High Frequency Decentralized Cryptocurrency Price Analysis — Using Sentiment Score of Bitcoin

May 2021

Chang Cai
Yuqing Han
Yifan Peng

1.Introduction

When mentioning Satoshi Nakamoto's name, most people may not know who he is, but what she/he created the first decentralized cryptocurrency—Bitcoin[1]— changed our world dramatically, and its price surpasses the new high at $50,000[2] for the first time on February 16, 2021. Bitcoin was the first released decentralized cryptocurrency (crypto), and it led the global crypto market to hit $2 trillion for the first time in April 2021. Cryptos were built using "blockchain" technology that distributed digital ledger with records called blocks to record transaction information. On the one hand, including Bitcoin, Ethereum, and Dogecoin are the representatives of the three different types of popular cryptos. On the other hand, a data science technique called sentiment analysis has been used in various domains. It has become a factor of trading strategies in nowadays financial markets. Also, most of the empirical analysis on the cryptocurrency price use daily or hourly level data. We found a lack of literature on the interdisciplinary topic of high-frequency crypto price analysis using the sentiment score of major

[1] The original document introduced Bitcoin at https://Bitcoin.org/Bitcoin.pdf

[2] According to the news at https://www.cnbc.com/2021/02/16/Bitcoin-btc-price-hits-50000-for-the-first-time.html

crypto. To fill the gap in this topic, we conduct an empirical time series analysis for the price variation of three representative cryptos mentioned above, including Bitcoin, Ethereum, and Dogecoin, using the sentiment score of the major crypto-Bitcoin.

2. Background and Literature Review

Bitcoins are generated in the "mining" process (Ciaian, Rajcaniova, and Kancs 2016), and mining is considered as driven by the return of Bitcoins given the number of Bitcoin is pre-determined. However, the mining itself does not impact our research, given no considerable change in the difficulty of mining and the advancement of technology on the mining during our study period. Therefore, we can exclude the factor of "difficulty of mining" from our topic. Many empirical studies in this field regarding crypto price have increased exponentially in recent years in the fields of applied econometrics, and most of them used daily or monthly level data.

Bitcoin is identified as behaving like a speculative investment rather than a currency (Yermack 2013), and therefore we consider Bitcoin price might be driven by an index from the market reaction. A well-cited study using daily-level data finds supply and demand fundamentals significantly impact Bitcoin price but not from the macroeconomic indicators, such as the stock market index (Ciaian, Rajcaniova, and Kancs 2016). In recent years, data science techniques on text analysis help us show a potential new channel for analyzing the price of cryptos. A study in 2017 shows a positive relationship between the price of Bitcoin and sentiment analysis of tweets using daily basis data for the month-length research (Perry-Carrera 2018). Although a certain number of works have been done in either field mentioned above, it is rare to find the attempts that combined those two fields together as well as using high-frequency data, and therefore it motivated our group to conduct this analysis using new approaches.

First, Bitcoin has the biggest market capitalization of approximately $1 trillion[3] and is the most widely traded and is one of the few cryptos accepted as a payment method by companies. The Bitcoin limit of 21 million was set when it was created in 2009. Second, Ethereum is the second largest crypto, with approximately half of Bitcoin's 1 trillion, reached at 450 billion, and there is no issuance limit like Bitcoin; Yet Ethereum functions more than crypto, such as Bitcoin, it works as a "smart contract" platform where hosts can run use the Ethereum blockchain for a marketplace of decentralized apps (Ranganthan et al. 2018). Third, Dogecoin is a crypto that began in a non-formal way. Given the fact that Dogecoin's core is a meme (Benaim 2018) that fundamentally relies on social media and the price is driven by memes, such as from Elon Musk. Therefore, it represents the type of cryptos with its characteristic for fun and the high likelihood to be driven by the market heat from social networks.

3 Data and Method

3.1 Data

To assess the Bitcoin price discovery variation and shed light on the changes in price discovery for high frequency, we use minute level data for all the variables. The critical index we used is the minute level sentiment score on the Bitcoin topic, given that algorithmic trading took over most equities in the U.S. Most importantly, high frequency played a critical role in that (Glantz and Kissell 2013). In other words, institutions who invest in the cryptocurrency market may use algorithmic trading strategies, such as algorithms of market heat of reaction from social networking. Therefore, we set the sentiment score as the key variable here.

[3] According to the news on April 12, 2021 at https://www.visualcapitalist.com/bitcoin-is-the-fastest-asset-to-reach-a-1-trillion-market-cap/

Sentiment Score

We explore using social media posts from Twitter to assess the relationship between Bitcoin price changes and public opinion on Bitcoin. We utilize the Twitter API to extract desired tweets about Bitcoin and a Python Library called "Tweepy" to connect to the Twitter API and download the data at UTC +0 time zone. Twitter's standard API allows us to retrieve a subset[4] of tweets from the last seven days for free[5]. We filter recent tweets that contain the keywords "Bitcoin" or hashtag "#Bitcoin". The search list does not include currency abbreviations such as "btc" for Bitcoin, because these search criteria can be ambiguous. The search query was set to stream only tweets in English and the search results include both original tweets and retweets.

Besides the 7-day restriction of historical archive, twitter API can only make a limited number of calls using a basic and free developer account. Currently, the standard API can only send 450 calls per application every 15 minutes. If we use up our 450 calls in the 15-minute window, a "rate limit reached" message will be returned, and Twitter will not provide any updates until 15 minutes is up. Each request can deliver up to 100 tweets in reverse-chronological order. To extract tweets in a 4-hour window, it usually takes approximately 2 hours to finish extraction that would yield 20,000 tweets. Thus, it is a time-consuming process to retrieve full-day volume of tweets on a specific date. For this project, we decided to let the script run 2 hours and collect 20,000 tweets for a day from May 3rd to May 6th. In total, the final tweet dataset consists of 80,000 tweets (see Figure 1).

[4] This means that the query is not exhaustive. Some Tweets and users may be missing from search results.
[5] Twitter's premium and enterprise search APIs enable users to access any publicly available Tweet as early as 2006.

To remove the noise from the tweet dataset, we first use regular expressions to clean the collected tweets. After preprocessing the collected tweets, we use VADER (Valence Aware Dictionary and sEntiment Reasoner) to evaluate each cleaned tweet and classify it as negative, neutral, or positive by looking at the compound score. To match the sentiment score to the financial data, the time series of sentiment is aggregated at the minute level and converted to UTC +12 time zone. Overall, we found all the tweets about Bitcoin provide positive sentiment and the score is well below 0.5 with large variability minute by minute (see Figure 2). Besides, Figure A.1 in the Appendix plots the sentiment scores of Bitcoin with the Bitcoin price ($).

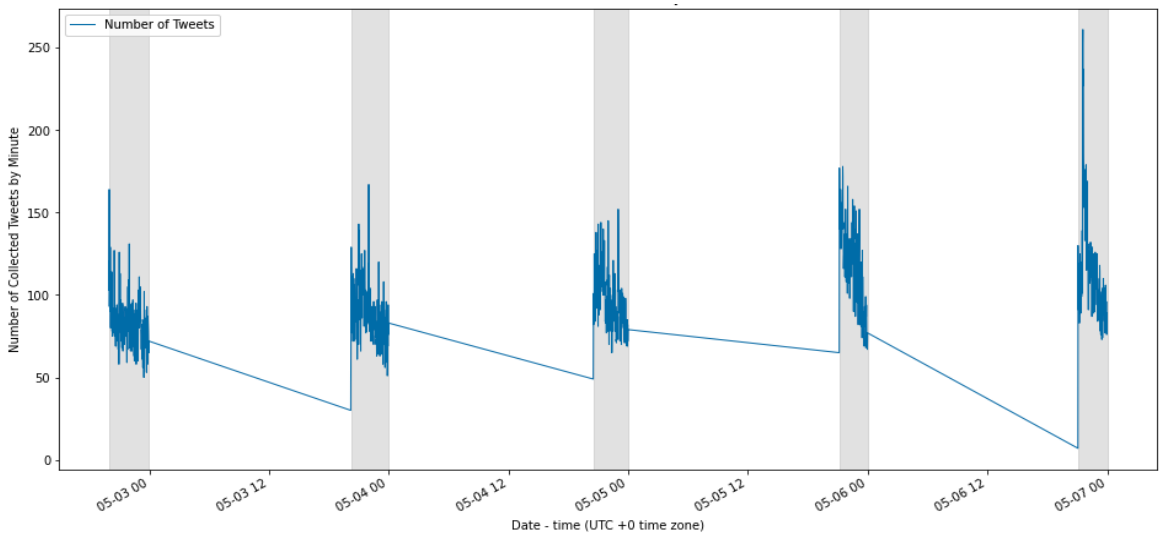

Figure 1 Volume of collected tweets during study period. Grey bars show the time windows of tweets retrieved.

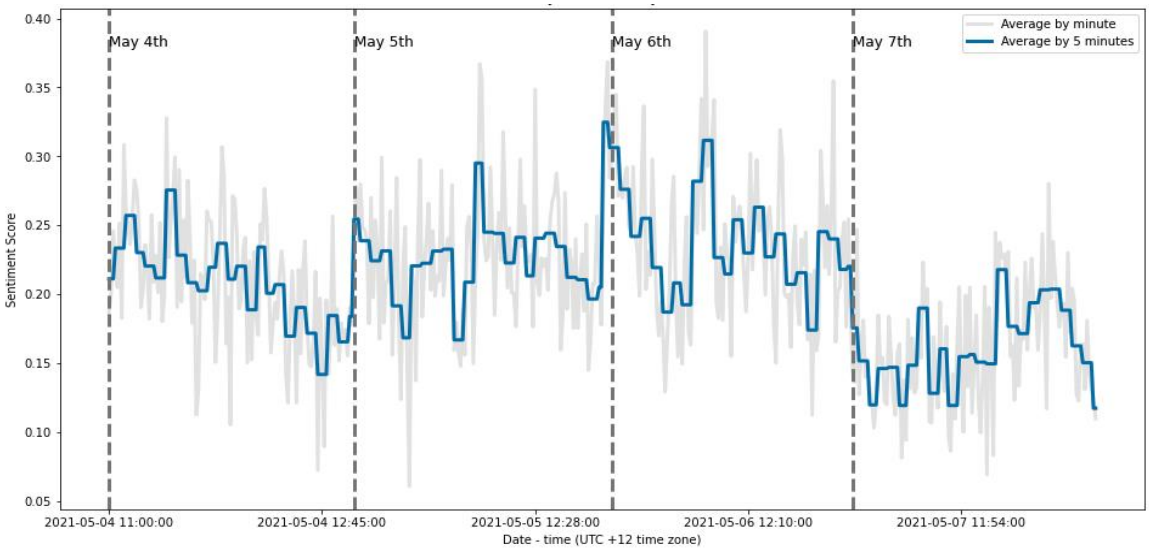

Figure 2 Time series plot of tweet sentiment by minute vs. 5-minute interval. Time periods with no data are not shown in this figure.

Financial Data

We use two types of financial data in this study we collected via Yahoo Finance API, which allows users to collect the most recent 7- day high frequency - one minute level - data. First, the variable for the prices in USD of three representative cryptos, Bitcoin, Ethereum, and Dogecoin, at UTC+1 time zone. We move all the crypto price data into UTC +0 time zone to match the sentiment score we created. Second, the country's stock index, where English is one of the official languages, matches the content of sentiment scores we created in English. We selected the main stock market index NZX 50[6] of New Zealand at UTC +12 for our study because the time zone of NZX 50 also matches all other variables and because the language matched. NZX 50 index contains the 50 biggest stocks through free-float capitalization trading on the New Zealand Stock Market (NZSX), and we move the index into UTC+0 time zone as well.

[6] For more information about NZX50 https://en.wikipedia.org/wiki/NZX_50_Index

Time zone Matching

To match the timestamp of the sentiment score for high frequency, we select the stock index and the price data of other cryptocurrencies that match the timestamps. Finally, to match all the variables at the same time zone in the study, we choose from 10 PM to 11:59 PM of four-day length data from May 3 to May 6 at UTC +0 time zone, which equivalent to New Zealand's 10:00 AM to 11:59 AM of four-day length from May 4 to May 7 at UTC +12 time zone. The reason we use four days data rather than five days due to the time zone difference between UTC +0 and UTC +12; for example, May 7 (Friday) at UTC +0 time zone is May 8 (Saturday) at UTC +12, where the stock market closed and thus no stock market information for the variable representing the index of New Zealand's stock market.

3.2 Time Series Methodology

Unit Root Test

Regressions of interdependent and non-stationary time series may lead to spurious regression (Engle and Granger 1987).To avoid spurious regression, we test the stationarity of dataset and applied the augmented Dickey-Fuller (ADF) test.

Optimal Lag of VAR Model

Whether performing the Granger-causality test, cointegration test, or building up the vector autoregressive model, it is necessary to determine the optimal lag order. VAR model could explain the dynamic characteristics of the model more comprehensively when with a larger value of lag order. However, the degrees of freedom would decrease as the lag order goes up. We use the optimal order according to the sequential modified LR test statistic criterion.

Co-integration Test

The ADF stationarity test shows that the original times series except the sentiment score are not stationary at the level, and therefore OLS cannot be performed directly for coefficient estimation, nor can the Granger-causality test be performed. However, we are able to test if there is a relatively long-term stable relationship among the variables, which can be analyzed if they pass the cointegration test.

VAR Model

As the sentiment score of Bitcoin might be endogenous with the price of other digital currency, we plan to apply a multivariate vector autoregressive (VAR) model as the following:

$$\begin{bmatrix} Ln(Bitcoin)_t \\ Sentiment_t \\ Ln(NZX50)_t \\ Ln(Ethereum)_t \\ Ln(Doge)_t \end{bmatrix} = a_0 + A_1 \begin{bmatrix} Ln(Bitcoin)_{t-1} \\ Sentiment_{t-1} \\ Ln(NZX50)_{t-1} \\ Ln(Ethereum)_{t-1} \\ Ln(Doge)_{t-1} \end{bmatrix} + \cdots + A_k \begin{bmatrix} Ln(Bitcoin)_{t-k} \\ Sentiment_{t-k} \\ Ln(NZX50)_{t-k} \\ Ln(Ethereum)_{t-k} \\ Ln(Doge)_{t-k} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \epsilon_{3,t} \\ \epsilon_{4,t} \\ \epsilon_{5,t} \end{bmatrix}$$

Where a_0 is a vector of interception terms, and A_1 to A_k are three by three matrices of coefficients. $Ln(Bitcoin)_t$ is the Bitcoin price in USD at time t, $Sentiment_t$ is the sentiment score of tweets that mentioned Bitcoin at time t, $Ln(NZX50)_t$ is the log value of NZX50 index at time t, $Ln(Ethereum)_t$ is the price of Ethereum in USD at time t, $Ln(Doge)_t$ is the price of Dogecoin. Then the Granger Causality test is applied to analyze the causality between variables.

Impulse response under VAR model

Cointegration analysis provides the non-short-term relationships between variables, while the impulse response reflects the reaction of endogenous variables to new information. In theory, the impulse response function describes the impact on the current and future values of

endogenous variables after a one-time shock is given to the disturbance term. The impact on a variable can be transferred to other endogenous variables via the dynamic structure of the model.

4. Results

We first conducted the unit root test (ADF) and found that all the variables except sentiment score are not stationary at the level but are all stationary after taking the first differences. According to Table 1, only the sentiment score variable is stationary, and other variables are stationary at the 1st order of differences. As we mentioned, we can only apply the Granger Causality test using stationary time-series data. As a result, we selected the 1st order differences of those variables in the estimation of the VAR model as well as the Granger Causality test.

Table 1: The results of ADF Stationary Test

Variable	No lag order	1st order
ln(Bitcoin)	-1.832	-21.028***
sentiment	-13.100***	
ln(NZX50)	-0.979	-23.797***
ln(Ethereum)	-1.36	-20.914***
ln(doge)	-1.554	-20.367***

Note: ***, **, * represents that significant at 1% level, 5% level,10% level, respectively. The null hypothesis of this test is that the time series has a unit root.

VAR, Granger Causality, and Impulse Response

As we mentioned in Method, impulse response can reflect the response of an endogenous variable to new information and describe the effect of a one-time shock on the current and future values of endogenous variables. Granger causality test in Table y shows that Dogecoin's return is the Granger causality of Bitcoin's sentiment score; The ETH rate of return is the Granger reason

for the Doge rate of return. Therefore, a VAR model can be established for the stable sequence after the 1st order difference according to the ADF test result, and impulse response and variance decomposition analysis can be carried out. The unit root test of the model shows that all the unit roots fall in the unit circle (see Figure A.2 at Appendix), which indicates that the model is stable and can be analyzed by impulse response. Impulse response analysis was performed on the VAR (2) model with multiple variables.

Table 2: The results of Granger Causality Test

Depvar	Lag	d_lbit	sentiment	d_lnzx50	d_leth	d_ldoge	R-sq
				Panel A			
d_lbit	1	-0.0544	-0.0007	0.2198	0.1113	0.0022	
	2	-0.0721	0.0012	0.0281	0.1263	-0.0028	0.0061
Granger Causality		No	No	No	No	No	
				Panel B			
sentiment	1	0.3945	0.3404***	-7.4696	-0.5170	0.4787**	
	2	-1.5250	0.2336***	-13.1825*	2.0581	0.1603	0.2660
Granger Causality		No	No	No	No	Yes	
				Panel C			
d_lnzx50	1	-0.0115	0.0002	-0.1551	0.0056	0.0007	
	2	0.0067	-0.0002	-0.0190	-0.0100	-0.0024	0.0258
Granger Causality		No	No	No	No	No	
				Panel D			
d_leth	1	-0.0639	-0.0026	0.4968	0.1179	0.0068	
	2	-0.0436	0.0039	-0.5045	0.0551	0.0057	0.0160
Granger Causality		No	No	No	No	No	
				Panel E			
d_ldoge	1	-0.5845*	-0.0087	1.8322	0.8963**	0.0160	
	2	0.2941	0.0027	-0.3798	-0.3710	-0.0332	0.0191
Granger Causality		No	No	No	Yes	No	

In Table 2, Panel A-E shows the estimated results of five equations in VAR respectively. Column 1 is the dependent variable in each equation, column 2 shows the lag of order, and columns 3-7 are the independent variables. Column 8 shows the R-square value of each equation.

The last row of each panel shows whether or not the independent variable passed the Granger Causality test. The result of Table 2 shows that none of the selected variables have granger causality with the return of Bitcoin, the return of NZX50, and the return of Ethereum. The only two Granger causality relationships are the return in Dogecoin and sentiment score of tweets, as well as the effect of return in Ethereum and return in Dogecoin. After 1 increase in return of Dogecoin, the sentiment score will increase about 0.48. The increase in Dogecoin return will bring positive comments on Bitcoin. Also, after the increase in the return of Ethereum, the return of Dogecoin will also increase. Figure 3 shows what will happen if there is an impulse in return of Dogecoin. The sentiment score of tweets in Bitcoin will respond after 1 minute and then the effect will stay about 10 minutes. While in Figure 4, the effect of an impulse in return of Ethereum on the return of Dogecoin only exists about 2 minutes. This result could support the hypothesis that return increase in Dogecoin could have positive effects on comments on Bitcoin in Twitter users, while the relationship between the return in Dogecoin and the return in Ethereum should not be described as a causal relationship.

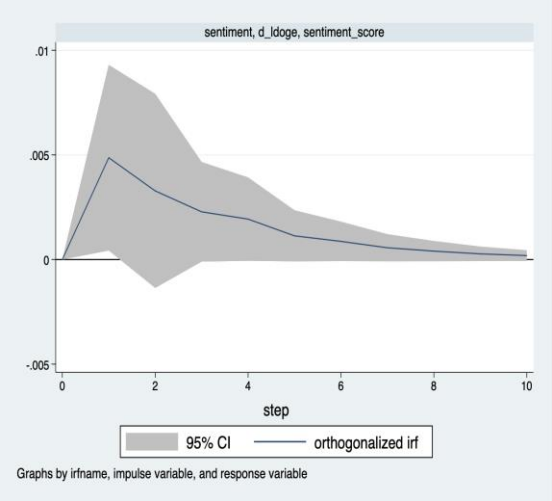

Figure 3: Impulse response of Dogecoin

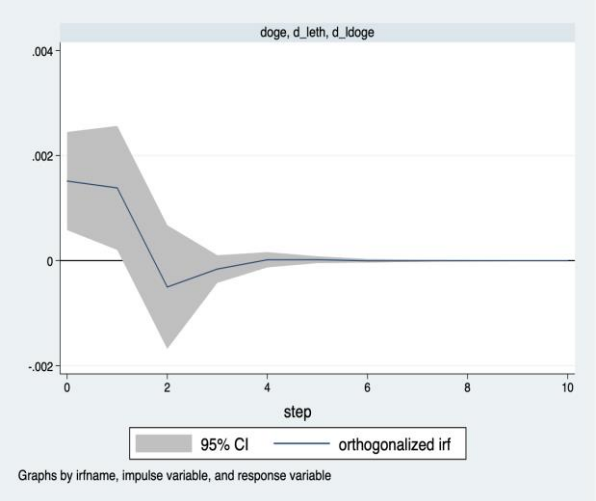

Figure 4: Impulse response of Ethereum

5. Conclusion

In this study, we applied VAR (2) in analyzing the short-term relationship between the cryptocurrency prices with sentiment analysis results in Bitcoin. We found that from May 3rd to May 6th in 2021, The return in Dogecoin has significant granger causality with the sentiment score in Bitcoin, when the return of Dogecoin increase by 1 dollar, the sentiment score in Bitcoin will increase 0.48 after 1 minute, and this effect will last for about 10 minutes. This result supports that the return in Dogecoin may have a positive relationship with the comments of Bitcoin. Also, after returns in Ethereum increase 1 dollar, the return in Dogecoin increases 0.016 dollars after1 minute. However, it only lasts for about 2 minutes. It is more reasonable to conclude that this relationship is just an order of time. And the relationship between returns in Bitcoin, in Ethereum, and in NZX50 do not have significant effects with each other.

In the future, it is valuable to analyze the mechanism of how the return in Dogecoin affects the comment on Bitcoin. Especially, on May 12th, the Bitcoin price dropped a lot after Elon Musk said Tesla Inc has suspended accepting Bitcoin as a form of payment for the purchase of its cars on Twitter. It is interesting to analyze what and how much will change in the return of cryptocurrency prices and sentiment scores.

References

Benaim, Mickael. 2018. “From Symbolic Values to Symbolic Innovation: Internet-Memes and Innovation.” *Research Policy* 47 (5): 901–10. https://doi.org/10.1016/j.respol.2018.02.014.

Ciaian, Pavel, Miroslava Rajcaniova, and d’Artis Kancs. 2016. “The Economics of BitCoin Price Formation.” *Applied Economics* 48 (19): 1799–1815. https://doi.org/10.1080/00036846.2015.1109038.

Engle, Robert F., and C. W. J. Granger. 1987. “Co-Integration and Error Correction: Representation, Estimation, and Testing.” *Econometrica* 55 (2): 251–76. https://doi.org/10.2307/1913236.

Glantz, Morton, and Robert Kissell. 2013. *Multi-Asset Risk Modeling: Techniques for a Global Economy in an Electronic and Algorithmic Trading Era*. Academic Press.

Perry-Carrera, Brian. 2018. “Effect of Sentiment on Bitcoin Price Formation.” Duke University Durham.

Ranganthan, Vishnu Prasad, Ram Dantu, Aditya Paul, Paula Mears, and Kirill Morozov. 2018. “A Decentralized Marketplace Application on the Ethereum Blockchain.” In *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, 90–97. Philadelphia, PA: IEEE. https://doi.org/10.1109/CIC.2018.00023.

Yermack, David. 2013. “Is Bitcoin a Real Currency? An Economic Appraisal.” w19747. National Bureau of Economic Research. https://doi.org/10.3386/w19747.

Appendix

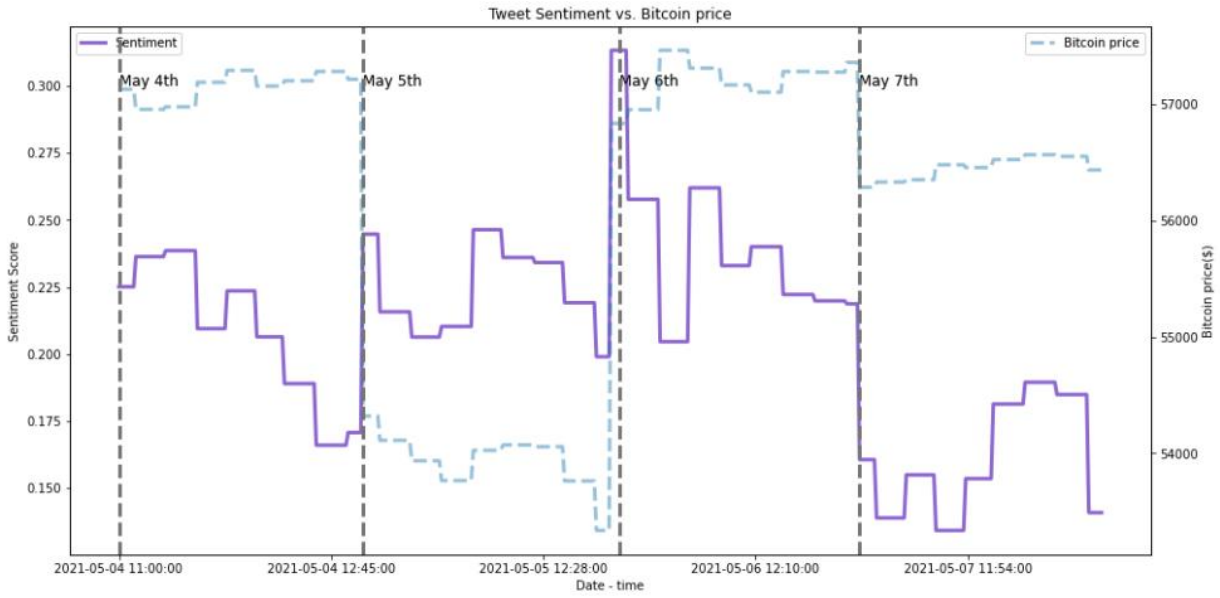

Figure A.1 plots the sentiment scores of Bitcoin with the Bitcoin price ($)

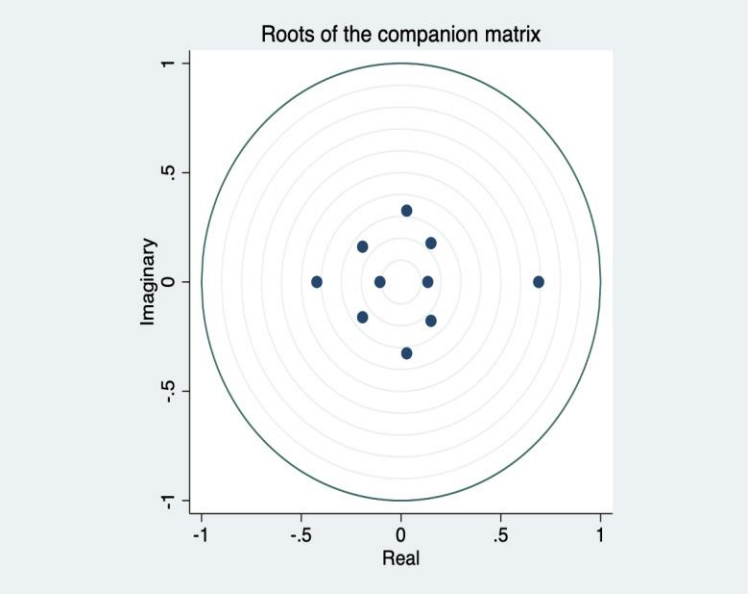

Figure A.2 Unit Root Circle