

Do Tweets Predict Cryptocurrency Price Movements?

Stephen Lee

2018

1 Introduction

In 2008, the Bitcoin white paper was published under the pseudonym Satoshi Nakamoto. The paper combined cryptography with a game theoretic incentive structure to provide secure peer-to-peer financial transactions without needing a trusted 3rd party. Since then, other projects have modified the Bitcoin rules to create new protocols for handling these transactions. Colloquially referred to as “cryptocurrencies”, these projects have captured the imagination of many. As of May 3, 2018, the three largest cryptocurrencies by market capitalization are Bitcoin, Ether, and Ripple. This paper explores the price movements of these coins in addition to a simple variant of Bitcoin called Litecoin.

Using trade level data from a European exchange over the dates August 2017 to April 2018, I previously explored the possibility of cointegration and Granger causality - however results so far are inconclusive. Interestingly in the dataset, a massive price bubble appeared from roughly December to February. While a more robust analysis is needed, the data does seem to be seperable into three distinct periods: pre-bubble, bubble, and post-bubble. Blah blah blah about how cointegration seems possible in post-bubble, signifying possible increase in investor awareness.

Here, I study if Twitter “tweets” can be predictive of price movements in the first two weeks of December (i.e. as the bubble was forming).

2 Data

2.1 Cryptocurrency Transactions

Price data comes from the European coin exchange Bitstamp. In it’s raw form, the data includes information about each transaction in the history of the exchange, including the unix-timestamp¹, trade price, and trade quantity. After

¹The unix-timestamp is given by the number of seconds that have passed since January 1, 1970 at 12:00:00 am GMT

cleaning, the final series consists of hourly periods ranging from Thursday, August 17, 2017 2:00 pm to Tuesday, April 24, 2018 4:00 am GMT.

Graphically, the final time series are as follows:

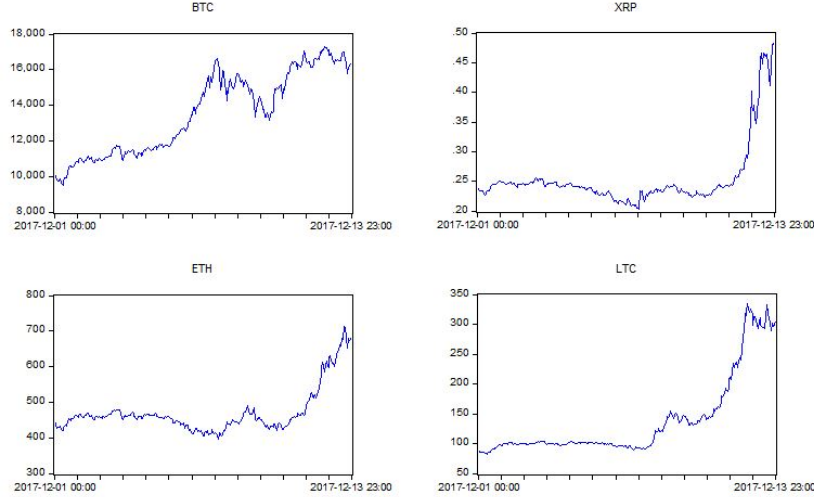


Figure 1: Time series of Bitcoin, Ether, Ripple, and Litecoin prices every hour.

2.2 Tweets

I scraped Twitter’s database for every “tweet” that mentioned the word Bitcoin from December 1, 2017 00:00, to December 13, 2017 23:00. This required overcoming several challenges: 1) Twitter’s official programming interface places strict limits on data accessibility - namely, you are restricted to 150 page requests every 15 minutes, and also you can only access tweets from within the last 7 days; 2) local storage and computational power (i.e. on my laptop) are insufficient for large scraping projects that may need to run for several days; and 3) this raw Twitter data needs to be manipulated and merged with my existing hourly price and volume data in a way that “best captures” the tweets.

Thus, to obtain usable data, I modified an open source project that bypasses the limitations of the official interface.² Specifically, this project reverse engineers the behavior of the official Twitter access interface by utilizing the same URL and query structure. Additional considerations were made to respect Twitter server’s capacity constraints by placing one second between subsequent data requests as per their specs.³ Note, however, that due to Twitter’s server protection, each web scraper I launched was essentially banned after scraping about a day’s worth of Tweets (i.e. about 72,000 tweets). Because of this, a new web scraper and IP address were needed for each day’s worth of tweets. Even after

²Code available here: https://github.com/slee981/TwitterSearch_API.

³Best practices for Twitter webscraping are described here: <https://twitter.com/robots.txt>.

trying to run a new scraper repeatedly over the same day and remove duplicates in the data cleaning phase, there are several stretches of missing information - presumably because of Twitters database blocking the requests.

Next, in order to scrape and store large quantities of data, I created new cloud based computational (Elastic Cloud Compute, EC2) and storage (Relational Database Service, RDS) instances using Amazon Web Services (AWS).

Finally, I re-grouped the data into an hourly timeseries with counts for the total number of tweets, the total retweets, and the total number of favorited tweets. The result was a “mostly complete” dataset that consisted of the number of tweets that mentioned the word Bitcoin every hour, corresponding to the same hour as the price and quantity series. As mentioned above, however, there is some missing data. The final series graph is as follows:

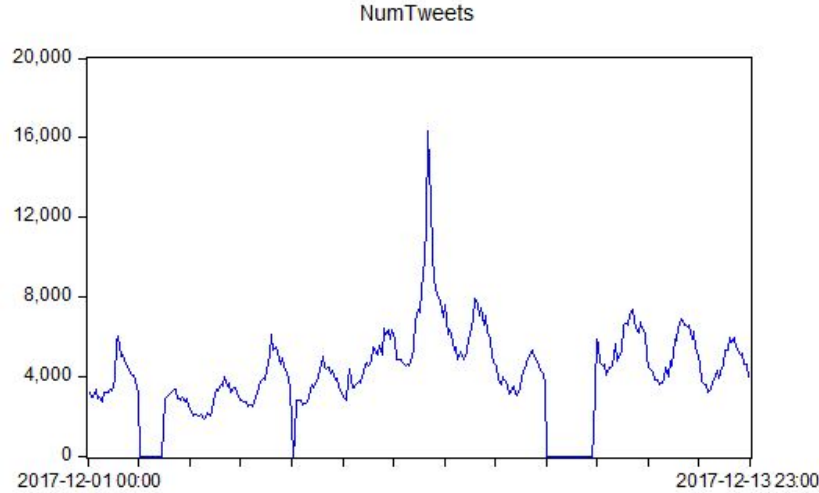


Figure 2: Time series of the number of tweets that mention Bitcoin per hour.

3 Analysis

The following is based on a first glance analysis of the data.

First, I regressed the change in BTC price from the previous hour on the number of tweets that mention Bitcoin and find the following results. Note that I add the previous hours change in price to a basic ARIMA(1,1) model based on the initial correlogram of the number of tweets series.

Dependent Variable: NUMTWEETS
Method: ARMA Conditional Least Squares (Gauss-Newton / Marquardt steps)
Date: 07/19/18 Time: 21:46
Sample (adjusted): 12/01/2017 02:00 12/13/2017 23:00
Included observations: 310 after adjustments
Failure to improve likelihood (non-zero gradients) after 9 iterations
Coefficient covariance computed using outer product of gradients
MA Backcast: 11/30/2017 23:00

Variable	Coefficient	Std. Error	t-Statistic	Prob.
BTC(-1)	0.280759	0.040046	7.010860	0.0000
AR(1)	0.913618	0.025248	36.18623	0.0000
MA(1)	0.159140	0.060297	2.639289	0.0087
R-squared	0.892019	Mean dependent var		4102.639
Adjusted R-squared	0.891315	S.D. dependent var		2231.102
S.E. of regression	735.5352	Akaike info criterion		16.04870
Sum squared resid	1.66E+08	Schwarz criterion		16.08486
Log likelihood	-2484.549	Hannan-Quinn criter.		16.06316
Durbin-Watson stat	1.983757			
Inverted AR Roots	.91			
Inverted MA Roots	-.16			

This yields the following ACF and PACF and suggests that, with statistical significance, the more the price of Bitcoin changes, the more tweets about it you can expect in the next hour.

Date: 07/19/18 Time: 21:47
Sample: 12/01/2017 00:00 12/13/2017 23:00
Included observations: 310
Q-statistic probabilities adjusted for 2 ARMA terms















































Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob*
		1	0.006	0.006	0.0102
		2	0.045	0.045	0.6379
		3	0.027	0.026	0.8639
		4	0.008	0.006	0.8853
		5	0.048	0.046	1.6249
		6	0.047	0.045	2.3262
		7	-0.029	-0.034	2.5912
		8	-0.000	-0.007	2.5912
		9	-0.056	-0.056	3.5944
		10	-0.066	-0.068	5.0085
		11	-0.140	-0.141	11.335
		12	-0.095	-0.091	14.287
		13	0.007	0.023	14.303
		14	-0.067	-0.050	15.767
		15	-0.040	-0.025	16.281
		16	-0.001	0.022	16.281
		17	0.005	0.031	16.288
		18	-0.016	-0.016	16.368
		19	0.031	0.023	16.679
		20	0.035	0.030	17.076
		21	0.077	0.051	19.051
		22	0.082	0.050	21.291
		23	-0.024	-0.058	21.490

Figure 3: ACF and PACF of residual series from above regression.

Similarly, based on the correlogram of the differenced BTC price series (i.e. the change in price), I added the number of tweets that to an AR(3) model and also find significance. This suggests that as more people talk about Bitcoin on twitter, the more the price increases.

Dependent Variable: D_BTC
Method: ARMA Conditional Least Squares (Gauss-Newton / Marquardt steps)
Date: 07/19/18 Time: 21:53
Sample (adjusted): 12/01/2017 04:00 12/13/2017 23:00
Included observations: 308 after adjustments
Convergence achieved after 6 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
NUMTWEETS	0.005874	0.002398	2.449519	0.0149
AR(3)	-0.134451	0.056598	-2.375554	0.0181
R-squared	0.025474	Mean dependent var		21.40503
Adjusted R-squared	0.022290	S.D. dependent var		224.6109
S.E. of regression	222.0936	Akaike info criterion		13.65055
Sum squared resid	15093618	Schwarz criterion		13.67477
Log likelihood	-2100.184	Hannan-Quinn criter.		13.66023
Durbin-Watson stat	1.866333			
Inverted AR Roots	.26+.44i	.26-.44i	-.51	

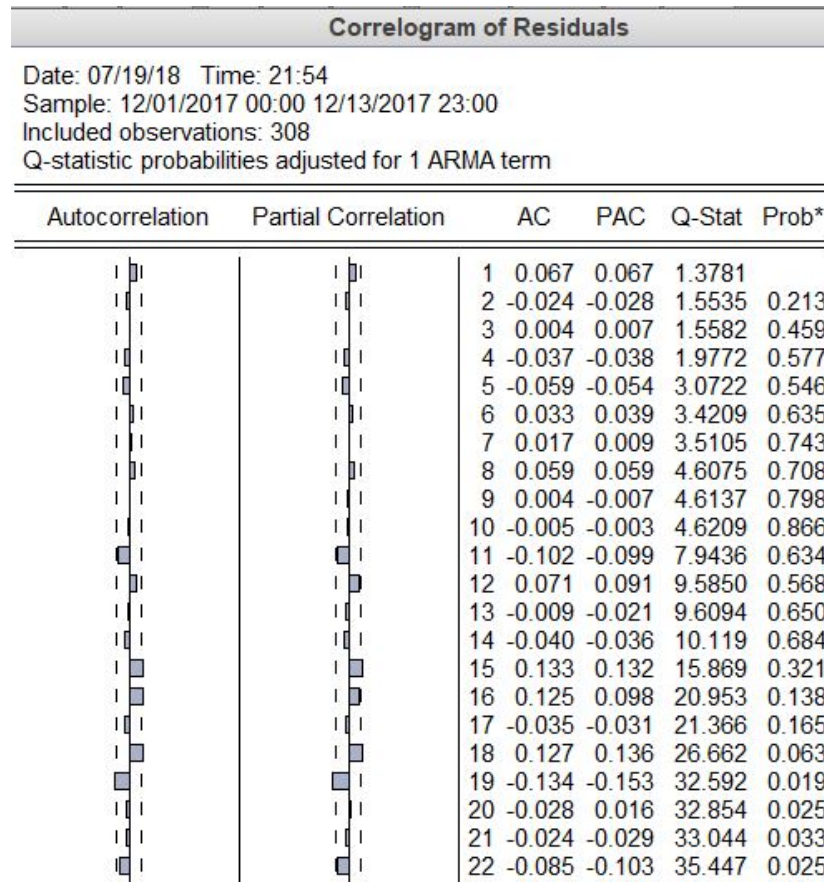


Figure 4: ACF and PACF of residual series from above regression.

4 Conclusion

While this is a VERY preliminary and crude look into the data - the appearance of statistical significance seems worth exploring further.