Reddit cryptocurrency comments analysis

Albin Aliu
albin.aliu@unifr.ch

François-Xavier Wicht
francois-xavier.wicht@unifr.ch

Grégoire Rebstein
gregoire.rebstein@unifr.ch

ABSTRACT

In this work, we make an attempt to analyse ten cryptocurrency subreddits to yield results and conclusions. We discovered that the discussion in those subreddits attract a long-term attention from different users. Also, individuals tend to gather together with an anchor in a particular subreddit but also spread over the other nine subreddits. Furthermore, there seems to be a correlation between the activity in said subreddits and Bitcoin price movement and we managed to highlight it over a comprehensive process.

Author Keywords

Reddit; Cryptocurrency; Analysis; Bitcoin Price; PageRank; Louvain; Correlation.

INTRODUCTION

Since the rise of cryptocurrencies, a lot of research have attempted to predict their price rates [6, 2, 8]. Due to the transparency of cryptocurrency transactions, researchers have considered using transaction information such as overall trends and cyclical changes to predict cryptocurrency prices [7].

In this project, we attempt to find a link between the activity in cryptocurrency subreddits and the price of Bitcoin. For that purpose, we scraped data directly from said subreddits, structured data in a network architecture and performed several analysis such as PageRank, Louvain community detection and correlation analysis.

This report is structured the following way. Firstly, we discuss the scraping methodology we used, secondly the data models we tried and considered for the analysis and thirdly we outline the analysis of our model. As a closing word, we highlight trials and errors during this project.

DATA COLLECTION

Reddit is a social news aggregation composed of different communities called subreddit, where people gather to discuss and comment on a specific subject. Each subreddit is composed of main posts which are in turn commented and voted either positively (upvoted) or negatively (downvoted). Each comment in turn can also be commented and voted upon which sometimes create a cascade of recursive comments as shown on figure 1. To capture these information and later work on it, we use the official Reddit API. To facilitate our work, we used the wrapper PRAW (Python Reddit API Wrapper) library.

To facilitate this data collection, we automated periodic calls to the API in form of a bot. Our bot runs on a linux server and checks all 20 minutes for new posts on specific subreddits. For this project, we used the following list of subreddits: Bitcoin, Ethereum, Dogecoin, BitcoinBeginners, CryptoCurrencies, CryptoTechnology, CryptoMarkets, Binance, CoinBase, btc. Using the multiprocessing package of python, we spawn a process for each subreddit for the scraping to happen in parallel. Also, since we care about actual and real time data, we consider the following sorting options provided by the Reddit API: hot posts, top posts and new posts. We don't set some time bound on when the post has been published, since we store the timestamps of each post and comment and filter what we need. This works really well and we didnt encounter any API limitations.

Everytime a new scraping run is started, we create a new JSON file with the timestsamp of the start of the scraping. All new posts that will be found in this run, will be stored inside the newly created JSON file. As time goes by, a post might receive many new comments. Therefore, if a previously crawled post has an increase in comments of more than 30%, we also refetch that post and update the JSON file it was originally stored in. Once we have a certain amount of raw data, we need to preprocess this data and create data models to ease the analysis. This the object of the next section.

DATA MODELS

In this section, we investigate the four data models we have designed. The main question that we tried to answer by modeling our data is what an edge should represent. Naturally, when an user opens some post on Reddit, they might start reading the first comment. If it doesn't catch their attention, they read the second top comment. They even might continue reading the reply tree to that comment. The nature of this is that the replies are very often context based. To understand the comment of depth 3, you might need to also have read the comments of depth 2 or 1. For this reason we assume that if an user replies to some comment, they have a connection to all the people in that chain of comments. We consequently opted for four different data models: the Unique Cartesian Link, the Deep Link, the Next Link and finally the Deep Link No Merge model. In the following subsections, we present subsequently these four models.

Unique Cartesian Link

The Unique Cartesian Link model is probably the most natural way to create an undirected graph from our data. The main concept is the following. Every user that commented on a post get linked together in a complete graph with edge weight 1. If there is already an edge between two users, the weight of the edge gets increased by 1. This creates an enormously large graph compared to our other data models. This makes computation on the graph very difficult and slow. Also, it doesn't really represent our idea of a relationship between two users. That's why we opted for other models.

User 0 (OP, depth=1, score=30)
User 1 (d=1, s=10)
User 2 (d=2, s=10)
User 3 (d=2, s=2)
User 4 (d=1, s=5)
User 5 (d=1, s=3)
User 6 (d=2, s=2)
User 3 (d=3, s=1)
User 7 (d=2, s=1)

Table 1. A typical Reddit post, where an indentation represents a reply to the above comment.

Deep Link

In the Deep Link model, we connect each user to the original poster (OP) and to all users who replied to the user's comment. The edges get a weight depending on the depth of the comment and the achieved upvote score. If there is already an edge between the two users, the weight gets increased. This is a directed graph. See Figure 1 and the corresponding table.

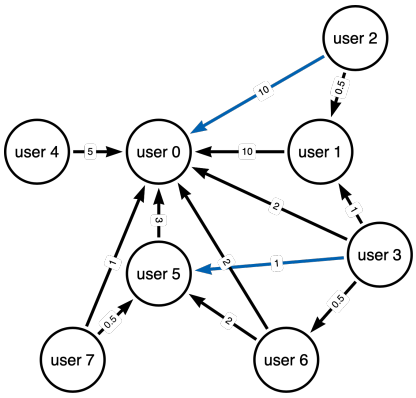

Figure 1. The graph representing the above post. The weight is computed as the score of the comment of user$_i$ divided by the depth of the user$_j$, who is above the user$_i$.

Next Link

Regarding the Next Link model it is highly similar to the Deep Link one, except that we only connect each user to the OP and the immediate users who commented directly on the user's comment. Thus the blue link in Figure 1 are missing in this data model. This model keeps the direction of the graph.

Deep Link No Merge

The Deep Link No Merge is pretty similar to the Deep Link one but here we don't merge the weights and allow multiple edges from one user to another. Thus, it gives rise to a directed multigraph. We found that this approach has still lots of information but is not too expensive to compute. Consequently, this is the privileged model for the analysis in the next section.

ANALYSIS AND VISUALISATION

In this section we pursue with the analysis of data scrapped on several cryptocurrency subreddits. More precisely, we use social media analytics on the network that we built in the previous section. Firstly, we investigate if our network follows a power law distribution as it should. Secondly, we discuss the results of the PageRank algorithm on the network. Thirdly, we run Louvain community detection algorithm and compare the results with the existing subreddits. And finally we analyse the price development of the Bitcoin and try to find a correlation with our data.

Power Law Distribution

The power law distribution in a network can be understood as "a few users are very popular and a lot of users are mildly or not popular at all". Intuitively, this relation holds in our network composed of vertices as users and edges representing the comment relation. Indeed, when browsing Reddit in general, we may notice that a few comments (users) are very popular (have a very high score) as others are not. This intuition reveals itself true when plotting the node degree distribution (see fig. 2). This is encouraging in relation to our network model and further analysis because natural networks tend to follow that law. It therefore means that our network is properly formed and that the analysis that follows is most likely consistent in regards to natural laws.

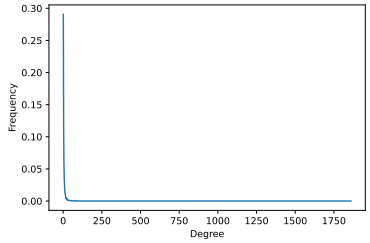

Figure 2. The network follows a power law distribution: a lot of users have few comments on their posts and a few users have a lot of comments on their posts.

Page Rank

Regarding PageRank (PR), it can be interpreted as the amount of random walks in the network that end up on a particular vertex. For our network, it means that the higher the rank of one user is, the higher the probability is of commenting one post of that particular user. We decide here to investigate the steadiness of PR over 15 days. We might observe unsteadiness on figure 5 but this is not entirely true with respect to the PageRank. First, this is true that the graph evolves quite rapidly over the days. The same users will not necessarily be active twice on the same comment over multiple days. This observation has been done by computing the set of users over the days and devising the similarity for each successive day. Let U_i be the set of users for day i, then $s(i,j)$ is the similarity between day i and j such that

$$s(i,j) = \frac{U_i \cap U_j}{U_i \cup U_j}.$$

We get the graph at figure 3. However, the PR itself seems to be pretty consistent over the days when we compute the sum of the squared error between the intersected PR values as shown in graph 4 where the error only spikes to a value of 0.005. We can conclude from these two observations that, although there are few overlapping users over the days, the

ones that overlap tend to be very popular for a long period. This also means that in the cryptocurrency subreddits, posts tend to have a lasting interest for the users. This conclusion is consistent with the paper "Characterizing Speed and Scale of Cryptocurrency Discussion Spread on Reddit." where they state that cryptocurrencies discussion are likely to trigger a longer lasting discussion.

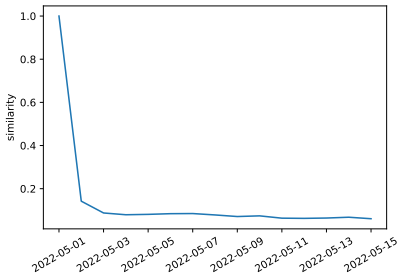

Figure 3. Similarity between user sets over the 1st to the 15th May.

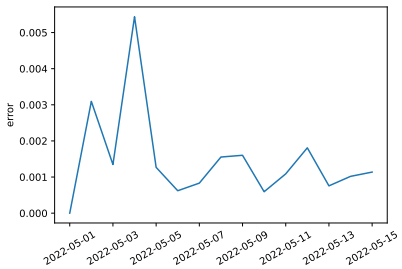

Figure 4. Sum of the squared error of PageRank values over intersected user sets between the 1st and the 15th May.

Louvain

Louvain method is a technique that find communities from large networks. In our case, we were interested to see the communities of our network and observe the links with the different subreddits. To do this, we first create a list of non overlapping communities corresponding to each subreddit (see figure 6). Each person is assigned to a single community, the one where she has interacted the most.

We have observed that 90% of the users post in a single subreddit. For the remaining 10%, on average they all have a main subreddit on which they are most active. So, building communities by associating each person where they interacts the most seems to be a good solution.

Then, we create a second list of communities by applying the Louvain method on the graph based on the *Deep Link No Merge* data model. After this, we get a list of 260 communities. Knowing that there are 10 subreddits, one subreddit can therefore have several inner communities. However, not all of these communities are relevant. Indeed, most of them are groups of 2 people. When we look at the graph, 30% of the nodes have only one edge and 50% have less than 3 edges. This may means that most of the users do not necessarily post a lot and this explains the number of small communities found by Louvain method. So, we filter the Louvain communities to

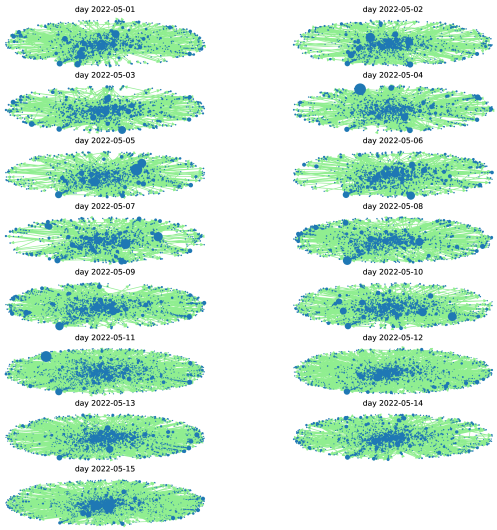

Figure 5. Networks between the 1st to the 15th May. The size of the nodes is relative to the PageRank values.

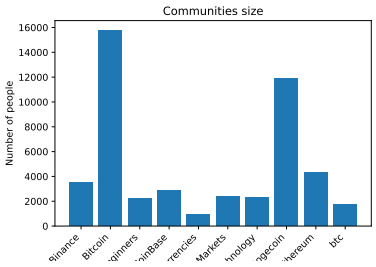

Figure 6. Community size for each subreddit.

keep those with more than 5 people. We then go from 258 to 53 communities.

After that, it can be interesting to know if some communities are formed by people from different subreddits. For this, we create the histogram of the number of different subreddits that people in a Louvain community belong to (see figure 7).

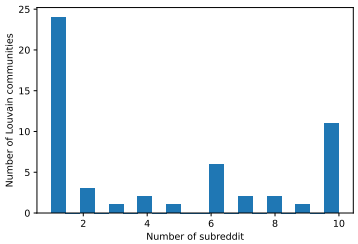

Figure 7. Repartition of Louvain communities in the subreddits.

With this histogram, we can observe that for the majority, the Louvain communities are formed by people from the same subreddit. However, there is also some Louvain communities that are not restricted to one subreddit but are even made up of people from all 10 subreddits. In addition, as we might expect, the larger a community is, the more it seems to interact on several subreddits (see figure 8).

Consequently, communities tend to spread over multiple subreddits but the majority of each community has a foot in one unique subreddit. This can be explained by the tree-like topology of our graph. Indeed, the commenter forms the root and the subcommenters the children until the last commenters being leafs. Those trees have a foothold in one particular subreddit and only few elements are shared among other posts and even less among other subreddits.

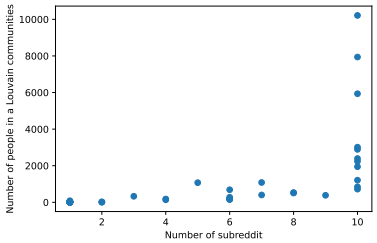

Figure 8. Repartition of Louvain communities in the subreddits according to size.

Correlation to price development

As for the correlation between subreddits activity and Bitcoin price development, we grouped our data by some time windows of one hour. We create a graph G_0 from the first time window. Then we create a second graph G_1 from the first union second time window. We continue until we have an inclusion sequence such that

$$G_0 \subset G_1 \subset ... \subset G_{n-1},$$

where n is the number of time windows, and G_i a graph based on the *Deep Link No Merge* data model. Then, we use the degree centrality to compare each graph with its subsequent graph. This results into a table where we have for each user (node) the development of its activity (degree). With this huge dataset, it is now possible to track users and also identify the most active users. However, since this requires lots of computation power and much more work to get something meaningful from this, we opted for the option to sum each time frame (e.g. 1 hour in our case).

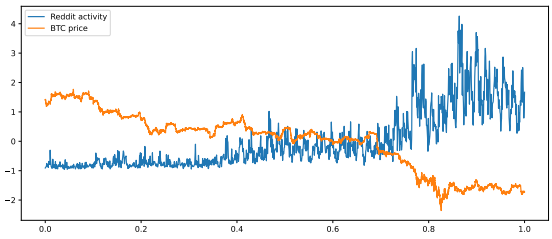

Figure 9. The Reddit activity and the BTC price, normalized, from 01.04 to 22.05.2022.

We're now at that stage where one could run some correlation method on our summed activity dataset against the price data. However, this will not yield good results as both datasets are very volatile. Therefore we needed to smoothen our data somehow. We opted for the moving window average method. We took both datasets and computed moving window averages with step 2, 3, 4, ..., 30. These steps represent the number of columns that will be averaged.

Then, we compared each activity moving window average for each price moving window average using two of pandas built-in correlation methods, namely Pearson, Spearman and we also tried the The Hilbert–Schmidt independence criterion (HSIC). The first two methods deliver strong correlation coefficients (around -0.85), while the HISC does not. Also, we noticed that the correlation differs depending on how much days (from when to when) we try to correlate. We've found that more than 10 days usually yields better results. Please refer to the interactive notebook for details.

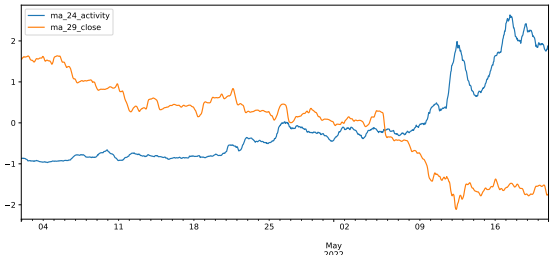

Figure 10. Moving averages of the BTC price and the Reddit activity. Pearson Correlation yields a value of -0.889.

However, it is important to remember that a correlation between two sets of data does not necessarily indicate a causal link. Here we rely on the results given by paper "Extracting Cryptocurrency Price Movements from the Reddit Network Sentiment." and assume that there is a causal link. Wooley et al. have indeed been able to prove a causal link between price movements and Reddit activity notably with bivariate Granger causality tests. Furthermore, the results shown above are consistent with "Cryptocurrency Return Prediction Using Investor Sentiment Extracted by BERT-Based Classifiers from News Articles, Reddit Posts and Tweets." where the authors are able to predict a future price changes with an accuracy

significantly better than random guesses [5]. Moreover, we only considered Bitcoin price movement to the detriment of other Cryptocurrency. This probably has an impact on our conclusions as other cryptocurrencies might behave differently. Wooley et al. shows us that this intuition is wrong by demonstrating that Cryptocurrencies follow more or less the same trends and therefore similar movements.

HOMEMADE IMPLEMENTATIONS

In this section, we discuss the two homemade implementations of the algorithm used in this project: PageRank and Louvain community detection.

PageRank is an iterative process that can potentially last forever. As a stopping condition we put a number of iterations as well as an error threshold. If the error does not exceed a certain threshold over one iteration, then the algorithm stops. If it does not, then the algorithm runs over a fixed number of iteration. This strategy is very common in the various implementations and can be seen in "NetworkX". This made it quite convenient to compare and benchmark our implementation. As a result, we discovered that our own implementation yields the same result but is much less optimised in terms of performance. Indeed, "NetworkX"'s implementation is more than 100 times faster and takes more than 400 times less RAM as it can be seen on the following table. This result is clearly surprising but only because the in-house implementation uses a dense matrix instead of a sparse matrix. This has the effect of using less RAM and significantly speeding up the calculations. With the use of a sparse matrix in our implementation, the computation takes 2s with a RAM consumption of about 0.2G. Other optimisations could then be made but the density of the matrix is really the core issue here. This is mainly explained by the Power Law distribution. Since a lot of vertices have very few edges, a lot of rows in the adjacency matrix have a lot of zeros. With a sparse matrix, we avoid storing all of these zeros and can therefore considerably speed up the computation.

implementation	nodes	edges	time	RAM
"NetworkX"	15537	58150	0.78s	0.128G
homemade (dense)	15537	58150	91.21s	5.81G
homemade (sparse)	15537	58150	2.18s	0.215G

Table 2. Benchmarks of PageRank algorithm on Fedora Linux 36 with Intel i7-8550U (8) @ 4.000GHz and 16GB of RAM

Louvain is an iterative algorithm that works in series of 2 steps until convergence. The first one is to assign every node to a community based on the modularity gain. The second step consists in building a new graph by transforming the communities into hypernodes. When implementing Louvain method, we took a small graph to ensure that the implementation was correct. However, when we tried to apply it on the graph based on *Deep Link No Merge*, our implementation was too slow and we stopped the process after 10min without having a result. When comparing with the implementation of "NetworkX", we see that it is much faster and ends in a few seconds with the graph *Deep Link No Merge*. Indeed, they do some optimizations and in particular they precalculate the gain of modularity. The computation of modularity gain is the most called instruction in the algorithm and it seems that it is the main reason why our implementation is too slow. Therefore, we used only a subset of the full graph to yield some results. Again, it shows significant difference bewtween our and "NetworkX"'s implementation. In any case, it is a successful implementation and still useful to understand the flow of the algorithm. On the other hand, other trials and errors may have been encountered during the project. This is the subject of the next section.

implementation	nodes	edges	time	RAM
"NetworkX"	1000	4852	0.37s	3.478MB
homemade	1000	4852	18.76s	3.235MB

Table 3. Benchmarks of Louvain algorithm on macOS Monterey with Apple M1 and 8Go of RAM

TRIALS AND ERRORS

This section briefly explains the original idea of this project and why and how we deviated from it. We begin by explaining why we switched from social network Twitter to Reddit and subsequently explain the consequences it has had.

The goal of this project was first to perform sentimental analysis (SA) in real time on cryptocurrencies. For that purpose, Twitter was taken as a network source both to discover communities centered around cryptocurrencies and to retrieve text data for SA itself. Influencer accounts (e.g. @Bitcoin, @ethereum, @BTCTN,etc.) as well as hashtags would have been taken as starting points to build the network and the analysis. SA would then have been done both inside and outside communities to deviate correlation and interesting results. Consequently, the end result would most likely have been the currencies' rates with key points (e.g. extrema) that could reveal both SA (in- and outside communities). However, we soon ran into a wall because of the Twitter API limitations, which made the scrapping nearly impossible. We soon thereafter switched to Reddit whose API has much friendlier limitations. The problem we had after that was to properly perform SA. Yet, Reddit is more comments-rich than posts-rich, which brings the problem of having off-topics comments that could bias the results. We soon realised that performing SA on Reddit was a tedious task and we therefore put the focus on a global analysis of the network we built.

CONCLUSION

In conclusion, during this work we have collected data from 10 cryptocurrency-related subreddits and have performed analysis on them. We first have modeled the data and have concluded that the one that best fits our need was the Deep Link No Merge model. We pursued with the analysis and proved that our data followed a Power Law distribution. With help of PageRank, we showed that the discussion on these subreddits tend to have a lasting interest. Furthermore, with Louvain, we showed that each community detected spread over multiple subreddits. However, the majority of each community has a foothold in one unique subreddit. Correlation between subreddits activity and Bitcoin price movement was later revealed with a value of -0.85. We also found that the more days were analysed the better results it yielded. All in all, the results that were produced in this work are consistent with related ones. Finally, a dedicated work could be done on this subject to further discover relations and relevant results.

References

[1] Andreas M. Antonopoulos. *Mastering Bitcoin. Programming the Open Blockchain*. 2nd ed. O'Reilly, 2017. ISBN: 978-1-4919-5438-6. URL: http://gen.lib.rus.ec/book/index.php?md5=a9bf133f45fdfd9bea6f792a66ca2a83.

[2] Luisanna Cocco, Roberto Tonelli, and Michele Marchesi. "Predictions of Bitcoin Prices through Machine Learning Based Frameworks." In: *PeerJ Comput. Sci.* 7 (2021), e413. DOI: 10.7717/peerj-cs.413. URL: https://doi.org/10.7717/peerj-cs.413.

[3] Maria Glenski, Emily Saldanha, and Svitlana Volkova. "Characterizing Speed and Scale of Cryptocurrency Discussion Spread on Reddit." In: *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. 2019, pp. 560–570. DOI: 10.1145/3308558.3313702. URL: https://doi.org/10.1145/3308558.3313702.

[4] Aric A Hagberg, Daniel A Schult, and Pieter J Swart. "Exploring Network Structure, Dynamics, and Function Using NetworkX". In: (2008), p. 5.

[5] Duygu Ider. "Cryptocurrency Return Prediction Using Investor Sentiment Extracted by BERT-Based Classifiers from News Articles, Reddit Posts and Tweets." In: *CoRR* abs/2204.05781 (2022). DOI: 10.48550/arXiv.2204.05781. URL: https://doi.org/10.48550/arXiv.2204.05781.

[6] Eunho Koo and Geonwoo Kim. "Prediction of Bitcoin Price Based on Manipulating Distribution Strategy." In: *Appl. Soft Comput.* 110 (2021), p. 107738. DOI: 10.1016/j.asoc.2021.107738. URL: https://doi.org/10.1016/j.asoc.2021.107738.

[7] Panpan Li et al. "Cross Cryptocurrency Relationship Mining for Bitcoin Price Prediction." In: *CoRR* abs/2205.00974 (2022). DOI: 10.48550/arXiv.2205.00974. URL: https://doi.org/10.48550/arXiv.2205.00974.

[8] Xiao Li and Linda Du. "A Multi-window Bitcoin Price Prediction Framework on Blockchain Transaction Graph." In: *Algorithmic Aspects in Information and Management - 15th International Conference, AAIM 2021, Virtual Event, December 20-22, 2021, Proceedings*. 2021, pp. 317–328. DOI: 10.1007/978-3-030-93176-6_27. URL: https://doi.org/10.1007/978-3-030-93176-6_27.

[9] Stephen Wooley et al. "Extracting Cryptocurrency Price Movements from the Reddit Network Sentiment." In: *18th IEEE International Conference On Machine Learning And Applications, ICMLA 2019, Boca Raton, FL, USA, December 16-19, 2019*. 2019, pp. 500–505. DOI: 10.1109/ICMLA.2019.00093. URL: https://doi.org/10.1109/ICMLA.2019.00093.

[10] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social Media Mining: An Introduction*. New York, NY: Cambridge University Press, 2014. 320 pp. ISBN: 978-1-107-01885-3.