The Impact of User Interaction on College Subreddits on Mental Health

Saksham Arora, Meher Kalra, Cody Theobald

Math 76.01: Final Project Report

Abstract

The prevalence of mental health issues among college students has been a growing concern in recent years. Platforms like Reddit, with its specific college subreddits, offer a unique lens to examine the intersection of user interaction and mental health. This study focuses on four college subreddits - 'r/uiuc', 'r/berkeley', 'r/harvard', and 'r/cornell' - to investigate the dynamics of these online communities. We collected data from these subreddits, including user details, scores, upvote ratios, number of comments, post text, and comment text. Leveraging network analysis and sentiment analysis techniques, we aimed to uncover patterns in user interactions and sentiments within these communities. The primary goal of this study was to understand the emotional landscape of these online communities and investigate its correlation with indicators of mental health. We hypothesized that more influential users tend to post more positive content, users who connect different communication channels are more likely to provide emotional support, and users with higher network centrality scores are more likely to receive emotional support. Through our analysis, we aimed to provide insights that could potentially help identify at-risk individuals within these communities and understand the role of user interactions in shaping the overall sentiment of these online spaces.

1 Introduction

Online social platforms have emerged as valuable sources of data for studying human behavior and mental health. With the growing concern over the mental well-being of college students, investigating the impact of user interaction on mental health within specific online college communities becomes crucial to identify what factors drive the mental health of college students.

Reddit has become one of the most prominent social media platforms on the web with over 50 million daily active users [2] and over 138,000 active topical communities or "subreddits."[12]. Reddit has been the epicenter of numerous significant and contentious incidents, demonstrating the potential volatility of networked communities. These include collective endeavors to identify the individuals responsible for the Boston city marathon bombing, a notable event that underscored the potential for collective intelligence and misinformation [18]. More recently, the platform was instrumental in orchestrating a coordinated financial maneuver against short-sellers of GameStop stock, demonstrating the potential for collective action in financial markets [16]. Nonetheless, Reddit has also functioned as a platiform for interactions that are less desirable, often reflecting darker aspects of online discourse, including instances of racism [14], sexism [5] and inflammatory political discourse [13]. These instances underscore the challenges of moderating large, diverse online communities.

Reddit has emerged as a significant platform for academic studies, fulfilling many criteria that make it a rich source of observational data. Its unique structure, which includes numerous subreddits, facilitates the process of locating relevant research data. Furthermore, unlike some social media platforms, Reddit does not impose strict character limits,

providing a more extensive dataset for researchers, both qualitatively and quantitatively. However, the use of Reddit data for academic purposes comens with its own set of complexities. The platform's diverse forms of media necessitate the application of variety of methodological approaches in research analysis. Each subreddit possesses its unique norms, cultures and moderation practice, implying the insights derived from social phenomena in one subreddit may not be universally applicable. Finally, Reddit's high degree of anonymity and the prevalence of one-time use accounts add another layer to complexity to research. The platform's relatively liberal content policies and the anonymity it provides may encourage users to engage in candid discussions. This openness can be beneficial for academic research. The candid nature of these discussions can provide researchers with a more authentic and unfiltered view of user perspectives, attitudes, and behaviors. This can lead to a deeper understanding of the phenomena under study, potentially revealing insights that might be obscured in a more controlled or formal environments. However, it also underscores the importance of ethical considerations when handling such data.

Paper Contribution

This project aims to harness data from the 'r/harvard', 'r/berkeley', 'r/uiuc', 'r/cornell' college subreddits on Reddit, to scrutinize the network dynamics and sentiment trends of user interactions, and their potential impact on mental health.

We selected the four most substantial college subreddits with the intention of enhancing the applicability of our results and identifying methods to extrapolate our findings accross other US colleges. Reddit's unique online environment encourages users to engage in dialogues, seek support, and share experiences relevant to college life.

Through the acquisition of this data, we were able to navigate the intricacies of user interactions and unearth valuable insights. Network analysis allowed us to understand the structure and dynamics of these online communities. We employed network measures such as centrality and clustering coefficient to identify influential users and understand the degree of interconnectedness within these communities.

In addition, sentiment analysis was employed to discern the emotional indicators and ascertain the dominant sentiment within posts and comments. This analsysis facilitated identification of emotional patterns and attitudes expressed by users, thereby quantifying emotions within these communities.

Finally, by pinpointing individuals who may be at risk, this project can contribute to the provision of resources, the cultivation of healthier online communitirs, and the enhancement of mental health outcomes among college students.

2 Data Collection

For our project, we collected Reddit network data using a custom script, developed in Python, in conjunction with the PRAW package [15] and Reddit API. The script navigates through the front page of a specified subreddit, which comprises the top posts at the time, and examines each post. For every post, the script traverses through each comment, recording all the users who commented on that post, along with additional variables about the users and comments. This information enables us to construct an edge list for users who post and users who comment on the same post as others.

Reddit's users, known as "redditors," have the ability to generate posts that may include a link or plaintext. Other users can then influence the visibility of these posts in the news feed by either "upvoting" or "downvoting" them. Upvotes contribute to a post's score and elevate its position in the feed, while downvotes have the opposite effect.

Beyond voting on a post, users can also contribute comments. These comments can

further be responded to by other users, resulting in a comment tree that adds an additional layer of structure to the site. An illustration of the Reddit homepage can be seen in Figure 1 while the comment structure for a Reddit post is depicted in Figure 2.

We programmed the script to execute continuously for upwards of two hours on four college subreddits 'r/harvard', 'r/berkeley', 'r/uiuc', and 'r/cornell'. We ultimately extracted data for 1,648, 2,015, 1,507 and 1,572 subscribers (users) for the aforementioned subreddits respectively.

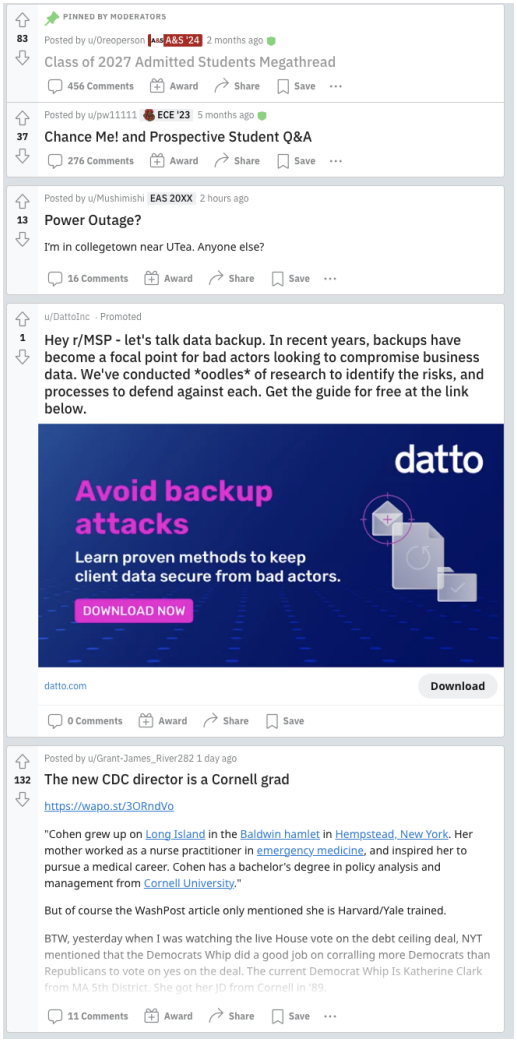


Figure 1: Subreddit Homepage

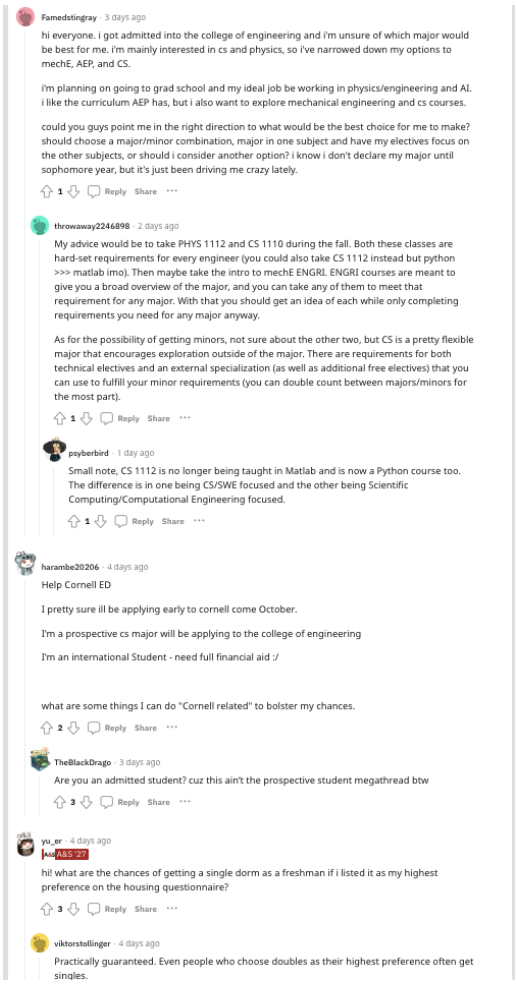


Figure 2: Subreddit Post & Comment Thread

In conclusion, we created four distinct dataframes, each representing a college subreddit. These dataframes encompass variables such as unique identifier for each post ('id'), the author of the post, the score of the post, the upvote ratio, the number of comments on the post, the name of the subreddit, the plaintext content of the post, and a compliation of all comments on the post as a list ('comments'). Figure 3 illustrates this dataframe for the 'r/cornell' subreddit.

3 Methodology

Online social platforms like Reddit provide millions of individuals unlimited access to information and connectivity. The content produced on Reddit has been shown to have a substantial impact on society, influencing social and political discussions, emergency and

[32]:

	id	author	score	upvote_ratio	num_comments	subreddit	selftext	comments
0	udjiue	zo_wee_mama	290.0	0.95	51	Cornell	I don't go to Cornell, I got to NYU but I somehow accidentally joined this group and I've been s...	[{'author': 'ConsistentScar9841', 'body': 'absolutely not. thankskska.', 'score': 213, 'created_...
1	lrg3fl	zikachu11	340.0	0.99	35	Cornell	In my FWS I was trying to be friendly, I remember asking this girl what her major was and she sa...	[{'author': 'BRF-or-bust', 'body': 'COE checks out', 'score': 183, 'created_utc': 1614182795.0},...
2	nf2k9z	zikachu11	263.0	0.90	12	Cornell	If you're interested in girls, there's really no more group of intelligent, ambitious, and compa...	[{'author': 'Itchy_Fudge_2134', 'body': 'Funny post. Totally upvoted my dude. Radtastic AND epic...
3	hdvn9a	you112233	918.0	0.99	68	Cornell	I summarized their summary with most of the bits that affect students but I would still encourag...	[{'author': 'pcsm12', 'body': 'Someone give this man mod status', 'score': 260, 'created_utc': 1...
4	kr3old	yikayika	288.0	0.97	61	Cornell	I'm super successful in the eyes of a lot of people. I go to an ivy, i'm studying a "valuable ma...	[{'author': 'cornellmanletboy', 'body': 'I certainly see where you're coming from. But personall...
...	...	...	...	...	...	...	...	...
302	qp29en	AgreeableAstronomer	254.0	0.97	43	Cornell	[The Greenest Pilgrim U'Know 📦🤝🍶 on Twitter: "Ok, so apparently @jiadagod is a camgirl who's be...	[{'author': 'None', 'body': '[deleted]', 'score': 211, 'created_utc': 1636333039.0}, {'author': ...
303	q0p56z	AGuineaHen	255.0	0.99	31	Cornell	Is it really stealing if he was on the Unlimited meal plan??? /s\n\nWhat is up with this freshma...	[{'author': 'AGuineaHen', 'body': 'UPDATE: Asked the front desk employee about it. She said it "...
304	a2ftbu	9986000min	358.0	0.99	18	Cornell	I was coming out of teagle and I spotted this bright yellow, almost blinding, object in the sky....	[{'author': 'cornellanon1998', 'body': 'This made me chuckle, may I offer you an egg in these tr...
305	k8ocrl	9986000min	268.0	0.98	18	Cornell	fuck	[{'author': 'a123bcer', 'body': 'SHIT', 'score': 83, 'created_utc': 1607375080.0}, {'author': 'S...
306	pg7tqv	4lokoluvr	273.0	0.93	18	Cornell	COVID. He was in the ICU for two weeks. He said he was proud of me for achieving my dreams (of g...	[{'author': 'FreeThinkingAmerican', 'body': 'Family comes first 100% of the time. Talk it over w...

307 rows × 8 columns

Figure 3: Generated dataframe for 'r/cornell'

disaster responses, and more, affecting offline world in tangible ways.

The central issue that motivates our methodology is understanding how the content produced and consumed on social media affects individuals' emotional states and behaviors. This is particularly relevant in the context of emotional contagion theory, which suggests that emotions can be transmitted via social networks and have long-term effects. For instance, a study conducted via Facebook [10] suggests that emotional contagion occurs online even in the absence of non-verbal cues typical of in-person interactions. Moreover, a recent neuroscience study found that engaging with a Twitter timeline stimulates emotional activity in the brain by 64% more than the typical web usage [11]. Furthermore, the act of tweeting and retweeting enhances this emotional stimulation to 74% more than the interaction with a standard website [11].

In this project, we use Reddit as a case study and explore the hypothesis of emotional contagion via network influence. We aim to understand how influence in a network modelling interactions between users, may influence the emotional states of others.

3.1 Graph Creation

In mathematics, a graph G is an ordered pair composted of a set V of vertices (or nodes) and a set E of edges: $G = (V, E)$. The set of edges E comprises of pairs $\{u, v\}$ of vertices from V, signifying that vertices u and v are connected by an edge. In this project, we use the terms network and graph interchangeably.

Graphs can be primarily categorized into two types: directed and undirected. In an undirected graph, the edges do not have a specific order (direction).

Consider, for instance, a graph representing a network of individuals. Each vertex could represent a person, and edges could be defined such that an individual A shares with another individual B if A and B have interacted on a subreddit. In this context, an interaction could be defined as commenting on the same post or replying to each other's comments. This results in an undirected graphh where each edge represents a connection or interaction between two individuals.

This undirected graph structure is particularly useful for our study as it allows us to analyze the network dynamics of user interactions within each college subreddit. By examining the structure and properties of these networks, we can gan insights into the patterns of user interactions and their potential influence on mental health conversations.

There are numerous potential methodologies for structuring a Reddit dataset into a network form. For our project, we were primarily interested in how redditors interact with each other within a specific subreddit. To facilitate this, we structured our network to

connect 1) users who have commented on the same post (poster - commenter interaction, see Fig 4) and 2) users who interact within comment threads on the same post (commenter - commenter interaction, see Fig 5). Our network is an undirected simple graph.

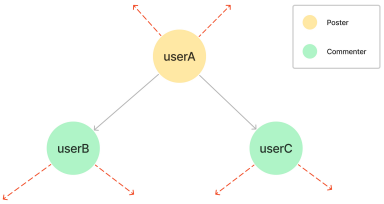


Figure 4: Schematic for Post-Commenter Graph

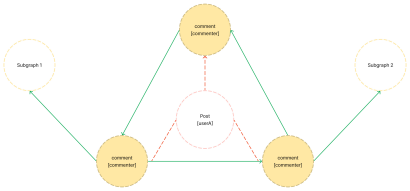


Figure 5: Schematic for Commenter-Commenter Graph

The rationale behind creating a poster-commenter network was to investigate the relationship between the influence of posters, as defined by network centrality measures, tend to post more positive content. This hypothesis is motivated by findings from a study by Ferrara et al. [6] which suggests that Twitter users who are exposed to more positive content tend to post more positive tweets themselves.

On the other hand, the commenter-commenter network was designed to capture the dynamics of user interactions within comment threads. This network structure provides a more granular view of user interactions, allowing us to understand how users engage with each other beyond the initial post. By examining these interactions, we can gain insights into the sentiment dynamics within comment threads and how they might relate to mental health of the users involved, contributing to fostering healthier online environments.

While there are alternative approaches we could have adopted, such as incorporating the nesting structure of comments where a comment can be a reply to another comment, we opted for a simpler structure for efficiency reasons. Incorporating reply relationships would have added an additional layer of complexity to our data scraping process, potentially leading to more API failures and significantly increasing the time required for data collection.

Our chosen structures allow us to collect data efficiently and relatively easily, while still capturing valuable information about how redditors interact within the subreddit. This approach aligns with our study's focus on understanding the network dynamics of user interactions within each college subreddit and their potential influence on mental health.

3.2 Sentiment & Text Analysis

We computed the sentiment scores for each comment and post to evaluate the relative positivity or negativity of the commenter's language. Sentiment analysis is a method used to quantify the positive or negative sentiment of a text by assigning each word/sentence a sentiment score and calculating an overall score for the entire text. We applied this technique to the posts and comments in our dataset.

Each post and comment's sentiment was calculated using the *vaderSentiment* [9] package in Python. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It is fast, computationally economical, and provides high accuracy [9]. VADER's performance is on par with more complex machine learning models, acheiving around 80% accuracy for binary (positive/negative) classification and around 60% for multiclass (positive/negative/neutral) classification on social media text [9].

Each post and comment's sentiment was calculated using VADER. This package assigns each word a sentiment score (a continuous value from -1 to +1), and an overall sentiment score is computed for the entire text. The overall sentiment score is outputted in the form of a dictionary with keys associated with positive ('pos'), negative ('neg') and compound sentiment having a continuous value from -1 to +1. Figure 6 illustrates the usage of *vaderSentiment* on five posts from data from 'r/cornell.'

	selftext	post_sentiment_dict
302	[The Greenest Pilgrim U'Know 🌰🤝🙏 on Twitter: "Ok, so apparently @jiadagod is a camgirl who's being targeted by some stalker, and doesn't have anything to do with the actual bomb threats. #wediditreddit #missionaccomplished" / Twitter](https://twitter.com/GregoriRomanova/status/1457502696697745409)\n\n[\\"Carter Morgan\\", most likely the fake \\"Jia Nakamura\\"](https://preview.redd.it/rvo9g26mu9y71.png?width=754&format=png&auto=webp&v=enabled&s=dc1e4feba5569b5b21f8bcece17e54837f3f6131)\n\nIn essence, some guy got mad that some e-girl didn't DM them, and instead of simply swatting them like how most of these stories end, they went a level up and called in bomb threats to numerous universities attaching their name to the threats, in an attempt to frame her?\n\n\n\n\n\nhttps://preview.redd.it/8z34pcbqu9y71.png?width=741&format=png&auto=webp&v=enabled&s=1cd7522031bcd2b1c1ed6a0e15e7b13255edee14\n\n\n\nhttps://preview.redd.it/jmt9qxfsu9y71.png?width=693&format=png&auto=webp&v=enabled&s=aef4ff36c290670a133d6d1a55ad00ac23c69786\n\n\n\nhttps://preview.redd.it/l9yosbmru9y71.png?width=678&format=png&auto=webp&v=enabled&s=087eb71cedb2e6b9ecddbd1ba223d6432602127d	{'neg': 0.182, 'neu': 0.798, 'pos': 0.02, 'compound': -0.9629}
303	Is it really stealing if he was on the Unlimited meal plan??? /s\n\nWhat is up with this freshman class tho	{'neg': 0.185, 'neu': 0.815, 'pos': 0.0, 'compound': -0.6739}
304	I was coming out of teagle and I spotted this bright yellow, almost blinding, object in the sky. It was spotted at approximately 1:34 PM Sunday December 2nd 2018. It made me feel warm (something I haven't felt in years) and made the environment around me brighter. Someone called it his "son," but idk how a child can get up in the sky. Also it seems to be moving from east to west. Anyone know what it is?\n\nEDIT: MY FIRST GOLD. It's almost as bright as whatever was in the sky	{'neg': 0.016, 'neu': 0.893, 'pos': 0.091, 'compound': 0.7407}
305	fuck	{'neg': 1.0, 'neu': 0.0, 'pos': 0.0, 'compound': -0.5423}
306	COVID. He was in the ICU for two weeks. He said he was proud of me for achieving my dreams (of getting into school) and as soon as he got out, he was going to come and visit me. My other uncle (on my mom's side) died in December, two days before I was accepted - also of COVID. I don't know what to do now. Should I go home and go to his funeral? Should I stay here and not make my family worry about having to pay to fly me back home and back to Ithaca? Should I tell my friends that I've only known for two weeks? I don't know. \n\nJust wear your mask. No matter how annoying you might think it is.	{'neg': 0.104, 'neu': 0.819, 'pos': 0.078, 'compound': -0.4047}

Figure 6: Sentiment Analysis output from a sample of 'r/cornell' posts

3.3 Regression Analysis

The hypotheses under consideration are of significant relevant and importance in the context of our project. They are centered around the influence of users, the positivity (or negativity) of their content, and the dynamics of emotional support within online communities. These hypotheses are particularly pertinent given the increasing role of online platforms in shaping discourse and providing support networks, especially in the context of college students' mental health.

Hypotheses

1. Our first hypothesis posits that **influential users or users with high centrality scores post more positive content**. In line with Goffman's [7] theory of self-presentation, studies have indicated that users of social media strive to project a positive self-image online [19]. This inclination is referred to as the 'positivity bias' in social media [17]. Given that teenagers and young adults devote a significant amount of time to social media, they are frequently exposed to seemingly flawless lives and looks of others [1]. Based on the conclusions of these studies, we can hypothesize that influential users or those with higher centrality scores may be more likely to post positive content. This hypothesis is grounded in the understanding that users, particularly those with significant influence, strive to project a favorable self-image online.

   On the contrary, investigating this hypothesis is crucial as it could also illuminate the potential influence of central users on the overall sentiment within the community. If influential users are indeed more likely to post positive content, they could play a pivotal role in cultivating a positive online environment. This could have significant implications for the overall emotional climate of the community, potentially promoting more positive interactions and contributing to the well-being of its members.

$$post\_sentiment_i = \beta_0 + \beta_1 \cdot ln\_centrality_i + X_i + \epsilon_i$$

We tested this hypothesis by using an OLS model by regressing post sentiment (both positive and compound sentiment) for a poster i on various centrality ($ln\_centrality$) measures (degree, eigenvector and betweenness) within the poster-commenter network. We added number of comments received on the post and the upvote ratio of the post as meaningful controls to reduce omitted variables bias (denoted by X_k).

2. Our second hypothesis posits that **users who connect different communication channels (comment threads) are more likely to provide emotional support to their peers.** This is relevant as it could highlight the importance of users who act as bridges in the network, connecting different sub-communities and facilitating the flow of emotional support.

$$comment\_pos\_sentiment_j = \beta_0 + \beta_1 \cdot btw\_centrality_j + X_j + \epsilon_j$$

We tested this hypothesis by using an OLS model by regressing comment positive sentiment for a commenter j on betweenness centrality measure within the commenter-commenter network. We added comment score as a control, and also included clustering coefficient as another meaningful independent variable (denoted by X_j).

3. Our third hypothesis proposes that **users with higher network centrality scores are more likely to receive emotional support from their peers**. This is significant as it could indicate that being more influential in the network increases the likelihood of receiving emotional support, which could have implications for mental health outcomes for using which are not well connected.

$$aggregated\_comment\_sentiment_k = \beta_0 + \beta_1 \cdot ln\_centrality_k + X_k + \epsilon_k$$

We tested this hypothesis by using an OLS model by regressing aggregated comment sentiment for a post k on various centrality measures (degree, eigenvector and betweenness) for a poster k within the poster-commenter network. We added number of comments received on the post and the upvote ratio of the post as meaningful controls to reduce omitted variables bias (denoted by X_k).

3.4 Prediction of Comment Sentiment using Network Embeddings

A study by Cheng et al [4] illustrated that network measures can form valuable features for investigating the predictability of information cascades in social media. Inspired by this result and in order to better understand the sentiment dynamics within Reddit data, we classified the compound sentiment of comments into five distinct classes : Very Negative $[-1, -0.65)$, Negative $[-0.65, -0.35)$, Neutral $[-0.35, 0.35)$, Positive $[0.35, 0.65)$, and Very Positive $[0.65, 1]$. This classification rubric, specifically created for VADER, was adapted from a study by Borg et al [3].

Following this classification, we utilized Node2Vec [8], a state-of-the-art algorithm for generating network embeddings, to create representations of the commenter-commenter graph. Node2Vec embeddings capture the network structure around each node, effectively summarizing a user's position and role within the network into a vector of features. Default parameters were used for this analysis.

With these embeddings as predictors, we then employed several machine learning models, including Logistic Regression (LogReg), Support Vector Machines (SVM), Random

Forest and Decision Trees, to train for a multi-class classification task and predict sentiment class of a given comment.

This approach is particularly relevant for several reasons. First, it allows us to understand user behavior in the context of their network position and interactions. By using network embeddings as predictors, we can investigate how a user's position within the network might influence the sentiment of their comments. Second, this methodology can potentially help identify at-risk individuals within the subreddit. If certain network positions or interaction patterns are consistently associated with negative sentiment, these could serve as indicators for users who may be struggling.

4 Results

Network Characteristics

In the poster-commenter networks (see Table 1), we observe a relatively sparse network structure, as indicated by the low average degree ranging from 2.6197 to 3.9915. The average clustering coefficient, a measure of the degree to which nodes in a graph tend to cluster together, is also low (ranging from 0.0110 to 0.0598), further indicating the sparseness of these networks. The transitivity[1] is also low, reinforcing the observation of sparseness in the poster-commenter networks. The attributes of the poster-commenter network indeed suggest a hub-style network structure. In a hub-style network, certain nodes (in this case, posters) serve as central points of connection for many other nodes (in this case, commenters). This is reflected in the relatively low average degree and low average clustering coefficient observed in the poster-commenter networks across all four subreddits.

This hub-style structure is characteristic of many social networks, where a small number of influential individuals or entities serve as the primary sources of information or interaction for a larger community. In the context of Reddit, this could reflect the role of popular posters in shaping discussions within a subreddit.

	cornell	**harvard**	**berkeley**	**uiuc**
Nodes	1572	1648	2015	1507
Edges	2585	3289	3050	1974
Connected Components	1	1	1	2
Nodes in largest component (%)	100.0	100.0	100.0	99.27
Average degree	3.2888	3.9915	3.0272	2.6197
Diameter	9	8	8	10
Average clustering coefficient	0.0390	0.0598	0.0232	0.0110
Transitivity	0.0204	0.0241	0.0112	0.0110

Table 1: Network Attributes for Poster-Commenter Network

On the other hand, the commenter-commenter networks (see Table 1) exhibit a denser structure. The average degree is significantly higher (ranging from 13.339 to 16.549), indicating that commenters tend to interact with a larger number of other commenters. The average clustering coefficient is also substantially higher (ranging from 0.8163 to 0.8799), suggesting a higher tendency for commenters to form tightly knit communities. The transitivity is also higher in these networks, indicating a higher likelihood of triangles, which is consistent with the denser structure of these networks.

[1]Transitivity (T) is a measure of the tendency of the nodes to cluster together. $T = 3 \cdot \frac{\#triangles}{\#triads}$

These differences in network attributes between the poster-commenter and commenter-commenter networks highlight the distinct patterns of interaction and connectivity among users in these different contexts. The sparser structure of the poster-commenter networks may reflect the more hierarchical nature of interactions between posters and commenters, while the denser structure of the commenter-commenter networks may reflect the more collaborative and interactive nature of discussions among commenters.

	cornell	**harvard**	**berkeley**	**uiuc**
Nodes	1572	1648	2015	1507
Edges	12280	13637	15301	10051
Connected Components	1	1	1	2
Nodes in largest component (%)	100.0	100.0	100.0	99.27
Average degree	15.623	16.549	15.187	13.339
Diameter	5	5	4	5
Average clustering coefficient	0.8375	0.8163	0.8426	0.8799
Transitivity	0.1106	0.1093	0.1039	0.1308

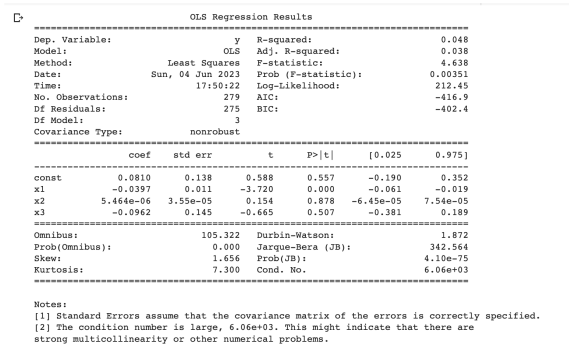
Table 2: Network Attributes for the Commenter-Commenter Network

Degree distribution histograms for both the poster-commenter andd commenter-commenter networks for all four subreddits are illustrated in Figures 20 and 21 in the Appendix.

Hypothesis 1

Our initial OLS regression analysis aimed to evaluate the relationship between sentiment of a post and the centrality of the post author. In this case of the Berkeley subreddit (see Fig 7), we found a statistically significant negative coefficient in relation to the natural log of eigenvector centrality (*ln_eig_centrality*). This suggests that more influential users, on average, tend to post content with a more negative sentiment. A similar trend was observed in the Cornell subreddit with respect to the compound sentiment of the post. However, for the Harvard and UIUC subreddits, the coefficients were not statistically significant, indicating that we did not observe a similar effect in these communities. It's noteworthy that the direction of the relationship between user influence and post sentiment remained consistent even when using the compound sentiment of the posts as the dependent variable. This consistency across different measures of sentiment further reinforces our findings. It suggests that in the Berkeley and Cornell subreddits, more influential users tend to post content with a more negative sentiment, regardless of whether we consider the overall compound sentiment of the posts or just the positive sentiment scores. Again, no such effect was observed in the Harvard and UIUC subreddits, indicating the potential influence of community-specific factors on this relationship.

In the case of the Berkeley and Cornell subreddits, our findings suggest the opposite of our initial hypothesis. More influential users, as measured by eigenvector centrality, were found to post content with less positive sentiment on average. This could imply that within these communities, influential users may be more inclined to discuss challenging or contentious topics, which could potentially lead to more net negative sentiment in their posts.



OLS Regression Results

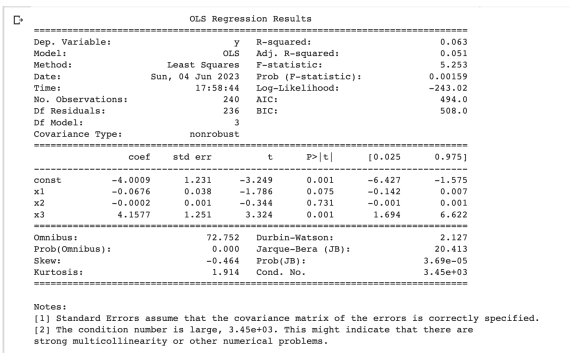
Dep. Variable:	y	R-squared:	0.048
Model:	OLS	Adj. R-squared:	0.038
Method:	Least Squares	F-statistic:	4.638
Date:	Sun, 04 Jun 2023	Prob (F-statistic):	0.00351
Time:	17:50:22	Log-Likelihood:	212.45
No. Observations:	279	AIC:	-416.9
Df Residuals:	275	BIC:	-402.4
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0810	0.138	0.588	0.557	-0.190	0.352
x1	-0.0397	0.011	-3.720	0.000	-0.061	-0.019
x2	5.464e-06	3.55e-05	0.154	0.878	-6.45e-05	7.54e-05
x3	-0.0962	0.145	-0.665	0.507	-0.381	0.189

Omnibus:	105.322	Durbin-Watson:	1.872
Prob(Omnibus):	0.000	Jarque-Bera (JB):	342.564
Skew:	1.656	Prob(JB):	4.10e-75
Kurtosis:	7.300	Cond. No.	6.06e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.06e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 7: Berkeley



OLS Regression Results

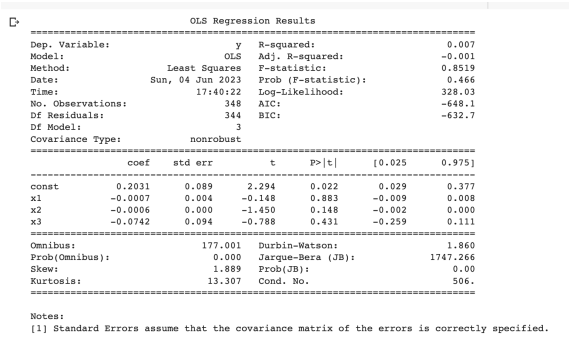
Dep. Variable:	y	R-squared:	0.063
Model:	OLS	Adj. R-squared:	0.051
Method:	Least Squares	F-statistic:	5.253
Date:	Sun, 04 Jun 2023	Prob (F-statistic):	0.00159
Time:	17:58:44	Log-Likelihood:	-243.02
No. Observations:	240	AIC:	494.0
Df Residuals:	236	BIC:	508.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.0009	1.231	-3.249	0.001	-6.427	-1.575
x1	-0.0676	0.038	-1.786	0.075	-0.142	0.007
x2	-0.0002	0.001	-0.344	0.731	-0.001	0.001
x3	4.1577	1.251	3.324	0.001	1.694	6.622

Omnibus:	72.752	Durbin-Watson:	2.127
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20.413
Skew:	-0.464	Prob(JB):	3.69e-05
Kurtosis:	1.914	Cond. No.	3.45e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.45e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 8: Cornell



OLS Regression Results

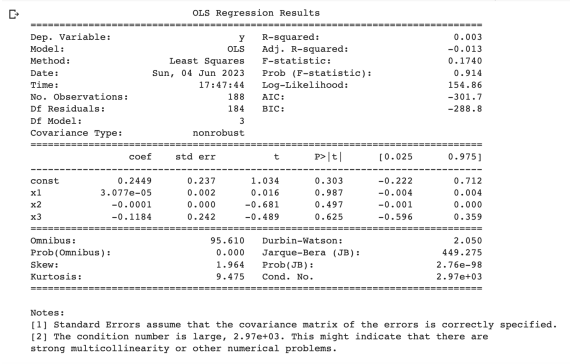
Dep. Variable:	y	R-squared:	0.007
Model:	OLS	Adj. R-squared:	-0.001
Method:	Least Squares	F-statistic:	0.8519
Date:	Sun, 04 Jun 2023	Prob (F-statistic):	0.466
Time:	17:40:22	Log-Likelihood:	328.03
No. Observations:	348	AIC:	-648.1
Df Residuals:	344	BIC:	-632.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2031	0.089	2.294	0.022	0.029	0.377
x1	-0.0007	0.004	-0.148	0.883	-0.009	0.008
x2	-0.0006	0.000	-1.450	0.148	-0.002	0.000
x3	-0.0742	0.094	-0.788	0.431	-0.259	0.111

Omnibus:	177.001	Durbin-Watson:	1.860
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1747.266
Skew:	1.889	Prob(JB):	0.00
Kurtosis:	13.307	Cond. No.	506.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 9: Harvard



OLS Regression Results

Dep. Variable:	y	R-squared:	0.003
Model:	OLS	Adj. R-squared:	-0.013
Method:	Least Squares	F-statistic:	0.1740
Date:	Sun, 04 Jun 2023	Prob (F-statistic):	0.914
Time:	17:47:44	Log-Likelihood:	154.86
No. Observations:	188	AIC:	-301.7
Df Residuals:	184	BIC:	-288.8
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2449	0.237	1.034	0.303	-0.222	0.712
x1	3.077e-05	0.002	0.016	0.987	-0.004	0.004
x2	-0.0001	0.000	-0.681	0.497	-0.001	0.000
x3	-0.1184	0.242	-0.489	0.625	-0.596	0.359

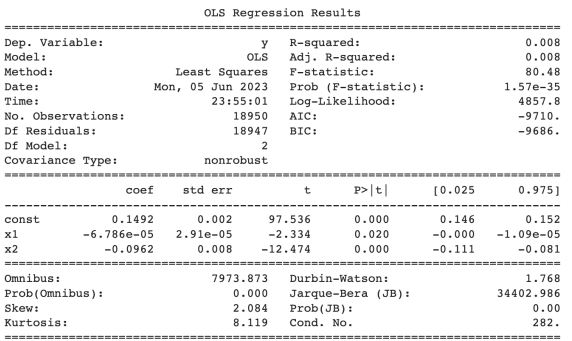
Omnibus:	95.610	Durbin-Watson:	2.050
Prob(Omnibus):	0.000	Jarque-Bera (JB):	449.275
Skew:	1.964	Prob(JB):	2.76e-98
Kurtosis:	9.475	Cond. No.	2.97e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.97e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 10: UIUC

Hypothesis 2

Having explored the relationship between user influence and post sentiment, we now turn our attention to the second hypothesis of our study. This hypothesis posits that users who connect different communication channels, essentially acting as bridges within the network, are more likely to provide emotional support to their peers. We tested this hypothesis by regressing comment sentiment against centrality of the commenter in our commenter-commenter graph.



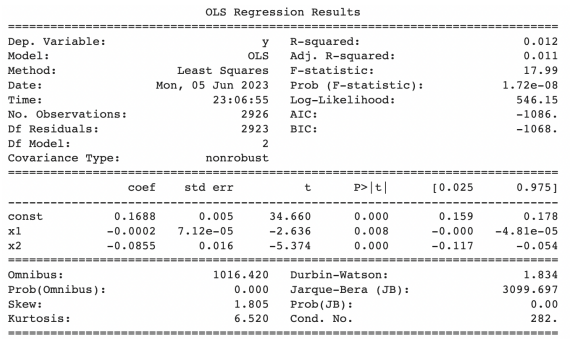
OLS Regression Results

Dep. Variable:	y	R-squared:	0.008
Model:	OLS	Adj. R-squared:	0.008
Method:	Least Squares	F-statistic:	80.48
Date:	Mon, 05 Jun 2023	Prob (F-statistic):	1.57e-35
Time:	23:55:01	Log-Likelihood:	4857.8
No. Observations:	18950	AIC:	-9710.
Df Residuals:	18947	BIC:	-9686.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1492	0.002	97.536	0.000	0.146	0.152
x1	-6.786e-05	2.91e-05	-2.334	0.020	-0.000	-1.09e-05
x2	-0.0962	0.008	-12.474	0.000	-0.111	-0.081

Omnibus:	7973.873	Durbin-Watson:	1.768
Prob(Omnibus):	0.000	Jarque-Bera (JB):	34402.986
Skew:	2.084	Prob(JB):	0.00
Kurtosis:	8.119	Cond. No.	282.

Figure 11: Berkeley



OLS Regression Results

Dep. Variable:	y	R-squared:	0.012
Model:	OLS	Adj. R-squared:	0.011
Method:	Least Squares	F-statistic:	17.99
Date:	Mon, 05 Jun 2023	Prob (F-statistic):	1.72e-08
Time:	23:06:55	Log-Likelihood:	546.15
No. Observations:	2926	AIC:	-1086.
Df Residuals:	2923	BIC:	-1068.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1688	0.005	34.660	0.000	0.159	0.178
x1	-0.0002	7.12e-05	-2.636	0.008	-0.000	-4.81e-05
x2	-0.0855	0.016	-5.374	0.000	-0.117	-0.054

Omnibus:	1016.420	Durbin-Watson:	1.834
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3099.697
Skew:	1.805	Prob(JB):	0.00
Kurtosis:	6.520	Cond. No.	282.

Figure 12: Cornell

OLS Regression Results

Dep. Variable:	y	R-squared:	0.020
Model:	OLS	Adj. R-squared:	0.020
Method:	Least Squares	F-statistic:	44.46
Date:	Mon, 05 Jun 2023	Prob (F-statistic):	7.71e-20
Time:	23:42:48	Log-Likelihood:	1529.5
No. Observations:	4345	AIC:	-3053.
Df Residuals:	4342	BIC:	-3034.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1589	0.003	47.643	0.000	0.152	0.165
x1	-0.0008	0.000	-4.057	0.000	-0.001	-0.000
x2	-0.0900	0.011	-8.525	0.000	-0.111	-0.069

Omnibus:	1851.704	Durbin-Watson:	1.831
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8817.240
Skew:	2.047	Prob(JB):	0.00
Kurtosis:	8.652	Cond. No.	66.4

Figure 13: Harvard

OLS Regression Results

Dep. Variable:	y	R-squared:	0.011
Model:	OLS	Adj. R-squared:	0.011
Method:	Least Squares	F-statistic:	12.37
Date:	Mon, 05 Jun 2023	Prob (F-statistic):	4.54e-06
Time:	23:49:46	Log-Likelihood:	680.83
No. Observations:	2136	AIC:	-1356.
Df Residuals:	2133	BIC:	-1339.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1560	0.005	33.231	0.000	0.147	0.165
x1	-7.696e-05	4.16e-05	-1.850	0.064	-0.000	4.62e-06
x2	-0.0691	0.015	-4.606	0.000	-0.098	-0.040

Omnibus:	794.003	Durbin-Watson:	1.897
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2853.654
Skew:	1.846	Prob(JB):	0.00
Kurtosis:	7.294	Cond. No.	428.

Figure 14: UIUC

Variables used for the above comment sentiment regression: - y: positive comment sentiment - $x1$: comment score - $x2$: betweenness centrality
Multicollinearity makes it challenging to assess the relative importance of predictors in explaining the dependent variable (positive comment sentiment). Highly correlated predictors can lead to inflated standard errors and undermine the significance of individual predictors, making it harder to identify which predictors are truly influential. Betweenness centrality was highly correlated with eigenvector centrality (≈ 0.9), degree centrality (≈ 0.95), and clustering coefficient (≈ 0.5). Removing these highly correlated predictors resulted in a clearer interpretation of the regression coefficients. Each predictor included in the model (x1 = comment score, x2 = betweenness centrality) can be more confidently attributed to its unique contribution to the positive comment sentiment with reduced standard errors of the coefficients, facilitating a better understanding of the relationships between predictors and the sentiment. For two subreddits (Harvard, Cornell) (see Fig 12 and 13), we found a statistically significant negative coefficient in relation to betweenness centrality (*btw_centrality*) with respect to comment sentiment. For the other two subreddits (UIUC, Berkeley) the coefficients were still negatively correlated. This finding suggests that as the betweenness of a commenter increases, the sentiment of their comments tend to become less positive.

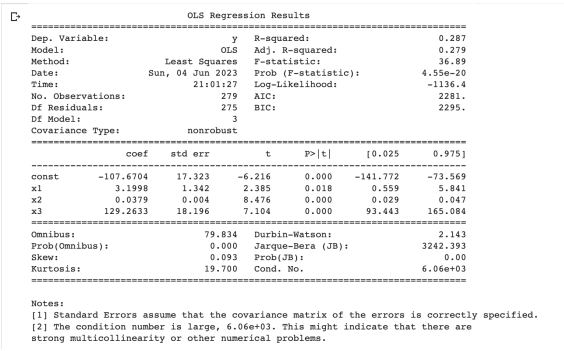
The observed negative relationship between betweenness centrality and comment sentiment has implications for our second hypothesis, which posits that users who connect different communication channels are more likely to provide emotional support to their peers. Given that influential users, as measured by betweenness centrality, tend to post comments with less positive sentiment, it suggests that these users might not be providing as much emotional support as we hypothesized. Instead, they might be more involved in critical discussions or expressing negative emotions.

Hypothesis 3

After examining the relationship between user influence and post sentiment, as well as the role of users who connect different communication channels, we now shift our focus to the third hypothesis of our project. This hypothesis suggests that users with higher network centrality scores are more likely to receive emotional support from their peers.

To test this hypothesis, we again utilized OLS regression analysis, this time focusing on the relationship between the aggregate sentiment of comments received by a poster and their centrality within the poster-commenter network. This analysis allows us to explore whether posters who are more central within the network, and thus potentially more influential or visible, tend to receive more positive comments, which we interpret as a form of emotional support.

The regression results for this hypothesis are presented in Figures 15 - 18. For the



OLS Regression Results

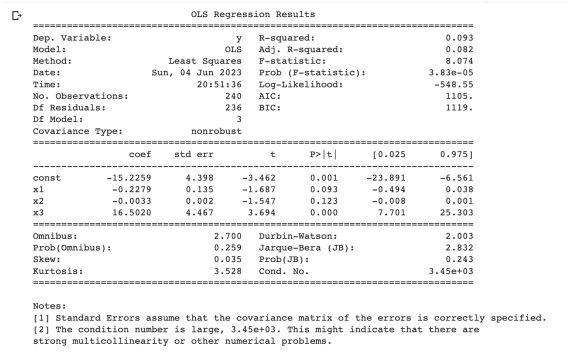
Dep. Variable:	y	R-squared:	0.287
Model:	OLS	Adj. R-squared:	0.279
Method:	Least Squares	F-statistic:	36.89
Date:	Sun, 04 Jun 2023	Prob (F-statistic):	4.55e-20
Time:	21:01:27	Log-Likelihood:	-1136.4
No. Observations:	279	AIC:	2281.
Df Residuals:	275	BIC:	2295.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-107.6704	17.323	-6.216	0.000	-141.772	-73.569
x1	3.1998	1.342	2.385	0.018	0.559	5.841
x2	0.0379	0.004	8.476	0.000	0.029	0.047
x3	129.2633	18.196	7.104	0.000	93.443	165.084

Omnibus:	79.834	Durbin-Watson:	2.143
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3242.393
Skew:	0.093	Prob(JB):	0.00
Kurtosis:	19.700	Cond. No.	6.06e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.06e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 15: Berkeley



OLS Regression Results

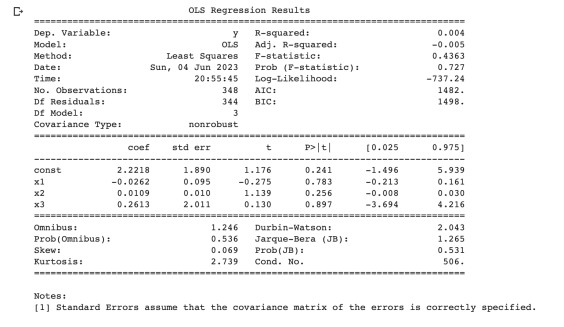
Dep. Variable:	y	R-squared:	0.093
Model:	OLS	Adj. R-squared:	0.082
Method:	Least Squares	F-statistic:	8.074
Date:	Sun, 04 Jun 2023	Prob (F-statistic):	3.83e-05
Time:	20:51:36	Log-Likelihood:	-548.55
No. Observations:	240	AIC:	1105.
Df Residuals:	236	BIC:	1119.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-15.2259	4.398	-3.462	0.001	-23.891	-6.561
x1	-0.2279	0.135	-1.687	0.093	-0.494	0.038
x2	-0.0033	0.002	-1.547	0.123	-0.008	0.001
x3	16.5020	4.467	3.694	0.000	7.701	25.303

Omnibus:	2.700	Durbin-Watson:	2.003
Prob(Omnibus):	0.259	Jarque-Bera (JB):	2.832
Skew:	0.035	Prob(JB):	0.243
Kurtosis:	3.528	Cond. No.	3.45e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.45e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 16: Cornell



OLS Regression Results

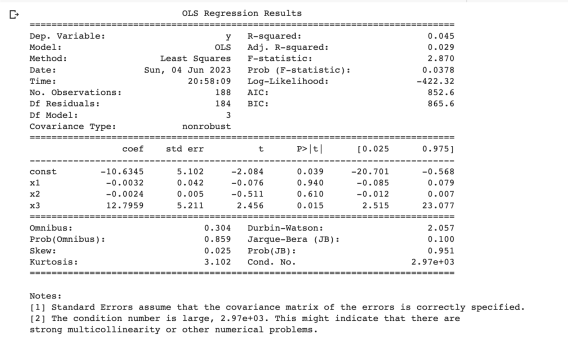
Dep. Variable:	y	R-squared:	0.004
Model:	OLS	Adj. R-squared:	-0.005
Method:	Least Squares	F-statistic:	0.4363
Date:	Sun, 04 Jun 2023	Prob (F-statistic):	0.727
Time:	20:55:45	Log-Likelihood:	-737.24
No. Observations:	348	AIC:	1482.
Df Residuals:	344	BIC:	1498.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.2218	1.890	1.176	0.241	-1.496	5.939
x1	-0.0262	0.095	-0.275	0.783	-0.213	0.161
x2	0.0109	0.010	1.139	0.256	-0.008	0.030
x3	0.2613	2.011	0.130	0.897	-3.694	4.216

Omnibus:	1.246	Durbin-Watson:	2.043
Prob(Omnibus):	0.536	Jarque-Bera (JB):	1.265
Skew:	0.069	Prob(JB):	0.531
Kurtosis:	2.739	Cond. No.	506.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 17: Harvard



OLS Regression Results

Dep. Variable:	y	R-squared:	0.045
Model:	OLS	Adj. R-squared:	0.029
Method:	Least Squares	F-statistic:	2.870
Date:	Sun, 04 Jun 2023	Prob (F-statistic):	0.0378
Time:	20:58:09	Log-Likelihood:	-422.32
No. Observations:	188	AIC:	852.6
Df Residuals:	184	BIC:	865.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-10.6345	5.102	-2.084	0.039	-20.701	-0.568
x1	-0.0032	0.042	-0.076	0.940	-0.085	0.079
x2	-0.0024	0.005	-0.511	0.610	-0.012	0.007
x3	12.7959	5.211	2.456	0.015	2.515	23.077

Omnibus:	0.304	Durbin-Watson:	2.057
Prob(Omnibus):	0.859	Jarque-Bera (JB):	0.100
Skew:	0.025	Prob(JB):	0.951
Kurtosis:	3.102	Cond. No.	2.97e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.97e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 18: UIUC

Berkeley subreddit, we observed a positive, statistically significant coefficient in relation to the natural log of eigenvector centrality ($ln_eig_centrality$). This suggests that more influential posters in this subreddit tend to receive, on average, more positive comments in response to their posts, even when controlling for the number of comments and upvote ratio. However, this effect was not observed for the remaining three subreddits, where no statistically significant correlation was found between centrality measures and the sentiment of comments received.

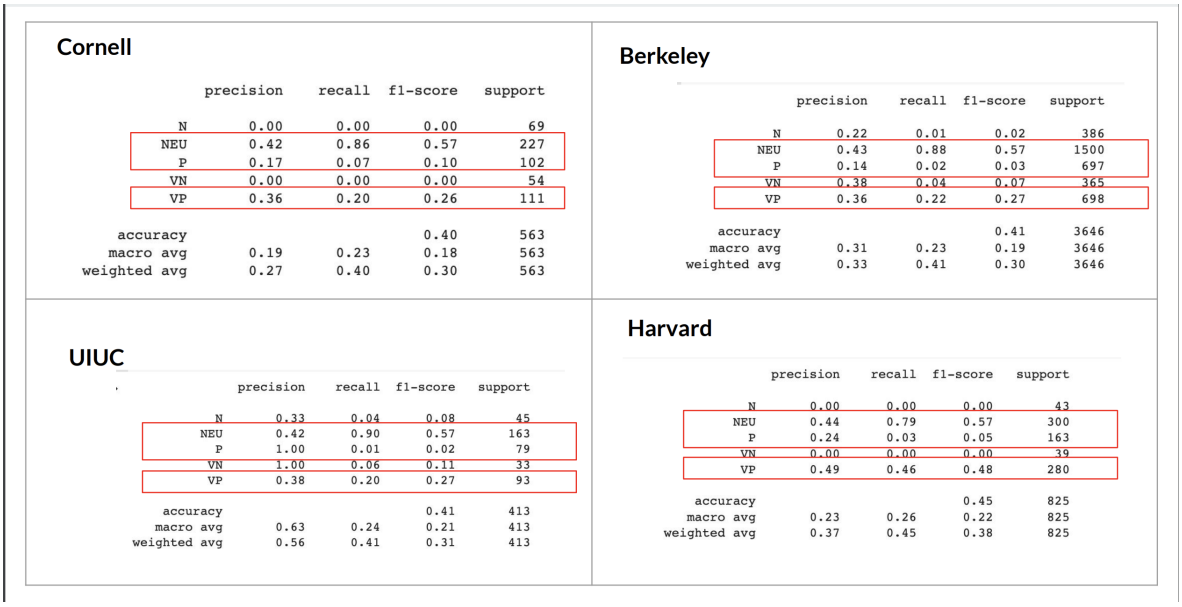
The results of our analysis provide partial support for the third hypothesis. In the case of the Berkeley subreddit, the data suggests that more influential users, as indicated by higher eigenvector centrality, do indeed receive more positive comments, which can be interpreted as a form of emotional support. This finding aligns with our hypothesis that users with higher network centrality scores are more likely to receive emotional support from their peers.

However, this pattern was not observed in the other three subreddits we analyzed. In these communities, there was no statistically significant correlation between a user's centrality and the sentiment of the comments they received. This suggests that the relationship between user influence and the receipt of emotional support may vary across different communities.

While the low R-squared values in our analysis suggest that the models do not have strong predictive power, the statistically significant coefficients still provide valuable insights. The sign of a statistically significant coefficient can be interpreted with confidence, as it indicates the direction of the relationship between the predictor and the outcome variable.

Prediction of Comment Sentiment using Network Embeddings

In line with our methodology, we conducted a series of machine learning analyses to predict the sentiment class of a given comment based on network embeddings. The results of these analyses provide valuable insights into the sentiment dynamics within Reddit data.



Cornell

	precision	recall	f1-score	support
N	0.00	0.00	0.00	69
NEU	0.42	0.86	0.57	227
P	0.17	0.07	0.10	102
VN	0.00	0.00	0.00	54
VP	0.36	0.20	0.26	111
accuracy			0.40	563
macro avg	0.19	0.23	0.18	563
weighted avg	0.27	0.40	0.30	563

Berkeley

	precision	recall	f1-score	support
N	0.22	0.01	0.02	386
NEU	0.43	0.88	0.57	1500
P	0.14	0.02	0.03	697
VN	0.38	0.04	0.07	365
VP	0.36	0.22	0.27	698
accuracy			0.41	3646
macro avg	0.31	0.23	0.19	3646
weighted avg	0.33	0.41	0.30	3646

UIUC

	precision	recall	f1-score	support
N	0.33	0.04	0.08	45
NEU	0.42	0.90	0.57	163
P	1.00	0.01	0.02	79
VN	1.00	0.06	0.11	33
VP	0.38	0.20	0.27	93
accuracy			0.41	413
macro avg	0.63	0.24	0.21	413
weighted avg	0.56	0.41	0.31	413

Harvard

	precision	recall	f1-score	support
N	0.00	0.00	0.00	43
NEU	0.44	0.79	0.57	300
P	0.24	0.03	0.05	163
VN	0.00	0.00	0.00	39
VP	0.49	0.46	0.48	280
accuracy			0.45	825
macro avg	0.23	0.26	0.22	825
weighted avg	0.37	0.45	0.38	825

Figure 19: Performance Metrics for SVM Model for Multi-Class Comment Sentiment Classification

Figure 19 showcases the effectiveness of the SVM model, which was trained for a multi-class classification task to predict the sentiment category of comments based on Node2Vec network embeddings across all four subreddits. We chose to highlight the performance metrics for the SVM model as it outperformed the other models.

Although the overall accuracy of the SVM model was relatively modest, ranging from 0.40 to 0.45, the performance metrics across different sentiment classes varied significantly. The SVM model demonstrated superior prediction capabilities for comments within the neutral, positive, and very positive classes, with the highest F1-score (0.57) achieved for neutral comments, followed by very positive and positive classes. We hypothesize that this may be attributed to the larger representation of neutral and positive comments within our datasets, as indicated by the support column in the performance reports above.

High recall but low precision for the neutral class in the model implies that the model is good at identifying true neutral comments (i.e., it correctly identifies a high proportion of the actual neutral comments), but it also incorrectly classifies a significant number of non-neutral comments as neutral. In other words, the model is sensitive to neutral comments and doesn't miss many of them (high recall), but it is not very precise in its classification (low precision). It tends to over-predict the neutral class, meaning it often labels a comment as neutral when it actually belongs to a different sentiment class. This could potentially lead to an overestimation of neutral sentiment in the data.

Similarly on the other hand, high precision and low recall for the positive class in the model implies that when the model predicts a comment to be positive, it is usually correct (high precision). However, the model also misses a significant number of actual positive comments (low recall). In other words, the model is very accurate in its positive predictions, but it is not very sensitive to positive comments. It tends to under-predict the positive class, meaning it often fails to label a comment as positive when it actually is. This could potentially lead to an underestimation of positive sentiment in the data.

Despite the relatively modest performance of our model, the fact that network embeddings demonstrated some predictive power in determining comment sentiment is a significant finding of our project. This suggests that a user's position and role within the network, as captured by the embeddings, do have some bearing on the sentiment of their comments. This finding is important because it opens up new avenues for understanding and predicting user behavior in online social networks. It suggests that network structure and user interactions can provide valuable insights into user behavior, beyond what can be gleaned from the content of their posts alone. Moreover, this finding lays the groundwork for the development of more sophisticated models that can leverage network embeddings along with other types of features to predict user behavior. For instance, future models could combine network embeddings with text-based features derived from the content of the comments, or with user-level features such as their posting frequency or longevity on the platform.

5 Conclusion

In conclusion, this study aimed to explore the relationship between user influence, as measured by network centrality, and sentiment dynamics within four college subreddits. We formulated three hypotheses to guide our analysis, each focusing on a different aspect of user influence and sentiment.

Despite the low predictive power of our models, the statistically significant coefficients still provide valuable insights into the relationship between user influence and sentiment dynamics within online communities. These findings contribute to our understanding of how influential users might shape the emotional tone of these communities, which is an important consideration for studies of online social dynamics.

In interpreting these results, it's important to consider the context of our study. Our goal was not to predict sentiment with high accuracy, but rather to explore the potential relationships between user influence and sentiment. While our findings provide some support for our hypotheses, they also highlight the complexity of these relationships and the need for further research in this area.

Overall, our study underscores the potential of using network analysis and sentiment analysis to investigate user behavior in online communities. We believe that this approach holds promise for future research, particularly in the context of mental health, where understanding the dynamics of online interactions could provide valuable insights for prevention and intervention efforts.

6 Limitations

Our analysis, while providing valuable insights, is not without its limitations.

Firstly, there is a potential selection bias in our dataset. The users of the college subreddit may not be representative of the entire student population. This limits the generalizability and external validity of our analysis. The behaviors and sentiments expressed by users on the subreddit may not accurately reflect those of all students, and hence, our findings may not apply to the broader student population.

Secondly, desirability bias could also be a factor in our analysis. Users may be reluctant to express certain emotions or admit to mental health issues on a public forum. This could lead to an underrepresentation of certain types of posts and comments, skewing the sentiment distribution. This bias could potentially affect our findings, particularly those related to the negativity bias of influential users.

Lastly, dealing with noisy text data presents a significant challenge. The language used on Reddit often includes slang, abbreviations, and other informal language that may

not be well-suited to sentiment analysis algorithms. This could lead to inaccuracies in our sentiment scores, potentially affecting the results of our analysis. Future work could involve refining the sentiment analysis process to better handle the unique language used on Reddit.

Despite these limitations, our study provides a valuable starting point for understanding the sentiment dynamics within Reddit data and the potential role of network structure in influencing user behavior. Future research could address these limitations and further refine our understanding of these complex socio-technical systems.

7 Future Directions

Looking forward, there are several promising directions for further research. It would be interesting to explore link prediction techniques to predict future interactions between users. By combining our sentiment analysis with network measures, we could potentially predict not only who will interact, but also the likely sentiment of their interaction. These future directions could significantly enhance our understanding of user behavior within Reddit and other online communities.

Author Contribution Statement

Data scraping and graph creation was done by Saksham Arora. Meher Kalra and Cody Theobald conceived, planned and executed the experiments with respect to regressions for the commenter-commenter graph and contributed to the interpretation of the results. Saksham Arora planned and executed the experiments with respect to the regressions for the poster-commenter graph and the prediction of comment sentiment using network embeddings. All authors provided critical feedback and helped shape the research and analysis. Saksham Arora created the manuscript.

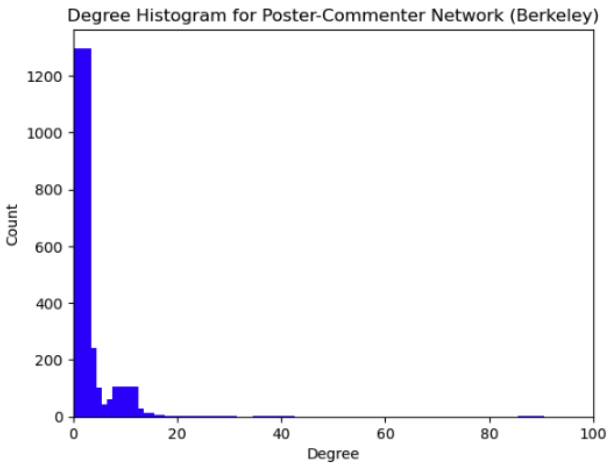
Data and code for the analysis have been made available via a Github repository https://github.com/meherkalra/M76.01_finalproject.

Acknowledgements

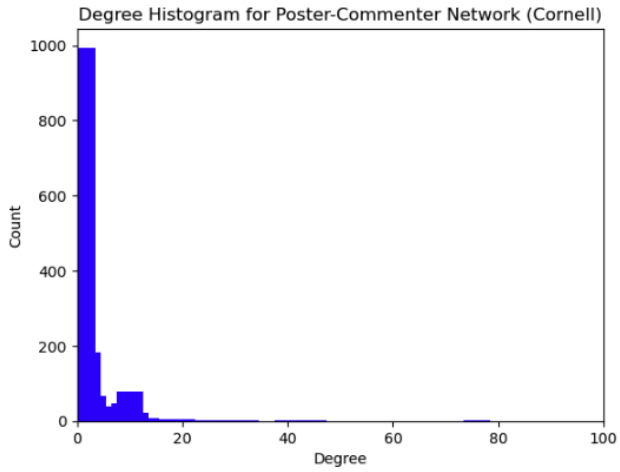
We would also like to thank Xie He for her constant support and guidance and for inspiring the idea for this project. Xie's comments and suggestions were extremely helpful in the conception and execution of this project and manuscript.

We would also like to thank our classmates and peers from the Math 76 course for their helpful suggestions via the peer review process, and for fostering a collaborative and engaging environment during lectures and presentations. We also thank Xie for facilitating the immensely helpful peer review process.

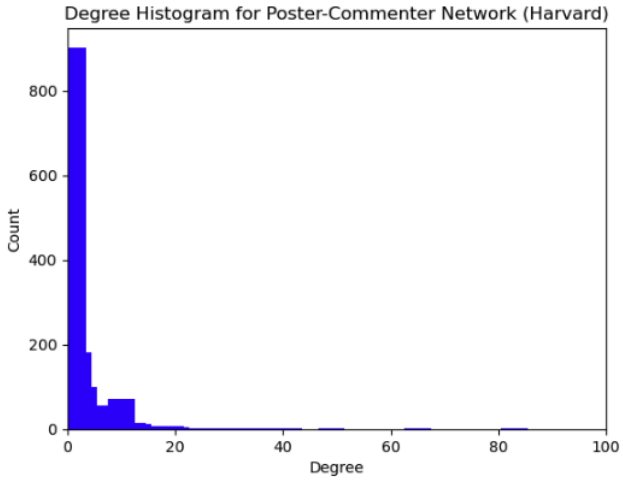
Appendix



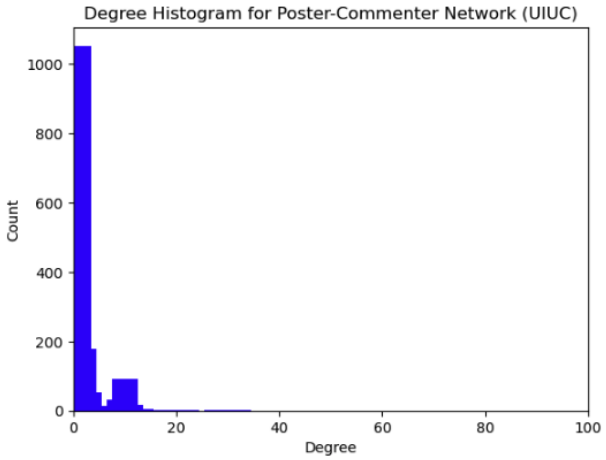
(a) Berkeley



(b) Cornell

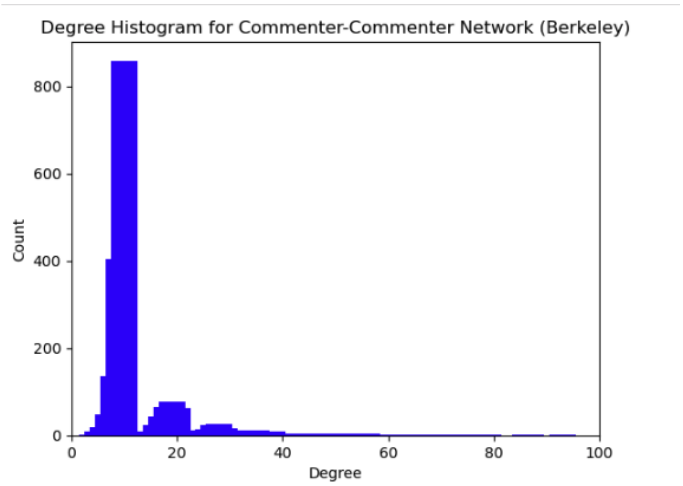


(c) Harvard

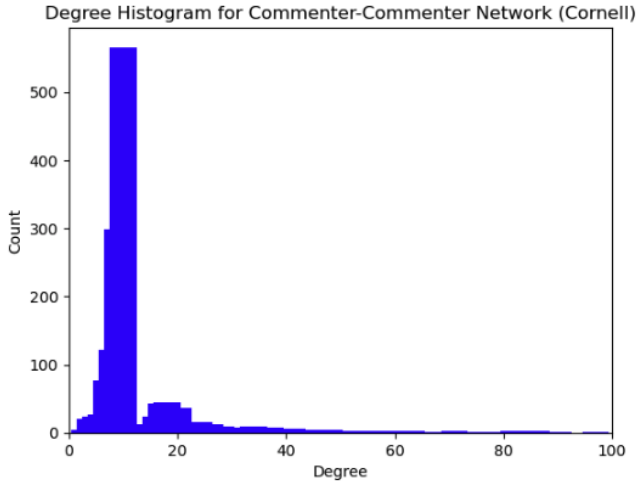


(d) UIUC

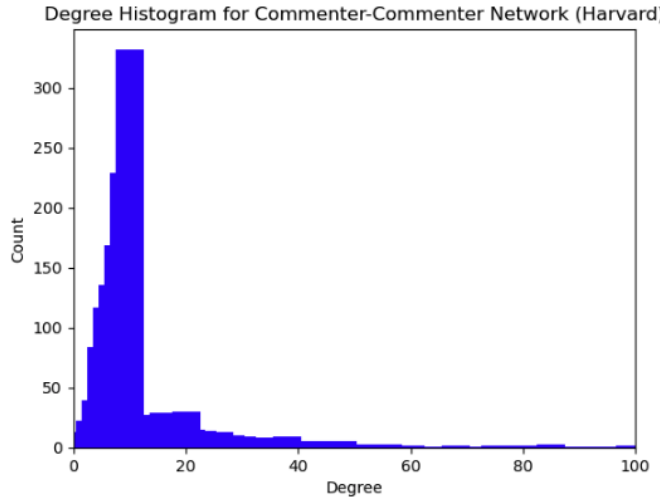
Figure 20: Degree Distribution Histograms for Poster-Commenter Network for all four subreddits



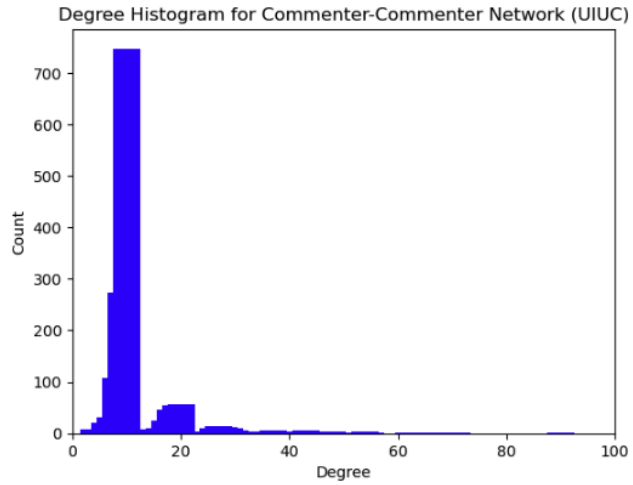
(a) Berkeley



(b) Cornell



(c) Harvard



(d) UIUC

Figure 21: Degree Distribution Histograms for Commenter-Commenter Network for all four subreddits

References

[1] B. T. Bell. ""You take fifty photos, delete forty nine and use one": A qualitative study of adolescent image-sharing practices on social media". In: *International Journal of Child-Computer Interaction* 20 (2019), pp. 64–71.

[2] T. Bianchi. *Reddit.com web visitor traffic 2022*. Jan. 2023. URL: https://www.statista.com/statistics/443332/reddit-monthly-visitors/#:~:text=Reddit%20usage&text=As%20of%20June%202021%2C%20Reddit,48%20million%20monthly%20active%20users..

[3] A. Borg and M. Boldt. "Using VADER sentiment and SVM for predicting customer response sentiment". In: *Expert Systems with Applications* 162 (2020), p. 113746.

[4] J. Cheng et al. "Can cascades be predicted?" In: *Proceedings of the 23rd international conference on World wide web*. 2014, pp. 925–936.

[5] T. Farrell et al. "Exploring misogyny across the manosphere in reddit". In: *Proceedings of the 10th ACM Conference on Web Science*. 2019, pp. 87–96.

[6] E. Ferrara and Z. Yang. "Measuring emotional contagion in social media". In: *PloS one* 10.11 (2015), e0142390.

[7] E. Goffman et al. "The presentation of self in everyday life. 1959". In: *Garden City, NY* 259 (2002).

[8] A. Grover and J. Leskovec. "node2vec: Scalable feature learning for networks". In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016, pp. 855–864.

[9] C. Hutto and E. Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1. 2014, pp. 216–225.

[10] A. D. Kramer, J. E. Guillory, and J. T. Hancock. "Experimental evidence of massive-scale emotional contagion through social networks". In: *Proceedings of the National Academy of Sciences* 111.24 (2014), pp. 8788–8790.

[11] S. Levy. *This is your brain on Twitter*. Feb. 2015. URL: https://medium.com/backchannel/this-is-your-brain-on-twitter-cac0725cea2b.

[12] A. Marotti. *Reddit to open Chicago office as part of Advertising Push*. Dec. 2018. URL: https://www.chicagotribune.com/business/ct-biz-reddit-chicago-office-20180418-story.html.

[13] R. A. Mills. "Pop-up political advocacy communities on reddit. com: SandersForPresident and The Donald". In: *Ai & Society* 33 (2018), pp. 39–54.

[14] A. Mittos et al. ""And we will fight for our race!" A measurement study of genetic testing conversations on Reddit and 4chan". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 452–463.

[15] *PRAW: The Python Reddit API Wrapper*. URL: https://praw.readthedocs.io/en/stable/.

[16] K. Roose. *The gamestop reckoning was a long time coming*. Jan. 2021. URL: https://www.nytimes.com/2021/01/28/technology/gamestop-stock.html.

[17] L. Schreurs and L. Vandenbosch. "Introducing the Social Media Literacy (SMILE) model with the case of the positivity bias on social media". In: *Journal of Children and Media* 15.3 (2021), pp. 320–337.

[18] K. Starbird et al. “Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing”. In: *IConference 2014 proceedings* (2014).

[19] J. C. Yau and S. M. Reich. ““It’s just a lot of work”: Adolescents’ self-presentation norms and practices on Facebook and Instagram”. In: *Journal of research on adolescence* 29.1 (2019), pp. 196–209.