Community Analysis of Subreddit Wallstreetbets

Xinyun Shen
xinyun@umich.edu
Computer Science

Xueming Xu
xueming@umich.edu
Computer Science

ABSTRACT

WallstreetBets, (r/wallstreetbets, also known as WSB), which is a subreddit where participants discuss stocks and option trading. It has become notable for its profane nature and allegations of users manipulating securities. Recently the community became mainstream again with its interest in Gamestop shares. In order to better understand the community, this project used techniques in Data Mining to explore the intrinsic relationships between users in the community. We crawled three months' posts in the Wallstreetbets subreddit and constructed a weighted directed user-to-user graph correspondingly. We used centrality measures to find users who are influential in the subreddit. We run Community Detction algorithms to detect communities where users share similar investment interests.

1 INTRODUCTION

Reddit is a popular social media platform where users can share links, stories, photos, videos, and various other forms of media. Each content submission is uploaded to a particular subreddit, a specialized community created and maintained by Reddit users devoted to a certain topic. Users then have the ability to leave comments and vote on posts submitted to these subreddits.

In this project, we analyze user activities in a large community named WallstreetBets, (r/wallstreetbets, also known as WSB), which is a subreddit where participants discuss stocks and option trading. It has become notable for its profane nature and allegations of users manipulating securities. Recently the community became mainstream again with its interest in Gamestop shares. In order to better understand the community, this project plans to use techniques in Data Mining to explore the intrinsic relationships between users in the community. We will answer two main questions as followed,

- Who plays an important role in WallstreeBets?
- Are there any small communities sharing similar investment strategies or specifically interested in certain stocks?

These questions can help us better understand this community and how it will affect the stock market. The first question helps us learn whether there are any influential users who have the ability to lead a trend and influence the market. The second question explores small communities who share similar interests in the investment. With these communities, we may further study their community behaviors and analyze how they influence the stock market.

To answer these two questions, we first collected millions of posts from the subreddit wallstreetbets, which was then used to construct a user-to-user network. We analyzed the network by various Data Mining techniques, like centrality, community detection and role detection.

We will discuss about how we collected data and the basic statistics of the graph in the section 2. In the Section 3, we will run experiments and analyze the results corresponding to each question we just raised. We will have a conclusion and discussion in the Section 4.

2 DATA

As mentioned above, we collected the comments and posts dating from 2021/2/1 to 2021/4/19 in the wallstreetbets subreddit.

After crawling, we got three files. In the file $post\_info.csv$, it contains postid, epoch time, post author name, post's url, post's title, and post content. In the file $comment\_info.csv$, it contains the post id where the comment is occurred, comment id, comment's author, and comment content. These two files, containing actual content of the posts, were not used for constructing graphs but for analyzing user behaviors later. The last file $post\_comment\_info.csv$, which is the most important one, shows the relation between the users. Each line of the dataset corresponds to one comment and contains the parent comment's ID, the comment's ID, the author of the parent comment, and the author of the comment. For our model, we constructed a directed weighted user-to-user network according to the $post\_comment\_info.csv$ file. An edge from user i to user j exists if and only if i comment on or replied to j. The edge weight of edge from user i to user j corresponds to the number of comments that user i commented on user j.

After construction, the basic information of the weighted directed graph is shown as followed,

#node	189825
#edge	924111
Avg in degree	4.8682
Avg out degree	4.8682
#Strongly connected components	125274
#Weakly connected components	3
node in each weakly connected components	[189819, 4, 2]
Avg cluster coeff.	0.1992

Table 1: Weighted graph info

We can see most nodes are weakly connected with each other, while there are several outliers.

The degree distributions are as followed,

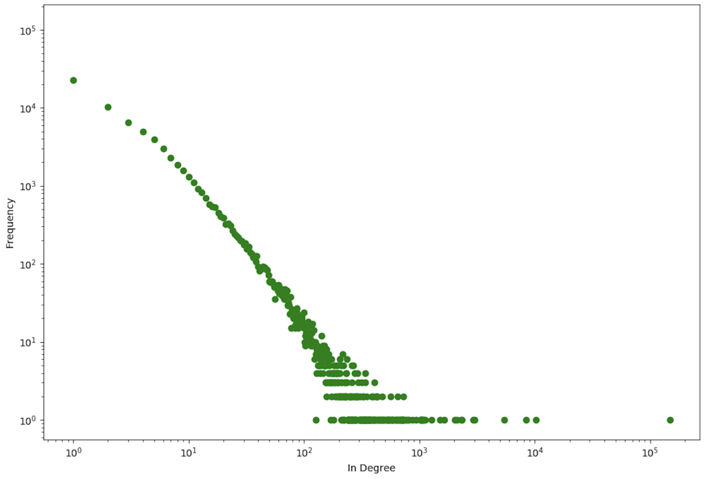


Figure 1: In-degree distribution of weighted graph

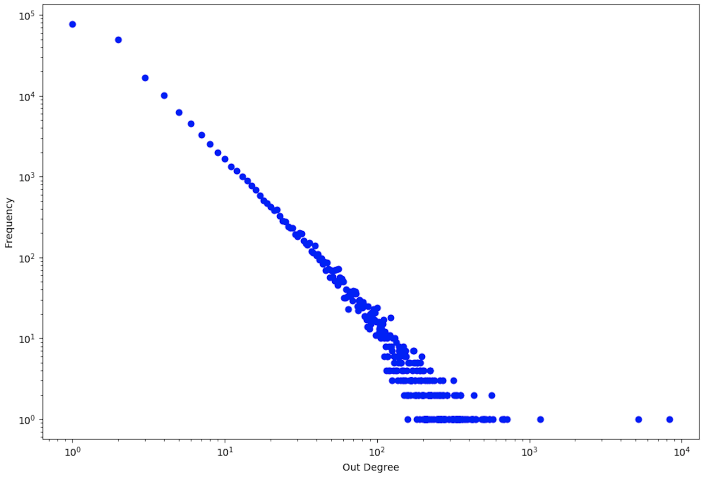


Figure 2: Out-degree distribution of weighted graph

3 DATA ANALYSIS

3.1 Q1: Who plays an important role in WallstreeBets?

3.1.1 Description. ....

For this question, we want to find whether there are users who are influential in the subreddit and is capable of leading a trend. This question is meaningful to study because if someone has the power to start a trend in the wallstreetbets, then their behavior can influence the stock market to a large extent. Also, it might also be of interest whether there is any investment agency behind the wallstreetbets inciting retail investors.

3.1.2 Data. ....

The data we used is the constructed graph described in the Section 2.

3.1.3 Technique & Challenges. ....

The main technique we use is the centrality measures. Centrality is a function $c : V \longrightarrow \mathbb{R}$ inducing a total order over the nodes. If the centrality of a vertex is higher, then the vertex is more important. There are several different centrality measures. We will use degree centrality, closeness centrality, betweenness centrality, pagerank centrality and HITS centrality. We analyzed top 15 users who had the largest centrality scores.

The challenge is that our dataset is very large, which involves in 2991439 edges between 189825 users. At the same time, the time complexity for calculating the closeness centrality and betweeness centrality is large. Therefore it takes huge amount of time to finish calculating them. Due to the limit of the computational resources, we failed to finish calculating these two measures of centrality.

3.1.4 Experimental Setup. ....

(1) In-Degree Centrality
Since it is a directed graph, in-degree centrality is more representative of a user's important. We use Python library networkx to find the in-degree centrality.

(2) Closeness Centrality
We used Python library networkx to calculate the Closeness Centrality algorithm for weighted directed graph. We didn't get any result for closeness centrality because of the limited computational resources.

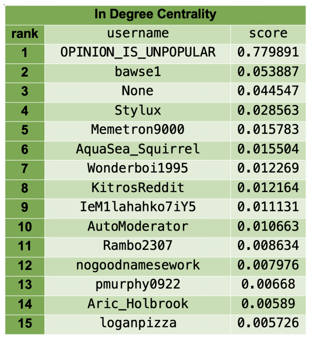
(3) Betweeness Centrality
We used Python library networkx to calculate the Betweeness Centrality algorithm for weighted directed graph. We didn't get any result for closeness centrality because of the limited computational resources.

(4) PageRank Centrality
We used Python library networkx to apply the pagerank algorithm for weighted directed graph.

(5) HITS Centrality
We used Python library networkx to apply the pagerank algorithm for weighted directed graph.

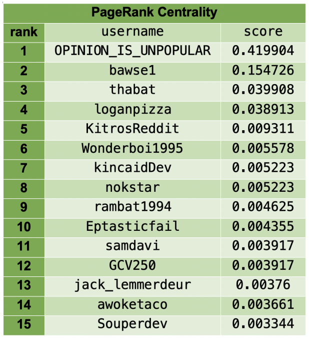
3.1.5 Observations. ....

We got four tables which contain the top 15 scores corresponding to in-degree centrality, pagerank centrality, hub and authority for hits centrality.



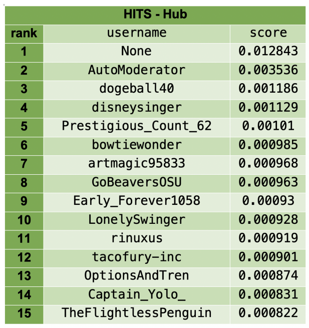
In Degree Centrality		
rank	username	score
1	OPINION_IS_UNPOPULAR	0.779891
2	bawse1	0.053887
3	None	0.044547
4	Stylux	0.028563
5	Memetron9000	0.015783
6	AquaSea_Squirrel	0.015504
7	Wonderboi1995	0.012269
8	KitrosReddit	0.012164
9	IeM1lahahko7iY5	0.011131
10	AutoModerator	0.010663
11	Rambo2307	0.008634
12	nogoodnamesework	0.007976
13	pmurphy0922	0.00668
14	Aric_Holbrook	0.00589
15	loganpizza	0.005726

Figure 3: Indegree Centrality



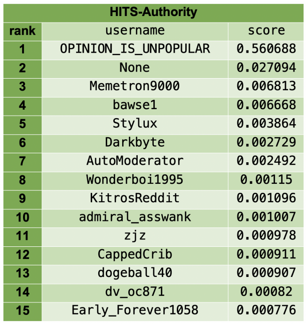
PageRank Centrality		
rank	username	score
1	OPINION_IS_UNPOPULAR	0.419904
2	bawse1	0.154726
3	thabat	0.039908
4	loganpizza	0.038913
5	KitrosReddit	0.009311
6	Wonderboi1995	0.005578
7	kincaidDev	0.005223
8	nokstar	0.005223
9	rambat1994	0.004625
10	Eptasticfail	0.004355
11	samdavi	0.003917
12	GCV250	0.003917
13	jack_lemmerdeur	0.00376
14	awoketaco	0.003661
15	Souperdev	0.003344

Figure 4: Pagerank centrality



HITS - Hub		
rank	username	score
1	None	0.012843
2	AutoModerator	0.003536
3	dogeball40	0.001186
4	disneysinger	0.001129
5	Prestigious_Count_62	0.00101
6	bowtiewonder	0.000985
7	artmagic95833	0.000968
8	GoBeaversOSU	0.000963
9	Early_Forever1058	0.00093
10	LonelySwinger	0.000928
11	rinuxus	0.000919
12	tacofury-inc	0.000901
13	OptionsAndTren	0.000874
14	Captain_Yolo_	0.000831
15	TheFlightlessPenguin	0.000822

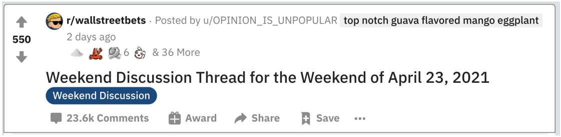
Figure 5: HITS hub score



HITS-Authority		
rank	username	score
1	OPINION_IS_UNPOPULAR	0.560688
2	None	0.027094
3	Memetron9000	0.006813
4	bawse1	0.006668
5	Stylux	0.003864
6	Darkbyte	0.002729
7	AutoModerator	0.002492
8	Wonderboi1995	0.00115
9	KitrosReddit	0.001096
10	admiral_asswank	0.001007
11	zjz	0.000978
12	CappedCrib	0.000911
13	dogeball40	0.000907
14	dv_oc871	0.00082
15	Early_Forever1058	0.000776

Figure 6: HITS authority score

After looking into these nodes with high centrality scores, we found some nodes are not true users but robots. For example, we can find that user None plays an important role in the in-degree centrality and Hits centrality. After figuring out what is the user None, we found that this user's name will occur every time a post or a comment is deleted and our crawling program didn't detect it, which leads to tons of relation with a node None, which is actually meaningless. Besides None, AutoModerator is also generated by the reddit system is a post is deleted by moderator, which our crawling program didn't detect. We can alse find that user OPINION_IS_UNPOPULAR is the another important user in the subreddit wallstreetbets. However, after looking into what the user OPINION_IS_UNPOPULAR posted, we found that this user posted a lot of 'daily discussion' and 'weekend discussion', which attracts users to comment below his posts as shown in the below picture.

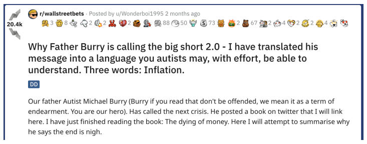


r/wallstreetbets · Posted by u/OPINION_IS_UNPOPULAR top notch guava flavored mango eggplant
2 days ago

Weekend Discussion Thread for the Weekend of April 23, 2021

Figure 7: Opinition_is_unpopular's posts

However, we are more intersted in actual users who are influential on the community. We do find some users with high centrality scores who have lots of followers and provide valuable contents to the community. For example, Wonderboi1995 is a user who shares the investment opinion of themself or other analysts. Here is one of his sample posts.



r/wallstreetbets · Posted by u/Wonderboi1995 2 months ago

Why Father Burry is calling the big short 2.0 - I have translated his message into a language you autists may, with effort, be able to understand. Three words: Inflation.

Our father Autist Michael Burry (Burry if you read that don't be offended, we mean it as a term of endearment. You are our hero). Has called the next crisis. He posted a book on twitter that I will link here. I have just finished reading the book: The dying of money. Here I will attempt to summarise why he says the end is nigh.

Figure 8: Wonderboi1995's posts

3.2 Q2: Are there any small communities sharing similar investment interests?

3.2.1 Description. ....

For this question, we want to find whether there are any small communities sharing similar investment interests in the Wallstreebets subreddit. For example, there might be small groups of people who are interested in a certain stock. This question is meaningful to answer because detecting these groups can have many applications. For example, we can study how the sizes of these groups evolve over time and its association with the stock price.

3.2.2 Data. ....

The data we used is the constructed graph described in the Section 2.

3.2.3 Technique & Challenges. ....

We mainly used community detection algorithms to find these tiny communities. We used Louvain algorithm in our experiments, which is a community detection algorithm based on Modularity Optimization. The Louvain algorithm performs multiple passes. Each pass has two phases. One is the modularity optimization and the other is community aggregation. Then the algorithm creates hierarchies of communities.

Since Louvain algorithm is only suitable for undirected graph, we have to transform our graph to undirected graph. Therefore, we transformed our weighted, directed graph to a weighted undirected graph. If there is any edge between node u and v in the original directed graph, then there will be an edge between node u and v and the weight will be the sum of the weights of the original edges between node u and v. For the undirected graph, the edge can be interpreted as an interaction between two users and the edge weight is the interaction intensity.

3.2.4 Experimental Setup. ....

After transforming tmnJ@1he directed graph to undirected graph, we used Python library networkx to run the Louvain algorithm for weighted undirected graph. The function we called is "community.best_partition" that computes the partition of the graph nodes which maximises the modularity using the Louvain heuristics.

3.2.5 Observations. ....

The algorithm returned 739 clusters. But we observed that most of them were only composed of several nodes, which might be outliers. Hence, we decided to filter out clusters that were too small. After filtering out the clusters consisted of less than 20 nodes, we got only 32 clusters. Since it takes huge amount of time to generate a diagram of the whole graph, we sampled a subgraph. We sampled 50 nodes from the three largest clusters and plotted it to have an overview of how the clusters interacting with each other. The plot is shown in figure 9.

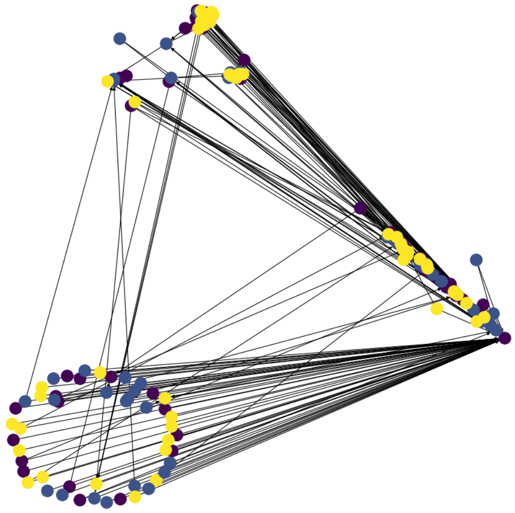


Figure 9: Community plot of sampled network

Comment
Stocks only go up
Got Apple calls and Disney call wish me well
So royal back of Canadian dropped 64% after hours any one known why? I can find info
AMD go!
TSM makes some good potato chips
It's weird that no broker offers 24/7 trading, they could just load up on most popular stocks and let idiots trade it at made up price on weekends between each other
Buy puts on Tesla. A million of them

Table 2: Exemplar comments in Cluster2

As you can see, the clusters are not very pure. There are still many edges across different clusters. But the bad result may also caused by sampling a small portion of the graph.

After having these clusters, we can analyze whether these communities share some common interests by analyzing their posts and comments. For each cluster, we looked at their inter-community comments, i.e. comment from a user in the cluster to another user in the cluster. Table 2 is some exemplar comments in Cluster 2.

As you can see, only judging by eyes and examples, it is still very hard to see what topics the comments in a community is talking about. The contents are very diverse. For now, we haven't observed any community with specific interests in certain type of investment.

But for future works, to better understand what these communities are interested in, applying Natural Languag Processing techniques like topic analysis will be helpful to study the contents in each community.

4 CONCLUSIONS & DISCUSSION

....

Discuss the following:

(1) What are the key observations from your analysis?

Using centrality measures, we are able to find those users who have a lot of followers in the community and therefore have the power to influence the community and the stock market implicitly. Currently, we didn't find any small groups who are specifically interested in certain topics only based on community detection algorithm. For future analysis, NLP techniques may help with identifying the intrests of these detected communities.

(2) What challenges did you face?
  (a) When crawling the data from reddit using reddit api, we can only get at most 1000 posts, which is too less for us. Therefore, we try to use another python package called PushshiftApi. However, this api only provide us the post and we need the relation between users. Finally, we combine the use of praw, which is reddit api and PushshiftApi to first get post id and then use this post id to get all the comments under this post.
  (b) The graph size is too large to draw any plot of the graph. We need some way of sampling or aggregation, which lost some information.

(3) What did you learn by doing this project?

We didn't have a project related to graph mining. So doing this project gives our team a deeper understanding of different methodologies and concepts related to graph mining.

(4) What did you like most about your project?

We have always been curious about how a subreddit can cause the GME event at the end of January. In this project, we explore the power of the internet and social network that can even have a strong impact on the real financial world.

(5) What did each member contribute to the project?
  (a) Xinyun Shen:
    (i) Project report: Abstract/ Introduction/ Data/ Conclusion Discussion
    (ii) Project report: Q1-data/techniquechallenges/experimental setup/observation for all centrality meastures(pagerank/degree)
    (iii) Coding: crawl all the data from subreddit and set up network for wallstreetbets;
    (iv) Coding: get all the useful graph information
    (v) Coding: centrality measures(pagerank/degree/hits)
    (vi) Readme
  (b) Xueming Xu:

(6) Brainstroming the methods and structures of the project
(7) Project report: Abstract/ Introduction/ Data/ Conclusion Discussion
(8) Project report: Q1-data/techniquechallenges/experimental setup/observation
(9) Project report: Q2-data/techniquechallenges/experimental setup/observation
(10) Coding: community detection