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-timestamp[1], trade price, and trade quantity. After 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.

---

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

2.2 Tweets

I scraped Twitter's database for "tweets" that mentioned the word Bitcoin from December 1, 2017, to December 14, 2017. 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]

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.

3   Analysis

4   Conclusion

---

[2]Code available here: https://github.com/slee981/TwitterSearch_API.
[3]Best practices for Twitter webscraping are described here: https://twitter.com/robots.txt.