Reddit Cryptocurrency Comments Analysis

Albin Aliu, François-Xavier Wicht & Grégoire Rebstein

May 31, 2022

Social Media Analytics

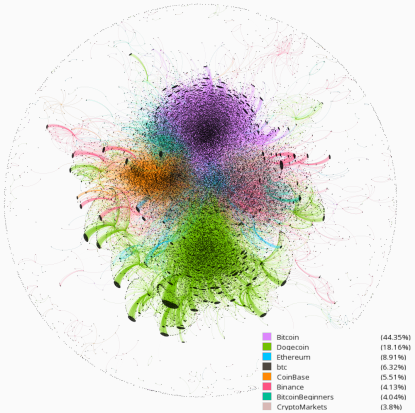
Agenda

1. Overview
2. Trials and Errors
3. Data collection
4. Data Models
5. Analysis
6. Homemade implementation
7. Discussion

Overview

Overview

- The main idea is to find a link between the activity in cryptocurrency subreddits and the price of **Bitcoin**.
- For that we modeled the data in an appropriate way and ran several algorithms seen in class.



Trials and Errors

Trials and Errors

The original idea was to perform Sentimental Analysis (SA) on Tweets.

- The Twitter API is unfortunately very limited.

$\Longrightarrow$ We took Reddit as a Social Network instead.

- Reddit is less suitable to perform Sentimental analysis.

$\Longrightarrow$ We decided to analyse the network and find interesting results as well as correlation between subreddit's activity and Bitcoin price.

Data collection

Data collection

- We used Reddit API for data collection.
- API calls are performed by a bot running every 20 minutes on several Cryptocurrency related subreddits.
- Storage is optimal in a JSON format because of Reddit comment structure.

Data Models

Data Models

We have tried four types of data models.

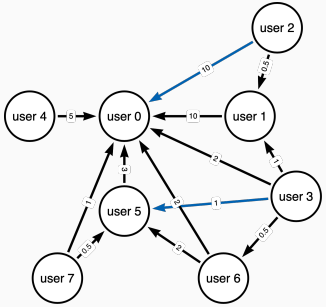
1. Unique Cartesian Link: all commenters from a specific post are linked together with an undirected edge.
2. Next Link: only main commenters of the post are linked to the Original Poster (OP) with a directed edge.

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)

Data Models

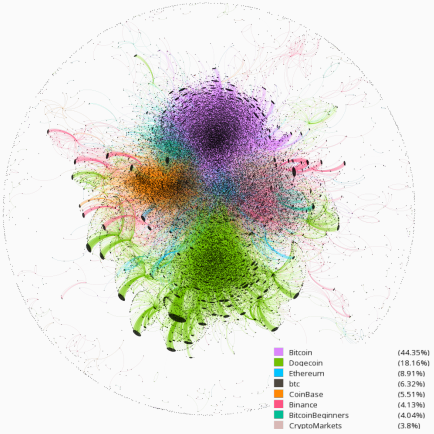
3. Deep Link: commenters are recursively linked together with a directed edge (i.e. capturing the relation "has commented") with a weight of relevance according to the depth of comments.
4. Deep Link No merge: same as Deep Link without merging the different edges. This results in a directed multigraph.

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)



Data Models

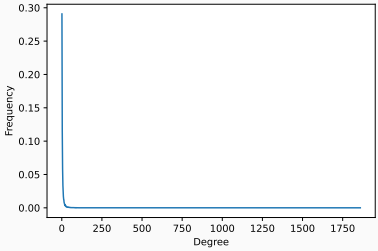
- Deep Link No merge model represented on the right.
- Tree-like structure may be observed for large comments.
- Most imposing subreddits are Bitcoin, Dogecoin and Ethereum.



Analysis

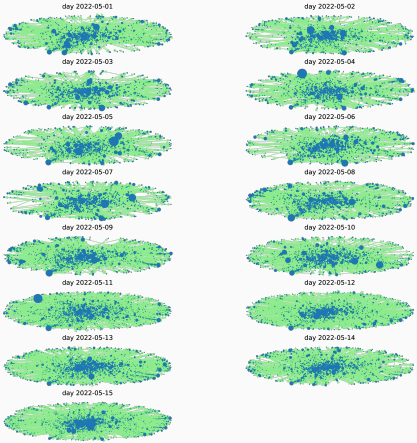
Power law distribution

- The graph data model follows a Power law distribution.
- This most likely means that the data model is well thought out.



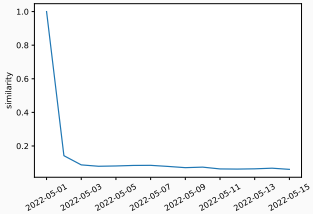
PageRank

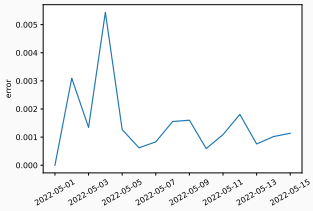
- The PageRank in the network can be interpreted as being the probability of commenting one user's post when browsing the given subreddits.
- Yet, the PageRank alone does not yield very surprising results: users with the most upvoted and commented comments have the highest PageRank.



PageRank

- We therefore took interest in evaluating the **steadiness** of the PageRank along the days.
- We concluded two things.
  1. There is a tiny similarity between user bases along the days. Two days do not have a lot of users in common.
  2. There is a tiny error between PageRank intersections. Two days share about the same PageRank values for the same users.
- These two observations help us conclude that the topics and posts are lasting in time and this is consistent with other works [1, 4].



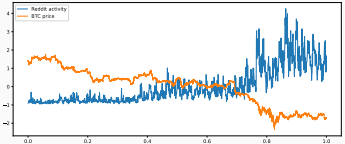


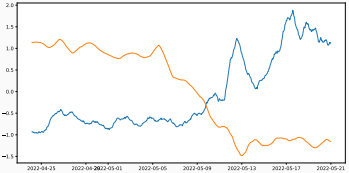
Correlation to price development

- We made an inclusion sequence of graph with a time window of 1h.

$$G_1 \subset G_2 \subset G_3 \subset \ldots \subset G_n$$

- We used degree centrality to compare each Graph i with its subsequent graph $i + 1$.
- We match these data with the ones of the Bitcoin price rate.
- We applied some smoothing with the window average method.





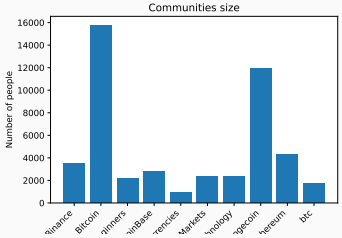
- The correlation we've found is of -0.85.

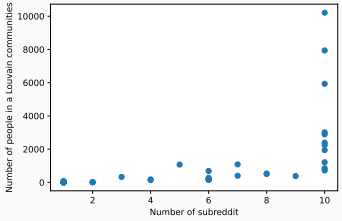
Causal link

- Yet, correlation does not capture causal link.
- Causal link is difficult to assess between activity in subreddits and price rate.
- "Extracting Cryptocurrency Price Movements from the Reddit Network Sentiment." shows with Granger causality tests, that there exists indeed a causal link between the two.
- Correlations have also been proven in "Cryptocurrency Return Prediction Using Investor Sentiment Extracted by BERT-Based Classifiers from News Articles, Reddit Posts and Tweets."

Louvain community detection method

- We observed that Louvain communities are formed from people belonging to the same subreddit.
- Some communities spread over multiple subreddits.
- The larger the community the more it spreads over different subreddits.
- Communities have a foothold in one unique subreddit but also spread on the nine other subs.





Homemade implementation

PageRank

1. “NetworkX” and homemade implementation yields the same results.
2. Yet, the benchmarks show a night and day difference.
3. Using a dense matrix is the main reason behind that.
4. It is a direct consequence of the Power Law distribution.

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

PageRank

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

Louvain

- Our implementation was very slow but yielded the same result as the one from "NetworkX".
- Cause of that is the computation of modularity gain, which could have been optimised.
- We only ran the benchmark on a subset of the total graph.

Louvain

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

Table 2: Benchmarks of Louvain algorithm on macOS Monterey with Apple M1 and 8Go of RAM

Discussion

Questions

Any questions?

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.