences, through YouTube's and Google's intrinsic suggestion mechanism or sometimes chance. As found out in previous research studies [1], the time evolution of the number of views or the number of Google searches per day is mostly dependent on collective interest of the viewers/searches and the quality of the topic discussed.

3.1 Power-law

In this section, we present the epidemic model that we apply to analyze the dynamics of viewing behavior. In order to be consistent with the model described in [1], we reduce the explanation only to YouTube videos. However, a similar analysis can be conducted to describe the dynamics of searching behavior in Google.

The first ingredient of this epidemic model is a power law distribution of waiting times describing the human activity that expresses the potential impact of these various factors by using a response function, which, from previous work, we take to be a long-memory process of the form.

$$\phi(t) \approx \frac{1}{t^{1+\theta}}; \text{with } 0 < \theta < 1 \tag{3.1}$$

By definition, the memory kernel $\phi(t)$ describes the distribution of waiting times between “cause" and “action" for an individual.

The cause can be any of the factors mentioned in Chapter 3. The action is for the individual to view the video in question after a time t since he was first subjected to the cause without any other influences between 0 and t, corresponding to a direct (or first-generation) effect.

In other words, $\phi(t)$ is the “bare" memory kernel or propagator, describing the direct influence of a factor that triggers the individual to view the video in question. Here, the exponent θ is the crucial parameter of the theory that is determined empirically from the data.

The second ingredient is an epidemic branching process that describes the cascade of influences on the social network. This process captures how previous attention from one individual can spread to others and become the cause that triggers their future attention. In a highly connected network of individuals whose interests make them susceptible to the given video content, a given factor may trigger action through a cascade of intermediate steps.

Such an epidemic process can be conveniently modeled by the so-called self-excited Hawkes conditional Poisson process.

This gives the instantaneous rate of views $\lambda(t)$ as

$$\lambda(t) = V(t) + \sum_{i,t_i<t} \mu_i \phi(t - t_i) \tag{3.2}$$

where μ_i is the number of potential viewers, who will be influenced directly over all future times after t_i by person i who viewed a video at time t_i. Thus, the existence of well-connected individuals can be accounted for with large values of μ_i. Lastly, V(t) is the exogenous source, which captures all spontaneous views that are not triggered by epidemic effects on the network.

3.2 Time Series Analysis

In this Section, we will further explore the relationships between Bitcoin price with YouTube views and Bitcoin price with Google searches. Conducting a time series analysis for both cases is convenient.

As a start, we use min-max normalization for re-scaling the total views/day, searches/day and price time series. The reason for this is to put all time series into one standard scale. Equation 3.3 represents the min-max re-scaling. Thus, we constrain the values of the time series between [0,1]. The rest of the analysis from this point is conducted on the re-scaled time series. Nevertheless, a more stable approach would be to standardize the time series by removing the mean and dividing by the standard deviation. In future projects, the latter approach will be used to prevent biases from the large fluctuations of the min and max.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.3}$$

It is necessary that the three time-series fulfill the stationarity conditions. The results of the three checkups are presented in Subsection 3.2.1.

The intensity of demand and supply sets the Bitcoin price. Since the amount of Bitcoins is finite (21 million), demand plays a major role in setting the price. Buyers believe that in the future the demand for Bitcoins will increase. Therefore, speculators buy Bitcoins in anticipation of that future demand that would drive the price up. However, speculative buyers tend to be sensitive to current events and news. As a result, speculative buyers can quickly turn into speculative sellers. This change of speculator buyers into sellers is one main reason why the price of bitcoin varies so wildly making it so volatile. Since we are interested in the evolution of Bitcoin price and its fluctuations, it comes in handy to do a volatility analysis over different rolling windows. The volatility model is presented in Subsection 3.2.2.

In Subsection 3.2.3 we present the VAR model that we use to find the optimal lag values for each time series. Likewise the volatility analysis, we apply several rolling windows to the VAR model.

In the last Subsection 3.2.4 of this Chapter, we introduce the Granger-Causality concept to test whether one time series Granger causes the other time series.

3.2.1 Stationarity

In order to check for stationarity, we use three different methods.

Correlogram

The first method we use to test stationarity is a correlogram plot. The plots shown in Section 4.3.1 give a visual hint to see whether the time series are stationary.

Dickey-Fuller test

The second method used is the Dickey-Fuller test. There we test the null hypothesis H_0 that a unit root is present in an autoregressive model, which would violate the stationarity condition. The alternative hypothesis H_A holds that the time series is stationarity. If the p-value is above a critical size, then we cannot reject that there is a unit root.

We define a simple autoregressive model of order 1 AR(1) as follows:

$$y_t = cy_{t-1} + \epsilon_t \tag{3.4}$$

where y_t is the variable of interest, t is the time index, c is a coefficient, and ϵ_t is the error term. Moreover, a unit root is present if $c = 1$ for the AR(1).

For the statistic test we take the first difference. Equation 3.5 shows the first difference of the autoregressive model shown in Equation 3.4.

$$\Delta y_t = (c-1)y_{t-1} + \epsilon_t = \delta y_{t-1} + \epsilon_t \tag{3.5}$$

where Δ denotes the first difference operator. Moreover, for H_0 the unit root is equivalent to testing $\delta = 0$. In other words, $\delta \equiv c - 1$.

Tables including the results of the Dickey-Fuller test are presented in Section 4.3.2.

Johansen test

In the time series of interest, A is the coefficient matrices for each lag. The test checks for the situation of no cointegration, which occurs when $A = 0$. The rank of the matrix A is denoted by r. The Johansen test tests whether r is equal to zero or equal to one. The H_0 of $r = 0$ means that there is no cointegration. Rank $r > 0$ implies a cointegrating relationship between the time series.

While testing for for cointegration for the time series $log(price)$, $log(views)$ and $log(trends)$, we can conclude whether it is possible that the time series are stationary.

Tables showing the results of the Johansen test are presented in Section 4.3.3.

3.2.2 Volatility Analysis of Returns

Measuring volatility is a very common method to see the degree of variation of the different time series. Since we are interested to analyze different time slots and especially bubble periods, we create rolling window with different lengths and compute the standard deviation of each rolling window. By computing the standard deviation, we know directly the volatility over those rolling windows. These rolling windows go through the whole time series in weekly (7 days) and monthly intervals (30 days).

In Section 4.4 the results of these volatility rolling windows are shown.

3.2.3 VAR Analysis of Returns

In order to find the linear interdependencies between the two time series, we perform Vector autoregression with p lags (VAR(p)) on the stationary time series. The VAR(p) model describes the evolution of a set of variables over a time period as a linear function of their past values. The variables are stored in vector y_t, where $y_{i,t}$ denotes the i-th element of vector y_t at observation time t. In other words, it is the observation at time t of the i-th variable. Equation 3.6 shows the VAR(p) model.

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t \tag{3.6}$$

where c denotes a vector of constants, A_i is a time-invariant coefficient matrix of the i-th lag of y, and e_t the error term.

In order to determine the optimum lag value (p) for the VAR analysis, we use Akaike information criterion (AIC) and Bayesian information criterion (BIC) in VAR model.

Equations 3.7 and 3.8 define the two model selection criteria.

$$AIC = 2k - 2\ln(\hat{L}) \tag{3.7}$$

$$BIC = \ln(n)k - 2\ln(\hat{L}) \tag{3.8}$$

where k is the number of estimated parameters in the model, n the number of observations and $\hat{L}$ is the maximum value of the likelihood function for the model. For more details see [10].

The results of this analysis are depicted in in Section 4.5 for the VAR analysis with rolling windows and Section 4.6 for the overall VAR analysis without rolling windows.

3.2.4 Granger Causality

In order to find the causal dependencies between the two time series, we are calculating the Granger causality. Granger causality means that past values of a second time series have a statistically significant effect on the current value of the first time series, taking past values of the first time series into account as regressors.

The null hypothesis of the test is that the second time series does not Granger cause the first time series. We reject the null hypothesis that the second time series does not Granger cause the first time series if the p-values are below a desired size of the test. The null hypothesis for all four tests is that the coefficients corresponding to past values of the second time series are zero [11].

Overall we perform several tests. In the first one, we check whether the price is Granger causing YouTube views as well as the opposite. Moreover, for the second test, we check whether the price is Granger causing Google searches. We perform these tests over different rolling windows similar to the volatility and VAR analysis.

The results of these tests are presented in Subsection 4.7.

Chapter 4

Results

4.1 Behaviour of Data

Figure 4.1 depicts the absolute values of Bitcoin price, the total amount of daily YouTube videos and Google searches related to Bitcoin from 2012 till the present time. It is clear from Figure 4.1 that the interest of people searching for YouTube videos and in Google increased especially during the bubble period in late 2017. Figure 4.2 represents the scaled *log* values. From this plot, we can recognize that there was a bubble period in 2012 that finished with a crash in August 18th [12]. Figure 4.3 illustrates the first order difference of the *log* scaled values. From these plots, we can see that all plots have similar trajectories.

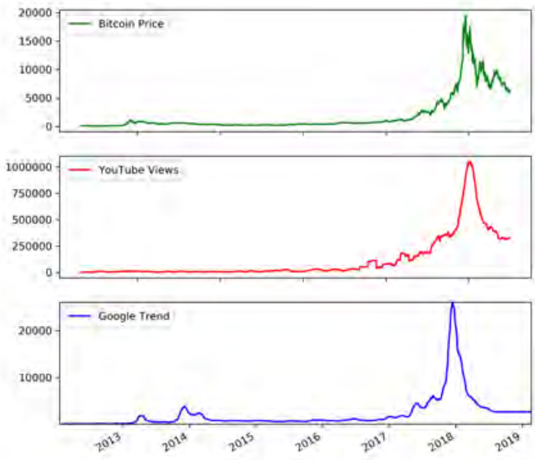


Figure 4.1: Top plot: Bitcoin price, middle plot: total daily YouTube views related to Bitcoin, bottom plot: total daily Google trends related to Bitcoin.

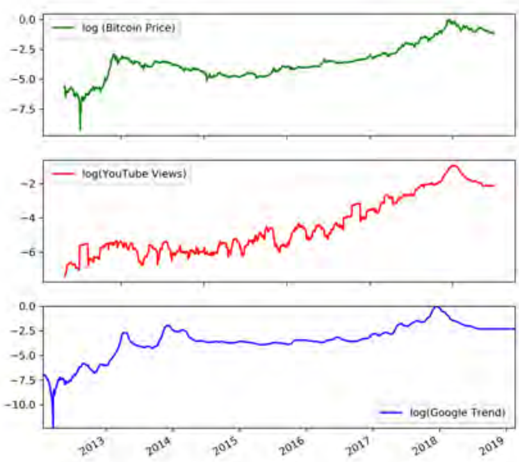


Figure 4.2: Scaled log(values). Top plot: Bitcoin (price), middle plot: total daily YouTube views, bottom plot: total daily Google trends.

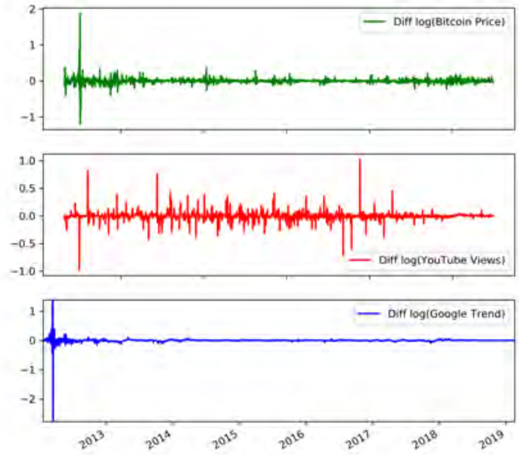


Figure 4.3: First order difference scaled log(values). Top plot: Bitcoin (price), middle plot: total daily YouTube views, bottom plot: total daily Google trends.

4.2 Fitting the Power-law

For each publication day t_c, we have a time series. For each time series, at day $t \geq t_c$ we aggregate daily views of YouTube videos related to Bitcoin. As a result, videos that were published on an arbitrary day in 2011 have more data points than videos that were published in 2018. For each time series, we fit the exponential decay curve with equation 4.1 using Least-Square fit to find the regression value of θ_t and A_t (amplitude). The coefficient c is arbitrary chosen to be 0.8.

$$f(t) = \frac{A_t}{(t-c)^{\theta_t}} \tag{4.1}$$

Figure 4.4 depicts the values of the θ_t exponents after fitting Equation 4.1 for each time series. The range of values of θ_t is between 0.4 and 1. These values are compatible with the predictions of the epidemic model previously done in [1].

Figure 4.5 depicts the value of the amplitude A_t and $log(A_t)$ after fitting Equation 4.1 for each time series. We observe that the amplitude has a similar trajectory as the total aggregated daily views. As time progressed and we reach closer to the bubbles, we get the super exponential growth of the amplitude values, and after the burst the amplitude decays accordingly.

Figure 4.6 represents an example of the continuous decay with the power-law fit for videos published on May 25th, 2013. This and the majority of the trajectories for the other time series fit well with the power-law function.

Figure 4.7 represents an example for videos published on May 27th, 2013. It depicts a burst of activity at the end shifting upward the power-law fit towards the tail, biasing it. In this case, the fit is adequate to represent the data up to t 10^2 and then the big dragon-king structure is present [13]. Only a minority of the of the time series display this behaviour, and for the time being we are considering them within power-law decay domain as the overall behavior is following a power-law decay pattern. However, more pre-processing would be done in future to get an estimate of the exact exponent and the amplitude value.

Taking a closer look at some bubble periods we can observe that the YouTube views and Bitcoin price behave very similar during bubble periods. Figure 4.8 depicts this behavior. That discovery acted as a major motivation to take a closer look to analyze the volatility and find the lag values over different rolling windows.

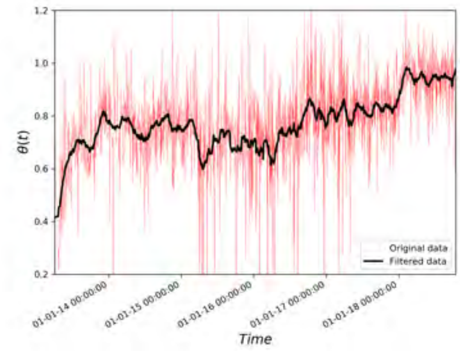


Figure 4.4: Time evolution of the power-law decay exponents obtained through the fitting procedure. The red line represents the original data and the black line represents the filtered data after Savitzky-Golay Smoothing with window length 37, and order of polynomial 1.

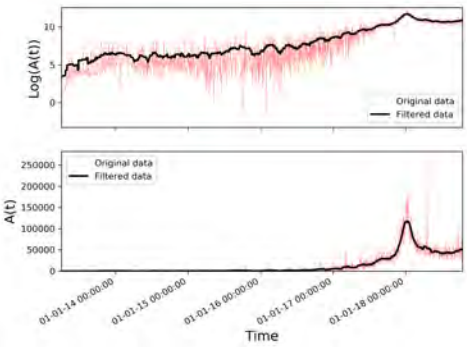


Figure 4.5: Time evolution of the power-law decay amplitude obtained through the fitting procedure. The red line represents the original data and the black line represents the filtered data after Savitzky-Golay Smoothing with window length 37, and order of polynomial 1. Top panel is log of amplitude and the bottom panel is amplitude over time.

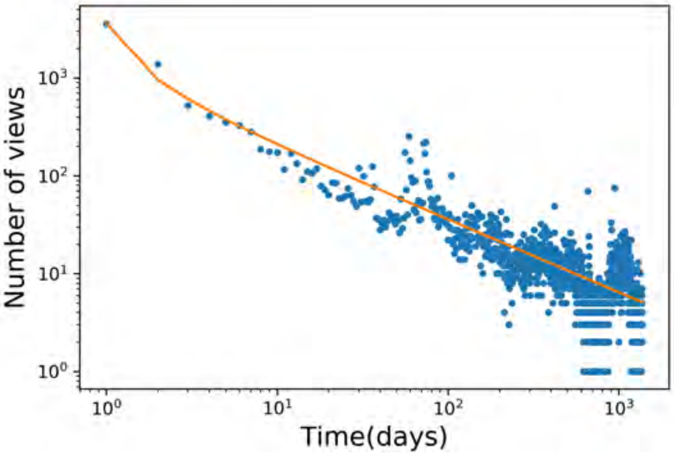


Figure 4.6: Least square fit of the total aggregated daily views of YouTube videos published on May 25th, 2013. The blue dots represents the total number of daily views. The orange curve depicts the fitted values of the regression of function 4.1. The figure shows a continuous decay. Range of x-axis is 1000 days.

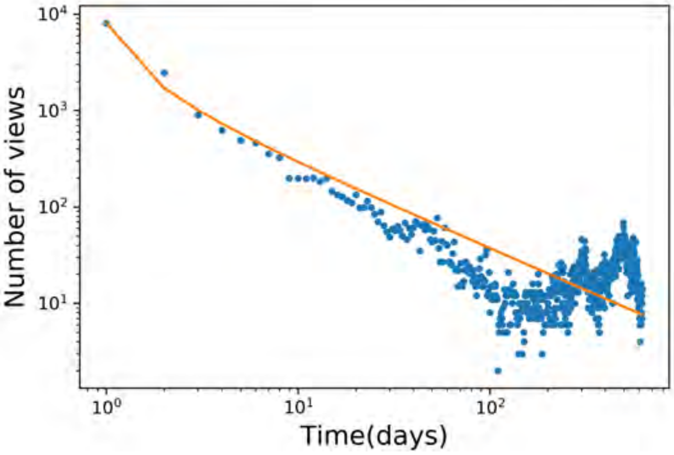


Figure 4.7: Least square fit of the total aggregated daily views of YouTube videos published on May 27th, 2013. The blue dots represents the total number of daily views. The orange curve depicts the fitted values of the regression of function 4.1. The figure shows the burst of activity towards the tail. Range of x-axis is 1000 days.

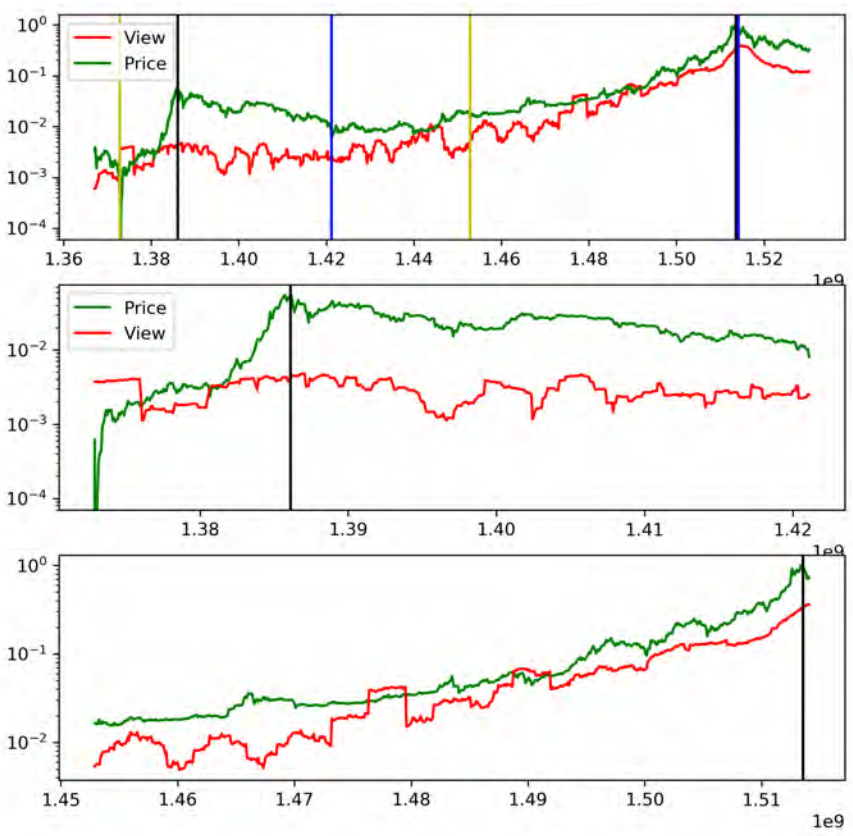


Figure 4.8: Figure representing the full time series data, along with two long bubbles [14]. First subplot: Total number of YouTube views per day (in red) and Bitcoin price (in green)(full length data). (yellow vertical line: starting point of bubble, black vertical line: bubble peak, blue vertical line: bubble crash). Second subplot: representing the first bubble (03-07-2013 to 14-01-2015). Third subplot: representing the second long bubble (15-01-2016 to 25-12-2017)

4.3 Stationarity

In order to check the stationarity of the time series, we checked the correlogram (autocorrelation values of the time series) and Dickie-Fuller test.

4.3.1 Correlogram

A correlogram is an autocorrelation plot. Its purpose is to show autocorrelations versus time lags. From Figure 4.9 we observe long-range dependencies in the first three autocorrelation subplots. These long dependencies suggest non-stationarity of the time series. Delta autocorrelation behavior in bottom autocorrelation plots indicates the stationarity.

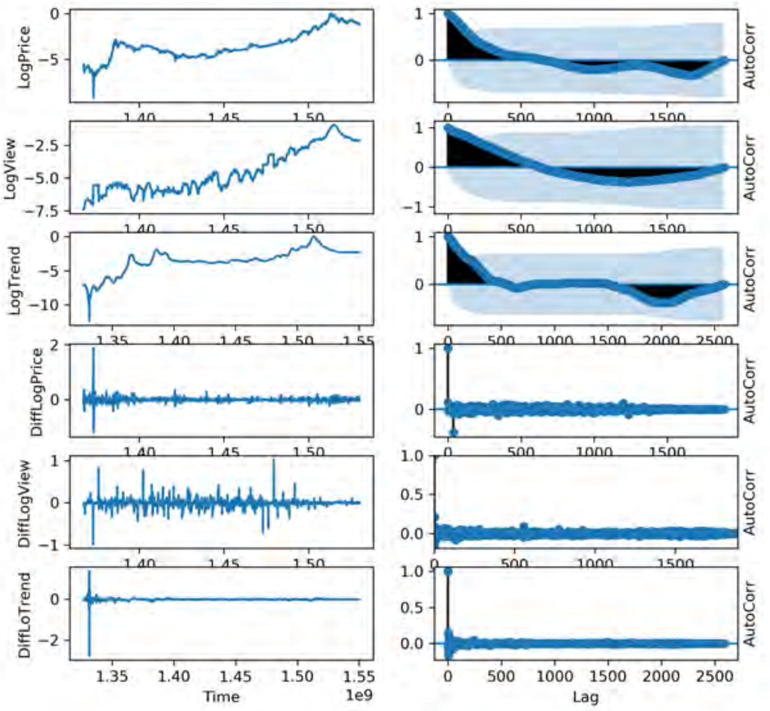


Figure 4.9: Correlogram plot for the time series, Left panel: From top to bottom: Log price over time, Log daily views over time, Log daily trends over time, $\Delta Log(Price)$, $\Delta Log(Views)$,$\Delta Log(Trends)$. Time in epoch format in x-axis. Right panel: Corresponding correlograms. Auto correlation values in y-axis and lag values in x-axis.

4.3.2 Dickey–Fuller test

We performed the Augmented Dickey-Fuller test on the $Log(Timeseries)$.

Tables 4.1, 4.2 and 4.3 show that the time series are not stationary. Therefore, we compute the first order difference of the time series and do the test once again. Tables 4.4, 4.5 and 4.6 show that $\Delta Log(Timeseries)$ are stationary.

Test statistic	P value	Test Critical Value	Critical Value test statistic
-1.057	0.732		
		1%	-3.433
		5%	-2.863
		10%	-2.567

Table 4.1: Augmented Dickey-Fuller test for Log(Price) time series. t-statistics is greater than the critical values. Thus, time series is non-stationary.

Test statistic	P value	Test Critical Value	Critical Value test statistic
-2.254	0.187		
		1%	-3.433
		5%	-2.863
		10%	-2.567

Table 4.2: Augmented Dickey-Fuller test statistics for Log(Views). p-value is significant, hence, the time series is non stationary.

Test statistic	P value	Test Critical Value	Critical Value test statistic
-2.517	0.111		
		1%	-3.433
		5%	-2.863
		10%	-2.567

Table 4.3: Augmented Dickey-Fuller test statistics for Log(Trends). p-value is significant, hence, the time series is non stationary.

Test statistic	P value	Test Critical Value	Critical Value test statistic
-12.793	7.028		
		1%	-3.433
		5%	-2.863
		10%	-2.567

Table 4.4: Augmented Dickey-Fuller test statistics for $\Delta Log(Price)$. t-statistics is much less than the critical values, hence the time series is stationary.

Test statistic	P value	Test Critical Value	Critical Value test statistic
-29.186	0.0		
		1%	-3.433
		5%	-2.863
		10%	-2.567

Table 4.5: Augmented Dickey-Fuller test statistics for $\Delta Log(View)$. t-statistics is much less than the critical values, hence the time series is stationary.

Test statistic	P value	Test Critical Value	Critical Value test statistic
-8.461	1.568		
		1%	-3.433
		5%	-2.863
		10%	-2.567

Table 4.6: Augmented Dickey-Fuller test statistics for $\Delta Log(Trend)$. t-statistics is much less than the critical values, hence the time series is stationary.

4.3.3 Johansen Test

For the Johansen test, we use the trace statistics method and 2 as the number of lagged differences in the model.

The first null hypothesis, $r = 0$, tests for the presence of cointegration. Table 4.7 depicts the result of the Johansen test for the *log* values of the YouTube views and the Bitcoin price. From Table 4.7 we see that the test statistic exceeds the 1% level significantly ($105.65 > 23.52$). Thus, we have strong evidence to reject the null hypothesis of no cointegration versus the $r > 0$ alternative hypothesis.

In the case of Google trends, Table 4.8 depicts the result of the Johansen test for the *log* values. The tests statistic exceeds as well the 1% level significantly ($32.01 > 19.93$). Thus, we have strong evidence to reject the null hypothesis of no cointegration versus the $r > 0$ alternative hypothesis.

Next, when we carry out the $r \leq 1$ null hypothesis versus the $r > 1$ alternative hypothesis, we conclude from 4.7 that we can not reject that null hypothesis 4.02 <= 11.65. This means is that it may be possible to form a linear combination to create a stationary time series for the YouTube data.

The same effect happens to the Google trends. From Table 4.8, we conclude that we can not reject that null hypothesis, since 4.48 <= 6.663. This means is that it may be possible to form a linear combination to create a stationary time series for the Google trends data as well.

Null hypothesis	test statistics	10%	5%	1%
$r = 0$	105.65	15.66	17.95	23.52
$r \leq 1$	4.02	6.50	8.18	11.65

Table 4.7: Johansen cointegration test statistics for Log(Price) and Log(Views) time series. Test type: trace statistic, with linear trend. Results show that it is possible to form a linear combination to create a stationary time series.

Null hypothesis	test statistics	10%	5%	1%
$r = 0$	32.01	13.43	15.49	19.93
$r \leq 1$	4.48	2.71	3.84	6.63

Table 4.8: Johansen cointegration test statistics for Log(Price) and Log(Trends) time series. Test type: trace statistic, with linear trend. Results show that it is possible to form a linear combination to create a stationary time series.

4.4 Rolling Window Volatility

In this Section, we conduct a volatility analysis of our YouTube and Google data with weekly and monthly rolling windows. Figure 4.10 corresponds to the YouTube views and Figure 4.11 to the Google trends. From both plots we observe that weekly and monthly volatility behave very similar for the YouTube views and the Bitcoin price. Furthermore, the highest volatility appears in the 3rd quarter of 2012 precisely at the bubble period of summer 2012. Other high volatility changes happened at the end of 2017 when the last bubble was present.

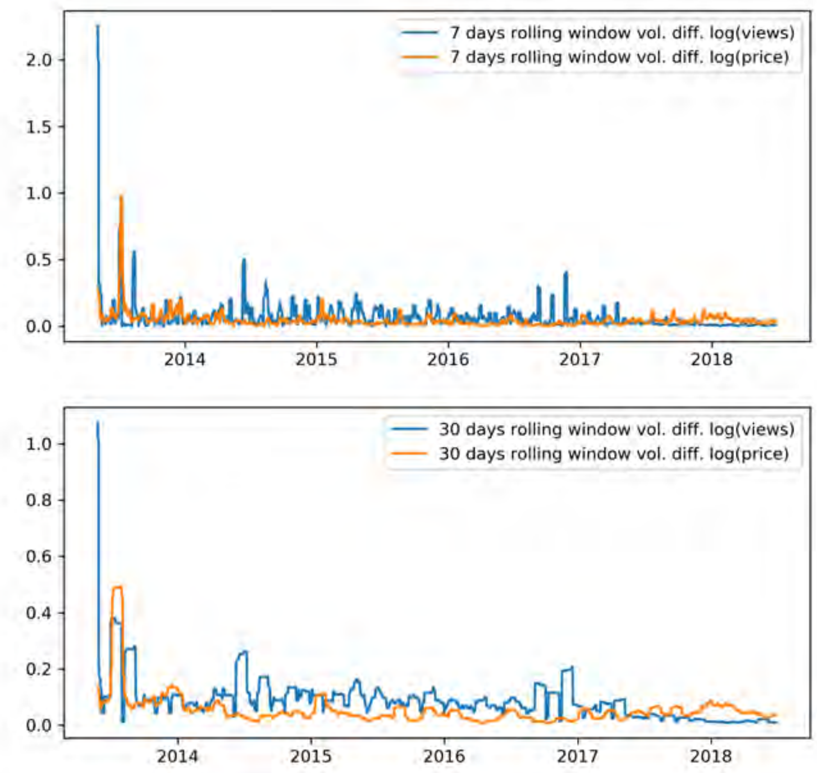


Figure 4.10: Rolling window volatility for YouTube views. Top subplot: 7 days rolling window. Bottom: 30 days rolling window.

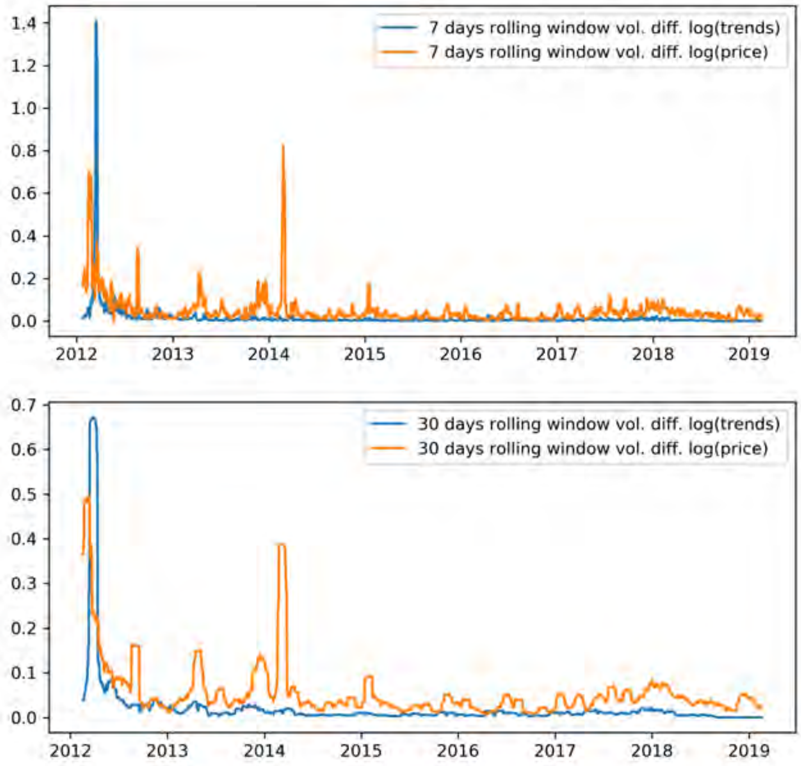


Figure 4.11: Rolling window volatility for Google trends. Top subplot: 7 days rolling window. Bottom: 30 days rolling window.

4.5 Rolling Window VAR Analysis

In this Section, we conduct a VAR analysis of the YouTube and Google data with rolling windows of 350 samples. First, we calculate the total amount of lag values (lag length) using AIC criterion which is affecting the other times series at time $t = 0$. Figure 4.12 depicts the lag length for YouTube views. Figure 4.14 depicts the lag length for Google trends. During the bubble periods, the lag length tends to be minimized towards 0 for the YouTube views and towards 1 for the Google trends. Second, we calculate the optimal lag value of each time series by taking the minimum p-value over the rolling windows. The smallest p-values tells us which are the most statistical significant coefficients. Figure 4.13 depicts the optimal lag values for YouTube views and Bitcoin price. Figure 4.15 depicts the optimal lag values for Google trends and Bitcoin price.

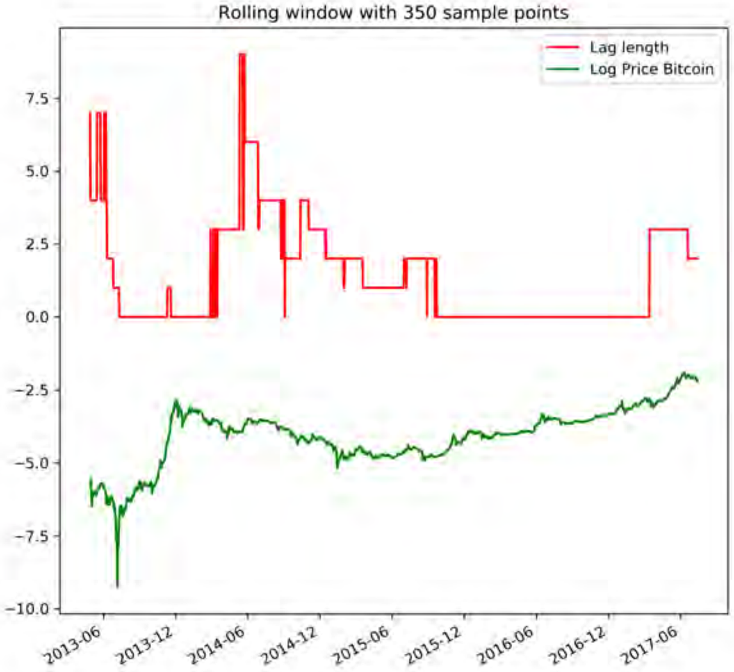


Figure 4.12: Rolling window VAR analysis for YouTube views. Each window has 350 samples (dates). Red line depicts the lag length at every window. Green curve depicts the scaled $log(price)$ of Bitcoin. The larger the lag length, the stronger the correlation with the price, since it has more positive feedback. During the bubble periods the lag length tends to go towards 0.

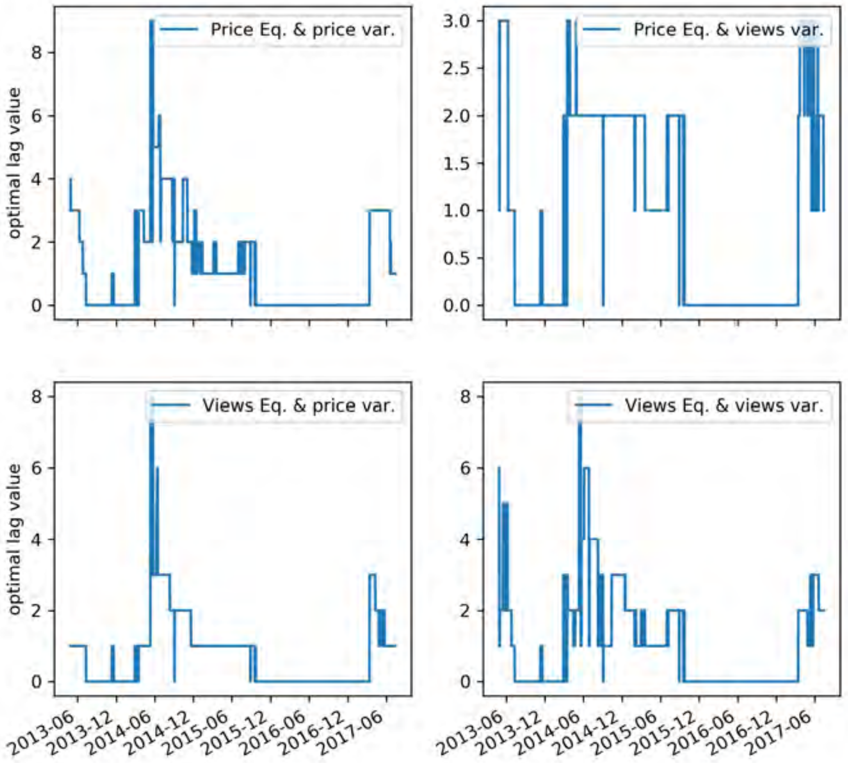


Figure 4.13: Optimal lag value of rolling window VAR analysis for YouTube views and Bitcoin price. Optimal lag for one variable corresponding to one equation in y-axis for each subplot. Starting date of rolling window in x-axis. Each window has 350 samples (dates). Top left subplot depicts optimal lag for $\Delta Log(Price)$ equation with $\Delta Log(Price)$ variables. Top right subplot depicts optimal lag for $\Delta Log(Price)$ equation with $\Delta Log(Views)$ variables. Bottom left subplot depicts optimal lag for $\Delta Log(Views)$ equation with $\Delta Log(Price)$ variables. Bottom right subplot depicts optimal lag for $\Delta Log(Views)$ equation with $\Delta Log(Views)$ variables.

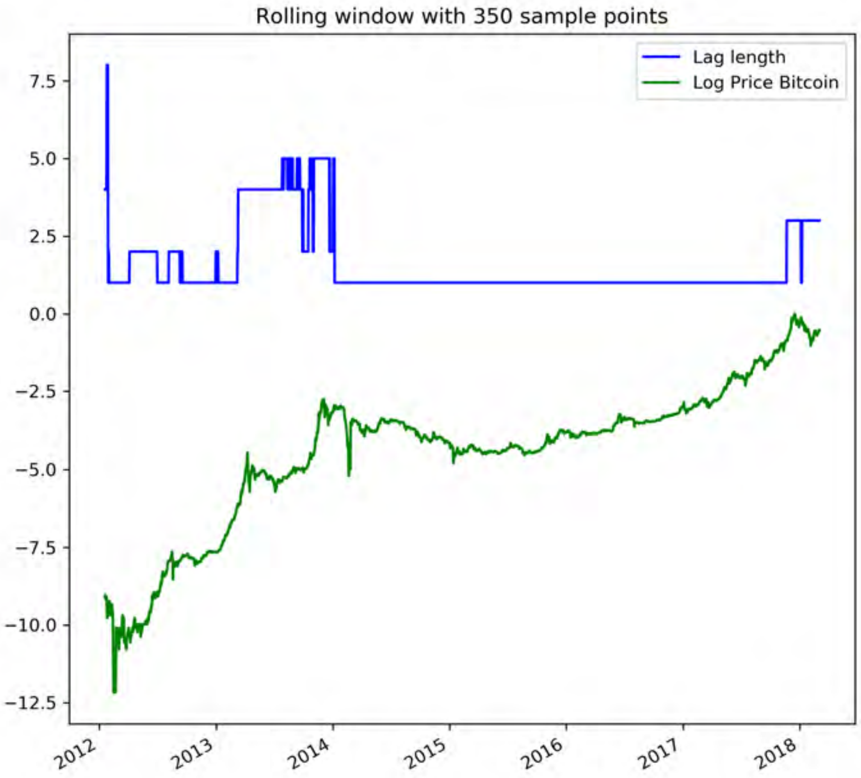


Figure 4.14: Rolling window VAR analysis for Google. Each window has 350 samples (dates). Red line depicts the lag length at every window. Green curve depicts the scaled $log(price)$ of Bitcoin. The larger the lag length, the stronger the correlation with the price, since it has more positive feedback. During the bubble periods the lag length tends to go towards 1.

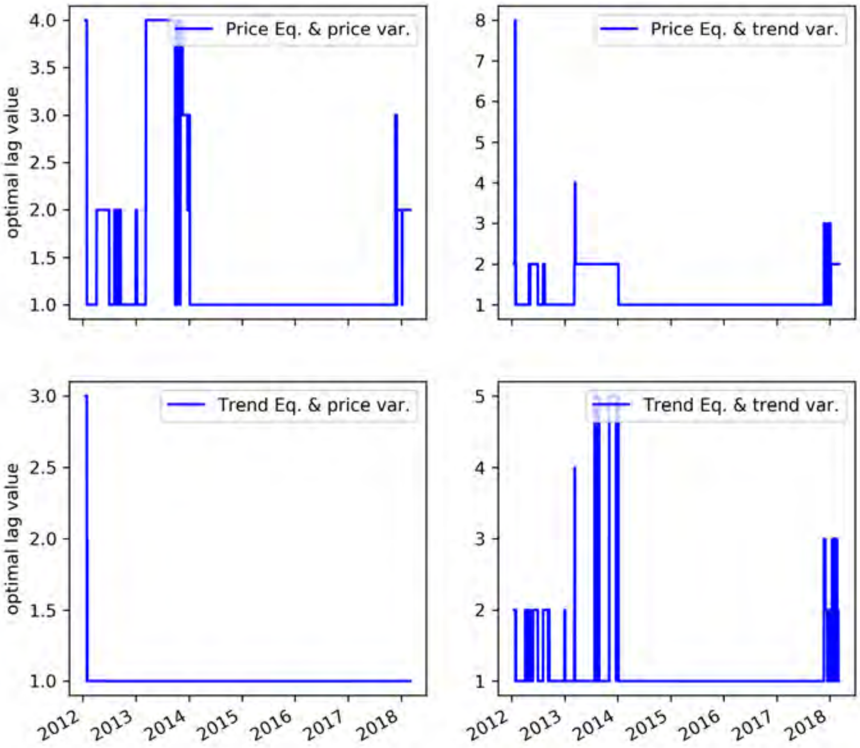
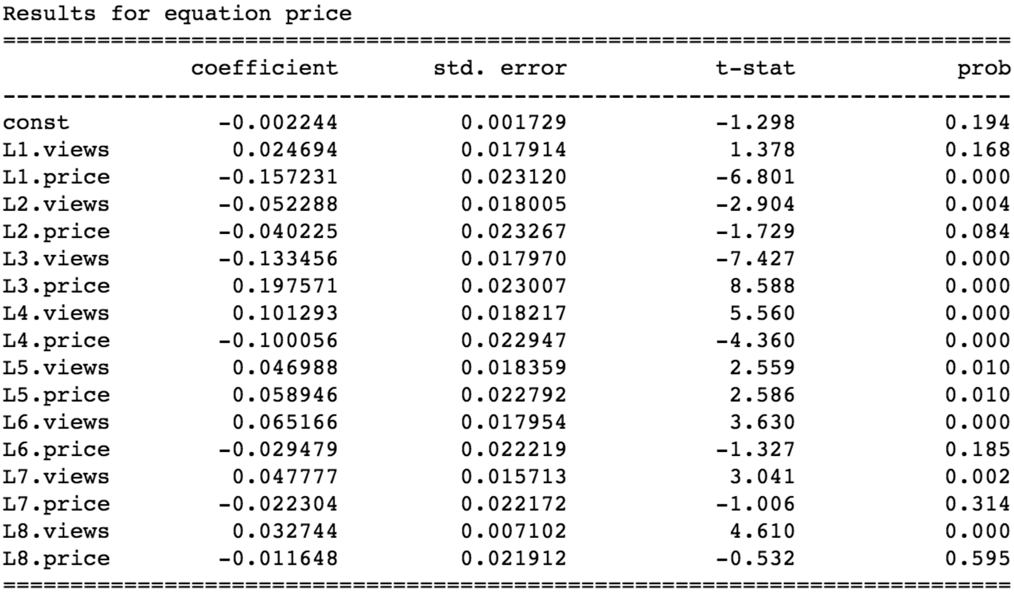


Figure 4.15: Optimal lag value of rolling window VAR analysis for Google trends and Bitcoin price. Optimal lag for one variable corresponding to one equation in y-axis. Starting date of rolling window in x-axis. Each window has 350 samples (dates). Top left subplot depicts optimal lag for $\Delta Log(Price)$ equation with $\Delta Log(Price)$ variables. Top right subplot depicts optimal lag for $\Delta Log(Price)$ equation with $\Delta Log(Trends)$ variables. Bottom left subplot depicts optimal lag for $\Delta Log(Trends)$ equation with $\Delta Log(Price)$ variables. Bottom right subplot depicts optimal lag for $\Delta Log(Trends)$ equation with $\Delta Log(Trends)$ variables.

4.6 Overall VAR

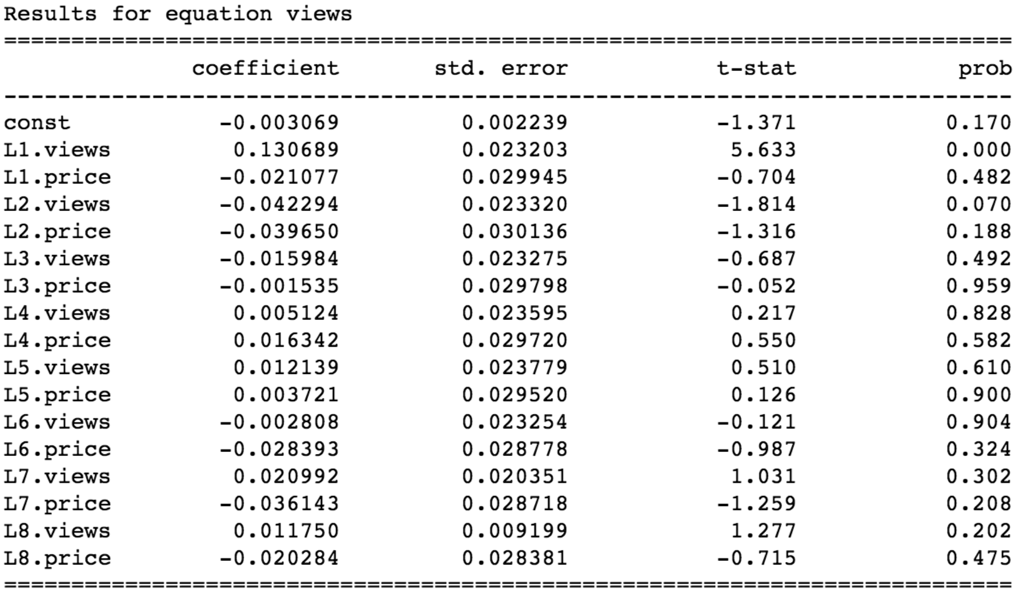
In this Section, we conducted a VAR analysis for the YouTube views data with the Bitcoin price data. Another VAR analysis was conducted for the Google trends data with the Bitcoin price data. The AIC criterion was used in the VAR analysis to determine the lag length. We determined the statistically significant coefficients by looking at the smallest p-values.

The first analysis was conducted for the time series with $\Delta Log(Price)$ and $\Delta Log(Views)$ data. In this analysis, the largest coefficients for both equations (price and views) are located within the first three lags and are negative. The second analysis was conducted for the time series with $\Delta Log(Price)$ and $\Delta Log(Trends)$ data. In this analysis, the largest coefficients for both equations (price and trends) are located within the first four lags and are negative again. The following Tables show the results obtained in the VAR analysis.



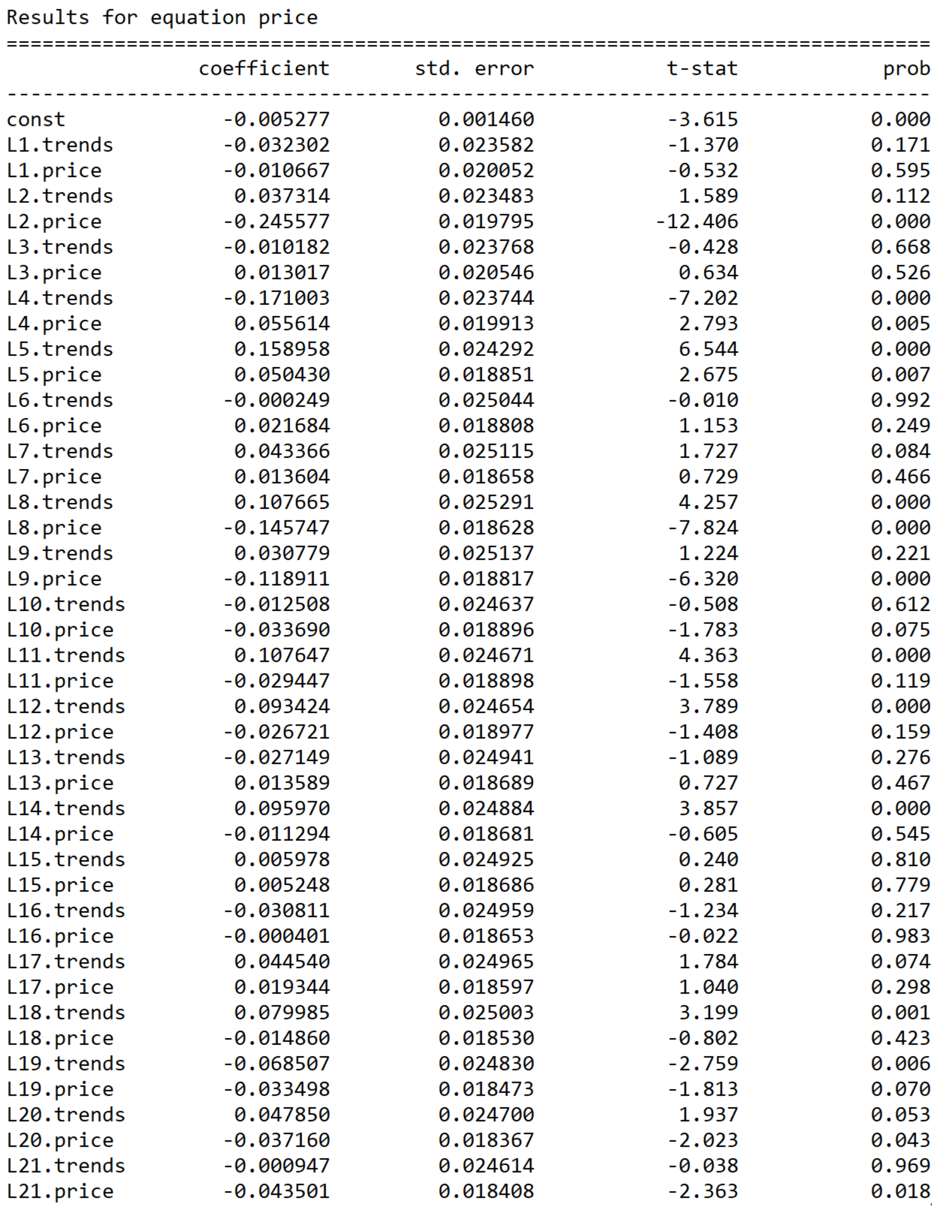
```
Results for equation price
===========================================================================
               coefficient       std. error           t-stat            prob
---------------------------------------------------------------------------
const            -0.002244         0.001729           -1.298           0.194
L1.views          0.024694         0.017914            1.378           0.168
L1.price         -0.157231         0.023120           -6.801           0.000
L2.views         -0.052288         0.018005           -2.904           0.004
L2.price         -0.040225         0.023267           -1.729           0.084
L3.views         -0.133456         0.017970           -7.427           0.000
L3.price          0.197571         0.023007            8.588           0.000
L4.views          0.101293         0.018217            5.560           0.000
L4.price         -0.100056         0.022947           -4.360           0.000
L5.views          0.046988         0.018359            2.559           0.010
L5.price          0.058946         0.022792            2.586           0.010
L6.views          0.065166         0.017954            3.630           0.000
L6.price         -0.029479         0.022219           -1.327           0.185
L7.views          0.047777         0.015713            3.041           0.002
L7.price         -0.022304         0.022172           -1.006           0.314
L8.views          0.032744         0.007102            4.610           0.000
L8.price         -0.011648         0.021912           -0.532           0.595
===========================================================================
```

Figure 4.16: VAR results for $\Delta Log(Price)$ equation with $\Delta Log(Price)$ and YouTube $\Delta Log(Views)$ variables. The first column corresponds to the different lag values. The second column represents the coefficients of the lagged values. The last column corresponds to the p-values. The majority of the coefficients are significant. Coefficients with the largest absolute value are the 1st for the $\Delta Log(Price)$ variable and the 3rd for the $\Delta Log(Views)$ variable.



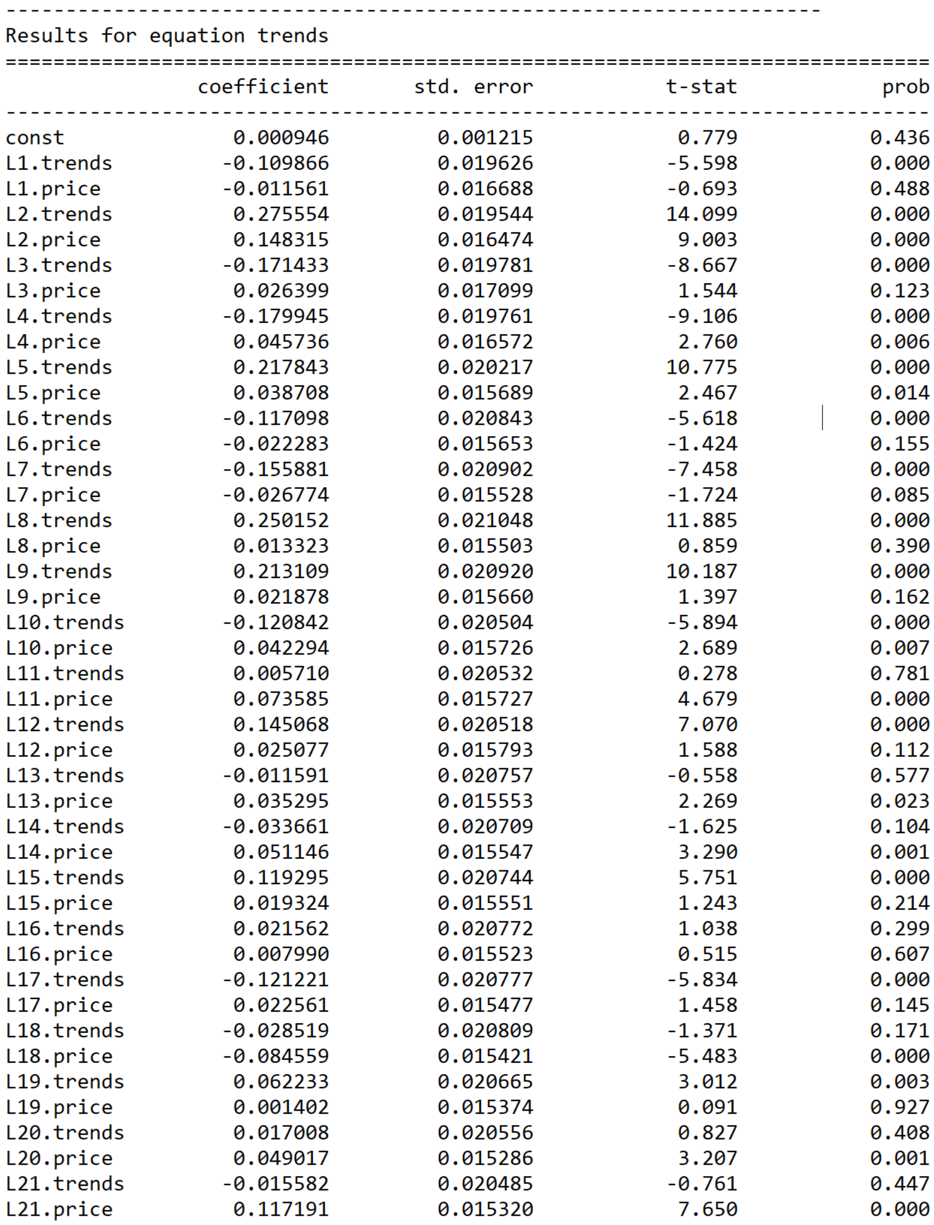
```
Results for equation views
==============================================================================
              coefficient       std. error           t-stat              prob
------------------------------------------------------------------------------
const           -0.003069         0.002239           -1.371             0.170
L1.views         0.130689         0.023203            5.633             0.000
L1.price        -0.021077         0.029945           -0.704             0.482
L2.views        -0.042294         0.023320           -1.814             0.070
L2.price        -0.039650         0.030136           -1.316             0.188
L3.views        -0.015984         0.023275           -0.687             0.492
L3.price        -0.001535         0.029798           -0.052             0.959
L4.views         0.005124         0.023595            0.217             0.828
L4.price         0.016342         0.029720            0.550             0.582
L5.views         0.012139         0.023779            0.510             0.610
L5.price         0.003721         0.029520            0.126             0.900
L6.views        -0.002808         0.023254           -0.121             0.904
L6.price        -0.028393         0.028778           -0.987             0.324
L7.views         0.020992         0.020351            1.031             0.302
L7.price        -0.036143         0.028718           -1.259             0.208
L8.views         0.011750         0.009199            1.277             0.202
L8.price        -0.020284         0.028381           -0.715             0.475
==============================================================================
```

Figure 4.17: VAR results for $\Delta Log(Views)$ equation with $\Delta Log(Price)$ and YouTube $\Delta Log(Views)$ variables. The first column corresponds to the different lag values. The second column represents the coefficients of the lagged values. The last column corresponds to the p-values. Only two coefficients are significant, the 2nd lag for the $\Delta Log(Price)$ variable and the 1st lag for $\Delta Log(Views)$ variable.



```
Results for equation price
==============================================================================
                coefficient       std. error           t-stat            prob
------------------------------------------------------------------------------
const             -0.005277         0.001460           -3.615           0.000
L1.trends         -0.032302         0.023582           -1.370           0.171
L1.price          -0.010667         0.020052           -0.532           0.595
L2.trends          0.037314         0.023483            1.589           0.112
L2.price          -0.245577         0.019795          -12.406           0.000
L3.trends         -0.010182         0.023768           -0.428           0.668
L3.price           0.013017         0.020546            0.634           0.526
L4.trends         -0.171003         0.023744           -7.202           0.000
L4.price           0.055614         0.019913            2.793           0.005
L5.trends          0.158958         0.024292            6.544           0.000
L5.price           0.050430         0.018851            2.675           0.007
L6.trends         -0.000249         0.025044           -0.010           0.992
L6.price           0.021684         0.018808            1.153           0.249
L7.trends          0.043366         0.025115            1.727           0.084
L7.price           0.013604         0.018658            0.729           0.466
L8.trends          0.107665         0.025291            4.257           0.000
L8.price          -0.145747         0.018628           -7.824           0.000
L9.trends          0.030779         0.025137            1.224           0.221
L9.price          -0.118911         0.018817           -6.320           0.000
L10.trends        -0.012508         0.024637           -0.508           0.612
L10.price         -0.033690         0.018896           -1.783           0.075
L11.trends         0.107647         0.024671            4.363           0.000
L11.price         -0.029447         0.018898           -1.558           0.119
L12.trends         0.093424         0.024654            3.789           0.000
L12.price         -0.026721         0.018977           -1.408           0.159
L13.trends        -0.027149         0.024941           -1.089           0.276
L13.price          0.013589         0.018689            0.727           0.467
L14.trends         0.095970         0.024884            3.857           0.000
L14.price         -0.011294         0.018681           -0.605           0.545
L15.trends         0.005978         0.024925            0.240           0.810
L15.price          0.005248         0.018686            0.281           0.779
L16.trends        -0.030811         0.024959           -1.234           0.217
L16.price         -0.000401         0.018653           -0.022           0.983
L17.trends         0.044540         0.024965            1.784           0.074
L17.price          0.019344         0.018597            1.040           0.298
L18.trends         0.079985         0.025003            3.199           0.001
L18.price         -0.014860         0.018530           -0.802           0.423
L19.trends        -0.068507         0.024830           -2.759           0.006
L19.price         -0.033498         0.018473           -1.813           0.070
L20.trends         0.047850         0.024700            1.937           0.053
L20.price         -0.037160         0.018367           -2.023           0.043
L21.trends        -0.000947         0.024614           -0.038           0.969
L21.price         -0.043501         0.018408           -2.363           0.018
```

Figure 4.18: VAR results for $\Delta Log(Price)$ equation with $\Delta Log(Price)$ and Google $\Delta Log(Trends)$ variables. The first column corresponds to the different lag values. The second column represents the coefficients of the lagged values. The last column corresponds to the p-values. Close to half of the coefficients are significant. Coefficients with the largest absolute value are the 2nd for the $\Delta Log(Price)$ variable and 4th for the $\Delta Log(Trends)$ variable.



```
-----------------------------------------------------------------------
Results for equation trends
===============================================================================
                coefficient       std. error           t-stat            prob
-------------------------------------------------------------------------------
const              0.000946         0.001215            0.779           0.436
L1.trends         -0.109866         0.019626           -5.598           0.000
L1.price          -0.011561         0.016688           -0.693           0.488
L2.trends          0.275554         0.019544           14.099           0.000
L2.price           0.148315         0.016474            9.003           0.000
L3.trends         -0.171433         0.019781           -8.667           0.000
L3.price           0.026399         0.017099            1.544           0.123
L4.trends         -0.179945         0.019761           -9.106           0.000
L4.price           0.045736         0.016572            2.760           0.006
L5.trends          0.217843         0.020217           10.775           0.000
L5.price           0.038708         0.015689            2.467           0.014
L6.trends         -0.117098         0.020843           -5.618           0.000
L6.price          -0.022283         0.015653           -1.424           0.155
L7.trends         -0.155881         0.020902           -7.458           0.000
L7.price          -0.026774         0.015528           -1.724           0.085
L8.trends          0.250152         0.021048           11.885           0.000
L8.price           0.013323         0.015503            0.859           0.390
L9.trends          0.213109         0.020920           10.187           0.000
L9.price           0.021878         0.015660            1.397           0.162
L10.trends        -0.120842         0.020504           -5.894           0.000
L10.price          0.042294         0.015726            2.689           0.007
L11.trends         0.005710         0.020532            0.278           0.781
L11.price          0.073585         0.015727            4.679           0.000
L12.trends         0.145068         0.020518            7.070           0.000
L12.price          0.025077         0.015793            1.588           0.112
L13.trends        -0.011591         0.020757           -0.558           0.577
L13.price          0.035295         0.015553            2.269           0.023
L14.trends        -0.033661         0.020709           -1.625           0.104
L14.price          0.051146         0.015547            3.290           0.001
L15.trends         0.119295         0.020744            5.751           0.000
L15.price          0.019324         0.015551            1.243           0.214
L16.trends         0.021562         0.020772            1.038           0.299
L16.price          0.007990         0.015523            0.515           0.607
L17.trends        -0.121221         0.020777           -5.834           0.000
L17.price          0.022561         0.015477            1.458           0.145
L18.trends        -0.028519         0.020809           -1.371           0.171
L18.price         -0.084559         0.015421           -5.483           0.000
L19.trends         0.062233         0.020665            3.012           0.003
L19.price          0.001402         0.015374            0.091           0.927
L20.trends         0.017008         0.020556            0.827           0.408
L20.price          0.049017         0.015286            3.207           0.001
L21.trends        -0.015582         0.020485           -0.761           0.447
L21.price          0.117191         0.015320            7.650           0.000
```

Figure 4.19: VAR results for $\Delta Log(Trends)$ equation with $\Delta Log(Price)$ and Google $\Delta Log(Trends)$ variables. The first column corresponds to the different lag values. The second column represents the coefficients of the lagged values. The last column corresponds to the p-values. More than half of the coefficients are significant. Coefficients with the largest absolute value are the 2nd for the $\Delta Log(Price)$ variable and the 2nd for the $\Delta Log(Trends)$ variable.

4.6.1 Response functions

After conducting the VAR analysis for the time series, we calculate their corresponding impulse response functions. The impulse response functions are the estimated responses to a unit impulse in one of the variables. The impulse response function are computed using the moving-average model MA∞ representation of the VAR(p) process.

$$y_t = \mu + \sum_{i=0}^{\infty} \Phi_i u_{t-i} \tag{4.2}$$

where y_t is the time series of interest. Furthermore, the expectation of y_t is defined as $\mu = (I_k - A_1 - ... - A_p)^{-1}c$. Having the coefficient c and the coefficient matrix A_i the same definition as in Equation 3.6. Moreover, u_t is zero mean white noise with non-singular covariance matrix and Φ_i the parameters of the model. Figures 4.20 and 4.22 depict the impulse response functions and Figures 4.21 and 4.23 depict the cumulative impulse response functions.

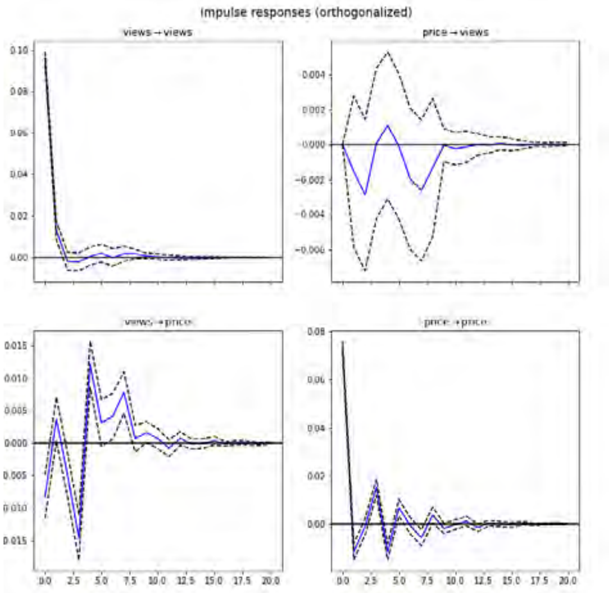


Figure 4.20: Impulse response functions for YouTube views and Bitcoin price, y-axis expresses standard deviation change that one variable cause to the other. It is computed using the MA(∞) representation in Equation 4.2.

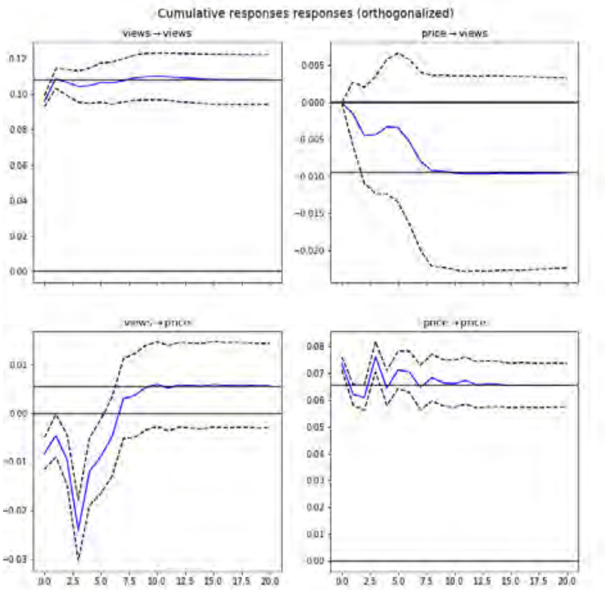


Figure 4.21: Cumulative impulse response functions for YouTube views and Bitcoin price, y-axis expresses standard deviation change that one variable cause to the other. From the figure it can be observed that the effect of impulse in price is effecting the number of views more than the impulse in views effecting the price.

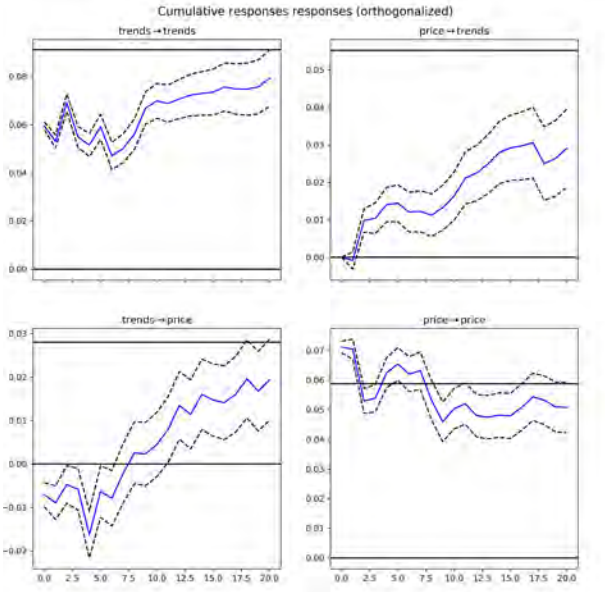


Figure 4.22: Impulse response functions for Google trends and Bitcoin price, y-axis expresses standard deviation change that one variable cause to the other. It is computed using the MA(∞) representation in Equation 4.2.

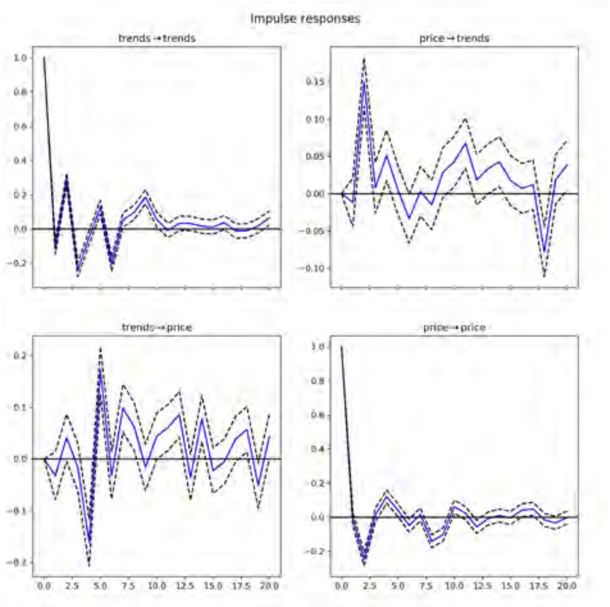


Figure 4.23: Cumulative impulse response functions for Google trends and Bitcoin price, y-axis expresses standard deviation change that one variable cause to the other. It can be observed that the effect of impulse in price is effecting similar the number of trends like the impulse in trends is effecting the price.

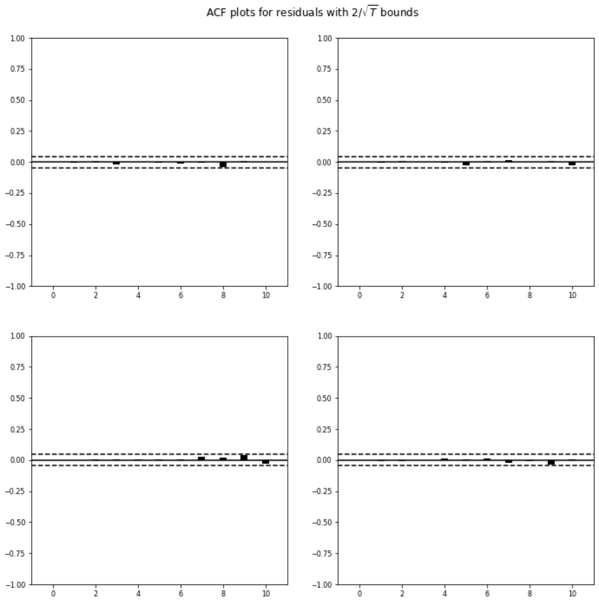


Figure 4.24: Autocorrelation of the residuals after VAR analysis of YouTube views and BTC price. This shows the goodness of fit and in fact the fit is done correctly (from low autocorrelation values. Dashed line is significance level.)

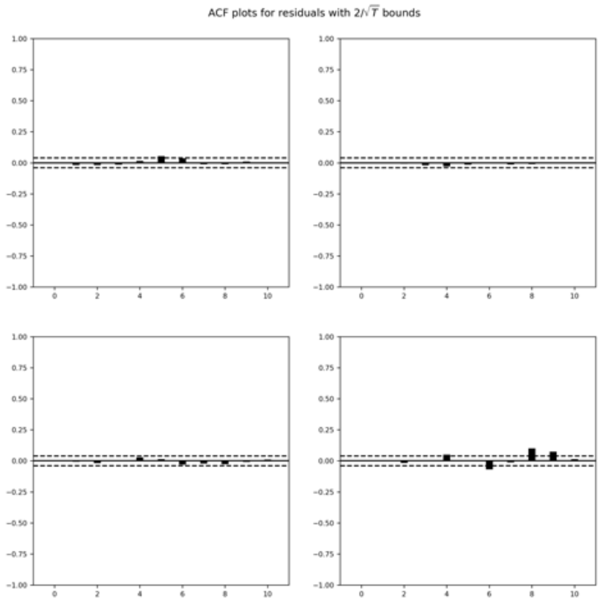
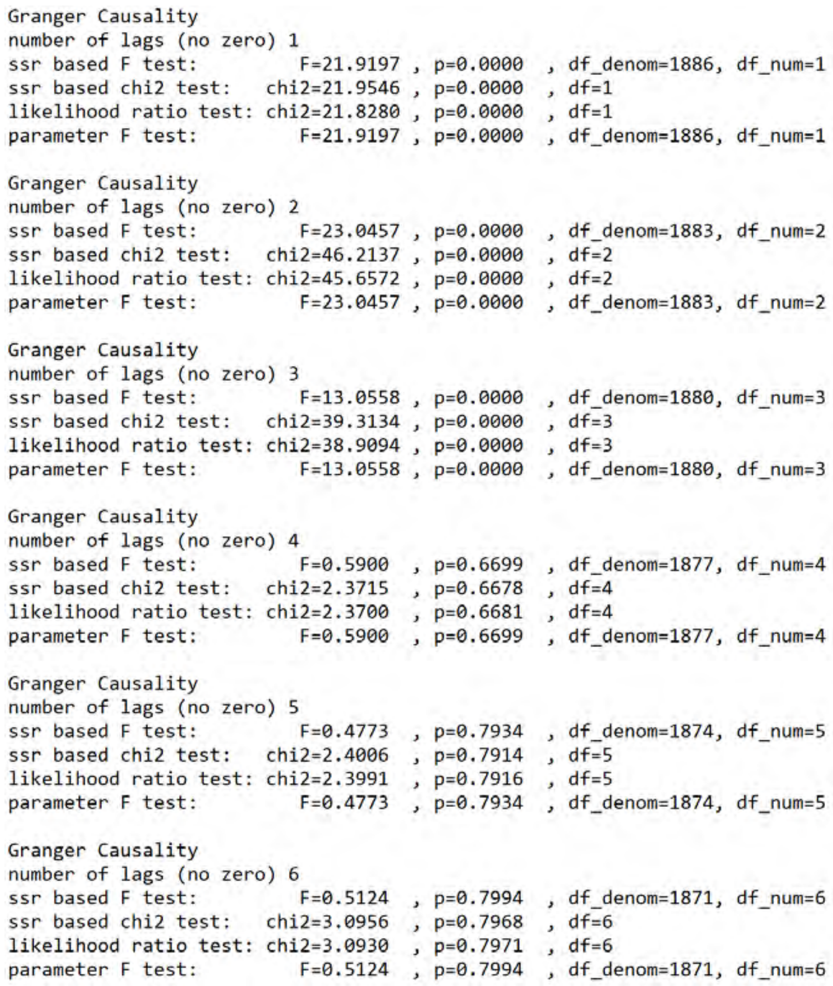


Figure 4.25: Autocorrelation of the residuals after VAR analysis of Google trends and BTC price. This shows the goodness of fit and in fact the fit is done correctly (from low autocorrelation values. Dashed line is significance level.)

4.7 Granger Causality

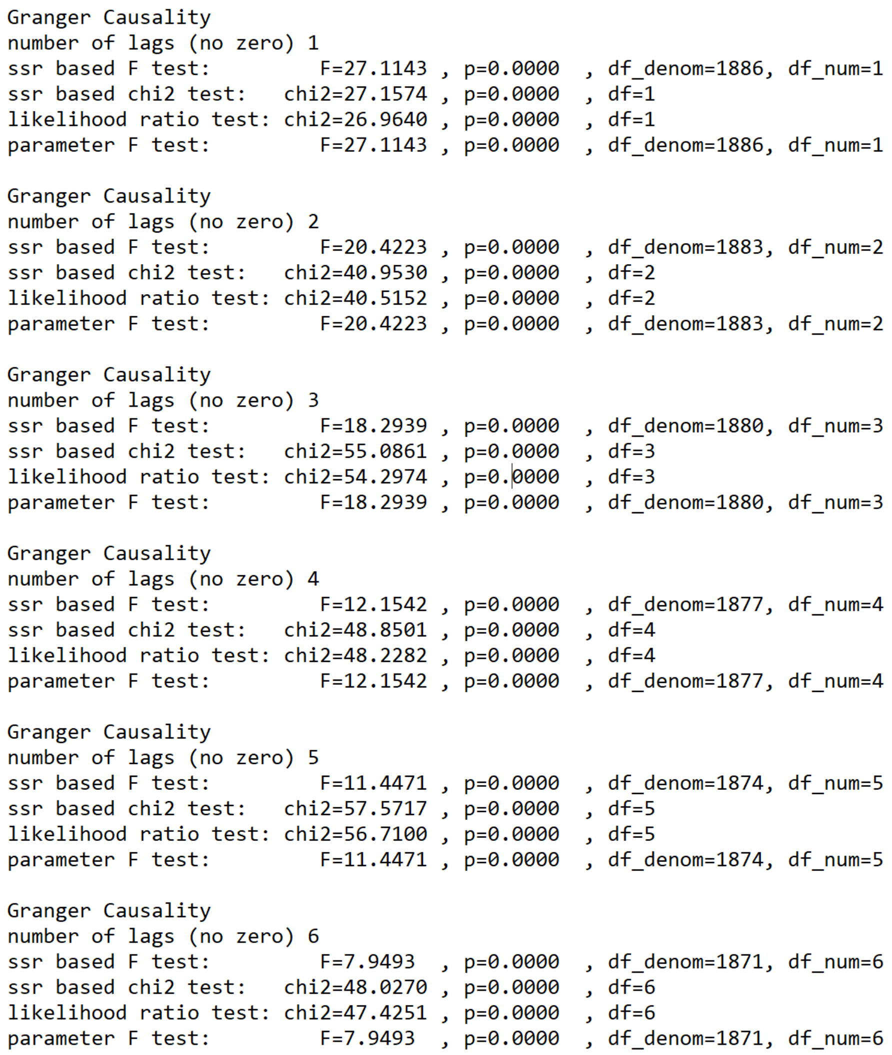
We want to find the causal dependencies between two time series. Thus, we calculate the Granger causality. The null hypothesis of the test is that the second time series does not Granger cause the first time series. We reject the null hypothesis that the second time series does not Granger cause the first time series if the p-values are below a desired size of the test.

Overall we perform four tests. In the first test, we check whether the Bitcoin price is Granger causing YouTube views. Figure 4.26 shows that Bitcoin price might Granger cause YouTube views for the first three lags. For the second test, we check whether YouTube views is Granger causing Bitcoin price. In the third test, we check whether Bitcoin price is Granger causing Google trends. Figure 4.28 shows that after the second lag Bitcoin price might Granger cause Google trends. In the last test, we check whether Google trends is Granger causing Bitcoin price. Figures 4.27 and 4.29 show that the collective public interest for YouTube views and Google trends might Granger cause Bitcoin price.



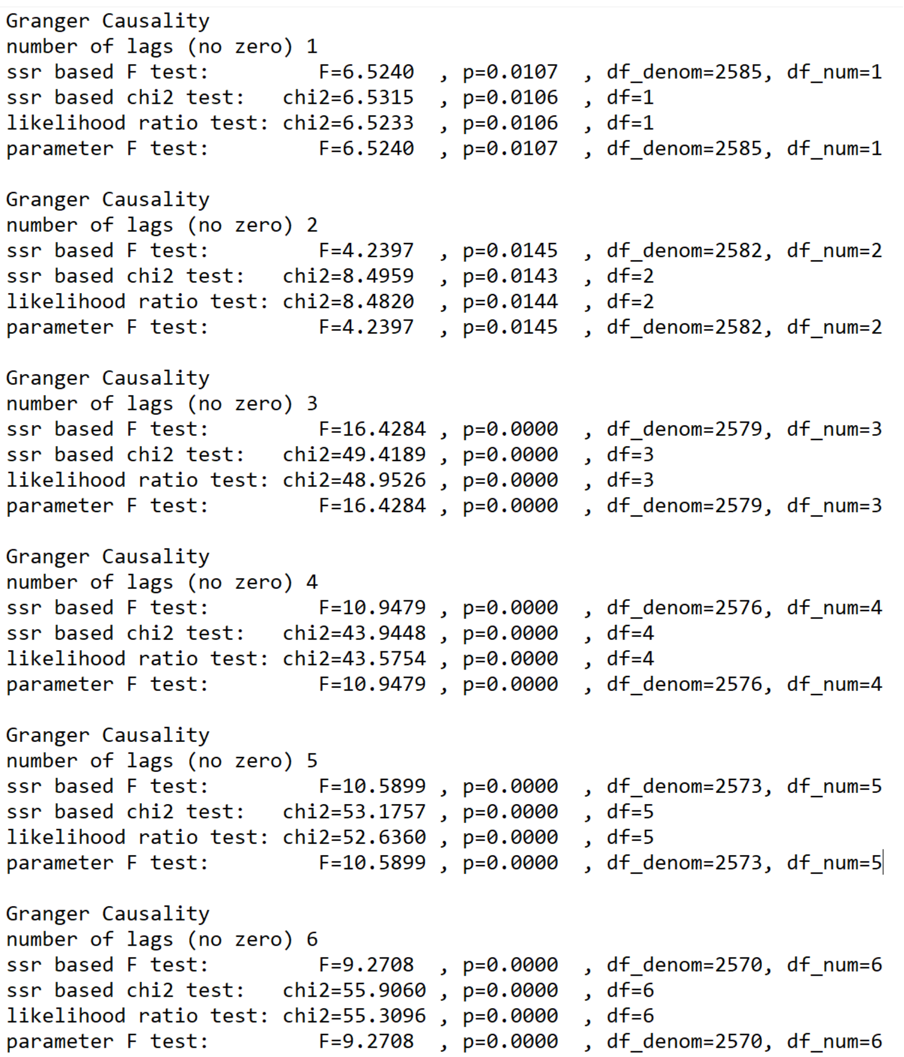
```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=21.9197 , p=0.0000  , df_denom=1886, df_num=1
ssr based chi2 test:   chi2=21.9546 , p=0.0000  , df=1
likelihood ratio test: chi2=21.8280 , p=0.0000  , df=1
parameter F test:         F=21.9197 , p=0.0000  , df_denom=1886, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=23.0457 , p=0.0000  , df_denom=1883, df_num=2
ssr based chi2 test:   chi2=46.2137 , p=0.0000  , df=2
likelihood ratio test: chi2=45.6572 , p=0.0000  , df=2
parameter F test:         F=23.0457 , p=0.0000  , df_denom=1883, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=13.0558 , p=0.0000  , df_denom=1880, df_num=3
ssr based chi2 test:   chi2=39.3134 , p=0.0000  , df=3
likelihood ratio test: chi2=38.9094 , p=0.0000  , df=3
parameter F test:         F=13.0558 , p=0.0000  , df_denom=1880, df_num=3

Granger Causality
number of lags (no zero) 4
ssr based F test:         F=0.5900  , p=0.6699  , df_denom=1877, df_num=4
ssr based chi2 test:   chi2=2.3715  , p=0.6678  , df=4
likelihood ratio test: chi2=2.3700  , p=0.6681  , df=4
parameter F test:         F=0.5900  , p=0.6699  , df_denom=1877, df_num=4

Granger Causality
number of lags (no zero) 5
ssr based F test:         F=0.4773  , p=0.7934  , df_denom=1874, df_num=5
ssr based chi2 test:   chi2=2.4006  , p=0.7914  , df=5
likelihood ratio test: chi2=2.3991  , p=0.7916  , df=5
parameter F test:         F=0.4773  , p=0.7934  , df_denom=1874, df_num=5

Granger Causality
number of lags (no zero) 6
ssr based F test:         F=0.5124  , p=0.7994  , df_denom=1871, df_num=6
ssr based chi2 test:   chi2=3.0956  , p=0.7968  , df=6
likelihood ratio test: chi2=3.0930  , p=0.7971  , df=6
parameter F test:         F=0.5124  , p=0.7994  , df_denom=1871, df_num=6
```

Figure 4.26: Results of Granger causality test. H_0: Bitcoin price does not Granger cause YouTube views. For the first four lags, the p-value equals zero. Thus, we reject the null hypothesis that Bitcoin price does not Granger cause YouTube views for the first three lags. However, upon the 4th lag we accept the null hypothesis that Bitcoin price does not Granger cause YouTube views.



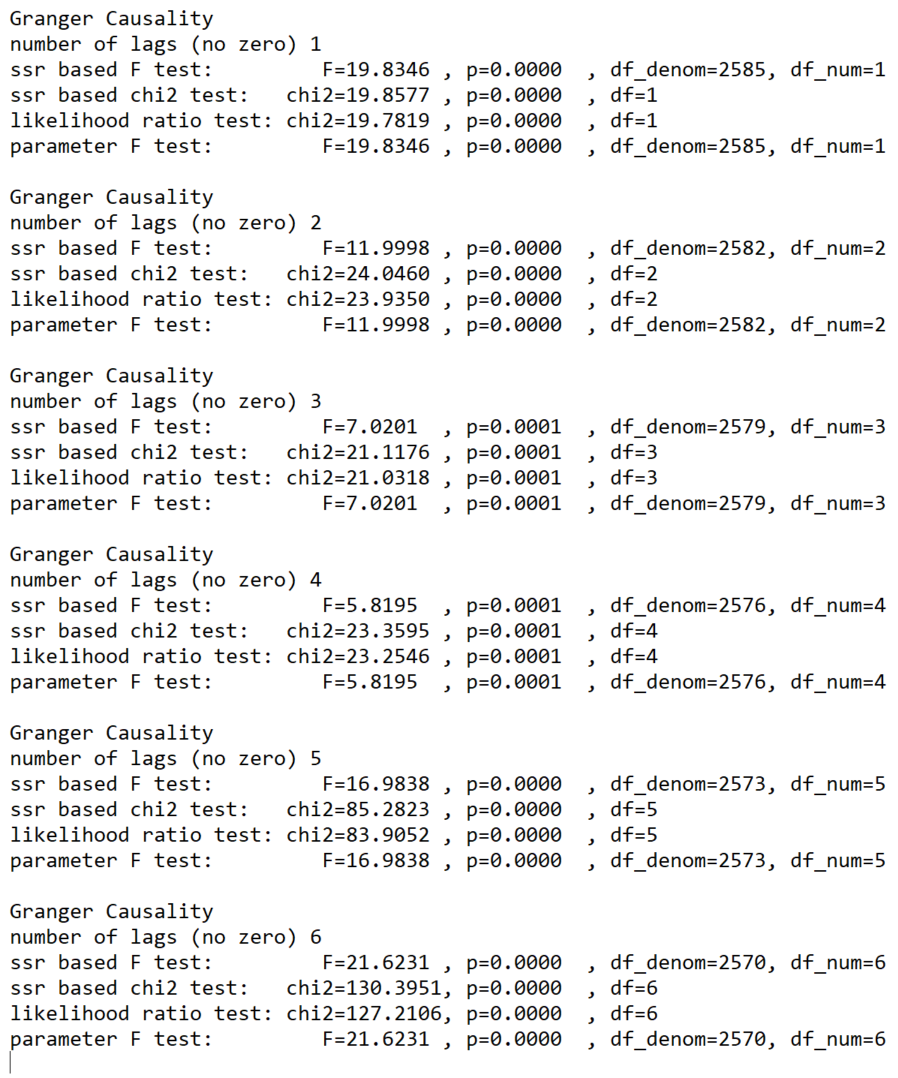
```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=27.1143 , p=0.0000  , df_denom=1886, df_num=1
ssr based chi2 test:   chi2=27.1574 , p=0.0000  , df=1
likelihood ratio test: chi2=26.9640 , p=0.0000  , df=1
parameter F test:         F=27.1143 , p=0.0000  , df_denom=1886, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=20.4223 , p=0.0000  , df_denom=1883, df_num=2
ssr based chi2 test:   chi2=40.9530 , p=0.0000  , df=2
likelihood ratio test: chi2=40.5152 , p=0.0000  , df=2
parameter F test:         F=20.4223 , p=0.0000  , df_denom=1883, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=18.2939 , p=0.0000  , df_denom=1880, df_num=3
ssr based chi2 test:   chi2=55.0861 , p=0.0000  , df=3
likelihood ratio test: chi2=54.2974 , p=0.0000  , df=3
parameter F test:         F=18.2939 , p=0.0000  , df_denom=1880, df_num=3

Granger Causality
number of lags (no zero) 4
ssr based F test:         F=12.1542 , p=0.0000  , df_denom=1877, df_num=4
ssr based chi2 test:   chi2=48.8501 , p=0.0000  , df=4
likelihood ratio test: chi2=48.2282 , p=0.0000  , df=4
parameter F test:         F=12.1542 , p=0.0000  , df_denom=1877, df_num=4

Granger Causality
number of lags (no zero) 5
ssr based F test:         F=11.4471 , p=0.0000  , df_denom=1874, df_num=5
ssr based chi2 test:   chi2=57.5717 , p=0.0000  , df=5
likelihood ratio test: chi2=56.7100 , p=0.0000  , df=5
parameter F test:         F=11.4471 , p=0.0000  , df_denom=1874, df_num=5

Granger Causality
number of lags (no zero) 6
ssr based F test:         F=7.9493  , p=0.0000  , df_denom=1871, df_num=6
ssr based chi2 test:   chi2=48.0270 , p=0.0000  , df=6
likelihood ratio test: chi2=47.4251 , p=0.0000  , df=6
parameter F test:         F=7.9493  , p=0.0000  , df_denom=1871, df_num=6
```

Figure 4.27: Results of Granger causality test. H_0: YouTube views does not Granger cause Bitcoin price. All p-values equal zero among the four tests. Thus, we reject the null hypothesis that YouTube views does not Granger cause Bitcoin price.



```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=6.5240  , p=0.0107  , df_denom=2585, df_num=1
ssr based chi2 test:   chi2=6.5315  , p=0.0106  , df=1
likelihood ratio test: chi2=6.5233  , p=0.0106  , df=1
parameter F test:         F=6.5240  , p=0.0107  , df_denom=2585, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=4.2397  , p=0.0145  , df_denom=2582, df_num=2
ssr based chi2 test:   chi2=8.4959  , p=0.0143  , df=2
likelihood ratio test: chi2=8.4820  , p=0.0144  , df=2
parameter F test:         F=4.2397  , p=0.0145  , df_denom=2582, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=16.4284 , p=0.0000  , df_denom=2579, df_num=3
ssr based chi2 test:   chi2=49.4189 , p=0.0000  , df=3
likelihood ratio test: chi2=48.9526 , p=0.0000  , df=3
parameter F test:         F=16.4284 , p=0.0000  , df_denom=2579, df_num=3

Granger Causality
number of lags (no zero) 4
ssr based F test:         F=10.9479 , p=0.0000  , df_denom=2576, df_num=4
ssr based chi2 test:   chi2=43.9448 , p=0.0000  , df=4
likelihood ratio test: chi2=43.5754 , p=0.0000  , df=4
parameter F test:         F=10.9479 , p=0.0000  , df_denom=2576, df_num=4

Granger Causality
number of lags (no zero) 5
ssr based F test:         F=10.5899 , p=0.0000  , df_denom=2573, df_num=5
ssr based chi2 test:   chi2=53.1757 , p=0.0000  , df=5
likelihood ratio test: chi2=52.6360 , p=0.0000  , df=5
parameter F test:         F=10.5899 , p=0.0000  , df_denom=2573, df_num=5

Granger Causality
number of lags (no zero) 6
ssr based F test:         F=9.2708  , p=0.0000  , df_denom=2570, df_num=6
ssr based chi2 test:   chi2=55.9060 , p=0.0000  , df=6
likelihood ratio test: chi2=55.3096 , p=0.0000  , df=6
parameter F test:         F=9.2708  , p=0.0000  , df_denom=2570, df_num=6
```

Figure 4.28: Results of Granger causality test. H_0: Bitcoin price does not Granger cause Google trends. For the first two lags, the the p-value rise above the significance level among all four tests. Thus, we accept the null hypothesis that Bitcoin price does not Granger cause Google trends for the first two lags. After the second lag the p-value equals zero. Thus, we reject the null hypothesis that Bitcoin price does not Granger cause Google trends after the second lag.



```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=19.8346 , p=0.0000  , df_denom=2585, df_num=1
ssr based chi2 test:   chi2=19.8577 , p=0.0000  , df=1
likelihood ratio test: chi2=19.7819 , p=0.0000  , df=1
parameter F test:         F=19.8346 , p=0.0000  , df_denom=2585, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=11.9998 , p=0.0000  , df_denom=2582, df_num=2
ssr based chi2 test:   chi2=24.0460 , p=0.0000  , df=2
likelihood ratio test: chi2=23.9350 , p=0.0000  , df=2
parameter F test:         F=11.9998 , p=0.0000  , df_denom=2582, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=7.0201  , p=0.0001  , df_denom=2579, df_num=3
ssr based chi2 test:   chi2=21.1176 , p=0.0001  , df=3
likelihood ratio test: chi2=21.0318 , p=0.0001  , df=3
parameter F test:         F=7.0201  , p=0.0001  , df_denom=2579, df_num=3

Granger Causality
number of lags (no zero) 4
ssr based F test:         F=5.8195  , p=0.0001  , df_denom=2576, df_num=4
ssr based chi2 test:   chi2=23.3595 , p=0.0001  , df=4
likelihood ratio test: chi2=23.2546 , p=0.0001  , df=4
parameter F test:         F=5.8195  , p=0.0001  , df_denom=2576, df_num=4

Granger Causality
number of lags (no zero) 5
ssr based F test:         F=16.9838 , p=0.0000  , df_denom=2573, df_num=5
ssr based chi2 test:   chi2=85.2823 , p=0.0000  , df=5
likelihood ratio test: chi2=83.9052 , p=0.0000  , df=5
parameter F test:         F=16.9838 , p=0.0000  , df_denom=2573, df_num=5

Granger Causality
number of lags (no zero) 6
ssr based F test:         F=21.6231 , p=0.0000  , df_denom=2570, df_num=6
ssr based chi2 test:   chi2=130.3951, p=0.0000  , df=6
likelihood ratio test: chi2=127.2106, p=0.0000  , df=6
parameter F test:         F=21.6231 , p=0.0000  , df_denom=2570, df_num=6
```

Figure 4.29: Results of Granger causality test. H_0: Google trends does not Granger cause Bitcoin price. All the p-values equal or are very close to zero. Thus, we reject the null hypothesis that Google trends does not Granger cause Bitcoin price.

Chapter 5
Discussion

This project has focused on the understanding of the dynamics of Bitcoin prices taking into account external news like YouTube videos and Google trends during bubble periods. Our initial guess was that the price movements are reflected in social media news. In other words, our initial hypothesis was that the Bitcoin price acts as a one-directional causality to the news. Thus, the total amount of daily views and searches are influenced by the Bitcoin price.

After analyzing the aggregated total number of daily YouTube views, we found out that the public interest of the YouTube videos published at an arbitrary day follow a power-law behavior introduced in [1]. In other words, videos that were published on an arbitrary day reach the maximum number of total views on the publication day then as time goes by the interest of those videos published in the past decreases which cause the daily aggregated number of views to become smaller. Moreover, we verified that the returns (first-order difference of the time series) are stationary. This checkup allowed us to perform a volatility analysis over different time windows to observe similar behavior among all the time series. Additionally, we conducted a VAR analysis over different rolling windows on the returns to obtain the lag length and the optimal lag value between Bitcoin price with YouTube views, and Bitcoin price with Google trends. The obtained results from the Granger causality test also showed us that there exists a bidirectional causality between Bitcoin price and public interest. Meaning that not only price causes trends and views, but also public interest which is reflected by the trends and views cause the price changes at specific lag values.

The present analysis could be extended to be more powerful. Co-integration tests could be implemented instead of taking the first order differences of the time series, since taking the first order difference might lead to enhancing the noise in the time series. Moreover, more analysis could be conducted primarily in the bubble windows. Additionally, more cryptocurrency news and social media could be incorporated into the analysis. An exciting simulation would be the implementation of a trading strategy based on the results obtained in this project. It would be possible and interesting to test whether it is profitable to hedge a position knowing the optimal lag values.

Bibliography

[1] R. Crane and D. Sornette, "Robust dynamic classes revealed by measuring the response function of a social system," *Proceedings of the National Academy of Sciences*, vol. 105, no. 41, pp. 15 649–15 653, Oct 2008.

[2] D. Sornette and P. Cauwels, "Financial bubbles: mechanisms and diagnostics."," *Review of Behavioral Economics*, vol. 2, no. 3, pp. 279–305, Oct 2015.

[3] D. Kahneman, *Thinking, fast and slow.* New York: Farrar, Straus and Giroux, 2011.

[4] G. Soros, *The Alchemy of Finance: Reading the mind of the Market.* Wiley Investment Classics, 1987.

[5] G. Soros, "Theory of reflexivity and the methodology of economic science," *Journal of Economic Methodology*, vol. 20, no. 4, pp. 342–351, 2013.

[6] B. Roehner and D. Sornette, "Thermometers of speculative frenzy," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 16, no. 4, pp. 729–739, Aug 2000.

[7] "Coinmarketcap," https://coinmarketcap.com/currencies/bitcoin/historical-data/, 2019, [Online; accessed 14-November-2018].

[8] "Coincheckup," https://coincheckup.com/, 2019, [Online; accessed 14-November-2018].

[9] "Google trends," https://github.com/bspammer/rebuild_trends, 2019, [Online; accessed 19-February-2019].

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R.* Springer Publishing Company, Incorporated, 2014.

[11] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis.* Springer, 2005.

[12] "An illustrated history of bitcoin crashes," https://www.forbes.com/sites/timothylee/2013/04/11/an-illustrated-history-of-bitcoin-crashes/, 2013, [Online; accessed 19-February-2019].

[13] D. Sornette, “Dragon-kings, black swans and the prediction of crises,” *Swiss Finance Institute, Swiss Finance Institute Research Paper Series*, vol. 2, Jul 2009.

[14] J.-C. Gerlach, G. Demos, and D. Sornette, “Dissection of bitcoin’s multiscale bubble history from january 2012 to february 2018,” *SSRN Electronic Journal*, Apr 2018.

[15] Z. Wu, N. E. Huang, S. R. Long, and C.-K. Peng, “On the trend, detrending, and variability of nonlinear and nonstationary time series,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 38, pp. 14 889–14 894, Sep 2007.

Appendix A

Appendix I

A.1 VECM

As the $Log(Price)$ and $Log(Views)$ were not stationary, we had to perform VAR(p) on $\Delta Log(Price)$ and $\Delta Log(Views)$. Hence, we only got the information about the short term dependencies and did not get any information about the long term trends. In order to get the long-range dependencies between the time series, we are using the Vector Error Correction Model. A VECM has the following form:

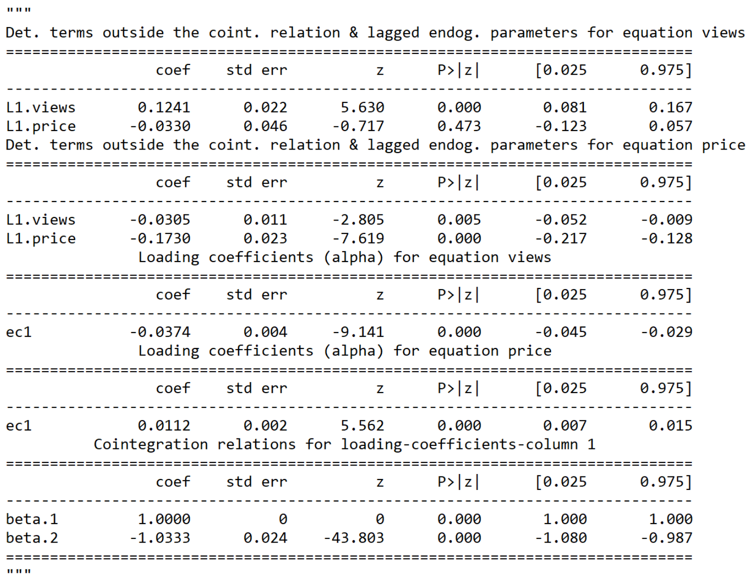
$$\Delta y_t = \Pi y_{t-1} + \Gamma_1 \Delta y_{t-1} + \ldots + \Gamma_{k_{ar}-1} \Delta y_{t-k_{ar}+1} + u_t \tag{A.1}$$

where y_t is K-dimensional, Π is a $(K \times K)$ matrix of rank r, $0 < r < K$, α and β are $(K \times r)$ with rank r, and u_t is K-dimensional white noise with mean zero. Furthermore, $\Pi = \alpha\beta'$ as derived in [11].

If the cointegration ranks lays at the one of the boundaries $r = 0$, Δy_t is stable and for $r = K$, y_t is stable. We know from the results of the Johansen test 4.3.3 that $r = 1$.

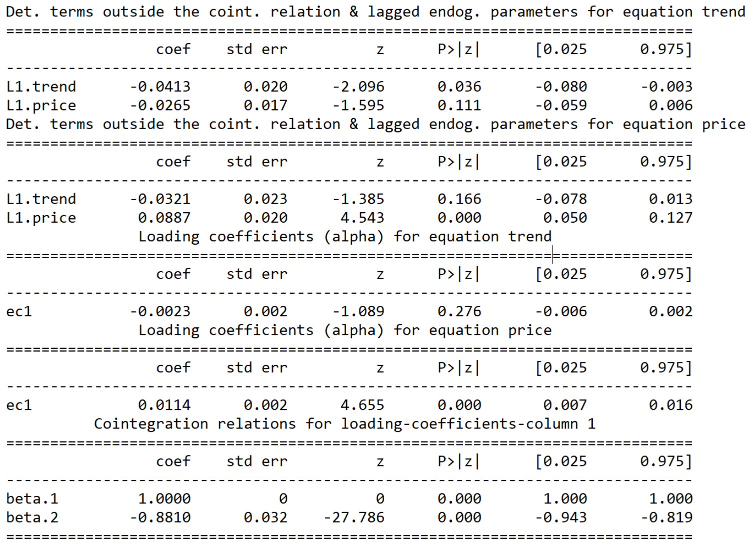
From Figure A.1, we see that the coefficients for the YouTube data are $\alpha = [-0.0374, 0.0112]^\top$ and $\beta = [1, -1.0333]^\top$.

From Figure A.2 we see that the coefficients for the Google data are $\alpha = [-0.0023, 0.0114]^\top$ and $\beta = [1, -0.8810]^\top$



```
"""
Det. terms outside the coint. relation & lagged endog. parameters for equation views
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
L1.views       0.1241      0.022      5.630      0.000       0.081       0.167
L1.price      -0.0330      0.046     -0.717      0.473      -0.123       0.057
Det. terms outside the coint. relation & lagged endog. parameters for equation price
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
L1.views      -0.0305      0.011     -2.805      0.005      -0.052      -0.009
L1.price      -0.1730      0.023     -7.619      0.000      -0.217      -0.128
               Loading coefficients (alpha) for equation views
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ec1           -0.0374      0.004     -9.141      0.000      -0.045      -0.029
               Loading coefficients (alpha) for equation price
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ec1            0.0112      0.002      5.562      0.000       0.007       0.015
          Cointegration relations for loading-coefficients-column 1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
beta.1         1.0000          0          0      0.000       1.000       1.000
beta.2        -1.0333      0.024    -43.803      0.000      -1.080      -0.987
==============================================================================
"""
```

Figure A.1: Output after the VECM is applied to Log(Price) & Log(Views).



```
Det. terms outside the coint. relation & lagged endog. parameters for equation trend
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
L1.trend      -0.0413      0.020     -2.096      0.036      -0.080      -0.003
L1.price      -0.0265      0.017     -1.595      0.111      -0.059       0.006
Det. terms outside the coint. relation & lagged endog. parameters for equation price
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
L1.trend      -0.0321      0.023     -1.385      0.166      -0.078       0.013
L1.price       0.0887      0.020      4.543      0.000       0.050       0.127
               Loading coefficients (alpha) for equation trend
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ec1           -0.0023      0.002     -1.089      0.276      -0.006       0.002
               Loading coefficients (alpha) for equation price
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ec1            0.0114      0.002      4.655      0.000       0.007       0.016
          Cointegration relations for loading-coefficients-column 1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
beta.1         1.0000          0          0      0.000       1.000       1.000
beta.2        -0.8810      0.032    -27.786      0.000      -0.943      -0.819
==============================================================================
```

Figure A.2: Output after the VECM is applied to Log(Price) & Log(Trends).

A.2 Empirical Mode Decomposition

So far we are dependent on the differenced time series to get the information about the interdependencies between the time series, as the VAR model is only applicable on stationary time series. Moreover, by only analyzing the differenced time series, we are not getting any information about the long term/seasonal dependencies.

In order to get the interdependency information on multiple scales, we are breaking the original time series into multiple Intrinsic Mode Functions (IMFs) using the Empirical Mode Decomposition (EMD)[15]. Each of the IMF is a stationary time series and contains seasonal information/trends. The sum of the IMFs will give us back the original time series.

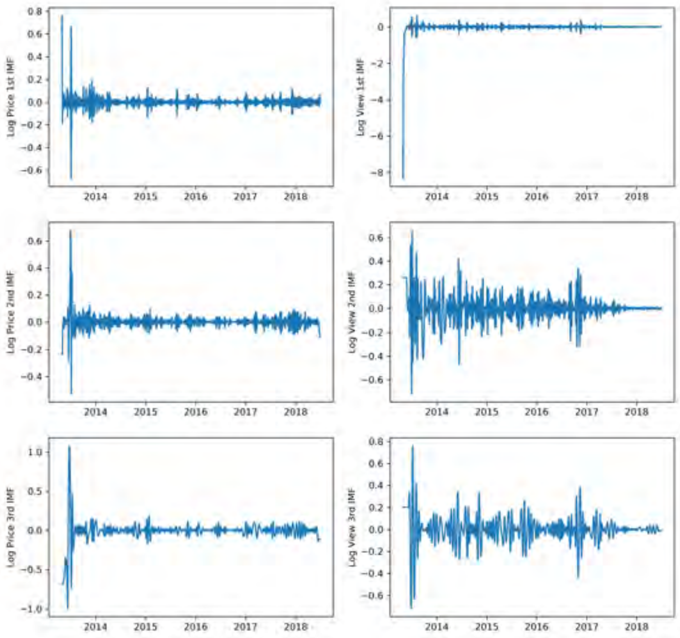


Figure A.3: Intrinsic Mode Functions of Log(price) and Log(Views).

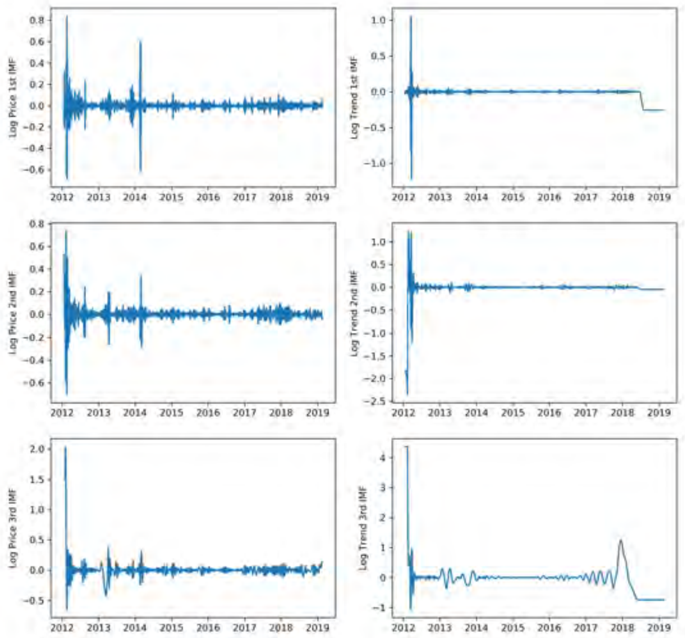


Figure A.4: Intrinsic Mode Functions of Log(Price) and Log(Trends).

Appendix B
Declaration of Authorship

I hereby declare that the semester project submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.

I am aware that the semester project in digital form can be examined for the use of unauthorized aid and in order to determine whether the semester project as a whole or parts incorporated in it may be deemed as plagiarism.

This paper was not previously presented to another examination board and has not been published in anywhere else yet.

Pedro Daniel Partida Güitrón