Reddit Data Analysis

Siddhartha Pratim Dutta

School of Computing Science

Sir Alwyn Williams Building

University of Glasgow

G12 8RZ

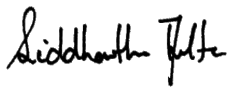
28th February 2025

Abstract

This report on Reddit data analysis explores the idea of performing network analysis on social networks to reveal patterns of user engagement, specifically within the r/InvestmentClub subreddit. The two graph visualisations present the usage of directed weighted graphs of author interactions and directed hierarchical graphs of comment structures, establishing the network structures to be further analysed. The evaluation metrics such as the rich club coefficient, z-score values and virality score help quantify the different roles among the users and identify common patterns of interaction. The data can also be used to answer additional research questions based on temporal and content analysis, which has been introduced in this report to provide a summative overview on the r/InvestmentClub subreddit community.

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic form.

Name: Siddhartha Pratim Dutta Signature: 

Acknowledgements

I would like to express my gratitude towards Dr. Joemon M Jose, the course instructor of the COMPSCI5107 Web Science for MSc course, for providing a comprehensive understanding of the various analytical approaches applicable to web data that enabled and facilitated the study and work underpinning this report.

I would additionally like to thank the course lab tutors - Jie Wang and Junchen Fu for providing their valuable help and guidance navigating particularly challenging parts of the coursework.

Finally, I would like to acknowledge the use of generative AI via OpenAI's ChatGPT to confirm my understanding of topics, bug fixing in the associated code and refining the wording in limited portions of the text in this report.

Contents

Introduction

The democratisation of the internet has shifted the paradigm of webpages from content to users. A prime example of this model is Reddit - a popular forum-based social network started in 2005 [1] where users post their content via text, images, videos, links, etc., allowing other users to view, vote or comment on such posts. The unique selling point of Reddit at the time was the organisation of its site into topic-specific communities or "subreddits". With over 100,000 active subreddits in 2025 ranging from topics based on news and politics to niche hobbies like woodworking or hiking, Reddit has a very diverse user base involving a variety of discussions.

r/InvestmentClub [2] is one such subreddit focused on discussions about investing, primarily in stocks. Users create posts on topics such as market analysis, stock recommendations, investment strategies, and news. It serves as a space for investors to discuss their thoughts by exchanging ideas and insights to make informed decisions. InvestmentClub is currently a restricted community where only approved users can make posts, though, any user can comment on these posts.

This report aims to analyse the interactions within this subreddit, particularly from a network perspective. Users' submissions and their comments with each other reveal a pattern of interactions which can be studied to understand user behaviour, topics of discussion, ultimately providing an overview about the community. Chapter 1 focuses on processing the data extracted from r/InvestmentClub - preparing the data for further analysis and performing some exploratory data analysis on it. Chapter 2 discusses the creation and visualisation of graph-like structures that were identified in this data. Further analysis of these networks by using specific metrics showed interesting insights that are summarised in Chapter 3. Finally, Chapter 4 explores some additional research areas that were found to be interesting in the data.

Chapter 1 Data Preparation and Analysis

This chapter outlines the process of preparing and analysing the raw data extracted from the r/InvestmentClub subreddit. Data preparation steps involved rectifying formatting errors, constructing schema definitions and mapping the appropriate keys from the JSON objects to Python dataclasses. A hierarchical object creation was also performed to recreate post objects representing Reddit's hierarchical structure. Furthermore, exploratory data analysis was conducted to investigate the distribution of key properties and identify patterns in community engagement. This groundwork lays the foundation for additional analysis in subsequent chapters.

1.1 Data Pre-Processing

Raw data from the r/InvestmentClub subreddit was provided in two files - InvestmentClub_submissions.json and InvestmentClub_comments.json, containing submission and comment objects respectively. Upon initial inspection, the files revealed that while the primary object in the form of a submission or a comment had a JSON structure, there were some errors in how they were collected within the JSON array. Rectifying these errors involved replacing a pair of close and open brackets }{ with },{ indicating a separation of objects and finally wrapping all the objects in box brackets [,] to collect them as a valid JSON array.

1.2 Schema & Object Creation

The corrected JSON files had 18,971 submission objects and 22,863 comment objects. To understand the underlying structure of these JSON objects, a custom script was developed to create a sort of JSON schema. This script recursively traversed each object's key-value structure yielding the key name and the data type of the value, representing a schema object (see: *schemas/*).

- **Submission Objects:** Contain 128 root-level keys, including attributes like title, author, created timestamp, score, etc.
- **Comment Objects:** Contain 78 root-level keys, including body, author, parent ID, etc.

The schema objects facilitated an easy exploration of each object type's structure, where supplementary information was filtered out and enabled the development of Data Transfer Object (DTO) classes to map each JSON object to a Python object, keeping only the important information that would be required in the analysis (see: *dtos/*).

- **Submission Dataclass Properties:** `id_`, `subreddit`, `title`, `selftext`, `url`, `author`, `ups`, `score`, `permalink`, `created_utc`, and `num_comments`.
- **Comment Dataclass Properties:** `id_`, `parent_id`, `submission_id`, `body`, `author`, `created_utc`, `ups`, and `score`.

1.3 Constructing Post Objects

Post the initial processing and additional filtering of the required keys, the data remained split into two separate types - submissions and comments. A quick count from the comment objects revealed that the comments data was from 6,372 unique submissions. Since there were a total of 18,971 submissions in the data, it was concluded and verified that 12,599 submissions had no comments. To merge the two submission and comment objects, an additional `Post` class was created, which contains the submission object and a nested `CommentNode` data structure representing all comments in a hierarchy tree. This tree-based representation reflects Reddit's layout, with a submission having root-level comments, and each comment having its own replies. An interaction between two users involving a back-and-forth discussion can go several levels deep in the tree structure (Figure A-1). The construction of the Post objects first required grouping the comments by their submission IDs and then performing a breadth-first-search traversal approach to reconstruct the tree hierarchy. The created objects were saved as `Pickle` files (see *objects/*) so that they could be used for further analysis.

1.4 Exploratory Data Analysis

The pre-processed data were available in three types of objects - `Submission`, `Comment` and a combined `Post` object. This allowed an exploratory data analysis to be performed on the data. First, submissions were analysed over a few properties to examine the distribution of created timestamps, scores and number of comments.

Property	Minimum Value	Maximum Value	Mean Value	Standard Deviation
created_utc	2012-02-01	2022-12-31	2019-02-22	-
score	0	541	3.74	9.51
num_comments	0	80	1.12	3.42

Table 1: Distribution of Submission Properties

There was a total of 18,971 submissions made between the 1st of February 2012 and the 31st of December 2022. The 75th percentile value of the score and number of comments properties being 2 and 1 respectively indicated that the distribution was heavily skewed by submissions with low user interactions. Next, submissions were grouped by users to explore the distribution of unique users and how many submissions were made. This revealed the user *Zurevu* to have made the highest number of submissions (2064), accounting for approximately 11% of all submissions. Completing the top three users with most submissions were *The-Techie* and *Ituglobal* with 551 and 141 submissions respectively. As this distribution was widespread, some particular bins of the number of submissions and the number of users are summarised below.

Number of Submissions	Number of Users
1	5433
2	747
3	264
10-50	150
2000+	1

Table 2: Some Distribution of Number of Submissions by User Count

Table 2 shows some of the distribution of submissions made by the 6,985 unique users. Approximately 78% of users made just one submission, and overall, 97.5% of users made less than 10 submissions to this subreddit. A similar skewed distribution was observed when analysing the 22,863 comment objects. Among the 7,225 unique users who commented on submissions, where the user *Zurevu* with 2,388 comments accounted for approximately 10% of the total comments.

Finally, an analysis was performed on the post objects based on the number of comments per post, and the depth, i.e. the maximum level of the comments tree. The two post submissions titled "[TSLA] Tesla Motors: 5 reasons we should vote for Tesla" [3] and "Bubble is bursting" [4] with the maximum number of comments of 80 have a depth of 14 and 16 each. However, the submission post titled "Amazon is paying its workers up to $5,000 to quit and it's a brilliant strategy" [5] with 22 comments has the maximum depth of 17, indicating a very long discussion thread on a particular comment.

An overview of the analysis on submissions, comments and posts already shows that there are a few "super" users driving the community discussion, with a few posts involving greater engagement as compared to most. The 75th percentile for metrics like number of comments can serve as a useful threshold filter to identify such high value users or submissions to determine the number of nodes and visualise their influence in a network-like structure such as graph visualisations.

Chapter 2 Graph Visualisations

Networks are an underlying structure in social communities representing relationships and interactions among entities. Graphs are a powerful tool of data representation to visualise such networks within a dataset. The entities in a dataset are represented as nodes while edges represent the relationships between these entities. Edges can be directed or undirected and weighted or unweighted depending on the type of relationship between the nodes. In this chapter, graph visualisations are used to represent interaction patterns within the r/InvestmentClub subreddit capturing user behaviour and discussion dynamics. These visualisations allow for further analysis to identify influential users or popular topics within the community.

2.1 Author Interaction Graph

The author interaction graph (see: *graphs/author_interaction_graph.html)* is constructed to model the interactions between users based on comments and replies. The graph was constructed based on the following properties:

- **Nodes:** Entities in the graph represent post or comment authors, i.e. the user who makes a submission or a comment. Out of 12,919 nodes, 5,693 nodes are exclusively submission authors while 5,933 are exclusively comment authors. 1,293 authors have made both submissions and comments.
- **Edges:** Relationships between the entities (authors) in this graph are represented by weighted, directed edges. An edge having a direction from node A → to node B, represents that the author A has commented on a post or reply made by node B. The number of such interactions represents the weight of this edge, with higher weights indicating more frequent exchanges between the nodes.

Figure 2-1: An Example Graph for Author Interactions

In the example Figure 2-1, the user *beeteac* has interacted with user *financeoptimum* eight times in the given dataset. However, because the edge is unidirectional, it indicates that *financeoptimum* has not interacted with *beeteac*.

An author interaction graph is useful to visualise the nature of interactions between users in a particular community. Users with a high number of edges are central to the cohesion of the entire network, and these key power users can be identified with the help of such a graph. The weighted and directed nature of the edges also help identify how frequent these interactions are and whether the interactions go both ways between two users.

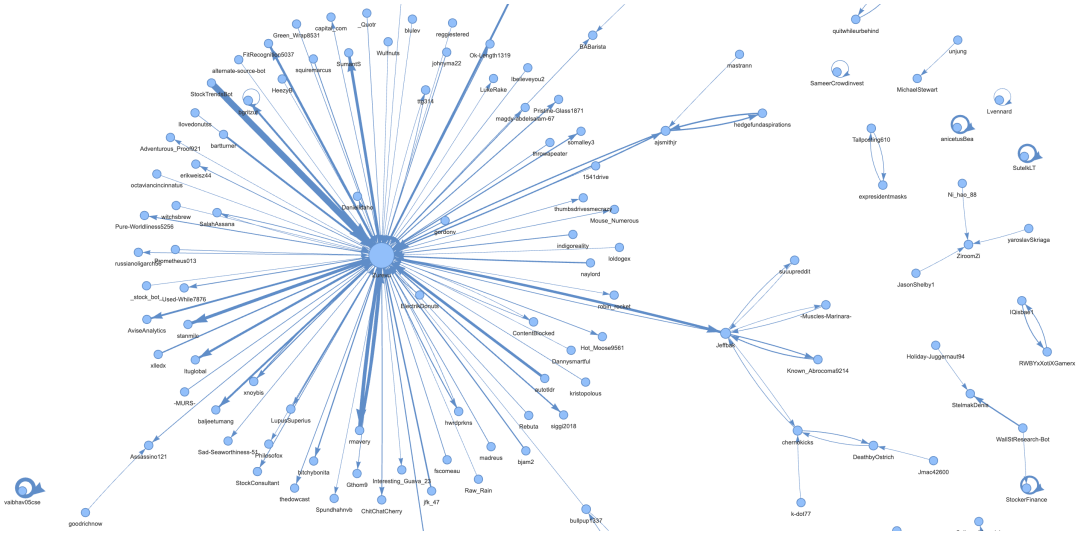


Figure 2-2: The Author Interaction Graph with Edge Threshold = 5

To visualise the network better, an edge threshold of 5 was determined to filter out those interactions with an edge weight lesser than the edge threshold. Hence, users that have only a few interactions, are not shown in the above graph. Nodes with a high outdegree represent a highly active commenter who engages with others, while nodes with a high indegree represent a central figure who receives many replies, likely being a submission author or discussion catalyst.

As seen in the author interaction graph Figure 2-2, several observations are made:

- The central user *Zurevu* is connected to multiple different users with varying levels of interactions. With an outdegree and indegree value of 47 and 48 respectively, the user not only comments on others' submissions but also receives a lot of comments on their submissions, making them a key node in this community. With interaction values of 16 and 20 to users *The-Techie* and *Jeffbak*, they are well connected to other influential nodes as well.
- The edge between *StockTrendsBot* and *Zurevu* has the highest weight of 48, although the username suggests that it is likely a bot account commenting on *Zurevu*'s submissions.
- Another such interaction is visible between the nodes of *evilbot666* and *Grumblecaaaakes* with an edge weight of 39.
- Oddly, there are some nodes with a highly weighted self-edge such as *vaibhav05cse* with a self-edge weight of 30 indicating that they frequently reply to their own submissions.
- Smaller independent graph components having two or three nodes in them most likely represent an active discussion between users in a comment thread, such as the one between *Tigen13* and *243james*, each having an edge weight of 6 to each other.

A macro-level perspective of this graph makes it evident that the user *Zurevu* is a super user with multiple interactions to other users and other users interacting with them as well, driving discussions within the community. The high number of independent components in this graph, however, suggests that most of the users do not have strong connections to other users, unless they engage in a

discussion thread on a particular submission. An example [6] between the previously mentioned user *243james* and a deleted user shows how quickly the weights of the edge add up when there is a back and forth in commentary or discourse.

Overall, the author interaction graph reveals key community members and distinct interaction patterns, highlighting both central users like *Zurevu* and smaller, isolated discussions. These observations lay the groundwork for deeper exploration of user engagement and network dynamics within the subreddit.

2.2 Post Discussion Graph

The post discussion graph (see: *graphs/posts_comments_graph.html)* is constructed on a post-by-post basis, representing the hierarchical layout of a Reddit discussion. This network structure captures the relationships between submissions and their associated comments, along with the reply chains within the comment thread. The graph has the following properties:

- **Nodes:** Entities in the graph represent a central node for the submission and additional nodes representing the comments made to that submission. Further nodes represent the replies to the previous comment in the reply chain. The size of the node represents the number of comments in reply to that particular submission or comment node.
- **Edges:** Directed edges capture the relationships between submissions and comments or comments themselves. An edge from node A → to node B indicates that node B is a direct reply to node A. This tree structure of the graph mirrors Reddit's nested comment format, with replies branching off from the original submission or subsequent comments.

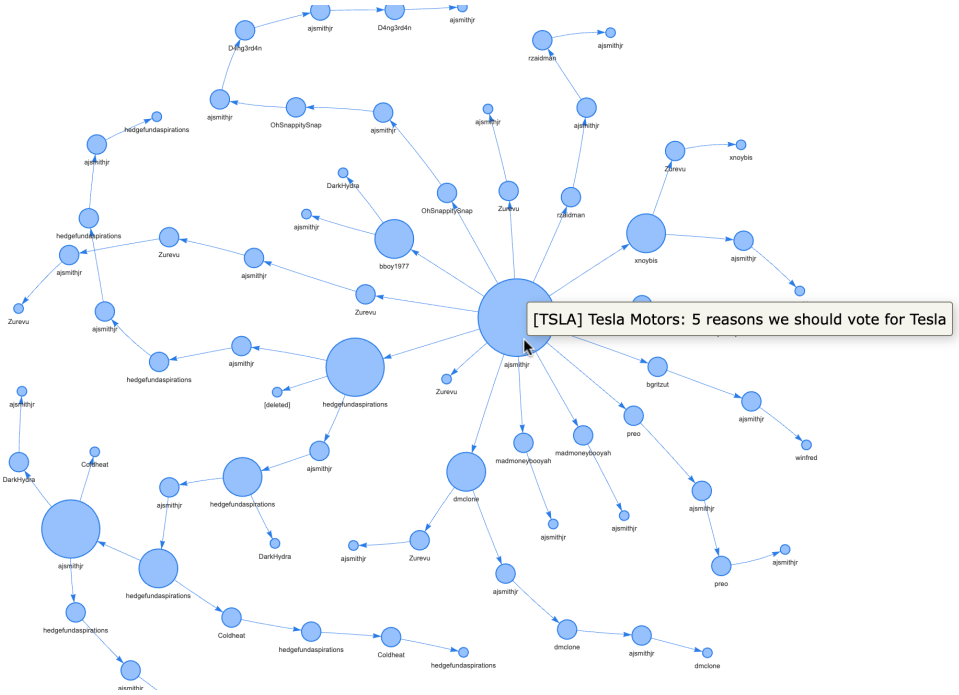


Figure 2-3: The Posts Comment Graph with Number of Posts = 1

The above Figure 2-3 shows an example of the post discussion graph illustrating how comments and replies are hierarchically organised. The submission in this case "[TSLA] Tesla Motors: 5 reasons we should vote for Tesla" [3] forms the root node with direct connections indicating root level replies and subsequent replies creating additional connections.

In this example, the submission receives 14 direct replies. Most of these direct comments elicit further discussion, notably the comment made by *hedgefundaspirations* shown in Figure 2-4, triggers a possible debate with long chains of replies being formed in multiple directions.

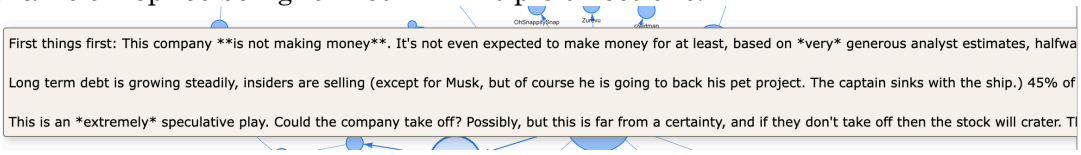


Figure 2-4: Comment by user *hedgefundaspirations* at Level = 1

The nature of the comment appears to have initiated a back-and-forth exchange between *hedgefundaspirations* and *ajsmithjr,* with the tone of the conversation becoming increasingly contentious and personal as the comment thread deepens.

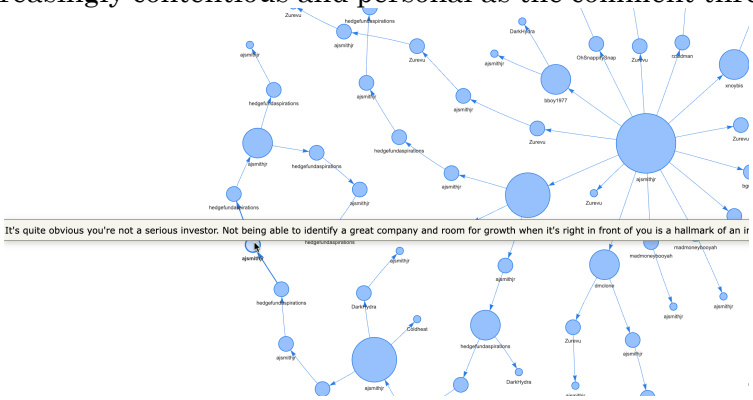


Figure 2-5: Change in Comment Tone as Tree Depth Increases

Posts with significant tree depths, such as in the example above, often indicate topics that are either controversial or highly engaging. These posts tend to foster dynamic discussions, where users are more likely to reply repeatedly and frequently, forming extensive reply chains. By identifying such posts with deep comment trees, the nature of the topics in the community can be gauged, and which topics generate meaningful or controversial conversations. Conversely, posts with minimal engagement would result in smaller graphs with little to no depth, indicating a failure to capture the community interest.

The post discussion graph is useful for visualising the discussion dynamics within a community. By visualising not just the number of comments but also the hierarchical depth of replies, it helps to understand user engagement patterns along with the intensity of discussions. Overall, the graph highlights the diversity in post engagement with nodes either representing active discussion hubs or generating limited responses by other users in the community.

Chapter 3 Network Analysis

The network analysis performed in this chapter uses several evaluation metrics to understand the dynamics of the graph structures introduced in the previous chapter. By applying the three metrics - superuser influence, questioner vs answerer z-score and virality score, it becomes possible to identify key patterns in user behaviour, the frequency of interactions and flow of information. As a result, these metrics capture different aspects of the subreddit community and provide a comprehensive overview of the network structure and reveal some important properties.

3.1 Superuser Influence - Connected Components and Rich Club Coefficient

The first metric evaluates the role and influence of "superusers" in the r/InvestmentClub subreddit from the 2.1 Author Interaction Graph visualisation. These users can be identified via their submission patterns and how well they connect with the rest of the community through their commenting patterns. Specifically, this method assesses how removing highly connected users, i.e. superusers, impacts the network structure. The two measures of indication used are:

- **Network Fragmentation:** By tracking how connected components evolve after removing top superusers one at a time.
- **Elitism in Interaction:** Using the Rich Club Coefficient [7], which identifies whether high-degree users preferentially interact with other high-degree users.

3.1.1 Network Fragmentation

A connected component is a subgraph where all nodes are reachable from one another. Superusers form the roof of such connected components as they are the users with the most interactions, i.e. high degree centrality. Removing them helps us analyse network resilience and whether the community is overly dependent on a few key users. The following steps are applied iteratively:

1. Compute degree centrality for each user: $\mathrm{C_D(v)} = \frac{\mathrm{degree(v)}}{|\mathrm{V}|-1}$, where v is the node (user) and $|V|$ represents the total number of nodes in the graph.
2. Remove the user with the highest centrality $C_D(v)$.
3. Recalculate connected components before and after removal.

3.1.2 Rich Club Coefficient

The rich club phenomenon occurs when high-degree nodes tend to connect with each other more than expected by chance. Mathematically, the rich club coefficient for a degree k is given by: $\phi(k) = \frac{2E_k}{N_k(N_k-1)}$ where,

- E_k is the number of edges among nodes with degree > k.
- N_k is the number of edges among nodes with degree > k.

A $\phi(k)$ value closer to 1.0 suggests that influential users form an exclusive group at degree k.

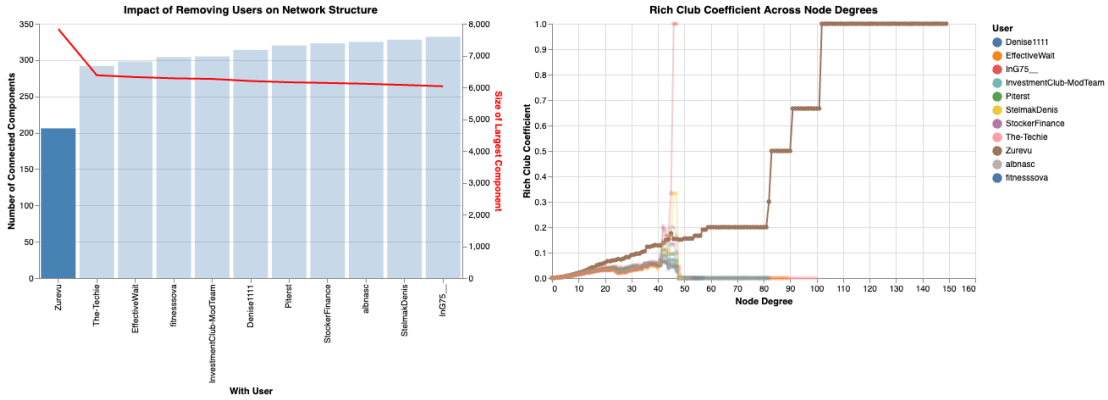


Figure 3-1: Network Structure & Rich Club Coefficient with user Zurevu

As shown in Figure 3-1, the user *Zurevu* is identified as the top superuser whose presence in the network limits the total number of connected components to 206 nodes as indicated by the left-hand bar chart, while the size of the largest connected component is 7,832 nodes indicated by the left-hand line chart. The right of the figure shows a line chart representing the rich club coefficient of various users (see *charts/Superuser-Influence.html*). The brown line representing user *Zurevu* dominates at higher degrees, suggesting that this user's presence is essential for maintaining a strong rich-club and the tendency to form a tightly connected elite group. If *Zurevu* is removed, the interconnectedness among the highest degree nodes weakens significantly, also affecting the network structure. This is evident by the steep change in the bar chart and line chart values after *Zurevu* is removed. The following 9 top superusers, after *Zurevu*, have similar rich club coefficient structures to each other, remaining close together indicating that removing these users does not drastically change the network's elite connectivity.

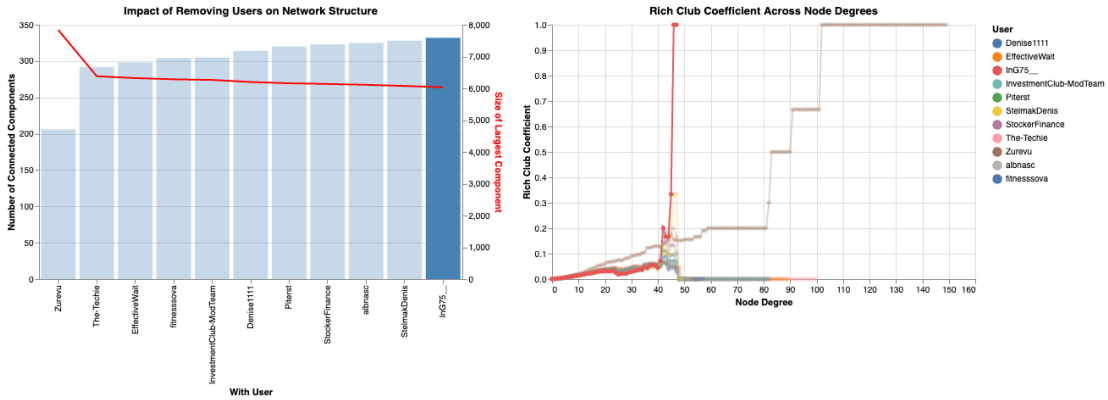


Figure 3-2: Network Structure & Rich Club Coefficient with InG75__

Interestingly, while there is not much variation in the interconnectivity among top users while removing the next top 9 superusers, after the removal of the tenth superuser, the red line for the rich club coefficient with user *InG75__* dominates the graph, albeit at lower degrees. This indicates that while the top users form an elite club with *Zurevu*, removing the top-level club allows the emergence of a new rich club led by the user *InG75__*. While the maximum node

degree of the second-tier rich club is at a k value of ~45, the pattern of users forming an exclusive club at different levels is evident within the community. Overall, within the r/InvestmentClub subreddit, the top superuser plays a key component in the network structure connecting several independent components. Furthermore, the occurrence of a second-tier rich club demonstrates the hierarchical nature of elite groups within the subreddit. These findings underscore the importance of key users in sustaining network cohesion and fostering discussions.

3.2 Questioner vs. Answerer Z-Score

The second metric relates to a measure of user posting behaviour in the r/InvestmentClub subreddit. In online discussion forum sites like Reddit, some users predominantly ask questions - initiating discussions or seeking information, while others primarily provide responses to these questions. The questioner vs. answerer z-score can quantify these behavioural differences by measuring the extent to which a user's behaviour deviates from an expected random behaviour under the null model assumption. Let:

- q_u = Number of questions asked by the user u i.e. the user's posts or comments that received replies.
- a_u = Number of answers given by the user u i.e. the user's replies to other posts or comments.
- n_u = Total number of interactions by user u i.e. $q_u + a_u$.

The probabilities for posting behaviour are calculated as follows:

$$P_q = \frac{\sum_u q_u}{\sum_u n_u}, \quad P_a = \frac{\sum_u a_u}{\sum_u n_u}$$

Where P_q and P_a are the probability of a post P being a question or answer respectively. Based on an analysis of the data, we find the probabilities of the Bernoulli process to be of equal probabilities, simplifying the null model to:

$$P_q = P_a = 0.5$$

Under this assumption, the expected number of answers for a user can be calculated as:

$$E[a] = P_a \cdot (q + a) = 0.5(q + a) = \frac{q + a}{2}$$

The variance follows from the binomial distribution as:

$$\sigma = \sqrt{nP_a(1 - P_a)} = \sqrt{(q + a) \cdot 0.5 \cdot 0.5} = \frac{\sqrt{q + a}}{2}$$

Hence, the modified z-score for questioner vs. answerer is defined as:

$$Z_u = \frac{a - E[a]}{\sigma} = \frac{a - q}{\frac{\sqrt{q + a}}{2}} = \frac{2(a - q)}{\sqrt{q + a}}$$

Where:

- $Z_u < 0$: The user asks more questions than expected (Questioner).
- $Z_u > 0$: The user provides more answers than expected (Answerer).

This metric helps in the categorisation of users as predominantly questioners or answerers determining their roles within the community.

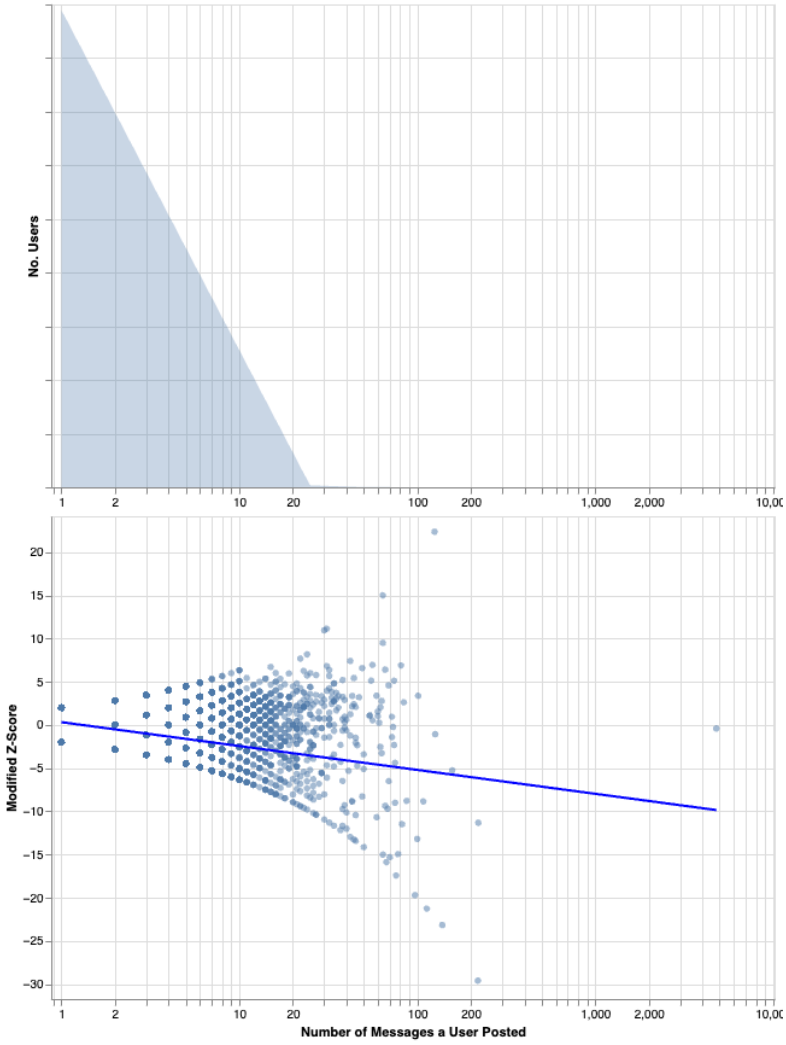


Figure 3-3: Distribution of Z-Scores Among Users

The visualisation in Figure 3-3 displays the distribution of the modified z-score among users in the subreddit (see *charts/Z-Score.html*). As expected, the distribution follows a power-law pattern, with most users posting very few messages and a small number of users posting extensively as shown by the area histogram at the top of the figure. As the number of messages increases, the number of users decreases sharply.

The scatterplot suggests that there are a lot of users with a z-score of around 0, indicating an equal level of questions and answers. Some users like *InvestmentClub-ModTeam*, *StockTrendsBot*, *autotldr*, and *WallStResearch-Bot* have z-scores of greater than 10, i.e. they answer more than they question, which seems consistent with behaviour of "bot" accounts that can be inferred from their usernames. The outlier with the maximum number of messages is the super user

Zurevu, having a z-score of -0.4, which indicates an almost equal questioner and answerer. The general trend displays a slightly negative slope, indicating that increased participation correlates with more questioning behaviour, such as by the user *The-Techie* with a z-score of -29.59.

The modified z-score metric for assessing the role of users as questioners or answerers reveals key dynamics in the r/InvestmentClub subreddit. User participation follows a power-law distribution with most users posting infrequently while a few users driving more content. While many users display a balanced behaviour in posting questions vs answers, a subset of users display outlier patterns. Bots and moderators have high positive z-scores while certain users display more question-oriented participation. The general trend suggests that increased participation correlates with more questioning, perhaps reflecting the subreddit's collaborative, information-seeking culture. Overall, this metric highlights the diverse roles in the subreddit's knowledge-sharing ecosystem, from bots to discussion leaders to information-seekers.

3.3 Virality Score based on Post Cascade Size & Lifespan

The third metric discusses virality, a key property from the 2.2 Post Discussion Graph that visualises the depth of the comment tree structure for a particular post. Virality in online discussions captures the extent to which a post generates engagement and sustains attention among the community over time. To quantify and evaluate this measure, two key metrics are used:

- **Cascade Size C_p:** The total number of comments in a submission thread, representing the breadth of engagement.
- **Lifespan L_p :** The time duration from the original post to the last comment, representing the discussion's longevity.

For a given post p,

- Let C_p represent the set of all comments in the submission's discussion tree cascade.
- $|C_p|$ is the cascade size, i.e. the number of comments.
- t_0 is the timestamp of the submission's creation.
- t_{max} is the timestamp of the last comment.

The lifespan of the post can be computed as:

$$L_p = t_{max} - t_0$$

The virality score V_p is then defined as:

$$V_p = \log\left(1 + \left|C_p\right|\right) \cdot \log\left(1 + L_p\right)$$

Logarithmic scaling is applied to prevent very large cascades or lifespans from dominating the virality score. An offset of +1 ensures that a score is defined for all posts, including posts with no comments or short lifespans. Therefore, a high V_p indicates a post with both substantial engagement and longevity, while a low V_p suggests limited or short-lived interactions.

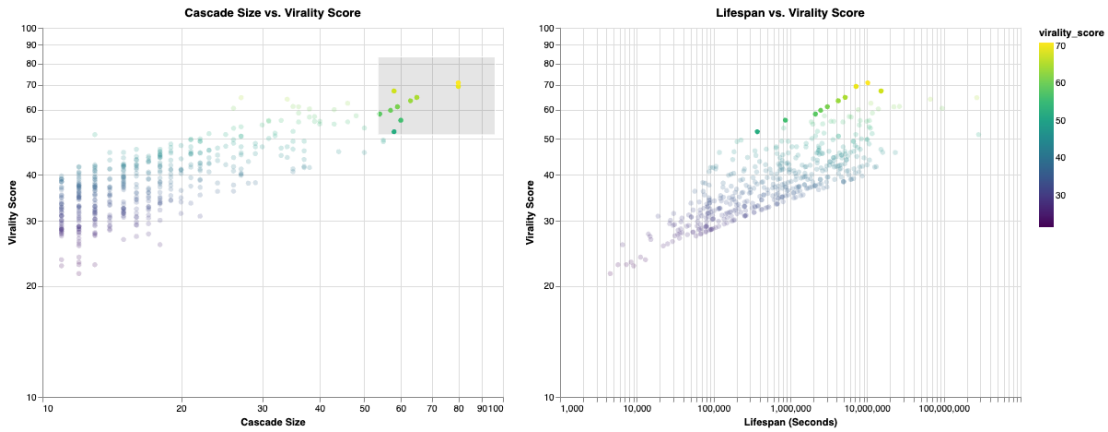


Figure 3-4: Comparing the Lifespan of Posts with Top Virality Scores

The visualisation Figure 3-4 of the virality metric (see *charts/Virality-Score.html*) compares two scatterplots - virality score against cascade size on the left and against lifespan on the right. While the cascade size vs virality score plot reveals a clear upward trend, i.e. larger cascades tending to have higher virality scores, the spread at each cascade size suggests that lifespan also plays a significant role. The lifespan vs virality score plot demonstrates that longer-lasting posts generally have higher virality scores, with some posts remaining active over 10 million seconds (~115 days). The most viral posts marked by yellow colour encodings correspond to the submissions with large depths [3], [4] discussed earlier, with prolonged discussions and sustained engagement.

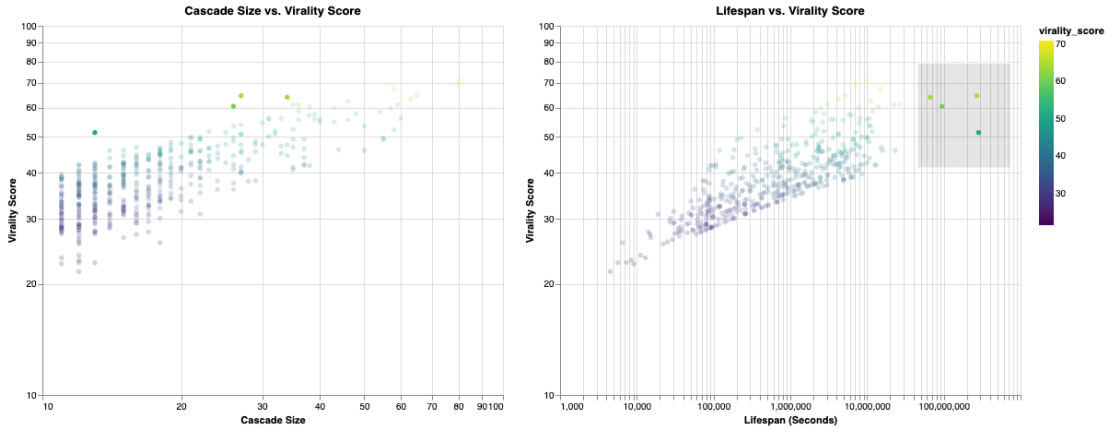


Figure 3-5: Comparing the Cascade Size of Posts with Highest Lifespans

While virality increases rapidly with small increases in cascade size, it exhibits diminishing returns at higher values of lifespan as seen in Figure 3-5. This indicates the presence of a saturation point beyond which extended lifespans contribute less to overall virality, suggesting an overloading effect.

Thus, the virality score effectively captures both the breadth and longevity of discussions in r/InvestmentClub. Posts with high virality tend to spark extended, high-volume discussions, serving as central hubs of engagement. The non-linear growth pattern highlights that both cascade size and lifespan contribute multiplicatively to virality, though the effect diminishes at extreme values. This metric provides valuable insights into which posts drive community interaction and sustain attention over time.

Chapter 4 Additional Research Tasks

This chapter on additional research tasks explores questions beyond the core metrics analysed in the earlier chapter. Broadly, it involves investigating temporal trends and content analysis to deepen our understanding of user behaviour and community topics within the r/InvestmentClub subreddit. Understanding such dynamics highlights opportunities for future work in data quality, discussion themes, and engagement patterns within social communities.

4.1 Temporal Trends Analysis

Temporal patterns in submissions or comments activity offer valuable insights into community behaviour and engagement within the subreddit. Exploring post distributions across different time scales enables the identification of peak activity periods and helps explain how discussion frequency and virality evolve over time. This analysis addresses the following two research questions:

1. When did submissions in r/InvestmentClub peak over time between February 2012 and December 2022?
2. What is the pattern of engagement in r/InvestmentClub across hours of the day and days of the week in terms of submissions and upvotes?

4.1.1 Temporal Distribution of Posts Over Months & Years

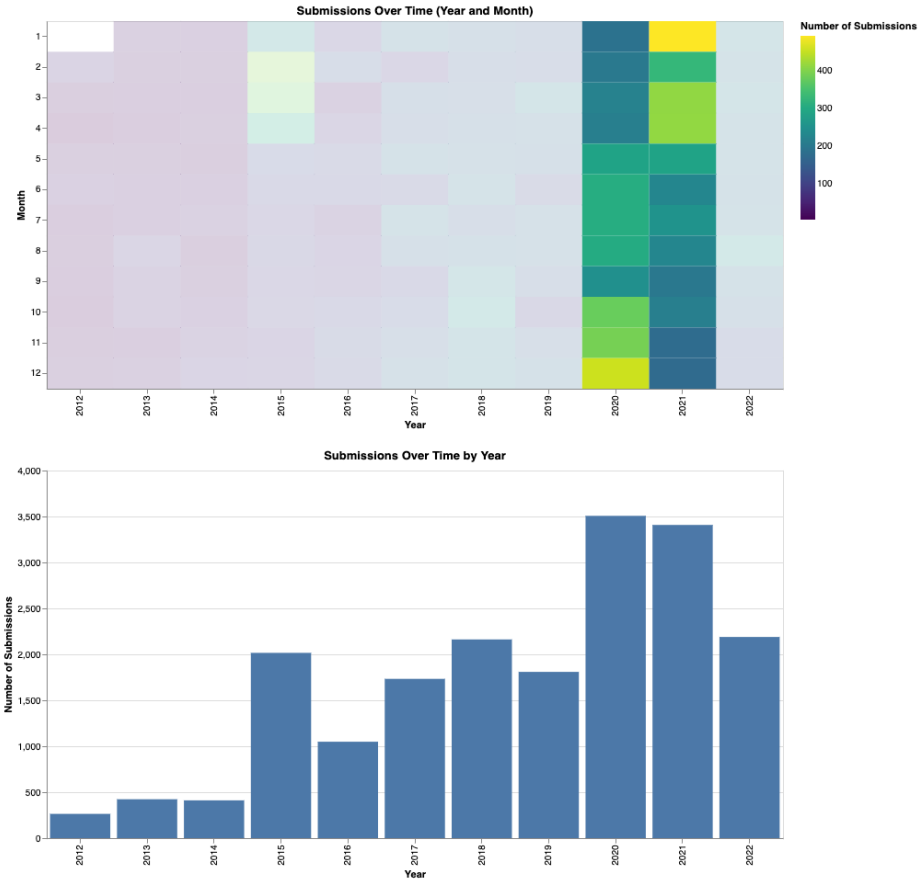


Figure 4-1: Number of Submissions over Months & Years

Figure 4-1 presents a 2D heatmap of submission frequencies over time, with the x-axis representing years and y-axis representing months (see *charts/Submissions-Year-Month.html*). A linked bar chart shows the cumulative number of submissions for each respective year. The bar chart indicates that the subreddit was relatively less popular in the early years with fewer than 500 annual submissions between 2012 and 2014. While the subreddit shows an increased number of submissions between 2015 and 2019, popularity peaked in 2020 and 2021 with around 3,500 submissions made annually. The heatmap reinforces this observation, highlighting increased activity between October 2020 and April 2021, with a peak of 493 posts in January 2021. This spike is most likely related to the GameStop short squeeze [8] driven primarily by retail investors coordinating on social networks like Reddit, aiming to challenge institutional short positions [9]. The volatility of the GameStop stock persisted until March 2021, after which the number of submissions tapered off as shown in the heatmap above.

4.1.2 Hourly & Weekly Post and Upvote Distribution

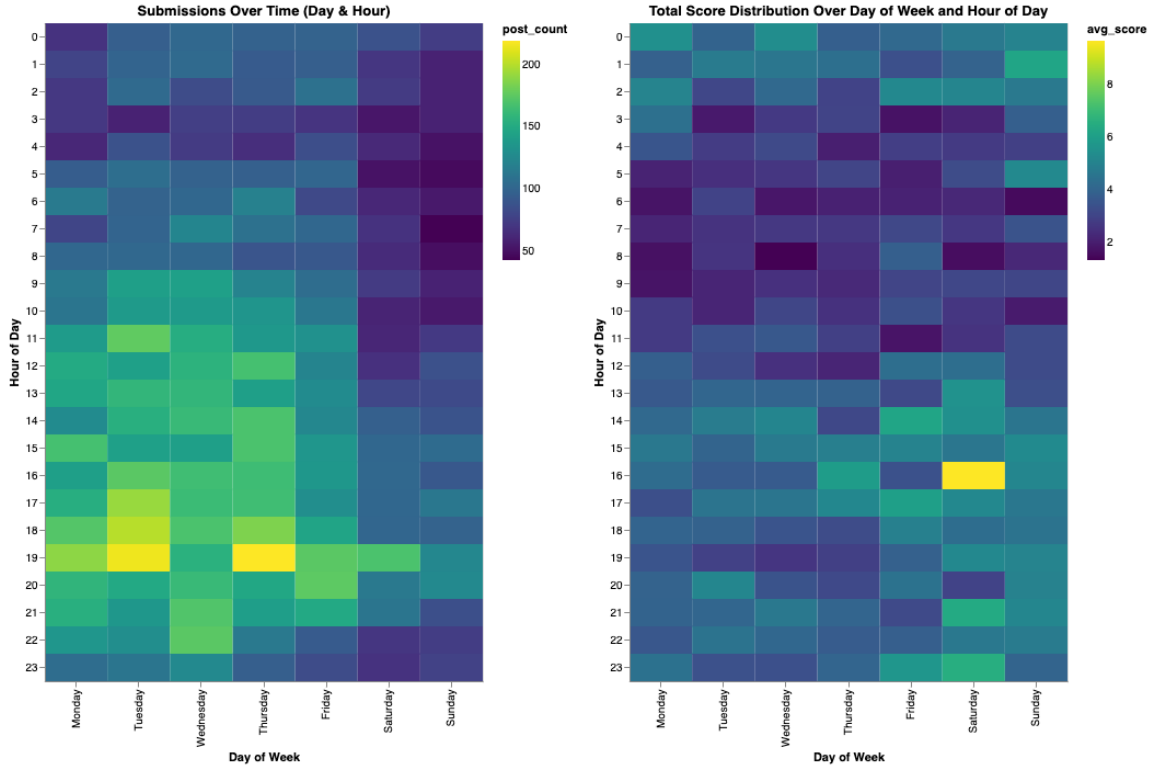


Figure 4-2: Distribution of Number of Submissions & Total Scores

Figure 4-2 illustrates the distribution of post frequency and upvote behaviour at finer temporal resolutions, showing the number of posts and total upvotes received during each hour and day combination (see *charts/Submissions-Day-Hour.html*). Submission activity was most frequent between the hours of 11 and 22 UTC during the weekdays, corresponding to the hours of 6 and 17 EST, which aligns with pre-market, market hours and post-market hours in the U.S. Particularly, 19 UTC shows higher submission frequencies across days, the hour corresponding to a typical lunchtime break in the U.S. However, in terms of the average distribution of upvotes, submissions at 16 UTC on Saturdays were most likely to receive more votes. This suggests that the number of submissions made at a given time is not necessarily correlated with the number of upvotes received.

4.2 Submission Patterns of Top Users

Earlier chapters discussed the presence of super users within the r/InvestmentClub subreddit - those users who dominate engagement and tend to have higher submission counts. Understanding the submission behaviour among such users can reveal whether the engagement is consistent over time or occurs for a short time and then disengages. Specifically:

- Do super users engage with the r/InvestmentClub community over extended periods or for short and intense periods?

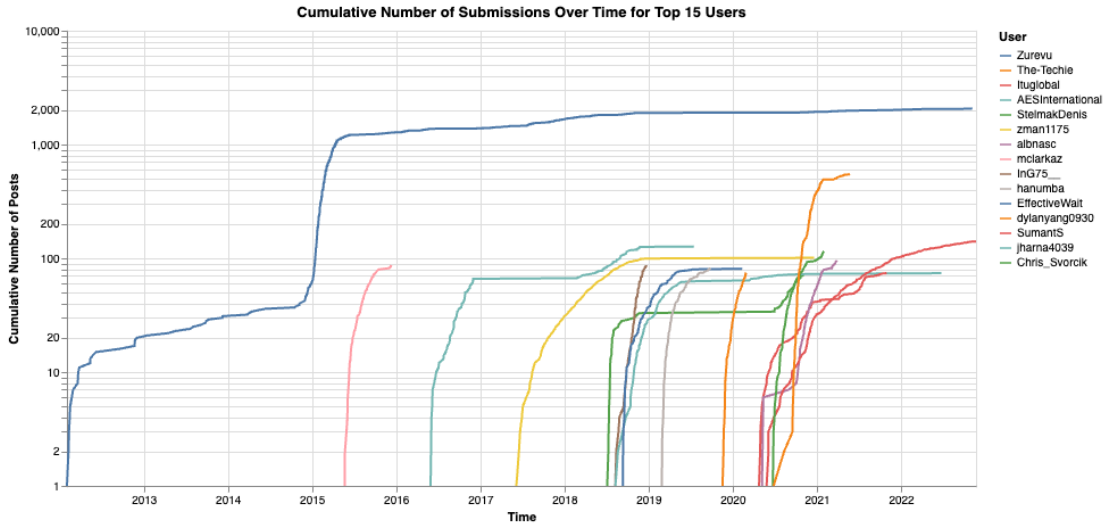


Figure 4-3: Submission Patterns of Top Users (n = 15) Over Time

Figure 4-3 presents the submission timelines of the top 15 users in the subreddit ranked by their total number of submissions (see *charts/Submissions-By-Users.html*). The x-axis represents time over years, while each coloured line indicates the cumulative number of submissions for a given user on a logarithmic scale to avoid dominance by the user *Zurevu*. The graph highlights two different behavioural patterns:

- **Consistent Engagement over Time:** Users like *Zurevu*, *Ituglobal, zman1175*, *jharna4039*, demonstrate steady and sustained activity over multiple years as suggested by the continuous growth in the cumulative submission lines. This pattern reflects active users in the community contributing continuously as discussions evolve.
- **Short-Term Bursts of Activity:** Several other users show rapid growth in cumulative submissions over a short period, followed by inactivity. The clusters of emergences of such users between 2018-2019 and 2020-2021 may reflect users joining the community during periods of heightened activity such as the GameStop squeeze discussed earlier and then disengaged after the conclusion of such significant market events.

A notable pattern among most users in Figure 4-3 reveals that the first 100 submissions are often over a relatively short timeframe as indicated by the steep vertical lines at the start of each user's cumulative submission line. Following this initial burst of submissions, activity tends to taper off, either with complete disengagement or sporadic submissions over extended periods, suggesting a

potential saturation effect, perhaps a pattern of a diminishing effect of an initial enthusiasm. These patterns make it easier to identify user roles in the evolution of a community in terms of stability and ongoing discourse.

4.3 Content Analysis Through Topic Clustering

Content analysis is a method used to systematically identify patterns, themes or topics within textual data. In the context of the r/InvestmentClub subreddit, content analysis on the submission titles helps reveal dominant discussion themes and offers insights into the discussion topics that drive engagement in the community. The following analysis involves applying topic modelling and clustering to group similar titles and extracting the key terms from each group to identify the themes of discussion, addressing the research question:

- What are the central discussion topics in the r/InvestmentClub subreddit and how are they distributed across submissions?

By leveraging natural language processing techniques such as tokenization and TF-IDF vectorization [10], titles of all 18,971 submissions were converted into their corresponding vector representations. The KMeans [11] clustering algorithm was applied to these vectors to cluster similar vectors. The elbow method was used to determine the number of clusters as 20 to capture diverse themes without overfitting. Dimensionality reduction using principal component analysis [12] was then performed to be able to visualise the topic clusters in two dimensions (see *charts/KMeans-Topic-Clusters.html*). The Figure 4-4 below highlights these clusters and their corresponding top terms.

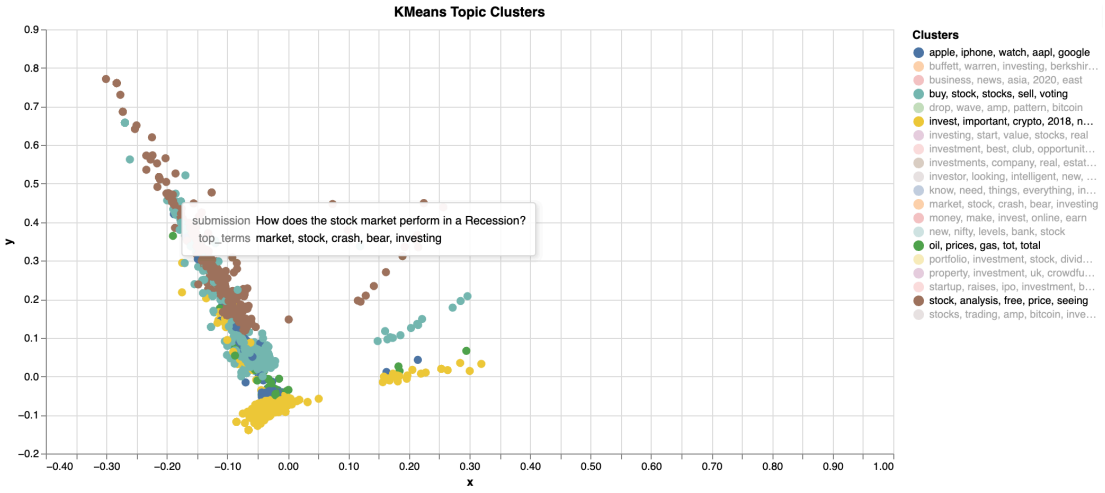


Figure 4-4: Selected Subset of Clusters using K-Means on Post Titles

For example, the selected post titled "How does the stock market perform in a Recession" [13] is clustered in a group having the top terms - "market, stock, crash, bear, investing" where the terms *crash* and *bear* while not mentioned in the title are perhaps closely related to the topics of *recession* as determined by the clustering algorithm. While the visual representation in two dimensions shows overlapping clusters, the analysis of top terms from the 20 clusters revealed some prominent themes and niche topics of discussion prevalent in the subreddit. Some highlighted clusters are discussed further:

- **Cluster (apple, iphone, watch, appl, google):** Discussions in this cluster focus on large tech stocks and the products launched by those companies, revealing a technology-based theme in certain submissions.
- **Cluster (oil, prices, gas, tot, total):** This group captures discussions around the energy sector in the market and how conflicts in the Middle East affect the volatility of oil and gas prices.
- **Cluster (buy, stock, stocks, sell, voting):** This cluster highlights transactional discussions, particularly about buy-sell decisions and corporate voting rights.

The structured breakdown of submission titles in clusters of relevant discussion themes enables a deeper understanding of community behaviour, interests and responsiveness to external events. The distribution of top terms of the twenty clusters showing a lot of overlap of terms suggests a general thematic consistency across submissions by the community. While subreddits are mainly thematic, several diverse sub-discussion themes can be identified through such a process of content analysis. This subreddit, for example, exhibits a broad range of investment-related conversations - from macroeconomic trends to stock selections for individual portfolios. Hence, the topic clustering provides a structured view of the community's core discussion themes, offering valuable insights into user interests and engagement patterns within the subreddit.

Chapter 5 Conclusion

The r/InvestmentClub subreddit on Reddit exemplifies an online community of users discussing various topics centred around a common theme. An analysis of the data collected over a decade reveals complex user dynamics and behavioural patterns. The subreddit itself has around 102K subscribers, however, the data revealed that only 12,919 users have engaged in the subreddit via submissions or comments. Despite 87.3% of Reddit users related to this subreddit being "lurkers", the data from the remaining active 12.7% of users provided highly valuable insights.

The visualisation of the author interaction and post discussion graphs was instrumental in understanding the way users engage amongst each other within a community. Throughout the various analyses, the user *Zurevu* dominated the metrics as an outlier with their massive contribution towards this community and thus exhibiting the role of a super user. Based on their contribution pattern, *Zurevu* was likely a moderator who has since deleted their account [14], leaving the subreddit inactive and hence restricted.

While a rich club dominated the community, additional lower-tier clubs were also apparent. Despite the rich club's dominance, the subreddit functioned as a supportive community. The majority of users had z-scores close to zero, indicating a balance between asking and answering questions. This pattern suggests that r/InvestmentClub operated more as a discussion forum than a one-way information channel. Finally, the virality score ranked the most discussed topics and how virality of submissions evolved over time.

Further analysis revealed that user contributions correlated with market hours and significant external events, highlighting temporal trends that influenced user behaviour. Through the implementation of natural language processing and machine learning techniques, important topics of discussion were extracted and displayed along with their submission distributions.

Ultimately, network analysis unveils the intricate dynamics of social networks, offering a profound understanding of how users interact and adapt within online communities.

References

[1] *Our 20 Year Rule: You can now ask questions about 2005!*, J-Force, r/AskHistorians, January 1, 2025. https://www.reddit.com/r/AskHistorians/comments/1hr05tu/

[2] *r/InvestmentClub*. https://www.reddit.com/r/InvestmentClub/

[3] *[TSLA] Tesla Motors: 5 reasons we should vote for Tesla*, ajsmithjr, r/InvestmentClub, February 11, 2012. https://www.reddit.com/r/InvestmentClub/comments/pl0ko/

[4] *Bubble is bursting*, Jeffbak, r/InvestmentClub, September 24, 2021. https://www.reddit.com/r/InvestmentClub/comments/pud5i4/

[5] *Amazon is paying its workers up to $5,000 to quit and it's a brilliant strategy*, [deleted], r/InvestmentClub, February 18, 2018. https://www.reddit.com/r/InvestmentClub/comments/7yfyxj/

[6] *"The Biggest Stock Market Crash Is On Us NOW" | Peter Schiff Latest interview*, biotonik25, r/InvestmentClub, 9 June, 2022. https://www.reddit.com/r/InvestmentClub/comments/v8grh4/comment/ijoms5v/

[7] *The rich-club phenomenon across complex network hierarchies*, Julian J. McAuley, Luciano da Fontoura Costa, and Tibério S. Caetano, Applied Physics Letters Vol 91 Issue 8, August 2007. https://arxiv.org/abs/physics/0701290

[8] *GameStop short squeeze*, Wikipedia.org. https://en.wikipedia.org/wiki/GameStop_short_squeeze

[9] *GME Institutional Ownership. If It doesn't drop below 100% it means something isn't adding up.*, [deleted], r/InvestmentClub, January 31, 2021. https://www.reddit.com/r/InvestmentClub/comments/l93gui/

[10] *TfidfVectorizer*, scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[11] *KMeans*, scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[12] *PCA*, scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[13] *How does the stock market perform in a Recession?*, [deleted], r/InvestmentClub, March 15, 2022. https://www.reddit.com/r/InvestmentClub/comments/tecbvy/

[14] Reddit user *Zurevu*. https://www.reddit.com/user/Zurevu/

Appendix A Additional Figures

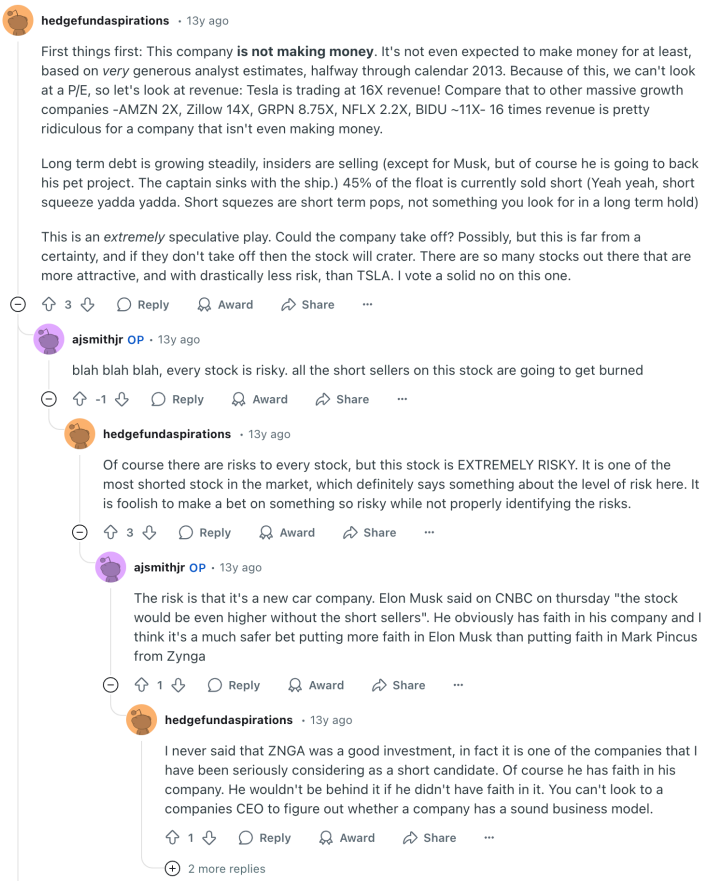


Figure A-1: Reddit Comments Hierarchical Tree Structure

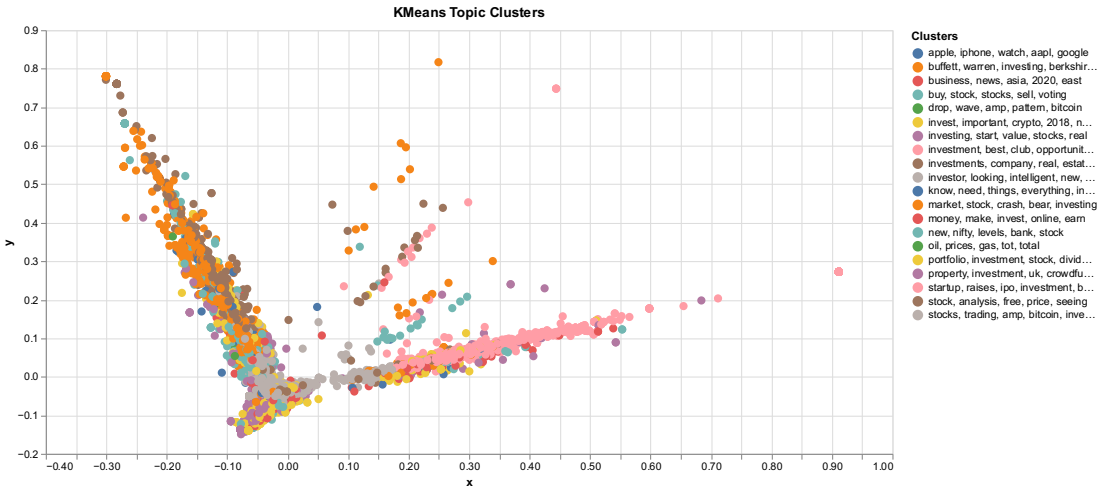


Figure A-2: Visualisation of Topic Clusters based on KMeans Clustering