Model, Analyze, and Comprehend User Interactions and Various Attributes within a Social Media Platform

Md Kaykobad Reza*, S M Maksudul Alam†, Yiran Luo‡ and Youzhe Liu§
Department of Computer Science, University of California, Riverside
Email: *mreza025@ucr.edu, †salam031@ucr.edu, ‡yluo147@ucr.edu, §yliu908@ucr.edu

***Abstract*—How can we effectively model, analyze, and comprehend user interactions and various attributes within a social media platform based on post-comment relationship? In this study, we propose a novel graph-based approach to model and analyze user interactions within a social media platform based on post-comment relationship. We construct a user interaction graph from social media data and analyze it to gain insights into community dynamics, user behavior, and content preferences. Our investigation reveals that while 56.05% of the active users are strongly connected within the community, only 0.8% of them significantly contribute to its dynamics. Moreover, we observe temporal variations in community activity, with certain periods experiencing heightened engagement. Additionally, our findings highlight a correlation between user activity and popularity showing that more active users are generally more popular. Alongside these, a preference for positive and informative content is also observed where 82.41% users preferred positive and informative content. Overall, our study provides a comprehensive framework for understanding and managing online communities, leveraging graph-based techniques to gain valuable insights into user behavior and community dynamics.**

***Index Terms*—User interaction modeling, Social media analysis, Post-comment relationship, Strongly Connected Component**

I. Introduction

SOCIAL media platforms have become integral components of our modern life [1], offering spaces for users to engage, share content, and interact with one another. Understanding the dynamics of user interactions within these platforms is important for various applications, including understanding user behavior [2], [3], community management [4], [5], content recommendation [6], [7] etc. In this study, we present a novel graph based approach on modeling, analyzing, and comprehending user interactions within a social media platform, focusing specifically on the post-comment relationship.

Availability of huge amount of social media data provides researchers with vast opportunities to explore user behavior and community dynamics. Previous studies have focused on analysing social media data from different aspects. This includes popularity prediction [8], [9], sentiment analysis [10], [11], user preference prediction [12], business decision making [13], public opinion analysis [14] etc. However, how people connect and interact with each other and what are the things they collectively prefer as a community is still under-studied. Effectively modeling and analyzing these interactions present significant challenges due to the sheer volume and complexity of the data and interaction among the users.

To address these challenges, we propose a novel graph-based approach to model and analyze user behavior and community dynamics. We construct a user interaction graph based on the post-comment relationships among users. Then we analyze different properties of the graph to understand the dynamics and preferences both from user and community perspectives. Our study aims to answer several key questions including:

- How does the interaction of a social media community look like based on post-comment relationship?
- If there exists any sub-group, how strongly are the members of the sub-groups connected to each other?
- Who are the most popular and active users of the community? Are they connected to each other?
- What kind of posts, comments and contents are preferred and go trendy throughout the community?

We investigate the structure of user interactions within the community, identifying clusters of users and examining the interconnectedness among them. One of the notable contributions of our work lies in characterizing user behavior within the social media platform. By analyzing post and comment patterns, along with user attributes such as upvotes, downvotes and comment counts, we gain insights into the preferences and tendencies of users. Additionally, we evaluate community preferences by looking into the trends in post and comment content, highlighting the types of content that resonate most with the community.

Throughout our analysis, several interesting findings emerge regarding the nature of social media interactions. We observe that about 55.44% of the community members are active. Among the active users, 56.05% are part of the largest user cluster. While there are a few very small user clusters, around 37.06% users are not part of any cluster, meaning they are not strongly connected to the community. Though we have 11,875 active members, only 0.8% of them significantly contributed to community dynamics. Moreover, we find that community activity varies over time, with certain periods, including admission and internship periods, witnessing increased engagement. Additionally, our analysis reveals a correlation between user activity and popularity which shows that more active users tend to be more popular. A preference for positive and informative

content among users is noticed as 82.41% of the total posts and comments reflected positive tones and attitudes.

In summary, our work offers a comprehensive framework for modeling, analyzing and understanding user interactions within social media platforms. By leveraging graph-based techniques, we provide insights into community dynamics, user behavior, and content preferences, paving the way for enhanced understanding and management of online communities.

II. Related Work

In social media analytics, particularly on reddit, numerous studies focus on understanding user engagement and content popularity. Kim et al. [15] developed a machine learning model to predict reddit post popularity, finding neural networks most effective. Glenski et al. [16] recorded the behavior of 186 reddit users, presenting statistics on their interactions and finding simple models that could predict user behavior. Research also focuses on content, like the work from Barnes et al. [17] analyzing 129,326 memes during COVID-19 [18], [19], predicting meme popularity using machine learning.

Apart from popularity, studies on reddit data cover different aspects. Balsamo et al. [20] investigate nonmedical opioid use, observing trends like synthetic opioids and rectal administration. Sawicki et al. [21] highlight reddit as a data source for science, analyzing 180 manually annotated papers. Melton et al. [22] conduct sentiment analysis on reddit discussions about COVID-19 vaccines, finding more positive than negative sentiments and focusing on side effects over conspiracy theories.

Reddit, a leading social platform, is a popular source for data mining and analysis. Extensive research has been conducted on popularity [8], [9], examining user behaviors [1], [3], content moderation [6], [7], and making business related decisions [13]. These studies reveal patterns in user engagement and community formation. However, how users are intereconnected to each other and how they interact to form a community preference is still not studied properly. Our project builds upon this foundation, utilizing graph based approach to model complex user interaction, analyze different attributes and reveal useful insights to understand different community dynamics.

III. Methodology

In this section, we will first outline data collection procedure. Subsequently, we will describe the methodology used to construct the user interaction graph, followed by an overview of the cluster analysis procedures. Finally, we will explain the procedures employed for user and post related analysis.

A. Data Collection

We collected our data from the r/ucr subreddit spanning a period of two years, specifically from January 2022 to December 2023. Due to restrictions imposed by the reddit API, we utilized the Arctic Shift[1] website as our tool for data collection form reddit. The summary of the dataset is shown in Table I. r/ucr subreddit community has 21,419 members, with 11,875 being active participants during the specified time frame. After cleaning and preprocessing the data, we found a total of 18,037 posts and 107,102 comments. This dataset provided a comprehensive overview of community interaction and engagement. Our collected data includes various attributes including post titles, post content, comments, upvotes, downvotes, thread-related information to capture post-comment hierarchy, timestamps and the usernames of post/comment authors. This dataset is the foundation for our subsequent analyses of user interaction, behavior and community dynamics within the r/ucr subreddit.

TABLE I: Summary statistics of our dataset.

Attribute	Value
Number of Users in the Subreddit	21, 419
Number of Active Users	11,875
Number of Posts	18,037
Number of Comments	107,102

[1]https://arctic-shift.photon-reddit.com/download-tool

B. Building the User Interaction Graph

In this section, we will first illustrate how we build the user interaction graph from post-comment relationship. Then we will introduce the methodology and details of how we build the our graph based on the r/ucr subreddit dataset, as well as a critical analysis of the structural features and communities in the subreddit. Besides, we also show multiple visualizations of the graphs we build, including different date ranges or applying random selection for the dataset, which gives a clearer view of the network structure and interesting features.

1) Building User Interaction Graph from Post-Comment Relationship: Our approach to build the user interaction graph is illustrated in Figure 1. We adhere to several key principles. Firstly, we represent each user in the platform as a node within the graph. Then we add directed edges between users based on their involvement in posts and comments. Specifically, for each comment made by a user, we draw a directed edge from the comment author to the author of the post as shown in Figure 1(a), ensuring that the flow of interaction is appropriately depicted. Furthermore, we connect each node to its immediate parent post/comment author rather than to the top-level post author as shown in Figure 1(b). This choice ensures that interactions are traced back to their original context and hierarchy within the platform. Additionally, to account for users interacting with their own content, we introduce self-loops to nodes where authors comment on their own posts or comments as shown in Figure 1(c). Finally, we do not consider explicit mentioning in comments; instead, we focus on the direct parent-child relationship as shown in Figure 1(d). This ensures that the graph accurately reflects the immediate interactions between users. By adhering to these principles, our constructed user interaction graph provides a comprehensive representation of user engagement and interaction dynamics within the social media platform.

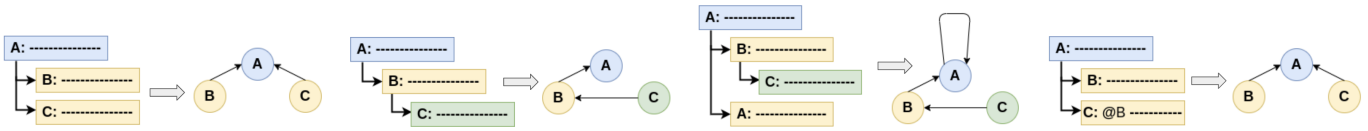


Fig. 1: We build the user interaction graph based on post-comment relationship. Here each node is a user. (a) We draw a directed edge from comment author to post author. (b) Nodes are connected to their immediate parent post/comment author, not to the top-level post author. (c) If any author comments on their own post/comment, we add a self loop to the corresponding node. (d) We do not consider explicit mentioning in any comment. Even if C is mentioning B in the comment, C is still connected to A as it is the immediate parent of C.

2) Building the Network from Dataset: The dataset is a CSV-format file with a header as follows:

Author , author_fullname , created , downs , ups , post_id , parent_id , permalink , Score , post , title , subreddit_subscribers , upvote_ratio , post_name , Parent_post_author , group_per_month , sentiment

As we can see, the crawled dataset includes detailed information about the users, the contents of their post/comments along with the parent post/comment information. Therefore, we can easily build a network based on the relations. We first build an adjacency list representation of the graph from the dataset using the parent information of each post/comment. Then we use the Python library NetworkX [23] to build the graph from the dataset and use gravis [24] for visualization.

3) Basic Network Visualization: Since the entire dataset is huge, we first draw a graph from a subset of data from the dataset. The graph includes several posts or comments. We can see from Figure 2 that the features of the connections are very clear. First, the users typically engage in sparse interactions, meaning most users are not influencer in a post or comment instance. Conversely, a smaller subset of the users demonstrated exceptionally high engagement levels, often acting as central nodes within the network due to their dense posting and commenting behaviors. We can view these dense sub-graphs as trending posts or comments. This dichotomy underscores the varied nature of user participation within the subreddit, ranging from passive observers to active contributors.

The graph contains nodes with two different colors. The Blue ones are commenters with only out-degree edges but not in-degree edges. While the red ones are both commenters and posters and they have both in-degree and out-degree edges. The red nodes make 24.4% of the total users.

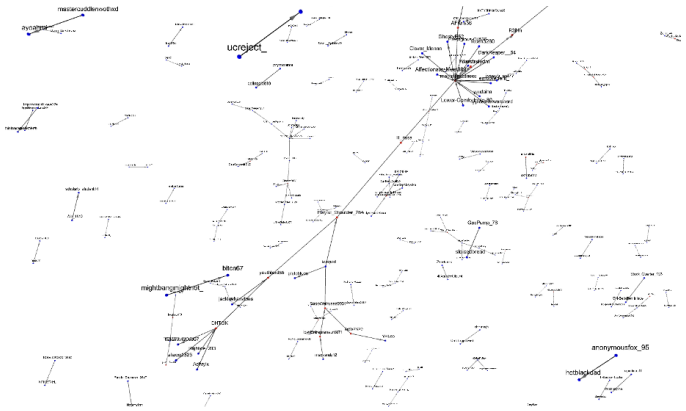


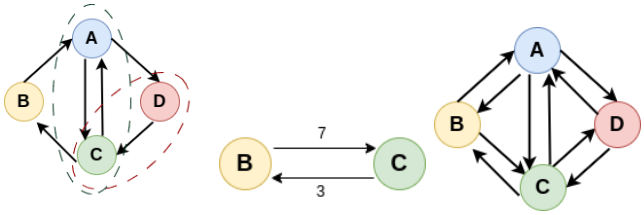
Fig. 2: User interaction graph on a subset of data.

C. Clustering

We analyzed the User Interaction Graph (UIG) built in III-B to find the clusters of users, i.e., sub-groups in the community. To build the clusters, we primarily focused on the user-to-user connection through post-comment relationship. According to our definition, each of the clusters are a strongly connected component in the UIG. Additionally, we have measured the close ties among the users of the community and redefined the the clusters with respect to the closely tied users. To analyse the clusters, we built an adjacency-list graph based on the User Interaction Graph and find out the strongly connected components (scc) in the graph. We have used *Tarjan's algorithm* [25] for finding the *ssc* in the graph. The the following sub-subsections, we will describe different parts of our clustering methods. The result cluster analysis has been shown in the subsection IV-B.

1) Weakly Formed Clusters (WC): In a Weakly Formed Clusters (WC), users in the User Interaction Graph are connected to their neighbor-users bidirectionally either by direct connection or by a indirect connection via other users in the clusters. We did not follow any restrictions such as close ties between adjacent users, minimum number of interactions and so on while finding the clusters. Fig 3(a) shows an example of WC. In this case, we can see that the adjacent nodes 'a' and 'c' are connected strongly with a direct connection to each others. On the other hand, the adjacent nodes 'c' and 'd' are connected bidirectionally with the help of node 'B' and 'A'. However, with a level of flexibility in our definition of WC, 'A', 'B', 'C' and 'D' are considered as the members of a single cluster.

2) Closely Tied User Pair (CTUP): We tried to measure the close ties among the users in the community with respect to their interactions, i.e., post-comment relationship to each others. For any pair of users in the community, if 3 or more comments were made by one user to the other one in the pair, we called the pair as a Closely Tied User Pair (CTUP). We are considering 3 as the threshold because, we know, repeating a task more than 2 times might not be incidental



(a) (b) (c)

Fig. 3: (a) WC (b) CTUP (c) SC

rather intentional. Fig 3(b) shows an example of CTUP where we can see that user 'B' got 3 comments from his adjacent user 'C'. On the other hand, user 'C' got 7 comments from the adjacent user 'B'. We developed a formula to calculate *Tie-Score* between any two users in the community.

$$\text{Tie-score} = \frac{\sum_{ij,ji}(ij, ji)}{0.4 \cdot \text{diff}(ij, ji)} \tag{1}$$

In Equation 1, 'ij' is the number of comments from user 'i' to 'j' and in the same way, 'ji' is the number of comments from user 'j' to 'i'. We ranked all the CTUPs according to the descending order of *Tie-score* to get top closely tied user pairs.

3) Strongly Formed Clusters (SC): Taking CTUPs into consideration, we implied a level of restriction to our definition of cluster. In Strongly Formed Clusters, each of the adjacent user-pairs must have to be a CTUP. So, we can say that the adjacent users in a SC are closely tied by the post-comment relationship. Fig 3(c) shows an example of SC. In this case we can see that all the adjacent users such as: ('A','B'), ('A','D'), ('B','C') and ('C','D') are directly connected bidirectionally and because of being CTUPs, they maintain close ties to each others.

D. User and Post Analysis

After collecting and pre-processing all the posts and comments from January 2022 to December 2023 from r/ucr subreddit, we analyzed them in-depth from both user and post-comment perspectives. We process the data further and sort it according to different attributes. Then we go through the top results manually to understand the topic, content, context and user behavior properly. We used YAKE [26] library to automate the keyword extraction process. Specifically we are interested to inspect the following things:

- Top 10 most upvoted users
- Top 10 most active users
- Top 10 most downvoted users
- Top 10 topics
- Top 10 most upvoted posts/comments
- Top 10 most commented posts/comments
- Top 10 most downvoted posts/comments

Our analysis has shown some interesting observations which we describe in Section IV.

IV. Result Analysis

In these section, we analyze our findings in detail. First we describe the outcome form the analysis of the user interaction graph. Then we discuss about clustering. Finally we discuss about the findings from user and post analysis.

A. Analysis of the User Interaction Graph

From the originally built network, as described in Section III-B2, we can draw many interesting conclusions. As a relatively anonymous forum, most users don't know each other. Most users only randomly reply to the posts they are interested in. However, some small communities actually exist with the same interests. For example, some users may comment multiple times on posts with similar topics.

With the basic findings, we build graphs with different features to find out more details. We build graphs within certain ranges of dates, spanning a month. We show the connections based on three different times in Figure 4. We can see that within a certain time period, the graph is more dense, and most of the users are connected. Since the r/ucr subreddit is a small community, there are not many posts/comments in a month. Also, in a short time period, the trending topics are usually consistent, which leads to a more dense network containing the users who are interested in the topics. Also, within each time period, we can find an influencer with the highest degree marked red in Figure 4.

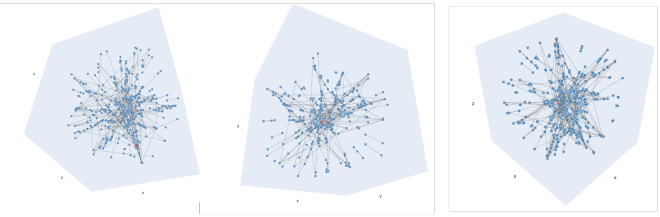


Fig. 4: User interaction graph in different months

Then, we add weights to the graph to see how many times a user comments on other user's post/comment; the graph is also sampled from subset of posts and comments, which can be viewed more clearly. We show the snapshot of one part of the graph in Figure 5.

Most of the edges have a weight of 1, which means the user only comment once in the selected posts/comments. Some of the active users may comment multiple times so the nodes have a weight of 2 or 3. This corroborates our earlier conclusions again, that is most of the users post or comment casually instead of building a community to make real connections with other users. The high-density sub-graphs are trending posts/comments that attract more users to comment under them.

Based on this fact, we try to use some community detection algorithms to see whether there are solid communities in the subreddit. We will describe the algorithms and visualizations we conduct in the following section.

1) Community Detection: Community detection is a classic task on social networks to find strongly connected groups or subgraphs in a large-scale network. We surveyed several

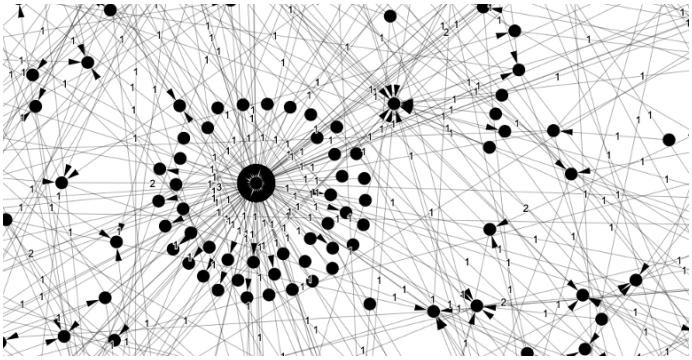


Fig. 5: Network with edge weights

different algorithms for community detection using different metrics, including centrality-based community detection [27], Leiden community detection [28], and walktrap community detection [29]. Finally, based on the precious findings from the visualizations, we select the simple centrality-based one, which is implemented in NetworkX library. Because there are not actually clear community structures in the whole network, the algorithm is relatively simple and fast enough to show the community structures.

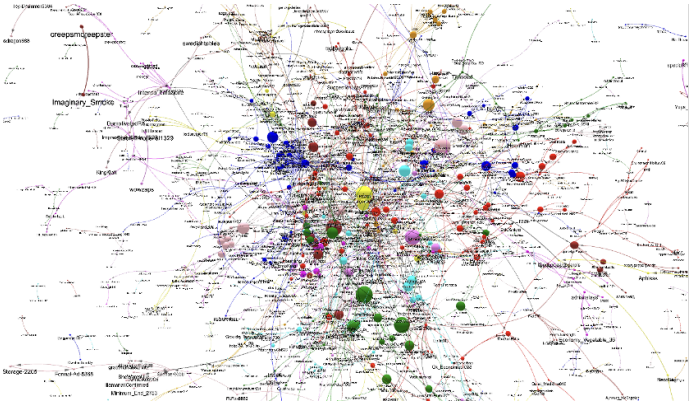


Fig. 6: Community detection visualization

We can see from Figure 6, the number of communities detected is very large, which means there is no clear boundary between different large groups. Most users post and comment on random activities instead of on a friend group, such as the connections on Twitter or Facebook. These sub-communities represent groups of users frequently interacting with one another, often around specific topics or shared interests. Such density indicates not only a higher level of engagement within these groups but also the potential for rich, substantive discourse.

Simultaneously, our analysis also suggests users who, despite the broader trend of limited engagement, maintain 1 or 2 active connections within the network. This pattern suggests a nuanced layer of interaction, where even less active users can significantly contribute to or benefit from specific threads of discussions.

2) Summary of the Network: Overall, our analysis of the r/ucr subreddit shows how people interact in interesting ways in anonymous platforms. We noticed that most people on this subreddit don't talk much. They might make a random post or comment here and there. However, there are a few users who are very active. They're like the stars of the subreddit, always posting and commenting a lot, which is a commonly found rule of online platforms [30], [31]. We also found small groups where people are talking a lot. These groups usually form around hot topics or really popular posts. It is like when something interesting happens, and everyone gathers around to talk about it.

This study tells us that even on an anonymous forum like this subreddit, where you might think people just come and go without really connecting, there are still some small, lively discussions happening. It's like finding little pockets of friends chatting at a big party where most people are just passing by. In the end, our investigation into the r/ucr subreddit helps us see all the different ways people can come together online, even if it's just for a short while or around something specific that everyone's excited about. It shows us how this subreddit is a place full of different stories and connections.

B. Cluster Analysis

We found different statistics based on the definition of WC and SC. All the statistics showed that the most of the users in the community are connected to others and a part of clusters. However, according to definition and theoretical proof, the clusters are non-overlapping for each of the cases.
We were able to identify 62 WCs and 135 SCs in our user interaction graph. Table II shows that the largest WC and SC contained 6,657 users and 624 users respectively in total. The data indicated that most of the users in the community are connected to each others either by a direct or indirect connection. At the same time, we can realize that most of the clusters in the community contain small number of users.

TABLE II: Summary statistics for WC and SC

# of WC	# of SC	# of users
1	-	6,657
-	1	624
1	2	7
-	2	5
-	4	4
9	19	3
51	107	2

The statistics shows that because of implying the CTUP restriction to the definition of cluster, the largest WC got split into multiple smaller SCs.

C. User Analysis

We did an in-depth analysis to understand who are the most influencing users in the community and why people love them. Besides that, we also analyzed who are the most disliked users in the community and why. In this section we will discuss our findings.

1) Top 10 Most Upvoted Users: We aggregated the data to find the top 10 most upvoted users in the subreddit. Then we analyzed the posts and comments made by them manually. We found out that people generally upvote and engage in comment to their post because they:

- Share informative posts and comments
- Answer question with positive attitudes
- Posts and comments are generally short, to-the-point and accurate
- Sometimes humorous

We also wanted to see if these users are connected to each-other. In other words, if they comment on each other's posts/comments. We show the interaction using a heatmap on Figure 7(a). It shows that the top 10 most upvoted users comment on each other often. Which means they are connected by post-comment relationship and are within the same cluster.

2) Top 10 Most Active Users: We listed down the top 10 most active users in the community. Here we consider post and comment as activities. People who make more posts and comments are considered to be more active. We found out an interesting trend. The people who are more active tend to be more popular in general. In other words, the more active a user gets, the more likely he is to get higher upvotes and comments. We found around 80% overlap between the top 10 most active users and top 10 most upvoted users. That suggests that 80% of the top active users are also top upvoted users in the community. Like the top upvoted users, top active users also like to comment on each others post very often as shown on Figure 7(b).

3) Top 10 Most Downvoted Users: We also analyzed the aggregated data to find the top 10 most downvoted users in the subreddit. Then we analyzed the posts and comments made by them manually to see why people dislike them so much. Our analysis showed that these users generally:

- Make uninformative or useless posts/comments
- Criticize people or show arrogant attitude
- Do political talks (which disliked by the students)
- Argue with people over simple things

We also wanted to see if these users are connected to each-other. As can be seen from Figure 7(c), these users comment on each other very very rarely. Which means they are not that connected by post-comment relationship.

D. Post Analysis

We did an extensive analysis of the posts and comments in our dataset. We are going to report our findings from the following four angles:

1) Top 10 Topics: The first thing we're interested in is what are the hot topics people of r/ucr love to discuss? If we can find the top topics, we will be able to have a good understanding of what the community is about. With the python library YAKE! (Yet Another Keyword Extractor), we managed to find the top 10 topics of r/ucr during the time frame.

From Table III, we can see the top 10 topics that the users of r/ucr discuss the most. The first two topics are about full-time/part-time students, which is reasonable because there is quite a huge difference between policies for part-time students and full-time students. People may have questions about how to convert between these two status or whether one should convert, so they seek suggestions by posting on r/ucr. The following topics can be grouped into four categories: class taking, work and job, school affairs, financial aid. These are indeed topics that university students usually care about and talk about most.

TABLE III: Top 10 topics from our dataset.

Serial	Topic
1	Full time student
2	Part time student
3	Taking summer classes
4	Work part time (part time job)
5	School year starts
6	High school students
7	Classes you're taking
8	Work full time
9	Financial aid office
10	Student account online

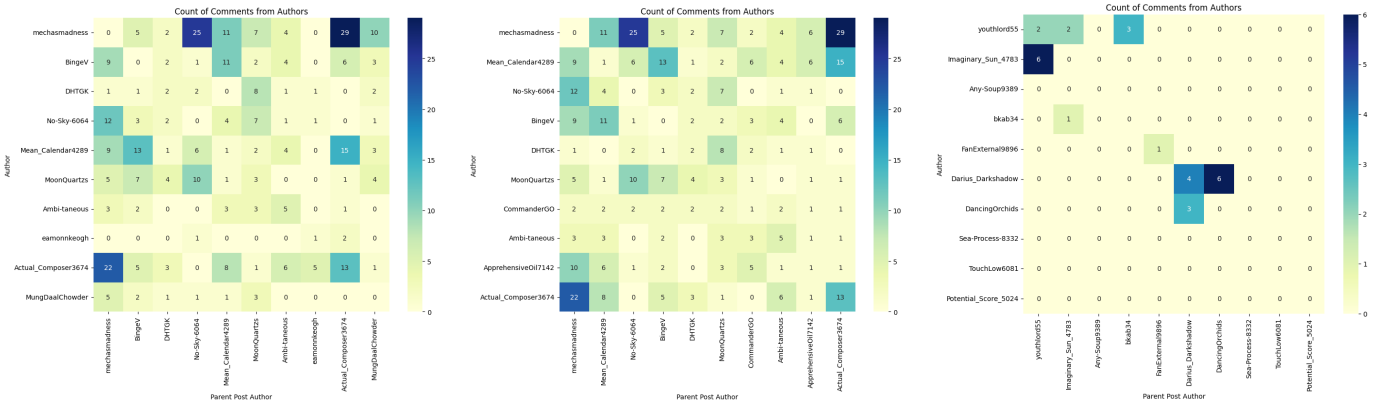
We also generated a word cloud from the top 50 topics as shown in Figure 8.

2) Top 10 Most Upvoted Posts: Next analysis is to find the top 10 upvoted posts and applied detailed analysis on their title, content, comments, and other important information, trying to understand what they have in common and what's special about them that insist people to upvote. We found the top 10 most upvoted posts can be roughly categorized into the following three kinds:

- **Funny memes.** This is very easy to understand. Users of r/ucr are mostly current students at UCR or alumni. They are young, active, humorous and love to make and spread funny memes. Consequently, posts about memes are likely to get many upvotes.
- **Sharing of inspiring news.** Examples are "Took me 7 years but I am a doctor y'all!" got 715 upvotes and "I GOT ACCEPTED INTO UCR!!" got 412 upvotes. This particular kind of posts usually get high upvotes because people love celebrating accomplishments, especially graduation and acceptance in university-related communities like r/ucr.
- **Safety concerns.** Examples are "Is this area safe?" got 561 upvotes and "The Botanical Garden was robbed of most of their equipment :(" got 385 upvotes. This category of posts usually get high upvotes because safety is a topic of universal interest, and they can raise awareness about the challenges faced by the university and students. Students spend most of their time on campus and they want the environment to be safe and secure. So they tend to engage in safety-related discussions, providing their experiences, suggestions and concerns.

3) Top 10 Most Commented Posts: We also managed to find the top 10 most commented posts. Even though this work seems similar to top 10 upvoted posts and there are indeed overlaps, we still found different categories and drew some interesting conclusions. Top 10 commented posts can be roughly categorized into the following categories:

- **Sharing of inspiring news.** This category is the same



(a) (b) (c)

Fig. 7: Interaction between (a) Top 10 most upvoted users. (b) Top 10 most active users. (c) Top 10 most downvoted users.

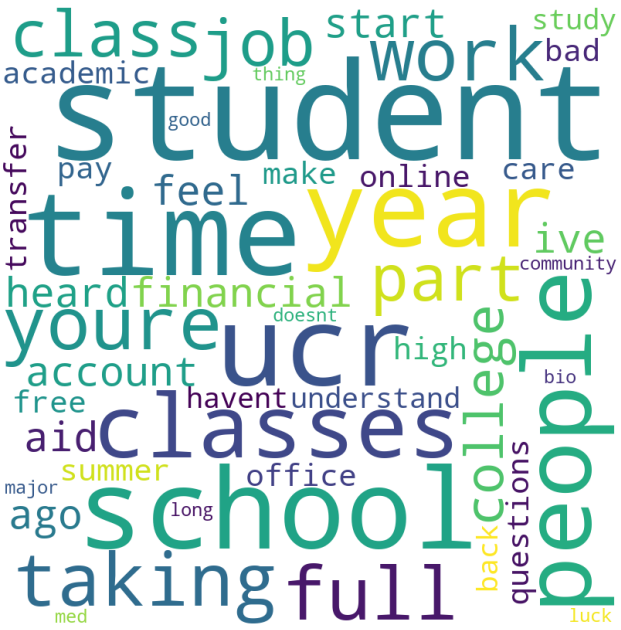


Fig. 8: Word cloud from top 50 topics

as the one in the top 10 upvoted posts. The examples "Took me 7 years but I am a doctor y'all!" and "I GOT ACCEPTED INTO UCR!!" from the top 10 upvoted posts also appear in the top 10 commented posts, and they both got 59 comments. This is not strange in a university-related community like r/ucr because these things are the most important for students and can resonate most with students.

- **Surveys and polls.** This is a category that usually gets many comments but not many upvotes. Examples are "What's your favorite professor(s)" got 40 upvotes and 59 comments, "What's your favorite songs at the moment?" got 32 upvotes and 69 comments, "GPA that got u in?" got 26 upvotes and 93 comments. The reason is that people love to comment and share their opinions on surveys, and they are very likely to find friends sharing mutual interests. They don't get a high number of upvotes because they can't stir up some kind of strong feeling in people's hearts.
- **Other Miscellaneous Topics.** There are some posts not from the two big categories mentioned above that got many comments. Examples are admission decision discussion threads or posts about finding friends on campus. People comment on these posts to exchange information and discuss. They don't get a high number of upvotes because of the same reason as surveys and polls: people just come here to find or share information, and they can't arouse people's feelings.

4) Top 10 Most Downvoted Posts: Apart from the previous two analyses on the top upvoted or commented posts, we also collected the top 10 most downvoted posts and analyzed them, trying to find what they had in common and why people dislike these contents. Unlike the posts people love, these posts are hard to categorize, and they are disliked due to different reasons:

- **Offensive speech.** Example is that "Some students got in UCR bc of our social mobility quota" got 98 downvotes.
- **Selfish speech.** Example is that "Covid is extremely low risk and UCR should reopen" got 86 downvotes.
- **Lack of empathy.** Example is that "Yeah, I hope you're also sorry when he commits suicide as well.Stalking is not acceptable, but treating people like TRASH is not acceptable either." got 103 downvotes because the author is speaking for the stalker instead of the girl victim.
- **Patronizing tone.**
- **Inappropriate humor.**
- **Oversimplification of facts and stereotypes.**

V. Discussion, Challenges and Future Directions

Though the community prefer mostly positive attitudes, we found around 17.59% posts and comments containing negative, aggressive or potentially inappropriate languages. The community as a whole disliked those contents and showed their protest by downvoting the contents. Since this is a university based community, most of the contents are related to study, job, courses and similar topics which is intuitive. Despite some

serious posts and comments, we found the overall vibe to be collaborative, helpful and sometimes humorous.

We faced several challenges while working on this project. One bigger challenge was to analyze the user interaction graph on the large volume of data. The graph was very dense and nearly impossible to analyze visually. We started with month by month analysis and later switched to data analysis methods to analyze different parts numerically. Also, inspecting the posts and comments to understand why people like or dislike some content was challenging. We used some python library like YAKE and NetrorkX to automate some part of the analysis. Other parts like understanding the context, content and user behavior were still manual.

Future research efforts could explore more nuanced aspects of community dynamics, such as the role of influential users in shaping community discourse and the impact of algorithmic recommendations on user engagement and content dissemination. Additionally, investigating the effectiveness of different community management strategies in promoting positive interactions and mitigating negative behaviors could provide valuable insights for platform administrators and moderators.

VI. Conclusion

Our project shed light on several interesting aspects of the community dynamics. Firstly, we observed strong interconnectivity among users, where aorund 56.05% of the active individuals are part of a larger cluster. However, it's notable that a very small subset of users, only around 0.8%, significantly influences the community's activity and dynamics, showing a potential power law distribution in user engagement. During the 2 years timeframe, around 55.55% of the total community members were active. Additionally, we found a correlation between user activity and popularity, indicating that more active users tend to be more interconnected and influential within the community. Furthermore, the preference for positive and informative content over negative, critical, or political discourse highlights the community's inclination towards constructive and uplifting interactions. Around 82.41% of the total posts and comments are positive which contributes to maintaining a positive atmosphere within the community, fostering a conducive environment for collaboration and exchange of ideas. In conclusion, our findings offer valuable insights into the dynamics of the community. It emphasizes the importance of understanding user behavior, interaction patterns, and content preferences. These insights can inform community management strategies aimed at fostering engagement, promoting positive discourse, and enhancing overall community cohesion.

References

[1] E. E. H. E. E. Hollenbaugh, "Self-presentation in social media: Review and research opportunities," *Review of communication research*, vol. 9, 2021.

[2] M. Garg, "Mental health analysis in social media posts: A survey," *Archives of Computational Methods in Engineering*, vol. 30, no. 3, pp. 1819–1842, 2023.

[3] A. M. A. Ausat, "The role of social media in shaping public opinion and its influence on economic decisions," *Technology and Society Perspectives (TACIT)*, vol. 1, no. 1, pp. 35–44, 2023.

[4] M. R. Ohara, "The role of social media in educational communication management," *Journal of Contemporary Administration and Management (ADMAN)*, vol. 1, no. 2, pp. 70–76, 2023.

[5] N. Proferes, N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer, "Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics," *Social Media+ Society*, vol. 7, no. 2, p. 20563051211019004, 2021.

[6] L. Madio and M. Quinn, "Content moderation and advertising in social media platforms," *Available at SSRN 3551103*, 2023.

[7] M. Singhal, C. Ling, P. Paudel, P. Thota, N. Kumarswamy, G. Stringhini, and S. Nilizadeh, "Sok: Content moderation in social media, from guidelines to enforcement, and research to practice," in *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2023, pp. 868–895.

[8] Q. Cao, H. Shen, J. Gao, B. Wei, and X. Cheng, "Popularity prediction on social platforms with coupled graph neural networks," in *Proceedings of the 13th international conference on web search and data mining*, 2020, pp. 70–78.

[9] S. Carta, A. S. Podda, D. R. Recupero, R. Saia, and G. Usai, "Popularity prediction of instagram posts," *Information*, vol. 11, no. 9, p. 453, 2020.

[10] K. Chakraborty, S. Bhattacharyya, and R. Bag, "A survey of sentiment analysis from social media data," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 450–464, 2020.

[11] N. K. Singh, D. S. Tomar, and A. K. Sangaiah, "Sentiment analysis: a review and comparative analysis over social media," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 1, pp. 97–117, 2020.

[12] E. Arrigo, C. Liberati, and P. Mariani, "Social media data and users' preferences: A statistical analysis to support marketing communication," *Big Data Research*, vol. 24, p. 100189, 2021.

[13] ——, "Social media data and users' preferences: A statistical analysis to support marketing communication," *Big Data Research*, vol. 24, p. 100189, 2021.

[14] X. Dong and Y. Lian, "A review of social media-based public opinion analyses: Challenges and recommendations," *Technology in Society*, vol. 67, p. 101724, 2021.

[15] J. Kim, "Predicting the popularity of reddit posts with ai," *arXiv preprint arXiv:2106.07380*, 2021.

[16] M. Glenski and T. Weninger, "Predicting user-interactions on reddit," in *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, 2017, pp. 609–612.

[17] K. Barnes, T. Riesenmy, M. D. Trinh, E. Lleshi, N. Balogh, and R. Molontay, "Dank or not? analyzing and predicting the popularity of memes on reddit," *Applied Network Science*, vol. 6, no. 1, p. 21, 2021.

[18] Y. Shi, G. Wang, X.-p. Cai, J.-w. Deng, L. Zheng, H.-h. Zhu, M. Zheng, B. Yang, and Z. Chen, "An overview of covid-19," *Journal of Zhejiang University. Science. B*, vol. 21, no. 5, p. 343, 2020.

[19] K. N. Hafiz and K. F. Haque, "Convolutional neural network (cnn) in covid-19 detection: A case study with chest ct scan images," in *2022 IEEE Region 10 Symposium (TENSYMP)*, 2022, pp. 1–6.

[20] D. Balsamo, P. Bajardi, A. Salomone, and R. Schifanella, "Patterns of routes of administration and drug tampering for nonmedical opioid consumption: data mining and content analysis of reddit discussions," *Journal of Medical Internet Research*, vol. 23, no. 1, p. e21212, 2021.

[21] J. Sawicki, M. Ganzha, M. Paprzycki, and A. Bădică, "Exploring usability of reddit in data science and knowledge processing," *arXiv preprint arXiv:2110.02158*, 2021.

[22] C. A. Melton, O. A. Olusanya, N. Ammar, and A. Shaban-Nejad, "Public sentiment analysis and topic modeling regarding covid-19 vaccines on the reddit social media platform: A call to action for strengthening vaccine confidence," *Journal of Infection and Public Health*, vol. 14, no. 10, pp. 1505–1512, 2021.

[23] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2008.

[24] "Gravis library, https://robert-haas.github.io/gravis-docs/."

[25] R. Tarjan, "Depth-first search and linear graph algorithms," in *12th Annual Symposium on Switching and Automata Theory (swat 1971)*, 1971, pp. 114–121.

[26] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "Yake! keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, 2020.

[27] L. C. Freeman *et al.*, "Centrality in social networks: Conceptual clarification," *Social network: critical concepts in sociology. Londres: Routledge*, vol. 1, pp. 238–263, 2002.
[28] V. A. Traag, L. Waltman, and N. J. Van Eck, "From louvain to leiden: guaranteeing well-connected communities," *Scientific reports*, vol. 9, no. 1, p. 5233, 2019.
[29] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Computer and Information Sciences-ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20*. Springer, 2005, pp. 284–293.
[30] D. Ye and S. Pennisi, "Analysing interactions in online discussions through social network analysis," *Journal of Computer Assisted Learning*, vol. 38, pp. n/a–n/a, 01 2022.
[31] S. Serpa, "Digital society and digital sociology: One thing leads to the other," *Science Insights*, vol. 38, pp. 314–316, 08 2021.