Mining Reddit to Identify Factors that Describe Prominent Links Between Different Communities

Adrian Cristea

4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh

2018

Abstract

This project attempted to identify factors that describe prominent links between online communities on Reddit. Data was mined and processed from an online storage service, called Google BigQuery. After processing this dataset, we created a social network graph, that identified nodes as subreddits, and shared users as edges. A weight was assigned to these links: whenever a user would post a comment in two subreddits, the weighted value for the edge would increase. After this process was completed, we chose four centrality measures, which are widely used in social networking analysis, to use as a measure for the popularity of subreddits.

We had three research objectives: replicate results from a previous study, find additional factors that are unique to Reddit, and use an advanced sentiment analysis tool to analyse the content of the comments in context. For each one of our three research objectives, we analysed different types of factors that we could correlate to the results obtained previously, and found several predictors that helped explain the variance obtained in the results. Finally, we provided detailed discussion and interpretation by comparing it to previous research done in this field, while suggesting future areas to explore.

Table of Contents

Chapter 1

Introduction

1.1 Social Networks and Reddit

The internet continues to be one of the biggest achievements in human communications, with continued advances in data transfer and storage ensuring that it has become the de-facto medium for communication and socialisation. The newest annual report published by the International Telecommunication Union (ITU) in 2017 shows that almost half the world's population is now online [Mumford, 2017]. What is even more impressive is that, in developed countries, 94% of young people aged 15-24 actively use the internet, with mobile broadband connections leading the way in adoption [Mumford, 2017].

The ubiquitous availability of the medium has created new ways for people to interact with each other: the ability to quickly create new content whenever the opportunity arises, and then share it with a large group of friends, or even anonymously on the web, has meant that there is a vast trove of interesting (and not so interesting) content online, available on-demand. In order to make sense of it all, Social Networks were created that allowed people to create and consume content which is more relevant to them, by joining others with simililar interests.

A new paper published by the Pew Research Centre [Smith and Anderson, 2018] shows that a majority of Americans routinely use Social Media Networks, with most accessing it on a daily basis. Websites (and applications) such as Facebook, Youtube, Instagram and Reddit, feature heavily in the lives of people, and have a profound influence on the relations they form with other people, the content they produce and consume, and the opinions they form about the world at large.

While Facebook, a more traditional type of online social network, is the most accessed such online service on the planet according to Alexa [Alexa Internet, 2018], it also seems to have hit a plateau in terms of growth of the number users (about 80% of the total adult population in the US [Smith and Anderson, 2018]), and time spent actively on their platform [Smith and Anderson, 2018], at least in the West. Competitors, such as Instagram (which is owned by Facebook) and Snapchat, have managed to attract a large number of younger users (between the ages of 18 and 24), with many confessing

to accessing them multiple times a day. Figure 2.1 shows the current trends in online social networks: data is taken from Google Trends, which creates these graphs by showing search interest in online platforms, not as absolute values of their total user numbers.

These networks are all based on the idea of users creating their own connections by following each others activities intentionally, or by adding "friends". Reddit is, in many ways, different: it is a social news aggregation site which styles itself as *"The Front Page of the Internet"*. People who access it do not have to create an account to view most content available on the website. Furthermore, accounts generally are not linked to the real-life names of the individuals who use them, and, while people are allowed to follow other accounts' activities, the focus is on the content, not on who posted it, and when.

Since its founding in 2005, Reddit has become one of the most visited platforms on the web, ranking 5th in the UK and 6th globally [Alexa Internet, 2018]. Registered users, called *"redditors"*, submit content which can range from articles, photos or text posts, which can then be voted on (up or down) by other users. Such content is usually called a *"post"*, and it has two aspects: the linked content, and the comments associated with it. What is interesting about the comments is that they employ the same voting mechanism as the posts themselved, ensuring a hierarchical view of the discussion taking place on a post: the most *"upvoted"* rise to the top, and the most *"downvoted"* are hidden from view. Both posts and comments have a score attributed to them, called *"karma"*, which represents the difference between the total number of upvotes, and the total number of downvotes. User accounts also have a karma score associated with them, that encompasses the totality of their activity on the platform. In order to allow users to directly support the platform financially, Reddit allows users to gift each other *"Reddit Gold"*, which is a feature reserved for situations in which they'd like to express their satisfaction with high quality content. Gold is displayed prominently on posts and comments, and it signifies that users of the community find them particularly thoughtful and engaging.

Posts are organised into *"subreddits"*, which are user-created boards based around a common topic, such as news, science, technology, politics, and even niche interests such as a specific video game or bonsai trees. If a post garners enough interest (and upvotes), it can then be highlighted on the *"front page"*, which is a constantly updated snapshot of some of the most popular content on Reddit. Most users who have not registered for an account can freely browse the front page (or a subreddit), but not comment or vote.

One other interesting aspect of Reddit is the average time spent on it, compared to other social networks: on average Redditors spend 15:47 minutes a day, as opposed to 11:08 on Facebook, and 6:23 on Twitter, as of April 2018 [Alexa Internet, 2018]. Due to its nature, Reddit is a platform where content is not only shared, but generated and managed by all of its users, and is a major driver of online communication and information distribution across the globe. With billions of people engaging with this social platform, it is intriguing to uncover the underlying patterns of interaction between users and subreddits.

1.1.1 Research Objectives

In the interest of extending the study of social networking communities to the online space, we would like to explore the character of different social groups present on Reddit by retrieving and analying the large caches of posts and comments which are freely available online. Thus we will explore the factors that best describe prominent links between subreddits, based on the work of Jialun Wun [Wu, 2017]. It will aim to replicate some of its findings by utilising the same network structure and four different selection criteria (described in chapter 3) to show which are the most important and popular subreddits. We will then try to find a correlation between these and various other measurements, including the total number of comments in a subreddit, and others given by a linguistic content analysis of the most representative comments.

As such, we pose the following research objectives:

- *RO1*: Given the same methodology employed by Wu [Wu, 2017], we will try to replicate findings that attempt to find factors which describe the popularity of subreddits.
- *RO2*: We will then try to explore other factors which could lead to even better results, given our knowledge of Reddit-specific features (gold, karma scores, controversiality factors, and so on).
- *RO3*: Finally, we will attempt to employ modern content and sentiment analysis tools that take into account the context present within conversations on Reddit, in order to improve our understanding of factors that describe prominent links between different subreddits.

1.1.2 Improvements on previous work

We attempted to improve upon previous projects by addressing some of their limitations:

- instead of only gathering data in a one month period (May 2017) of the top 1000 subreddits (by total number of comments), we have expanded that to include the latest data available in a six month period (July to December 2017) on the top 5000 subreddits. According to the online statistics portal Statista, there were over 1.5 billion unique monthly visitors to Reddit for each month in our selected timeframe [Statista, 2018], in which, instead of only getting a brief snapshot of about 80% of the comments [Wu, 2017], we have increased that to just over 95%. Additionally, we increased the number of custom features that rely on Reddit specific features from 2 to 15, a full list of which can be found in section 3.6
- In the interest of replicating previous findings, we initially employ the same linguistic content analysis tools used in the previous study. However, one of its limitations is that it ignores the context of comments and treats each subreddit as a bag of words, lacking insight into their structure [Wu, 2017]. We will use

higher level content analysis tools, which provide a more comprehensive evaluation of subreddit comments, as suggested in [Lowe, 2002].

1.2 Report Structure

The rest of the report is structured as follows:

- In Chapter 2 we will discuss some of the research done in the past that has given us a deep understanding of the subject we are trying to discuss. We will initially present some studies which specifically involve Reddit, then give the background that justifies the methodology used in this paper.
- Chapter 3 will offer an elaborate summary of the tools and methods we employed in order to obtain and analyse the data.
- In Chapters 4, 5, and 6 we will present the results for the three research objectives posed above.
- Finally, chapter 7 will provide detailed discussion and interpretation of the results, and future areas that remain to be explored.

Chapter 2

Literature Review

In this chapter we will introduce the reader to the concepts necessary to understand this paper, along with previous work done in this domain. Initially, studies involving Reddit generally will be presented, with details about mining online communities and social network analysis expanding on them. Then, we will broadly present content and sentiment analysis methods and how they can be used in relation to online social networks.

2.1 Previous studies involving Reddit and other similar platforms

Computer-mediated communication (CMC) has been articulated by Susan Herring and is defined as communication that occurs through the use of electronic devices [Herring, 2004]. Most CMC research focuses on the social aspects and patterns of online interaction within computer-mediated formats such as instant messaging, discussion forums, and chat rooms. A study which focused on Reddit revealed the difficulty in getting large datasets which included user information such as gender and age [Finlay, 2014]. Results indicate that the majority of Redditors fall into the age group of 12-27, but that people over those ages were more likely to have a higher karma score on their comments. Interestingly, the average female poster, while in the minority in terms of total users, was more likely to have longer comment lengths, but lower overall scores. The number of people who indicated that they were over the age of 36 was only 59, out of a total of 734, however, most of this data is self-reported [Finlay, 2014]. What is more, while there seemed to be a correlation between age and comment score, there seemed to be little difference in terms of the content posted via links, indicating that age has no bearing on the success of submitted content [Finlay, 2014]. It is of note that this study might show significantly different results in 2018, due to the vastly increased user base since [Alexa Internet, 2018] [Statista, 2018].

Steinbauer [Steinbauer, 2012] describes Reddit as a social network with user generated content that inherits functionality from social news websites Digg, StumbleUpon

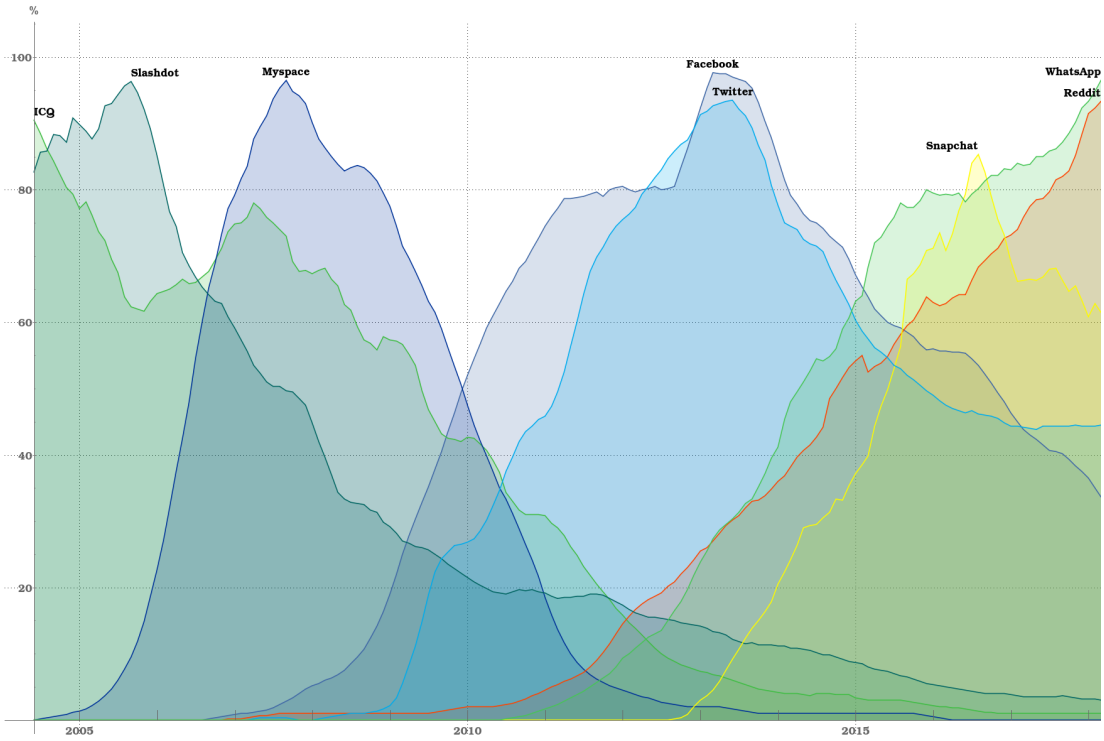

Figure 2.1: Trends in Social Networks [hnerixh (reddit user), 2018]

and Slashdot. In his paper, he shows that there is a significant gap between the largest subreddits (at the time having a maximum of 1,140,436 of subscribers) and the average subreddit, with only 1077 (in 2012). This provides insight into the idea that a large majority of subreddits are not very active, and this fact can be generalised for the average number of comments on a submission, which seems to be less than 10 [Steinbauer, 2012]. Furthermore, the way following another user on Reddit works is structured like Twitter, and is not bidirectional as in a classical Online Social Network (the followed user does not get a notification when another adds them as friends, and there is no requirement for them to follow back), which leads to a much looser type of user interaction, on average.

Lerman [Lerman, 2007] found that in the case of a more tightly connected network, such as Digg, user participation tends to encourage a small number of well-connected individuals to dominate the website. In 2006, the top 3% of the top 1000 users made 60% of the submissions that reached the front page. This has lead to the creation of *"voting rings"*, in which large number of connected users voted on each other's posts in order to have a chance of gaining visibility, a type of interaction which is specifically forbidden in Reddit's Terms of Service. It is however impossible to determine whether such interaction still persists, given that there is no way to tell which users added each other as friends using the official API.

The type of topical hierarchy of comments and evolution of a thread of discussion is described by Weninger [Weninger, 2014a]. During the evolution of a thread on a post, new comments are added and users vote on older comments, meaning that as time progresses the average comment depth increases. Furthermore, as the comment depth

increases, the discussion diverges significantly from the original topic of the post, the score of each comment decreases as a function of this this behaviour, and time elapsed since the start. A limitation of this study is that it relied on using the official Reddit API crawler, which restricted access to the data and retrieved only the 100 most popular posts from the 25 biggest subreddits, by number of comment. The way in which our work managed to overcome such limitations is described in section 3.2.

To understand the evolution of a content aggregator such as Reddit, Singer et al. [Singer et al., 2014] released one of the most comprehensive studies at the time, ecompassing all submissions in a period of about 4 years (between 2008 and 2012). It examined the evolution of user comments, and the perception and attention of the community as a whole. One interesting aspect is that they measured the popularity of subreddits by the total number submissions, which should lead to similar results as those which we analysed in our first research question (popularity by total number of comments) in chapter 4.1. They noted the diversification of subreddits, since at the beginning of 2008, the top 20 subreddits accounted for between 70 and 80% of the total submissions, while at the end of 2012 that number dropped to less than 40%. Of note is that there has been a large increase in *"self posts"*, which is simply textual content created by the user, as opposed to a link towards an external source, or image. In fact, this type of post has become the largest type of content by total number of submissions, with image posts being the second largest. In our analysis, we found that the biggest subreddit (both by the number of submissions, and comments), by far, is /r/AskReddit, in which users pose a question, and Redditors attempt to answer it. Singer et all. [Singer et al., 2014] concludes that large online communities with high degrees of freedom can often dramatically change their nature and focus over time, suggesting that *"The Front Page of the Internet"*, might be a more self-referencial, and content generating, community than it's initial direction might have led us to believe.

Another influence on this paper was a study [Stoddard, 2015] which analysed the popularity of posts on Reddit and Hacker News (a website which shares most of the traits of Reddit, but is focused on a niche of users that are technically inclined). It found that higher quality articles were more popular with the online communities than those of a lower quality, and it evaluated the popularity by the estimated number of user views. We used a similar metric: the total number of comments on a post, which should also be an indication of higher user activity (results described in section 4.1).

2.2 Online Communities

The concept of an *"online community"* is a natural extension of observed human interaction in the physical world, on the internet. The term was originally coined by [Hiltz, 1984], and is a framework that helps make sense of the previously mentioned idea of Computer Mediated Communication (CMC) [Herring, 2004], and the types of communities that users naturally form as a result of their use. The definition seems to have shifted over time, from a *"virtual community consisting of members distant with each other"* [Wu, 2017][Rheingold, 1994] to one which seems to describe modern online communities present in social media rigorously [Porter, 2004]:

> ***"a virtual community is defined as an aggregation of individuals or business partners who interact around a shared interest, where the interaction is at least partially supported and/or mediated by technology and guided by some protocols or norms"***

A useful concept in understanding the explosive success of online communities is that of *"the third place"*, coined by Ray Oldernburg [Oldenburg, 1999], which denotes a realm that is outside the home or work. While his book describes what makes such a third place successful and why they are important for civil society, it has been argued [Soukup, 2006] that CMC Systems, such as chat rooms and social networks, exhibit many of the traits commonly associated with a desirable third place: they are accessible and on neutral ground, conversation is the main activity, they emphasise local community, and they are social levelers. As these properties are often difficult to attain in the physical world, there has been a large shift towards the virtual in past decades.

Therefore, the difficulty in studying online communities is the scope and diversity of users, data, and interaction patterns. Thus, we will leverage concepts defined in the field of Computer-Supported Cooperative Work (CSCW), initially coined in [Kaufmann, 1988], and further expanded upon in [Wilson, 1991], in order to create a quantitative analysis of Reddit comments. It is briefly explained by Wilson as the term which combines *"[...] understanding of the way people work in groups with the enabling technologies of computer networking, and associated hardware, software, services and techniques."* In order to analyse patterns of behaviour that people exhibit when contributing to online communities such as Reddit, we will need to consider social network analysis, a method that is fundamental in social science research.

2.3 Social Network Analysis

In the interest of successfully replicating results presented by Jialun Wu [Wu, 2017], some of the background which led to decisions made in the methodology of that paper will be presented here. Social network analysis (SNA) investigates social structures by using graph theory and networks [Otte and Rousseau, 2002]. This has been widely applied to other social media networks, such as Twitter [Grandjean, 2016], and can be used to model collaboration, friendship acquisition, and romantic relationships [Abraham and Hassanien, 2009].

SNA is applied on network structures characterised by *nodes*, which represent individual people, or distinct things within the network, and *edges*, that determine the relationship between the nodes [Otte and Rousseau, 2002]. This structure is very versatile, and can be used to represent a large number of scenarios, including road maps where the cities are nodes and the roads are edges [Wu, 2017]. In the case of an online social network as complex as Reddit, serveral options could be viable, including assigning individual users, posts or subreddits as nodes, and their friendship, similarity (by some measure), or concurrent use as edges. In one study that tried to map the "digital humanities community" on Twitter [Grandjean, 2016], it identified the relevant users by

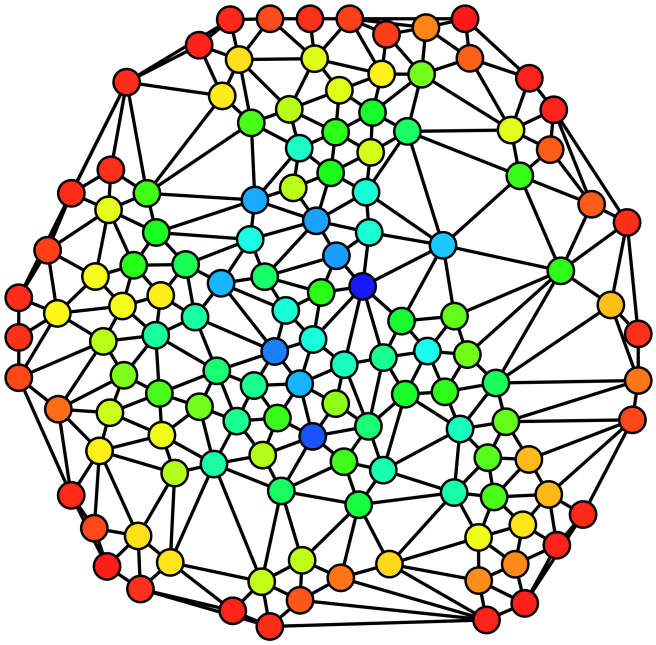

Figure 2.2: An example of a graph where the hue (from red=0 to blue=max) indicates each node's betweenness centrality. [Rocchini, 2007]

use of CSCW (i.e., picking out relevant users from a computer generated pool by hand, and adding relevant data), and assigned edges according to the follower connections the users had with each other.

Once this structure was created, the authors used *"centrality analysis"* to determine the significance of each node, then clustered users into disctinct language speaker communities (e.g., English, French, Spanish etc.) without ever performing a linguistic analysis of the content. Centrality measurements offer indications that a node has significant connections with other nodes [de Laat et al., 2007] [Liu, 2011], and are applied to describe each node in the whole network (by measuring aspects of their edges).

There is no single way to define the importance of a node, and each variant of centrality measurements takes different aspects into account [Wasserman, 1994]:

- *Degree centrality* is probably the easiest to understand from a conceptual point of view: it is simply defined as the total number of links for each node in the graph [Erdős and Gallai, 1960].
- *Closeness centrality* represents the average length of the shortest path between the node and all other nodes in the graph. The more central a node, the closer it is to all other nodes [Bavelas, 1950].
- *Betweenness centrality* quantifies the number of times a node acts as a bridge along the shortest path between two other nodes [Freeman, 1977].
- *Eigenvector centrality* assigns a relative score to all nodes on the network, based on the idea that a connection to another high-scoring node is more influential, than one to a lower scoring node. As such, a node with a high score is connected

to many other nodes who themselves have high scores [Langville, 2006].

How each of these aspects is computed, what they actually measure, and how they are both distinct and relevant indicators with respect to our study, is further explained in section 3.3.

An example of an application of the eigenvector centrality in the field of networking is that of Google Pagerank's algorithm, which is based on a variant of the measure we proposed above [Austin, 2006]. It assigns a weight based on the relative importance of each link that is indexed in a set of documents. In the study of sociology, it was found that centrality measures, combined with hierarchical clustering, were effective in showing which students were prominent within their social networks, and correlated with other factors such as academic competence, aggression, and popularity [Xie et al., 1999].

2.4 Linguistic Content Analysis

We will try to find prominent links between subreddits by measuring some of their aspects quantitatively, namely the number of comments in each subreddit, the average number of posts and authors per day, and so on. However, in order to gain insight into the content of these comments, we will have to rely on *"content analysis"* methods. It is used in social science to quantify patterns in communication, in a replicable and systematic manner [Bryman, 2015]. Content analysis methods have been used in the past on Reddit, but often relied on developing methods of classification by hand to identify eating disorders [Sowles et al., 2018a], or employed basic computerised systems that looked at the size and sharing of messages in order to determine possible interests and benefits to senior citizens [Nimrod, 2010].

We will generate statistical features from selected comments in subreddits in order to quantify their properties from linguistic data. A recent study has been done using these measurements to determine whether there are significant mood changes in the long term present in the population given some circumstances [Ethayarajh and Rudzicz, 2017]. The basic content analysis was implemented using a tool called, *"Linguistic Inquiry and Word Count"* (LIWC), which has been described as dictionary-based content analysis focusing on word count, and basic statistical tests of the raw content [Lowe, 2002] [Wu, 2017]. A limitation of this method is that it ignores context, and is unable to distinguish between different word senses [Ethayarajh and Rudzicz, 2017], although it has been successfully used in the past to measure Post Traumatic Stress Disorder levels on users of Twitter [Harman and Dredze, 2014]. It offers a feature vector of 95 dimensions for each analysed text, and we will try to see if there is any correlation between these and our centrality results.

2.5 Sentiment Analysis, Personality Traits, and Engagement

We will expand our content analysis tools by the use of more advanced tools which rely on *"sentiment analysis"*. Often regarded as *"opinion mining"*, it is a technique used in natural language processing to systematically derive and quantify subjective information from written text or speech excerpts [Cambria et al.,] [Ahmad, 2011]. It aims to extract the attitude of the author or speaker about a specific topic. The two basic tasks of sentiment analysis are emotion recognition and polarity detection [Cambria et al.,]. The former focuses on extracting a set of emotion labels, such as anger, joy or fear, while the latter is a form of binary classification task, defining two opposite emotional states, such as 'positive' vs 'negative' or 'good' vs 'bad' [Cambria et al.,].

Initially developed by Pang and Lee in the early 2000s [Pang et al., 2002], sentiment analysis has been extensively used in fields ranging between finance, marketing and social media analysis, to gain valuable insight about online social communities.

A study conducted by Keneshloo et al. [Keneshloo et al., 2016], successfully predicted the popularity of news using sentiment analysis and other language features, highlighting the importance of this text analysis technique. Furthermore, Horne et al. [Horne et al., 2017] use sentiment analysis in their study of social signals that influence Reddit comment popularity, showing that sentiment based features are more useful in accurate prediction than other categories.

According to the Big Five personality model [Digman, 1990] [Costa Jr and McCrae, 1994], the five main personality traits can be identified as Openness (to experience), Conscientiousness, Extraversion, Agreeableness and Neuroticism, which can be treated as bipolar items. To illustrate, extraversion can be rated on a range from 'extraverted, enthusiastic' to 'reserved, quiet'.

A study conducted by Selfhout et al. [Selfhout et al., 2010] aimed to identify links between the main five traits and the friendship selection process on social networks. Thus, a reliable model for social network friendship prediction has been built using the respective personality traits. A limitation of the study was that the five personality traits were self reported by the participants in the study. In section 3.8 we will use a tool which generates personality trait scores for each of the subreddits. These measurements are generated using a built-in prediction model created with the help of transfer-leaning based machine learning algorithm. For this purpose, we only use four of the five traits - Openness, Conscientiousness, Extraversion and Agreeableness.

A final metric in judging online social network Twitter is that of the *"social engagement"* factor that can be defined as the degree to which one participates in a community [Dwolatzky, 2012]. While it has been usefully applied in the past to rank a set of 264 Universities in the United States by analysing content on their official Twitter accounts, and finding a correlation between the user engagement with them, and their official ranking in a the Times Higher Education, Academic Ranking of World Universities [McCoy et al., 2017], it has not been used in a significant way to study Reddit. We will attempt to use this concept as a final factor that can help explain the popularity

of subreddits.

Chapter 3

Methodology

3.1 Overview of the methodology

In this chapter we will offer an overview of the tools and methods employed in the actual implementation of our work. We will initially present the way in which we collected and processed our dataset, the construction of our graph and the subsequent centrality measures applied to it, and finally the way in which we obtained various factors that we used to explain the popularity of subreddits.

3.2 Data collection and processing

Since our project requires that we analyze posts and comments on the online social network Reddit, we initially had to gather all the data necessary to run our experiments. There are two ways to do this which don't involve the time consuming task of web scraping: using the official Reddit API, which can be accessed through various API wrappers in different programming languages, and processing, then downloading, a freely available online data dump stored on Google BigQuery with Structured Query Language (SQL) Queries.

3.2.1 Official Reddit API

Reddit provides an official Application Programming Interface (API) which allows users to extract data from the website, including posts, comments, and their associated metadata (their scores, time of creation, number of gold received, and so on), however it limits its usage to a fixed number of requests per day. In the past, the size of the data was limited to just a few million comments in total, which made this approach possible [Singer et al., 2014], but recent studies have been limited in their size and scope recently because of it [Weninger, 2014a] [Autman, 2016]. Due to the massive amount of time required to download even a subset of the data, we chose not to use the API, but to resort to a more advanced approach, described in the next section.

3.2.2 Google BigQuery and Google Cloud Storage

Google BigQuery is an online web service which enables working with massive datasets in an interactive manner, by leveraging computing support from Google servers. Once the desired tables were created by using SQL Queries, we stored our data in Comma Separated Values (CSV) files on Google Cloud Storage, then downloaded it to our personal computer.

There is a freely accessible dataset on BigQuery, called *"fh-bigquery:reddit-comments"*, which contains all of the reddit comments from 2005 to 2017 (inclusive). Jialun Wu [Wu, 2017] analysed comments from May 2017 which is a table containing 79,810,360 comments with metadata, resulting in about 21.6GB of data. Table 3.1 gives an example of the schema used in these tables, with all the associated metadata provided for each comment.

While previous works were able to leverage pre-existing data that contained the top subreddits by total number of comments, we initially had to start our work by recre-

Field	Type	Example of data
body	STRING	"I'm so happy this is a thing."
score_hidden	BOOLEAN	null
archived	BOOLEAN	null
name	STRING	null
author	STRING	AeroRespawn
author_flair_text	STRING	Celtics
downs	INTEGER	null
created_utc	INTEGER	1493596801
subreddit_id	STRING	t5_2qh1i
link_id	STRING	t3_68dyyl
parent_id	STRING	t1_dgxwfwo
score	INTEGER	132
retrieved_on	INTEGER	1494514468
controversiality	INTEGER	0
gilded	INTEGER	5
id	STRING	dgyruj0
subreddit	STRING	AskReddit
ups	INTEGER	null
distinguished	STRING	moderator
author_flair_css_class	STRING	top_contributer

Table 3.1: Example of table schema for Google BigQuery dataset

ating this list (in order to have the most up to date information). We chose to analyse the top 5000 subreddits in the period from July to December 2017 (inclusively), which represents a big increase from previous work done in this domain. Once we created the list of top subreddits, we had to individually extract all the comments and associated metadata for each one. Finally, we did some pre-processing work which cleaned the data of comments that were *[removed]* or *[deleted]*, either by users, moderators, or website administrators. The final size of our large combined table was 103GB, and contained a total of 466,164,186 comments, which is a bit over a 5 times increase.

Since we didn't need all of the metadata for our project, we settled on the following fields:

1. body - The content of the comment.
2. subreddit - The name of the subreddit which contains the comment.
3. author - The user who wrote the comment.
4. score - The total score of the comment, after calculating the difference between the number of upvotes and downvotes, we will use this to only select the most relevant comments in each subreddit when doing content analysis.
5. gilded - The total number of Reddit gold received by the comment.
6. controversiality - Whether the Reddit algorithms consider this comment to be of a controversial nature in the community (which is a metric that isn't openly defined by Reddit, though observational data has lead us to assume that comments with large numbers of up-, and down-votes are classified as such, even if the overall score is positive).
7. created_utc - This helps us determine when the comment was created, which we will use to generate statistics, such as the average number of posts (or unique authors, gold, controversial comments, and so on) per day in a given subreddit.

The table below will give a comparison of the scope of this project, and previous data by Wu [Wu, 2017].

	Number of comments	Number of different users	Size of Data
Previous Work	79,810,360	2,937,308	21.6 GB
This project	466,164,186	8,558,522	103 GB

Table 3.2: Statistics generated from the data

3.2.3 Data Processing and Social Graph Construction

In section 2.3 of the background, we described how we would build a social network graph where we assigned a node to each subreddit and then created edges between them. A link was built between two subreddits whenever the same user commented in both subreddits, and we increased the weight for each subsequent user. As such, the

weight of a link between two nodes represented the total number of users who posted a comment in both subreddits.

Because our project was created in Python, building this graph was one of the most computationally expensive parts of the project. In fact, the script which created it ran for only a few minutes on the old dataset of 1000 subreddits in a one month period, but it increased to a few days when run on our larger dataset, a significant exponential increase. Even so, we improved performance by optimising the script where possible (using more efficient data structures, eliminating useless operations where possible), however, the biggest performance enhancement came from compiling our scripts with the use of Cython, a built-in library of Python, which built executable C files from our source. While we tried to create a graph using 15,000, then 10,000 nodes, it quickly became apparent that it was infeasible computationally, and hard to parallelise in spite of significantly more resources. We settled on 5,000 subreddits because this still represented about 95% of our dataset (441,820,314 comments out of the total of 466,164,186). Figure 4.1 shows how abruptly the total number of comments in each subreddit decreases, given that it follows a Power Law distribution. We will discuss this fact in section 4.1.

In creating our graph, we only took the data in the *"author"* and *"subreddit"* fields. We indexed the nodes and gave them a unique id, then created the links using the process described above. In order to carefully replicate the work described by Wu [Wu, 2017], we only kept the first 90% of links (ranked by weight), and removed those with weights under 2.

When we started out, we initially recreated these steps for the old data, to make sure that everything was working as expected, and that the results were replicable. Then, we tested a small subset of the new data, to gain some insight and correct any issues that arose. Finally, we used this streamlined workflow for the large dataset, a process that took several few days. Results can be found in section 4.1.

3.3 Centrality Measures

As described in section 2.3, centrality analysis was used to measure the significance of each node in our graph, in order to determine which subreddits were the most popular. Due to there being no single definition of popularity used in social network analysis, we have chosen four different centrality measurements, that take different aspects of our graph into account: degree centrality, closeness centrality, betweenness centrality and eigenvector centrality. These were measured using the standard formulas below, however, we leveraged the igraph library (in the R programming language) which contains tools that help with building graphs and calculating the scores for each node, described in section 3.5.

All of the following equations assume a graph $G := (V, E)$ with $|V|$ number of vertices, and $|E|$ edges.

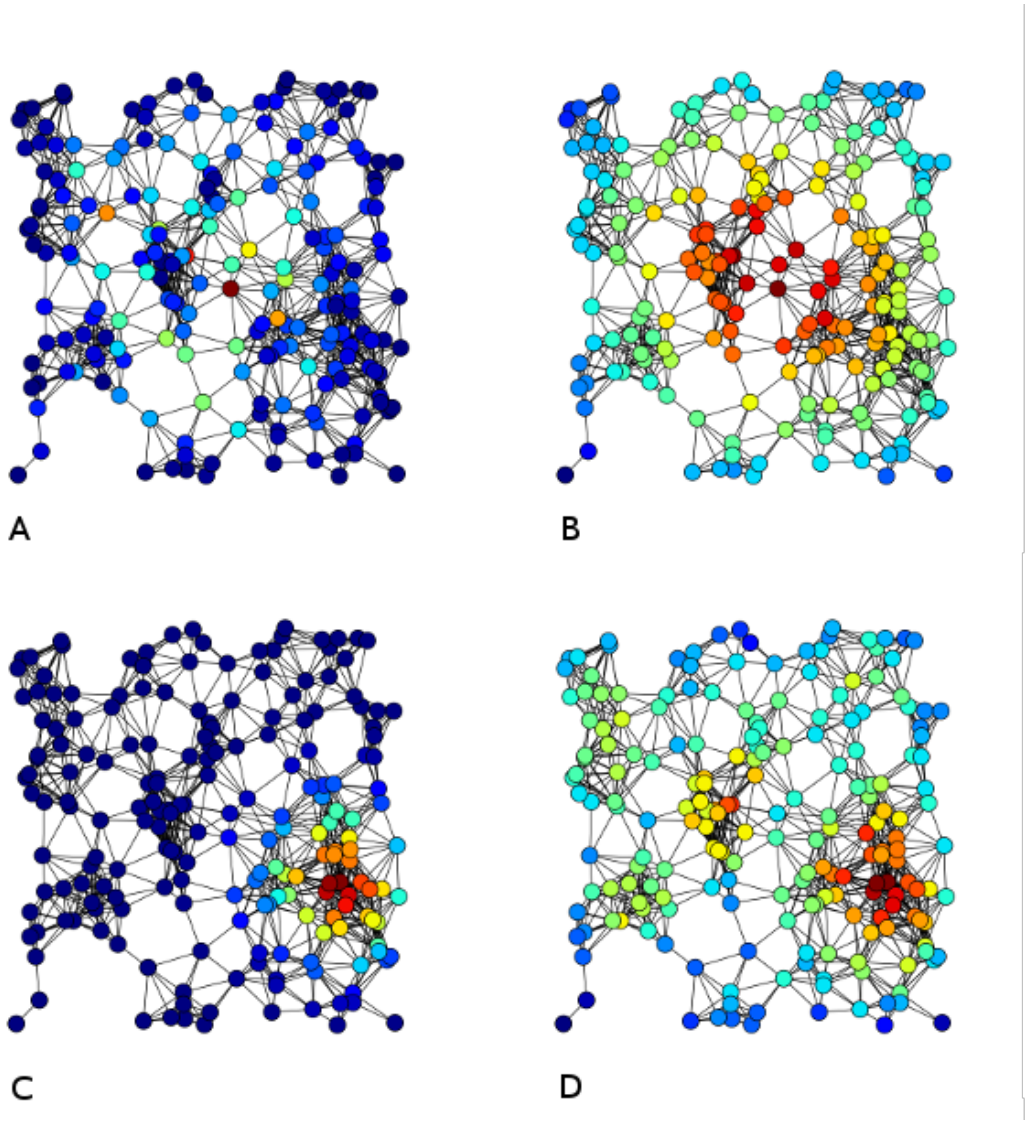

Figure 3.1: Examples of the four centrality measures we have used (A = Betweenness, B = Closeness, C = Eigenvector, D = Degree) (Inverted colours - blue=0, red=max). [Tapiocozzo, 2015]

3.3.1 Degree

Degree centrality is defined as the total number of links for each node in the graph [Erdős and Gallai, 1960]. The assumption is that a node which has a larger number of connections to others will be a central and important node, meaning that in our case we simply counted the number of subreddits that were connected to each other. The measurement is given by the following equation:

$$C_D(v) = \frac{1}{N-1} \deg(v) \tag{3.1}$$

where $C_D(v)$ is the degree centrality of a node v, $\deg(v)$ is the number of nodes that link

directly to it, and $1/(N-1)$ is the normalisation factor, where N is the total number of nodes in our graph. Figure 3.1 (D) shows an example of such a measurement in a sample graph.

3.3.2 Closeness

Closeness centrality represents the average length of the shortest path between the node and all other nodes in the graph [Bavelas, 1950]. It is defined as the reciprocal of farness, thus a node with high closeness centrality is one which is able to connect disparate parts of the graph, being closer to all other nodes. The measurement is given by the following equation:

$$C_C(v) = (N-1)\frac{1}{\sum_x d(x,v)} \tag{3.2}$$

where $C_C(v)$ is the closeness centrality of a node v, $d(x,v)$ is the distance between node v and node x, and $(N-1)$ is the normalisation factor, where N is the total number of nodes in our graph. Figure 3.1 (B) shows an example of such a measurement in a sample graph.

3.3.3 Betweenness

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes [Freeman, 1977]. It is usually given as an example in the case of telecomunication networks, where a node with high betweenness centrality would have more information pass through it, thus giving it more control over the network [Freeman, 1977], however, the argument is easily generalisable to transportation networks as well. The measurement is given by the following equation:

$$C_B(v) = \frac{(N-1)(N-2)}{2} \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3.3}$$

where $C_B(v)$ is the betweenness centrality of a node v, σ_{st} is is the total number of shortest paths from node s to node t, and $\sigma_{st}(v)$ is the number of those paths that pass through v. As usual, $(N-1)(N-2)/2$ is the normalisation factor where N is the total number of nodes in our graph. Figure 3.1 (A) shows an example of such a measurement in a sample graph.

3.3.4 Eigenvector

Eigenvector centrality assigns a relative score to all nodes on the network, based on the idea that a connection to another high-scoring node is more influential than one to a lower scoring node. As such, a node with a high score is connected to many

other nodes who themselves have high scores [Langville, 2006]. Let $A = (a_{v,t})$ be the adjacency matrix, i.e. $a_{v,t} = 1$ if vertex v is linked to vertex t, and $a_{v,t} = 0$ otherwise. Then,

$$C_E(x_v) = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t \tag{3.4}$$

where $M(v)$ is a set of the neighbors of v and λ is a constant. In practice, this value will be normalised over all N nodes. Figure 3.1 (C) shows an example of such a measurement in a sample graph.

3.4 Regression and Statistical Analysis

In order to find a correlation between our centrality measurements and the features given by our subreddit properties (described in section 3.6), the LIWC analysis (described in section 3.7), and the indico.io Sentiment Analysis API (described in section 3.8), we created a linear regression model that took into account all of our input features, and added a constat bias factor. The output was given by the four centrality measures, thus we had to run every experiment four times.

In his research, Wu [Wu, 2017] created two regression models for each experiment: it took all of the input features as predictor variables then ran it for each of the four output variables, and based on these preliminary results, created a second model with a two-step optimising method:

1. Evaluate the multicollinearity with the *"variance inflation factor"* (VIF), given that it indicates whether the input feature has a strong linear relationship with another feature [Fox and Monette, 1992]. The predictor with the largest VIF value was removed, and this process was repeated until all VIF values were below 10, a threshold suggested by Myers [Myers, 1990]

2. Use a stepwise regression, which is a method for fitting regression models by choosing the predictive variables by an automatic procedure [Ralston and Wilf, 1960]. Several ways in which to do this is possible, but Wu [Wu, 2017] chose the *"Akaike information criterion"* (AIC) as the optimising value, as it is widely used in measuring the fitness of the model [Akaike, 1998]. Once the AIC value stops decreasing, the model is stopped.

We will present the results for the full array of features in section 4.1.2, then for the stepwise regression in section 4.1.3. For the regression technique used, we will try to present the same statistical measures by which to judge our results, as in previous works:

- p-value, also called the asymptotic significance, is the probability that for a given statistical model, the statistical summary would be the same as, or of greater magnitude than, the actual observed results, given that there is no other relationship between the measured phenomena [Wasserstein and Lazar, 2016].

- Multiple R-squared, and Adjusted R-squared, to indicate the percentage of changes in the outcome variables that can be explained by input variables [Wu, 2017].
- F-ratio, which measures the level of improvement of the prediction of the model compared with the inaccuracy of the model [Field, 2012].

3.5 igraph Package

The igraph package (for the R programming language and Python) is freely available for download at:

`http://igraph.org/`

It was used in our analysis to conduct the centrality measurements and to fit a linear regression model (described in section 3.4), after we used built in functions to create the network model given our nodes and weights, calculated previously. Built-in functions were provided for all of the 4 centrality measurements, quickly speeding up the process of analysing our data.

3.6 Subreddit Properties Processing

In section 3.1, we presented the metadata fields we settled on for each comment in our selected timeframe. We used this information to create factors by using inherent properties of the comments such as the time they were created, the total karma score it received, and whether the Reddit algorithm considered it controversial or not. Some of these have been shown to have a direct link in explaining the popularity of subreddits, to an even larger degree than basic content analysis tools [Wu, 2017].

The results of these factors are presented in chapter 5, but we will summarise our final chosen properties here, with a brief explanation:

- AvgAuthor - The average number of unique users that posted in a subreddit, per day. This was a previously used factor by Wu [Wu, 2017].
- AvgPost - The average number of posts in a subreddit, per day. This was a previously used factor by Wu [Wu, 2017].
- AvgPostsPerAuthor - The average number of posts created by distinct users for every subreddit
- AvgScorePerDay - The average karma that users received in a subreddit, per day.
- AvgControversiality - The average number of controversial posts in a given subreddit, per day.
- TotalGold - The total gold received in a subreddit, in the whole timeframe.

- TotalGoldPerAvgAuthor - The total gold received in a subreddit divided by the number of unique users that posted in it, per day.
- TotalGoldPerAvgPost - The total gold received in a subreddit divided by the number of distinct posts created in it, per day.
- AvgGoldPerDay - The average number of gold received in the whole subreddit, per day.
- AvgGildedPostsPerDay - The average number of gold received per post in a given subreddit, per day.
- ScoreOver50Count - The total number of comments that had a karma score of over 50.
- ScoreOver100Count - The total number of comments that had a karma score of over 100.
- ScoreOver200Count - The total number of comments that had a karma score of over 200.
- ScoreOver500Count - The total number of comments that had a karma score of over 500.
- ScoreOver1000Count - The total number of comments that had a karma score of over 1000.
- ScoreOver2000Count - The total number of comments that had a karma score of over 2000.

The choice of these measures is based on our subjective understanding and use of Reddit in the past. Given that the AvgAuthor and AvgPost measures were shown to be an effective measure in describing the popularity of subreddits, we chose to leverage unique features inherent in the functionality of Reddit, that are not available on other online social networks. Results for this part can be found in chapter 5.

3.7 Linguistic Inquiry and Word Count

As previously described in section 2.4, the Linguistic Inquiry and Word Count (LIWC) program has been successfully used in research previously. It was created in the 1990s [Tausczik and Pennebaker, 2010], and has since been used to analyse the content of social networks, including Reddit [Ethayarajh and Rudzicz, 2017], by using a dictionary based analysis to track the number of occurences of specific words. The latest revision, LIWC2015, is available online. We will use this program to generate a vector of 92 dimensions for each subreddit, by analysing the top voted 500 comments in each of the 5000 subreddits we initially created the centrality measurements for, then run our regression models to find if there is any correlation. Results can be found in sections 4.1.2 and 4.1.3.

The latest revision of LIWC, uses a dictionary of 6,400 words, word stems and select emotions [Pennebaker et al., 2015]. While the system generated vectors of 93 dimensions for every analysed text, they can be largely sorted into the following categories:

- *Word Count* - a measure which analyses the total size of the text.
- *Summary Language Variables* - includes measures for Analytical thinking, clout, authenticity, and emotional tone .
- *Linguistic Dimensions* - measures pronouns, articles, prepositions, and so on.
- *Other Grammar* - counts the number of common verb, adjectives, comparisons, and so on.
- *Psychological Processes* - counts words which may indicate positive and negative emotions, social processes such as male and female references, cognitive processes (certainty, insight, etc.), biological processes (health, sexual), drives (achievements, risk, reward), relativity (motion, time, space), personal concerns, and informal language.

The justification for analysing only the top 500 comments (by total score) on each subreddit comes from the fact that the LIWC program is very computationally intensive on a personal computer. Given that Wu [Wu, 2017] did not specify how many were used in his research, we reached this number by trial and error, on the May dataset. Another limitation of LIWC is that it ignores context, and is unable to detect sarcasm or irony [Tausczik and Pennebaker, 2010]. To solve both of these issues, we used a more advanced cloud computing tool, described in the next section.

3.8 indico.io Sentiment Analysis

Indico.io is an online RESTful API that allows users to analyse text, image, and other types of comments using a transfer-learning based, pretrained machine learning algorithm [Indico Data Solutions, 2018]. Because the model is already trained, and the API makes calls to an online server, this reduces our computation time significantly, even though we analysed the same 500 comments, as we did previously with LIWC.

We used four different measures, some of which mirror categories available in LIWC. All of the returned results are probability measures:

- *Sentiment Analysis* - Determines whether a text is positive or negative. For example, the text of a 5 star review is positive, while the text of a 1 star review is negative. The API performs with 93% accuracy on the IMDB dataset [Indico Data Solutions, 2018].
- *Emotion* - Predicts the emotion expressed by an author in the following categories: anger, joy, fear, sadness, surprise.
- *Personality* - Predicts the personality traits of a text's author: extraversion, openness, agreeableness, and conscientiousness. Note that it lacks any measure of

neuroticism, which is one of the five big personality traits we mentioned in section 2.5.

- *"Twitter Engagement"* - A measure which is used to predict audience engagement on Twitter, but which we found useful in our testing even given large texts found on Reddit.

The results for these measurements can be found in chapter 6, and they can be contrasted with the replicated results in chapter 4.1.2 and 4.1.3.

Chapter 4

Research Objective 1: Replication of Previous Work

In this chapter, we will start by presenting our work done in replicating the results given by [Wu, 2017]. The aim of that project was to determine the relationship between several characteristics of subreddits, such as the total number of comments in a given timeframe or the average number of posts per day, and the popularity of subreddits, which was measured using centrality techniques, as described in 3.3. We will present our results for the larger dataset in the first research objective.

Based on those findings, we tried to find additional characteristics which could be relevant to Reddit, such as the average number of gold received on a given subreddit each day. The full list, and a description of these features can be found in section 3.6, and we will present the results for these in the second research objective.

Finally, we have tried to use more advanced content analysis tools, as suggested by [Lowe, 2002]. The features were generated using an online machine-learning based API that specialises in sentiment and content analysis, called *"indico.io"*. For reference, the categories for these features were presented in section 3.8. We aim to contrast these results with previous work, in our final research objective.

For each of these research objectives, we began by running a linear regression model of our full list of features against each of the four centrality measures. We then proceeded to eliminate the predictors with large VIF values, until the ones that remained had values below 10, as described in section 3.4. A final stepwise regression model was fitted for the measures, and these results will show which are the most significant factors that can best explain the popularity of subreddits.

4.1 Initial subreddit properties analysis

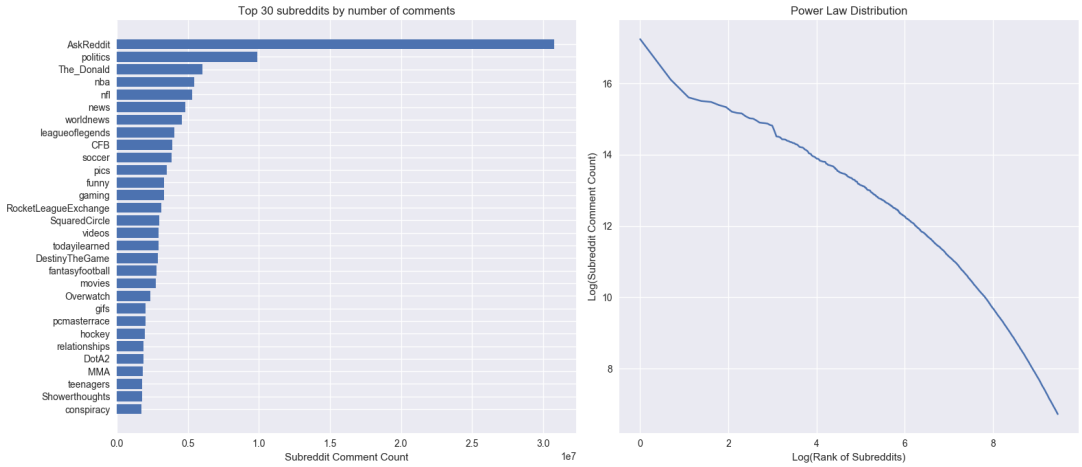

Figure 4.1: Distribution of comments among top subreddits

Figure 4.1 should offer some perspective on the large gap between subreddits, when taken by the total number of comments. We used this data to construct our social network graph, because we could only analyse a subset of the whole dataset, given computational and time constraints. It roughly corresponds to the Pareto Principle, also known as the 80/20, which states that roughly 80% of the effects come from 20% of the causes [Box and Meyer, 1986]. In our case, we chose to present the distribution of comments counts (for each subreddit) and the relative positions they have with each other by rank, in a log-log scale to show the proportional relationship between the two values.

In order to provide a reference for the centrality measures we obtained below, this table will show the ten most commented subreddits in the last 6 months of 2017.

Rank	Subreddit	Total Number of Comments
1	AskReddit	30,787,687
2	politics	9,885,668
3	The_Donald	5,999,631
4	nba	5,428,135
5	nfl	5,285,251
6	news	4,831,723
7	worldnews	4,579,878
8	leagueoflegends	4,022,852
9	CFB	3,884,399
10	soccer	3,832,341

Table 4.1: Top 10 subreddits with the highest number of comments between July and December 2017 (inclusive)

We found that *"AskReddit"*, the community with the most comments that only allows question-based text-posts, has over 30,000,000 comments, while the second largest subreddit, *"politics"*, has less than half that amount.

4.1.1 Centrality Results and number of comments

This section will show the four centrality results against the total number of comments in a subreddit. The measurements shown here are used subsequently in the rest of our discussion.

4.1.1.1 Degree

Since the degree centrality measurement is simply defined as the total number of links for each node in the graph, we expected to find large values for these, as similar results were found in previous studies [Steinbauer, 2012]. It seems that the the biggest subreddits are very interconnected, leading us to make the assumption that most users frequently interact with many different subreddits. The observed p-value is 0.001, with the R-squared value for degree centrality is being quite low, 0.035, which leads us to conclude that comment count is a significant predictor which doesn't fully explain the popularity distribution, as previously found by [Wu, 2017]. Of note is that all of the presented subreddits by [Wu, 2017] appear in the top 10 of our results, but not in the top 5, as were shown previously. Because of this, we decided to include larger tables, to show relative changes in results. Please refer to figure 4.2 for a graph of the degree score distributions with respect to the number of comments.

Rank	Subreddit	Degree	Comments
1	AskReddit	4,988	30,787,687
2	mildlyinteresting	4,984	1,390,705
3	Showerthoughts	4,983	1,764,335
4	funny	4,983	3,341,450
5	pics	4,983	3,522,976
6	movies	4,982	2,721,654
7	OldSchoolCool	4,981	705,974
8	todayilearned	4,980	2,913,802
9	worldnews	4,979	4,579,878
10	videos	4,979	2,934,422

Table 4.2: Top 10 subreddits ordered by their degree score

4.1.1.2 Closeness

Closeness centrality represents the average length of the shortest path between the node and all other nodes in the graph, and the results we found show a smoother curve than the one for degree centrality. In fact, the p-value for this measure was identical to the

one for degree centrality, but with an increased R-squared value of 0.232. Surprisingly, AskReddit does not appear in the top subreddits, with only 2 of the subreddits presented by [Wu, 2017] appearing in the top 10. Such changes suggest that taking into account a larger timeframe and more subreddits can significantly alter results in the ranking of subreddits, but the distributions remain similar, with *"RocketLeagueExchange"* having the largest score, just as it was found previously [Wu, 2017]. Please refer to figure 4.3 for a graph of the closeness distributions with respect to the number of comments.

Rank	Subreddit	Closeness ($*10^{-6}$)	Comments
1	RocketLeagueExchange	4.742168	3,136,838
2	relationships	3.841913	1,855,213
3	SquaredCircle	3.661756	2,964,882
4	CFB	3.648743	3,884,399
5	FireEmblemHeroes	3.613631	1,504,131
6	hockey	3.586916	1,956,251
7	MMA	3.541164	1,839,813
8	conspiracy	3.498975	1,723,041
9	fantasyfootball	3.498791	2,774,702
10	DotA2	3.376029	1,851,821

Table 4.3: Top 10 subreddits ordered by their closeness score

4.1.1.3 Betweenness

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. The table below shows that this measurement offers very little indication about the popularity of subreddits, given the comment count as the feature being analysed. Our p-value is equal to 0.398 and the R-squared is 0.0007, indicating that this is not a significant observation. It can be easily seen that the comment count of these subreddits is very low, and this is completely in line with what we expected to see [Wu, 2017]. Please refer to figure 4.4 for the graphs that summarise the results.

Rank	Subreddit	Betweenness	Comments
1	Makefile_dot_in_user	596,789.3	47,656
2	GothamElections	393,922.8	70,793
3	talesoftherays	367,224.4	18,409
4	subredditfortesting18	296,930.3	27,341
5	CasualPokemonTrades	240,932.5	59,825
6	dirtykikpals	221,423.2	265,098
7	porndiepio	218,688.9	78,236
8	PrivateFiction	214,060.9	15,318
9	friendsafari	201,598.7	138,051
10	YamakuHighSchool	194,990.9	9,158

Table 4.4: Top 10 subreddits ordered by their betweenness score

4.1.1.4 Eigenvector

Finally, the eigenvector centrality measure gives us the best p-value of 0.00001 and an R-squared measure of 0.397, indicating that the comment count is a significant factor in determining its value. Since this measure assigns a relative score to all nodes on the network, we can assume that the ones which appear in the top 10 below have many different connections to other nodes of high importance. Of note is that we received a lower R-squared value than the one presented by Wu [Wu, 2017] (0.4343), possiblty indicating that by using a larger dataset, the eigenvector is unable to explain the popularity of a node purely by comment count, and that we should add more features. Please refer to figure 4.5 for a visual summary of the results.

Rank	Subreddit	Eigenvector	Comments
1	AskReddit	1.0	30787687
2	pics	0.7492143	3522976
3	funny	0.706824	3341450
4	todayilearned	0.6483875	2913802
5	gaming	0.6130091	3305305
6	gifs	0.6092783	2009192
7	videos	0.6076977	2934422
8	worldnews	0.5866734	4579878
9	news	0.5777088	4831723
10	mildlyinteresting	0.4999678	1390705

Table 4.5: Top 10 subreddits ordered by their eigenvector score

In summary, we present previous results by Wu [Wu, 2017], and our own.

Type of centrality measurement	P-value (Old)	P-value (New)	R-squared (Old)	R-Squared (New)
Degree	<0.001	<0.001	0.0165	0.035
Closeness	**<0.0001**	<0.001	0.1674	0.232
Betweenness	0.356	0.398	0.0008	0.0007
Eigenvector	**<0.0001**	**<0.0001**	**0.4343**	**0.397**

Table 4.6: P-value and R-squared values for comment count centrality measurements

Due to the success of replicating these initial results, we will now proceed to the full linear regression, followed by the stepwise linear regression models.

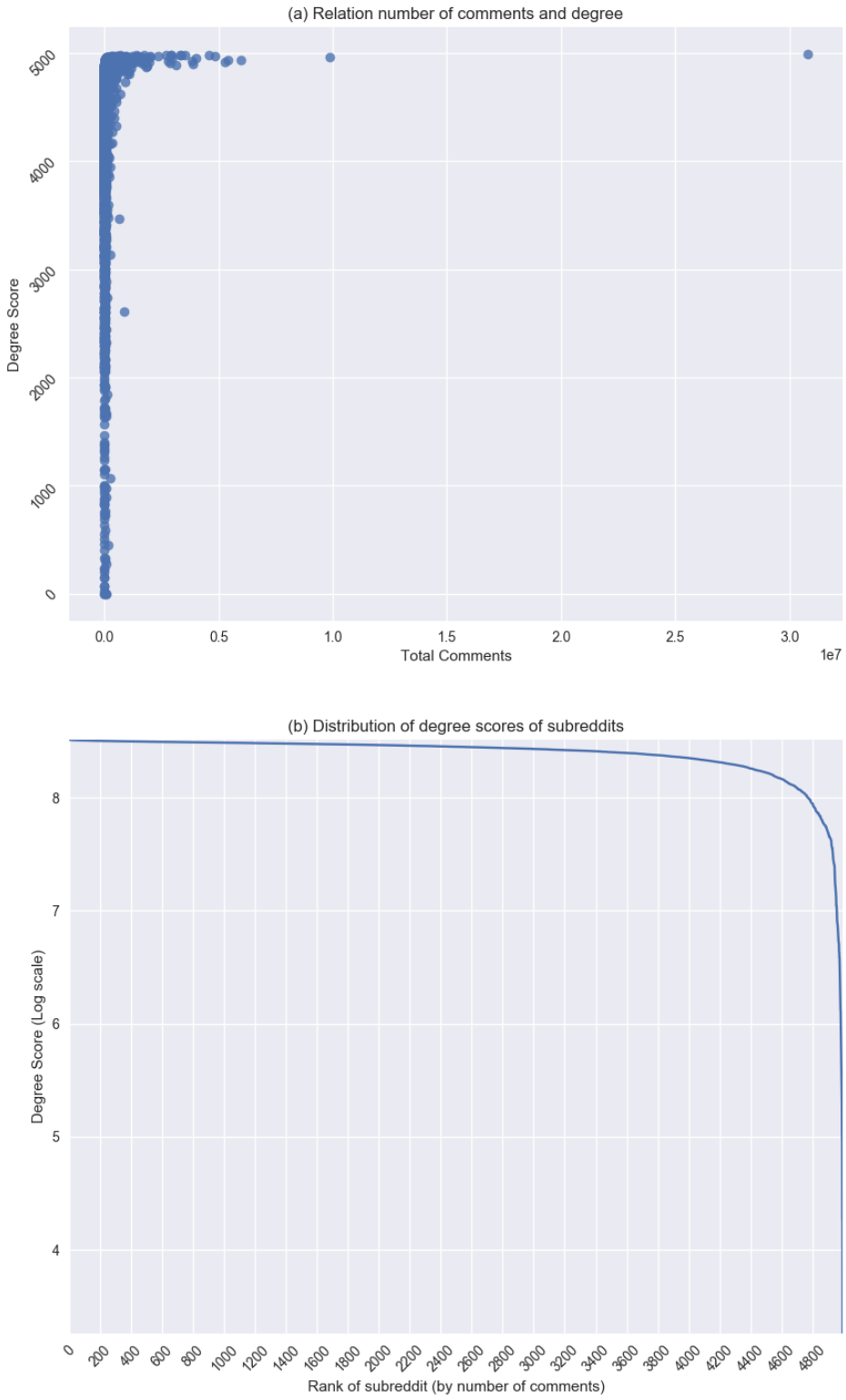

Figure 4.2: Results of Degree (with respect to comment count)

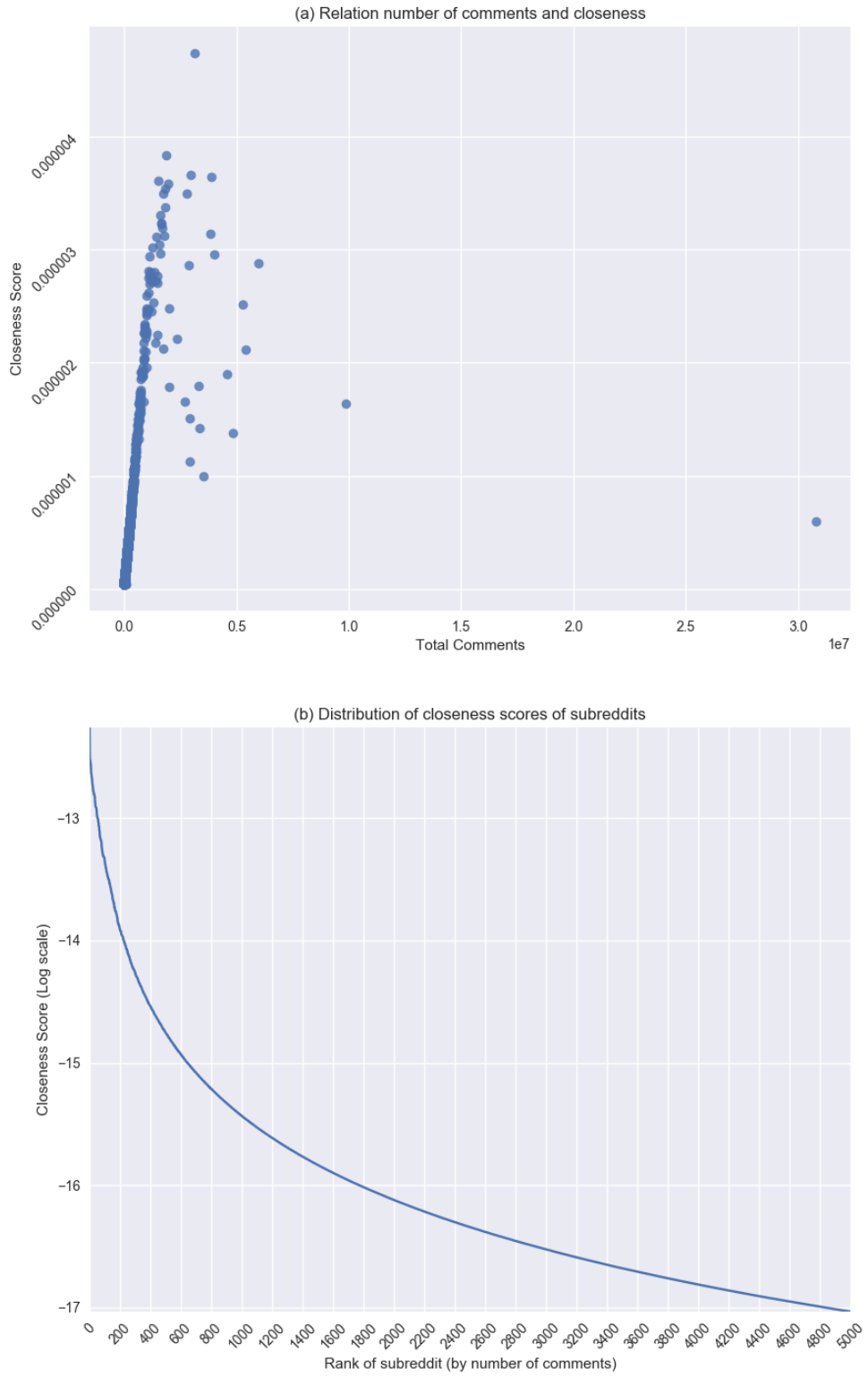

Figure 4.3: Results of Closeness (with respect to comment count)

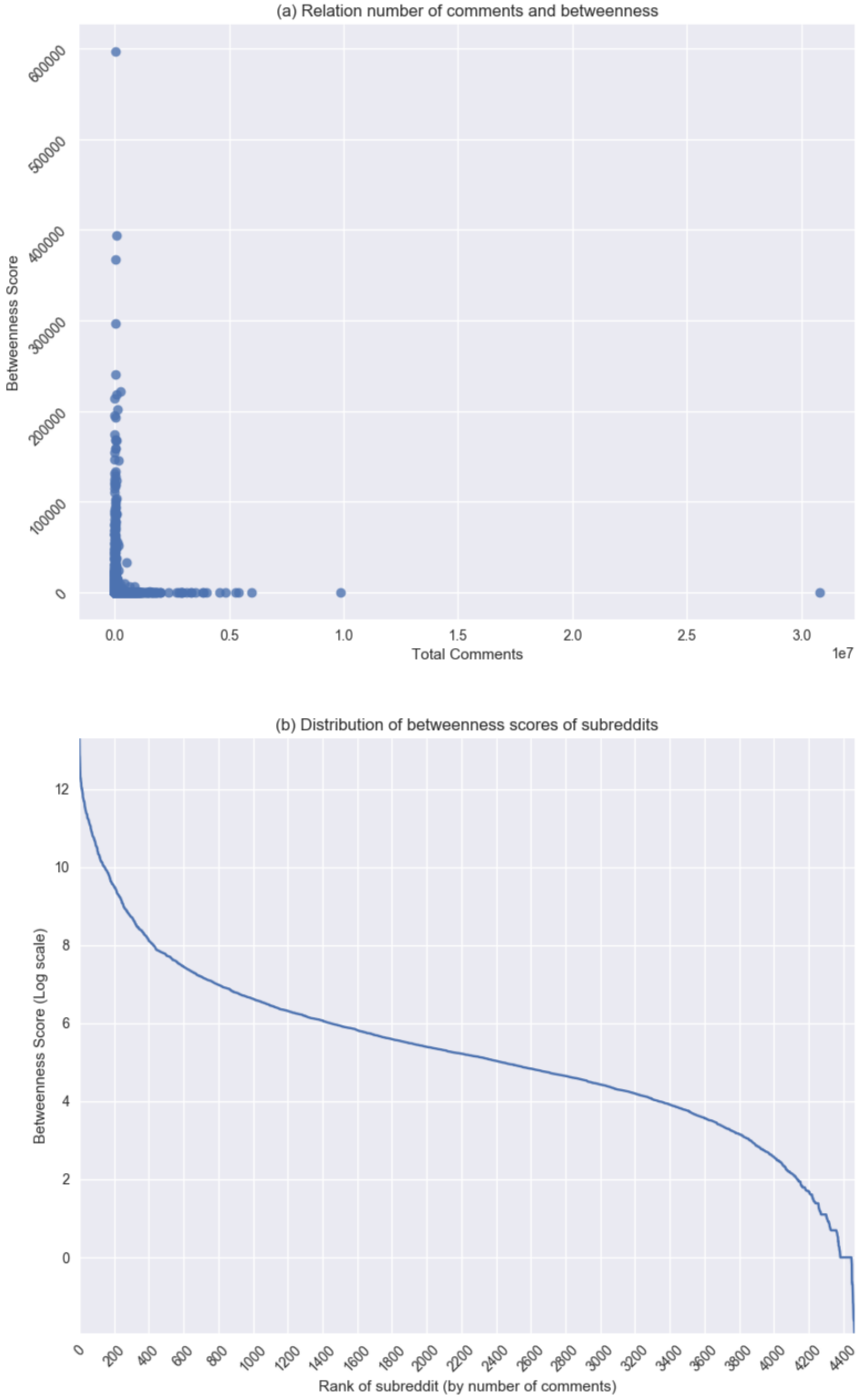

Figure 4.4: Results of Betweenness (with respect to comment count)

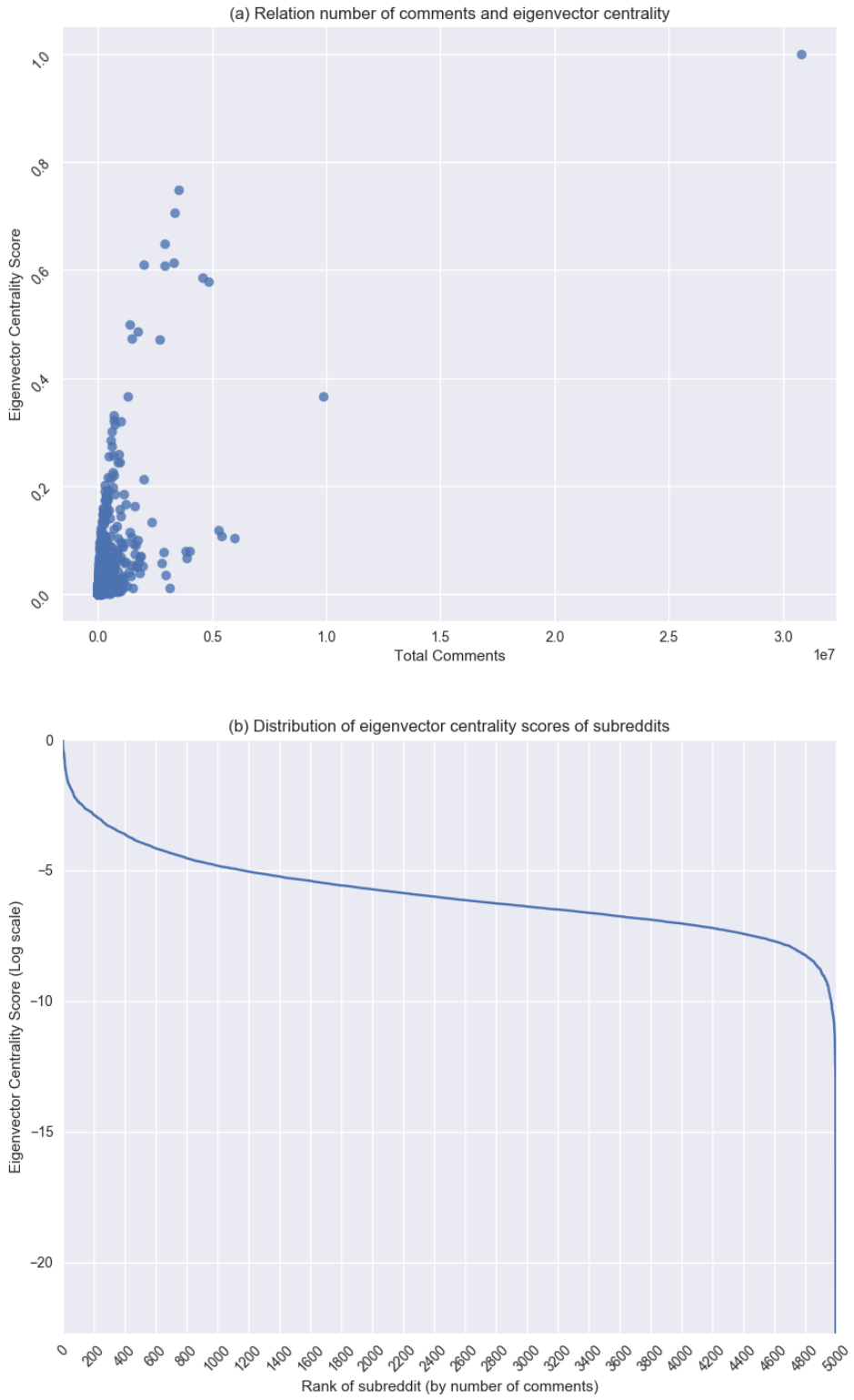

Figure 4.5: Results of Eigenvector Centrality (with respect to comment count)

4.1.2 Regression with all features included

The comment count seems to be significant for some of the centrality measures we adopted, and we managed to obtain very similar results to previous studies, even using a completely different dataset. The next step was to create a linear regression model using features obtained from LIWC. We ran it on our set of 5000 subreddits, from which we extracted 500 comments each: those that were deemed popular by the community (ranked by the highest scores overall, in our given timeframe). This resulted in a feature vector of 93 dimensions, described in section 3.7, and an additional two: the average number of posts per day, per subreddit (AvgPost), and the average number of unique users who posted in a subreddit per day (AvgAuthor).

Results generated by Jialun Wu [Wu, 2017], by running their full regression algorithm are presented below, in table 4.7. The eigenvector centrality provided the best results, due to the low Residual Standard Error and significantly higher F-ratio than the rest of the measurements. In fact, the Adjusted R-squared measurement is the highest observed such value, meaning that about 80% of the variance in the eigenvector centrality can be explained with the used feature vector [Wu, 2017].

	Degree	Betweenness	Closeness	Eigenvector
Residual Standard Error	0.69	0.80	0.85	**0.44**
R-squared	0.57	0.41	0.35	**0.82**
Adjusted R-squared	0.52	0.35	0.28	**0.80**
F-ratio	12.46	6.73	5.08	**43.87**
Degree of Freedom	904			
P-value	<0.0001***			

Table 4.7: Results for full regression on the limited dataset

The values in our table (4.8) are similar, however, in our case Closeness seems to be a better fit than Betweenness, with degree centrality the best out of the three. However, our eigenvector centrality measure remains the best, with an Adjusted R-squared of 0.77, and increased F-ratio's acrross the board, probably due to the increased dataset.

	Degree	Betweenness	Closeness	Eigenvector
Residual Standard Error	0.77	0.90	0.85	**0.48**
R-squared	0.42	0.21	0.30	**0.77**
Adjusted R-squared	0.41	0.19	0.29	**0.77**
F-ratio	37.72	13.35	22.08	**175.10**
Degree of Freedom	4904			
P-value	<0.0001***			

Table 4.8: Results for full regression of the expanded dataset

4.1.3 Regression with significant features

As mention previously, for our second suite of tests, we evaluated multicollinearity between input features and eliminated them on the basis of the variance inflation factor, calculated during our previous round of testing. Each time we ran the tests, we remove the predictor with the largest VIF value, and continued this process until those that were left presented VIF's smaller than our chosen threshold of 10. Afterwards, we ran a stepwise regression, which further reduced the number of statistically significant predictors by using the AIC value as an optimiser.

We present the previously removed predictors by Wu [Wu, 2017] in table 4.9. A total of 29 predictors were removed.

AllPunc, pronoun, ppron, affect, function, relativ, Dic, cogproc, Analytic, informal, social, verb, bio, drives, percept, negemo, Authentic, auxverb, Clout, shehe, differ, prep, AvgAuthor, adverb, Tone, focuspresent, OtherP, conj, adj

Table 4.9: Predictors that were removed in the limited dataset

The ones we removed in our testing can be found in table 4.10. As can be seen our list shrunk by 5 elements, leading to only 24 being removed. Our list contains only a single element which was not removed previously "posemo", which counts words categorised as representing positive emotions, such as "love", "nice", and "sweet".

AllPunc, ppron, pronoun, affect, function., informal, Dic, Analytic, cogproc, relativ, social, bio, verb, percept, negemo, drives, shehe, AvgAuthor, Authentic, focuspresent, Clout, posemo, auxverb, prep

Table 4.10: Predictors that were removed in the expanded dataset

The results for the stepwise regression aren't very surprising, but improvement can be noticed in all F-ratio's, except for the eigenvector measure. The model seems to be more confident in the results, since it can better explain the level of improvement of the prediction, compared to the innacuracy. Table 4.11 gives a summary of the results.

	Degree	Betweenness	Closeness	Eigenvector
Residual Standard Error	0.72	0.81	0.85	**0.69**
R-squared	0.50	0.36	0.29	**0.54**
Adjusted R-squared	0.48	0.34	0.27	**0.52**
F-ratio	23.47	18.81	15.02	**38.91**
Degree of Freedom	970	970	971	970
P-value		<0.0001***		

Table 4.11: Stepwise regression results on the limited dataset

In our testing we noticed a similar trend when compared to the full regression model results presented earlier. F-ratio's have improved across the board, with the exception of the eigenvector, which still have the best overall scores, but now the degree centrality measure is a close second. However, it is obvious that the larger dataset provides more confident results, it also seems to have reduced the adjusted R-squared values, meaning that even this full list of predictors is not able to completely, or even in majority, describe the variance seen in the results.

	Degree	Betweenness	Closeness	Eigenvector
Residual Standard Error	0.78	0.91	0.85	**0.76**
R-squared	0.40	0.17	0.28	**0.43**
Adjusted R-squared	0.40	0.16	0.27	**0.43**
F-ratio	66.24	21.87	56.19	**105.50**
Degree of Freedom	4949	4952	4965	4963
P-value		<0.0001***		

Table 4.12: Stepwise regression results on the increased dataset

Tables 4.13, 4.14, 4.15, and 4.16, give the full list of significant predictors for the stepwise regression model, in each of the four centrality measurement categories.

Table 4.13: RQ1 betweenness

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.220e-16	1.293e-02	0.000	1.000000
WC	-2.753e-02	1.881e-02	-1.463	0.143423
Tone	4.877e-02	2.390e-02	2.040	0.041373 *
WPS	-1.527e-01	1.749e-02	-8.728	<2e-16 ***
Sixltr	6.182e-02	2.092e-02	2.955	0.003143 **
we	-6.759e-02	2.311e-02	-2.924	0.003466 **
you	-4.352e-02	2.012e-02	-2.163	0.030612 *
ipron	-1.435e-01	2.535e-02	-5.662	1.58e-08 ***
article	-5.834e-02	1.982e-02	-2.944	0.003257 **
adverb	-6.552e-02	2.987e-02	-2.194	0.028317 *
conj	1.577e-01	3.567e-02	4.421	1.00e-05 ***
adj	-6.026e-02	2.841e-02	-2.121	0.033949 *
compare	-5.188e-02	2.858e-02	-1.815	0.069563 .
interrog	-6.923e-02	2.505e-02	-2.763	0.005742 **
number	3.497e-02	1.604e-02	2.181	0.029233 *
quant	3.750e-02	2.100e-02	1.786	0.074144 .
anger	6.221e-02	2.847e-02	2.185	0.028944 *
friend	-6.369e-02	1.787e-02	-3.565	0.000367 ***
insight	1.929e-01	2.174e-02	8.872	<2e-16 ***
discrep	-6.713e-02	2.503e-02	-2.682	0.007339 **
tentat	-6.289e-02	2.909e-02	-2.162	0.030658 *
differ	1.152e-01	3.441e-02	3.348	0.000819 ***
body	8.011e-02	2.126e-02	3.768	0.000167 ***
health	-3.781e-02	1.602e-02	-2.360	0.018318 *
affiliation	8.287e-02	2.589e-02	3.202	0.001376 **
achieve	-4.672e-02	2.406e-02	-1.942	0.052184 .
reward	8.912e-02	2.298e-02	3.877	0.000107 ***
focuspast	2.988e-02	1.887e-02	1.583	0.113378
focusfuture	4.831e-02	2.049e-02	2.358	0.018428 *
motion	9.404e-02	1.685e-02	5.581	2.51e-08 ***
work	-3.957e-02	1.932e-02	-2.048	0.040636 *
leisure	-5.333e-02	1.708e-02	-3.123	0.001800 **
money	5.627e-02	1.705e-02	3.300	0.000974 ***
relig	-4.043e-02	1.412e-02	-2.862	0.004227 **
death	-5.530e-02	1.597e-02	-3.462	0.000540 ***
swear	-4.010e-02	2.531e-02	-1.584	0.113237
netspeak	4.963e-02	3.029e-02	1.638	0.101386
filler	2.446e-02	1.420e-02	1.722	0.085170 .
Period	-2.127e-02	1.479e-02	-1.438	0.150515
Comma	3.562e-02	1.797e-02	1.982	0.047534 *
Colon	-5.882e-02	1.707e-02	-3.446	0.000574 ***
QMark	3.756e-02	1.777e-02	2.114	0.034577 *
Exclam	1.903e-02	1.334e-02	1.426	0.153888
Dash	3.003e-02	1.688e-02	1.779	0.075296 .
Quote	1.478e-01	1.946e-02	7.596	3.63e-14 ***
Apostro	-7.927e-02	2.217e-02	-3.576	0.000353 ***
Parenth	-4.608e-02	2.649e-02	-1.740	0.081972 .
OtherP	2.767e-01	3.184e-02	8.690	<2e-16 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Table 4.14: RQ1 closeness

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.441e-16	1.206e-02	0.000	1.000000
AvgPost	4.175e-01	1.226e-02	34.042	<2e-16 ***
Tone	-4.929e-02	1.816e-02	-2.714	0.006668 **
WPS	-2.284e-02	1.418e-02	-1.611	0.107294
Sixltr	-4.234e-02	1.791e-02	-2.363	0.018154 *
we	-6.526e-02	2.047e-02	-3.188	0.001439 **
ipron	7.152e-02	2.145e-02	3.334	0.000861 ***
conj	-1.632e-01	3.104e-02	-5.258	1.52e-07 ***
negate	-3.810e-02	1.588e-02	-2.400	0.016454 *
interrog	8.642e-02	2.139e-02	4.040	5.44e-05 ***
number	-2.113e-02	1.466e-02	-1.441	0.149543
anger	7.679e-02	2.384e-02	3.221	0.001287 **
family	3.341e-02	1.415e-02	2.362	0.018224 *
friend	3.336e-02	1.565e-02	2.131	0.033161 *
discrep	3.114e-02	2.170e-02	1.435	0.151333
tentat	-7.094e-02	2.553e-02	-2.779	0.005478 **
certain	4.331e-02	1.750e-02	2.475	0.013350 *
differ	-8.702e-02	3.137e-02	-2.774	0.005554 **
affiliation	6.437e-02	2.190e-02	2.939	0.003305 **
power	2.656e-02	1.581e-02	1.680	0.093040 .
focuspast	-4.028e-02	1.750e-02	-2.301	0.021410 *
focusfuture	5.729e-02	1.977e-02	2.899	0.003765 **
space	-4.775e-02	1.647e-02	-2.900	0.003752 **
time	5.919e-02	2.023e-02	2.925	0.003455 **
money	3.514e-02	1.372e-02	2.561	0.010474 *
swear	-5.760e-02	2.045e-02	-2.816	0.004876 **
netspeak	-7.049e-02	1.763e-02	-3.999	6.47e-05 ***
filler	-3.248e-02	1.311e-02	-2.477	0.013290 *
Period	-3.996e-02	1.352e-02	-2.956	0.003134 **
Colon	7.610e-02	1.574e-02	4.833	1.38e-06 ***
SemiC	-1.825e-02	1.279e-02	-1.427	0.153608
QMark	-7.770e-02	1.579e-02	-4.920	8.96e-07 ***
Dash	-5.175e-02	1.514e-02	-3.419	0.000634 ***
Quote	5.489e-02	1.548e-02	3.546	0.000394 ***
Apostro	-3.908e-02	1.952e-02	-2.002	0.045294 *
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Table 4.15: RQ1 degree

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8.650e-16	1.100e-02	0.000	1.000000
AvgPost	6.135e-02	1.117e-02	5.491	4.20e-08 ***
WC	-1.539e-01	1.547e-02	-9.943	<2e-16 ***
Tone	-3.764e-02	2.034e-02	-1.851	0.064254 .
you	-5.324e-02	1.643e-02	-3.239	0.001205 **
they	3.836e-02	1.510e-02	2.540	0.011104 *
ipron	1.622e-01	2.313e-02	7.012	2.67e-12 ***
article	1.220e-01	1.815e-02	6.722	2.00e-11 ***
adverb	-4.857e-02	2.601e-02	-1.868	0.061876 .
conj	-1.357e-01	3.145e-02	-4.316	1.62e-05 ***
negate	-9.465e-02	1.486e-02	-6.370	2.05e-10 ***
compare	7.708e-02	1.979e-02	3.895	9.95e-05 ***
interrog	1.237e-01	2.203e-02	5.613	2.10e-08 ***
number	-8.378e-02	1.298e-02	-6.454	1.19e-10 ***
quant	-9.782e-02	1.897e-02	-5.158	2.60e-07 ***
anger	5.265e-02	2.287e-02	2.302	0.021364 *
sad	6.009e-02	1.451e-02	4.142	3.49e-05 ***
family	8.333e-02	1.477e-02	5.640	1.79e-08 ***
friend	1.131e-01	1.544e-02	7.327	2.73e-13 ***
female	-7.794e-02	1.601e-02	-4.867	1.17e-06 ***
male	-2.892e-02	1.759e-02	-1.644	0.100342
insight	-4.672e-02	1.689e-02	-2.766	0.005701 **
cause	3.388e-02	1.584e-02	2.138	0.032551 *
tentat	-1.484e-01	2.482e-02	-5.977	2.43e-09 ***
certain	5.527e-02	1.692e-02	3.267	0.001093 **
differ	9.390e-02	3.033e-02	3.096	0.001972 **
feel	2.836e-02	1.602e-02	1.771	0.076700 .
body	3.482e-02	1.964e-02	1.773	0.076331 .
sexual	-4.400e-02	2.048e-02	-2.148	0.031723 *
ingest	3.723e-02	1.245e-02	2.991	0.002795 **
affiliation	-7.100e-02	1.564e-02	-4.540	5.75e-06 ***
achieve	1.013e-01	2.044e-02	4.954	7.52e-07 ***
power	-6.668e-02	1.749e-02	-3.812	0.000140 ***
reward	-8.658e-02	1.949e-02	-4.442	9.11e-06 ***
risk	-5.460e-02	1.674e-02	-3.262	0.001114 **
focuspast	-5.369e-02	1.686e-02	-3.184	0.001461 **
motion	-5.472e-02	1.496e-02	-3.657	0.000258 ***
space	4.053e-02	1.918e-02	2.113	0.034668 *
leisure	6.780e-02	1.414e-02	4.796	1.66e-06 ***
home	3.972e-02	1.343e-02	2.958	0.003112 **
relig	2.743e-02	1.210e-02	2.267	0.023452 *
death	2.925e-02	1.335e-02	2.191	0.028468 *
netspeak	-1.822e-01	2.514e-02	-7.249	4.85e-13 ***
assent	-3.797e-02	1.191e-02	-3.189	0.001438 **
nonflu	-4.565e-02	1.339e-02	-3.409	0.000657 ***
QMark	-7.712e-02	1.516e-02	-5.086	3.78e-07 ***
Exclam	-3.062e-02	1.138e-02	-2.691	0.007154 **
Dash	5.770e-02	1.277e-02	4.519	6.36e-06 ***
Quote	-1.470e-01	1.384e-02	-10.623	<2e-16 ***
Apostro	8.640e-02	1.918e-02	4.505	6.79e-06 ***
OtherP	-9.540e-02	2.409e-02	-3.960	7.61e-05 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Table 4.16: RQ1 eigen

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.309e-16	1.068e-02	0.000	1.000000
AvgPost	5.779e-01	1.086e-02	53.225	<2e-16 ***
WC	-2.524e-02	1.407e-02	-1.795	0.072767 .
Sixltr	-2.363e-02	1.563e-02	-1.512	0.130579
i	4.266e-02	1.761e-02	2.422	0.015462 *
you	2.227e-02	1.508e-02	1.477	0.139793
they	2.144e-02	1.417e-02	1.514	0.130188
ipron	9.789e-02	2.111e-02	4.636	3.64e-06 ***
article	7.598e-02	1.580e-02	4.810	1.56e-06 ***
conj	-8.698e-02	2.981e-02	-2.918	0.003535 **
negate	-2.898e-02	1.410e-02	-2.055	0.039895 *
adj	-6.898e-02	2.140e-02	-3.223	0.001275 **
compare	9.538e-02	2.260e-02	4.220	2.49e-05 ***
interrog	5.723e-02	1.965e-02	2.913	0.003601 **
quant	-3.486e-02	1.705e-02	-2.045	0.040946 *
family	7.280e-02	1.318e-02	5.522	3.53e-08 ***
friend	6.618e-02	1.435e-02	4.612	4.09e-06 ***
male	3.213e-02	1.735e-02	1.852	0.064140 .
cause	2.628e-02	1.515e-02	1.734	0.082906 .
differ	-1.177e-01	2.858e-02	-4.119	3.87e-05 ***
hear	-2.732e-02	1.258e-02	-2.171	0.029978 *
body	4.272e-02	1.648e-02	2.592	0.009576 **
ingest	4.760e-02	1.162e-02	4.096	4.27e-05 ***
affiliation	-4.148e-02	1.484e-02	-2.796	0.005190 **
achieve	-5.682e-02	1.699e-02	-3.345	0.000830 ***
focuspast	4.198e-02	1.609e-02	2.610	0.009095 **
focusfuture	-3.018e-02	1.500e-02	-2.012	0.044279 *
work	8.155e-02	1.416e-02	5.758	9.04e-09 ***
leisure	5.461e-02	1.395e-02	3.914	9.21e-05 ***
death	2.121e-02	1.195e-02	1.774	0.076072 .
swear	-4.765e-02	1.442e-02	-3.305	0.000955 ***
netspeak	-5.102e-02	1.746e-02	-2.921	0.003502 **
Colon	5.040e-02	1.374e-02	3.669	0.000246 ***
QMark	-3.789e-02	1.378e-02	-2.750	0.005981 **
Dash	-2.768e-02	1.362e-02	-2.032	0.042247 *
Apostro	-6.025e-02	1.829e-02	-3.294	0.000996 ***
tentat	3.293e-02	2.281e-02	1.444	0.148792
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Chapter 5

Research Objective 2: Reddit-specific features analysis

5.1 Overview of Research Objective 2

For our second research objective, we included all the features that could be mined from the metadata present in the increased dataset on Google BigQuery. As such, we intially included AvgPost and AvgAuthor, however, since we used these measures to build others, such as AvgPostsPerAuthor, and TotalGoldPerAvgAuthor, these predictors had a high multicollinearity with each other. The VIF revealed this, and we removed them by using the same process as in our first research objective. Although we initially started out by analysing all of the 15 predictors, it quickly became apparent that the stepwise regression model would more be suitable for this task, due to the similarity present in many of the factors we chose. Table 5.1 presents 9 out of the 15 that were removed.

ScoreOver200Count, ScoreOver1000Count, AvgScorePerDay, ScoreOver100Count, ScoreOver500Count, AvgGildedPostsPerDay, AvgAuthor, AvgPost, AvgGoldPerDay

Table 5.1: Predictors that were removed for the additional features

Descriptions for these predictors can be found in section 3.6 of the Methodology. The choice of different cut-off points for measuring the number of comments above a score, meant that a significant number of these had very similar results, leading to high multicollinearity. Of note is that the two extremes, namely karma scores above 50 (the lowest we measured), and karma scores above 2000 (the highest), were both very significant predictors, except for betweenness centrality, and had a low VIF score. This may lead us to conclude that both are necessary in order to gain a better understanding of how Reddit works in the high-, and tail-end of the distribution of subreddits by popularity.

5.2 Stepwise Regression Results

The F-ratio for the Closeness and Eigenvector values in this test reveals that with these predictors, the model presents confidence in the results. The R-squared value of the Eigenvector centrality measurement is able to explain 50% of the variance with the variance in the remaining 6 predictors, which is an increase on our results that included the content analysis results provided by LIWC.

	Degree	Betweenness	Closeness	Eigenvector
Residual Standard Error	1.00	0.98	0.77	**0.71**
R-squared	0.03	0.03	0.41	**0.50**
Adjusted R-squared	0.03	0.03	0.41	**0.50**
F-ratio	31.90	33.23	567.70	**986.90**
Degree of Freedom	4979	4979	4978	4979
P-value			<0.0001***	

Table 5.2: Summary of stepwise regression results for RQ2

Tables 4.13, 4.14, 4.15, and 4.16, give the full list of significant predictors for the stepwise regression model, in each of the four centrality measurement categories.

Table 5.3: RQ2 betweenness

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.246e-17	1.394e-02	0.000	1.0000
AvgPostsPerAuthor	1.731e-01	1.395e-02	12.411	<2e-16 ***
ScoreOver2000Count	3.913e-02	2.651e-02	1.476	0.1400
ScoreOver50Count	-5.342e-02	2.652e-02	-2.015	0.0440 *
TotalGoldPerAvgAuthor	3.533e-02	2.423e-02	1.458	0.1449
TotalGoldPerAvgPost	-5.381e-02	2.423e-02	-2.221	0.0264 *
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Table 5.4: RQ2 closeness

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.429e-17	1.092e-02	0.000	1.00000
AvgControversiality	3.037e-01	2.160e-02	14.058	<2e-16 ***
AvgPostsPerAuthor	4.430e-02	1.093e-02	4.054	5.12e-05 ***
ScoreOver2000Count	-7.550e-01	2.261e-02	-33.399	<2e-16 ***
ScoreOver50Count	7.932e-01	3.299e-02	24.043	<2e-16 ***
TotalGoldPerAvgAuthor	5.687e-02	1.899e-02	2.995	0.00276 **
TotalGoldPerAvgPost	-3.516e-02	1.898e-02	-1.852	0.06412 .
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Table 5.5: RQ2 degree

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.003e-17	1.395e-02	0.000	1.000000
AvgControversiality	4.873e-02	2.759e-02	1.766	0.077406 .
AvgPostsPerAuthor	-1.097e-01	1.395e-02	-7.863	4.57e-15 ***
ScoreOver2000Count	-1.351e-01	2.888e-02	-4.677	2.98e-06 ***
ScoreOver50Count	1.604e-01	4.214e-02	3.806	0.000143 ***
TotalGoldPerAvgPost	5.894e-02	1.396e-02	4.222	2.47e-05 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Table 5.6: RQ2 eigen

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.035e-17	1.004e-02	0.000	1.000000
AvgControversiality	4.838e-01	1.987e-02	24.355	<2e-16 ***
ScoreOver2000Count	1.561e-01	2.079e-02	7.509	7.02e-14 ***
ScoreOver50Count	1.185e-01	3.034e-02	3.905	9.55e-05 ***
TotalGoldPerAvgAuthor	-5.942e-02	1.745e-02	-3.405	0.000667 ***
TotalGoldPerAvgPost	1.165e-01	1.745e-02	6.679	2.67e-11 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Chapter 6

Research Objective 3: Sentiment, Personality and Engagement analysis of Reddit comments

6.1 Overview of Research Objective 3

For our final research objective, we used the *"indico.io"* API to generate sentiment, personality and engagement measures for the 5,000 subreddits we chose to include in our study. When fitting our regression model, we analysed the VIF's, and concluded that a singular removal would be necessary: all of the four sentiments included in our analysis showed very high values, and so either could have been removed, with the rest of the predictors attaining small values afterwards. Since "angry" had the highest overall value, it was removed. The results for the proceeding stepwise regression can be found in table 6.1. It should be noted that although we ran a full regression on this data as well, it was the first time most values increased across the board, including the F-ratio.

	Degree	Betweenness	Closeness	Eigenvector
Residual Standard Error	**0.88**	0.96	0.99	1.00
R-squared	**0.23**	0.08	0.03	0.01
Adjusted R-squared	**0.23**	0.08	0.03	0.01
F-ratio	**214.00**	76.88	27.95	9.78
Degree of Freedom	4992	4993	4994	4995
P-value		<0.0001***		

Table 6.1: Summary of stepwise regression results for RQ3

A surprise is that the Degree centrality measure is the most suitable for in this case. It might be possible that this result is a case of us not having included any predictors

which were based on gathered metadata, and thus a pure content analysis method is better suited when we take all of the links for every node into account. While similar results can be seen in [Steinbauer, 2012], who had success in using the degree centrality measure in the context of comment graph analysis to measure popularity of communities, it was not based on a content analysis method like we did here, but a more traditional social network analysis which relied on metadata (comment hierarchies, and number of votes). Moreover, we successfully showed that the twitter engagement measurement is a very significant predictor for all of the four centrality measures, giving some perspective about possible future extensions in this field.

Tables 6.2, 6.3, 6.4, and 6.5, give the full list of significant predictors for the stepwise regression model, in each of the four centrality measurement categories.

Table 6.2: RQ3 betweenness

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.369e-16	1.354e-02	0.000	1.0000
sentiment	-3.296e-02	1.482e-02	-2.224	0.0262 *
joy	6.333e-02	1.602e-02	3.953	7.82e-05 ***
surprise	-2.762e-02	1.572e-02	-1.757	0.0789 .
extraversion	-8.704e-02	2.213e-02	-3.933	8.50e-05 ***
agreeableness	9.101e-02	2.279e-02	3.994	6.59e-05 ***
twitter_engagement	-2.526e-01	1.453e-02	-17.379	<2e-16 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Table 6.3: RQ3 closeness

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.280e-16	1.396e-02	0.000	1
sentiment	-6.589e-02	1.498e-02	-4.400	1.11e-05 ***
openness	-1.206e-01	1.611e-02	-7.489	8.18e-14 ***
agreeableness	-7.536e-02	1.729e-02	-4.358	1.34e-05 ***
conscientiousness	-1.183e-01	1.648e-02	-7.180	8.00e-13 ***
twitter_engagement	8.643e-02	1.453e-02	5.947	2.91e-09 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Table 6.4: RQ3 degree

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.899e-16	1.241e-02	0.000	1.000000
joy	-5.351e-02	1.463e-02	-3.658	0.000256 ***
sadness	-3.547e-02	1.269e-02	-2.795	0.005210 **
surprise	2.203e-02	1.449e-02	1.520	0.128611
openness	-2.852e-02	1.440e-02	-1.981	0.047693 *
agreeableness	-6.415e-02	1.456e-02	-4.406	1.07e-05 ***
conscientiousness	-7.870e-02	1.462e-02	-5.382	7.68e-08 ***
twitter_engagement	4.515e-01	1.322e-02	34.145	<2e-16 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Table 6.5: RQ3 eigen

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.412e-17	1.409e-02	0.000	1.00000
joy	-2.681e-02	1.656e-02	-1.619	0.10548
surprise	4.684e-02	1.634e-02	2.867	0.00416 **
agreeableness	3.216e-02	1.414e-02	2.274	0.02301 *
twitter_engagement	6.757e-02	1.436e-02	4.705	2.61e-06 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Chapter 7

Discussion

This chapter presents a summary of the results obtained in this project and compares them with the expected ones, given previous research that has been made. We point out any differences and improvements, and try to guide the reader's understanding to the limitations inherent in the model. While it can be difficult to confidently interpret the results, we do make a note of some observations that might inspire additional research to be made in this area, or point to different fields that might provide us with better insights.

7.1 Research Objective 1:

We started our project by searching for background information on various techniques which were used to predict the popularity of online social communities. The work done by Jialun Wu [Wu, 2017] was novel in its approach, though limited in its scope. In order to gain a deeper understanding of the work done previously, and to be able to build upon it, we gathered the data and tools necessary to replicate findings. While we initially started out by analysing the same data from May 2017, it was clear that given our larger timeframe and computational resources, we could be able to significantly extend already existing capabilities. We proceeded by gathering new data on the online social network Reddit, by leveraging what was openly available on Google BigQuery. Though we initially restricted our replication efforts to a small subset (December 2017), it was readily apparent that optimisations to code and processes could lead to us being able to analyse the most recent six months that were available to us, from July to December 2017.

We created a list of the most popular 5,000 subreddits, since this meant being able to analyse over 95% of the activity in our timeframe. Tools provided by social network analysis are often used to identify interesting patterns in online communities [Grandjean, 2016] [de Laat et al., 2007]. We built a social networking graph by assigning each one of them a unique node in a graph, and created weighted links between them whenever the same user posted in several, an approach that helps determine the relationships between subreddits [Otte and Rousseau, 2002]. Then, we measured the

popularity of subreddits by calculating four centrality measures: degree, betweenness, closeness, and eigenvector.

In gaining an understanding of the measures being used, we chose to initially look at them, and the number of comments in each subreddit, to better understand what they were doing behind the scenes. Although our results are similar to those found previously, they are not identical, and it is clear that comment count, while a significant predictor for most centrality measures, is not able to explain the full variance in results. Thus, we proceeded to include LIWC measures, which is a basic, dictionary-based, content analysis tool that measured the appearance of significant words, sorted in several categories [Tausczik and Pennebaker, 2010]. With the inclusion of our large dataset, our results indicate that several of the features generated by LIWC, are very significant, but fail to explain more than 43% of the variance in the measurements (in the case of eigenvector centrality, and stepwise regression), which is not very surprising given the limitations of the tool [Pennebaker et al., 2015]. Since the increase in data resulted in having to eliminate fewer predictors than was previously necessary [Wu, 2017], it is possible that an even larger increase could lead to better results. Even so, previous studies which were successful in using LIWC for social content analysis [Ethayarajh and Rudzicz, 2017] [Harman and Dredze, 2014] indicate that improvements could be possible, given a better model to fit on centrality measurements.

As mentioned in chapter 2, even large communities like Reddit can significantly change their character over time [Singer et al., 2014], so running the same experiments on data from an earlier date could lead to different results. In our case, however, we were successfully able to replicate findings by [Wu, 2017], as hoped.

7.2 Research Objective 2:

Using reddit-specific measures yielded interesting results: a measure used in the first research objective, which proved to be significant predictor in the case of all, but one, centrality measure (AvgPost), provided the inspiration for the rest of the measures at this stage of the project. We created these measuremenets (the full list can be found in section 3.6) in order to achieve a high overall score in eigenvector centrality, which seemed to perform very well using predictors which grew linearly as the number of users increased, such as: the average score received in a day, or the average number of gold gifted by users in a particular subreddit each day. Although our initiall regression model performed very well, many of these measures presented high multicollinearity, which we eliminated using VIF and a stepwise regression model. As a result, the highest Adjusted R-squared and F-score for eigenvector centrality was found in this test.

A particularly well performing predictor was that of the total gold received per average post, which seems to scale very well even on less popular subreddits. Measures such as these, which use the inherent properties of Reddit have not been widely studied in the past, as most studies performed even on Reddit, measured popularity by standard social network analysis methods that can be applied to any other plat-

form (such as the total number of comments, and their hierarchies) [Weninger, 2014a] [Steinbauer, 2012]. Even comprehensive studies, such as those presented by Singer et al. [Singer et al., 2014], only took into account the different content types of posts, but not features which are inherently unique in the design of Reddit.

7.3 Research Objective 3:

Several suggestions have been made in the past in order to improve research done with the help of content analysis tools [Lowe, 2002]. Many of the ones proposed, such as Profiler Plus, and Visual Text [Wu, 2017] [Lowe, 2002], are proprietary software which is difficult to utilise, basic in functionality, and can be expensive. Though many offer visual representations of the data and measurements, they can be less useful than more modern approaches based on pre-trained machine-learning algorithms. Since these approaches have already been shown to be able to predict popularity of news content [Keneshloo et al., 2016], and even reveal hidden patterns in social interaction that can influence Reddit comment popularity [Horne et al., 2017], the decision was taken to apply these in the context of centrality measures.

Our approach provided us with several different measurements, including an overall sentiment score, the probability of different emotions dominating a subreddit, overall personality traits of users, and finally an engagement score. This final measure, which aimed to factor in the degree to which users participated in a community [Dwolatzky, 2012], proved to be the most significant predictor overall, which was surprising, given the usual application of such measures to significantly shorter texts than those analysed by us.

The overall results show promise, especially since they leverage the degree centrality measure best, which has previously been used only to correlate it with the number of comments [Steinbauer, 2012]. However, they are also least able to explain the variance observed, probably due to the fact that we analysed a very general population. These sentiment analysis tools are most useful when looking for specific traits, in a selected population [Selfhout et al., 2010], and it is not surprising that a very large dataset implies that each subreddit shows a wide range of personalities, and emotions.

7.4 Further Work and Limitations

We chose to categorise our work into three separate research objectives in order to give a clear picture of all the elements that we were looking at, and break them down into useful, and distinct categories. However, due to a lack of time, we were not able to combine all of the significant predictors into a single cohesive package. It is possible that pairing various measures, including "twitter_engagement", with useful predictors found in RO1 and RO2, we could have created a linear regression model that exceeded the best R-squared values and F-scores found previously.

Some of the results are surprising, and they give a better understanding of how a large online community, such as Reddit, works. In this sense, it's clear that no single measure could fully quantify what it means for a subset of a community to be "popular". If selected carefully, however, these could be used to predict specific behaviour, and large-scale changes on the platform.

Several approaches could be used in the future in order to leverage the utility of content and sentiment analysis software such as "indico": a topical clustering algorithm could be applied, in order to categorise subreddits into large groups, and instead of trying to create connections between single communities, it could be shown how these clusters interact with each other, and what kings of behavioural patterns they exhibit. For instance, another category of measurements provided by indico are those that are able show political affiliation, and these could be used to study the interaction (or lack of) between large clusters, potentially gaining important sociological information. The indico API (like LIWC) was not without limitations: it's still proprietary software, with most of the functionality running remotely. Since the source is not available, it is difficult to understand the underlying functionality, and to use it in rigorous studies which require accurate representations of data, and an understanding of the limitations.

While the use of centrality measures is promising, they do present several limitations: it is difficult to predict which centrality measure is the most useful in a given situation, although given a deep understanding of their underlying mechanism, it is possible to suggest uses and improvements, as in the case of eigenvector centrality [Austin, 2006]. While centrality measures are designed to produce a ranking which allows indication of the most important vertices [Bonacich, 1987], it is not designed to measure the overall influence of such a node [Borgatti, 2005]. Furthermore, while measures such as the PageRank link scoring algorithm are useful in predicting with high accuracy the most important interesting link (by some measure) when searching online, the quality degrades rapidly as users exhaust the first options presented [Lawyer, 2015].

Finally, although studies done in the past suffered from limitations in the quantity of data acquired, due to using the limited official Reddit API to gather the data [Weninger, 2014a], it could be interesting to explore not just comments, but to take into account data and metadata from posts, given the current ability to download and process large datasets from online data dumps. When creating a similar graph to the one in this study, weight could be added to the links that connect the nodes by taking into account the popularity of the post as well, to give a more accurate representation the interactions between online communities on Reddit.

Bibliography

[Abraham and Hassanien, 2009] Abraham, A. and Hassanien, A. (2009). *Computational Social Network Analysis: Trends, Tools and Research Advances*. Computer Communications and Networks. Springer London.

[Ahmad, 2011] Ahmad, K. (2011). *Affective Computing and Sentiment Analysis Emotion, Metaphor and Terminology*. Text, Speech and Language Technology, 45. Springer, Dordrecht ; New York.

[Akaike, 1998] Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.

[Alexa Internet, 2018] Alexa Internet (2018). The top 500 sites on the web (april 2018).

[Austin, 2006] Austin, D. (2006). How google finds your needle in the web's haystack. *American Mathematical Society Feature Column*, 10:12.

[Autman, 2016] Autman, H. (2016). A corpus study of ethnic slurs and derogatory language across reddit and youtube with sentiment considered.

[Bavelas, 1950] Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730.

[Bonacich, 1987] Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182.

[Borgatti, 2005] Borgatti, S. P. (2005). Centrality and network flow. *Social networks*, 27(1):55–71.

[Box and Meyer, 1986] Box, G. E. and Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics*, 28(1):11–18.

[Bryman, 2015] Bryman, A. (2015). *Business research methods*. Fourth edition.. edition.

[Cambria et al.,] Cambria, E., Das, D., Bandyopadhyay, S., and AntonioFeraco. volume 5 of *Socio-Affective Computing*.

[Choi et al.,] Choi, D., Matni, Z., and Shah, C. What social media data should i use in my research?: A comparative analysis of twitter, youtube, reddit, and the new

york times comments. *Proceedings of the Association for Information Science and Technology*, 53(1):1–6.

[Costa Jr and McCrae, 1994] Costa Jr, P. T. and McCrae, R. R. (1994). Set like plaster? evidence for the stability of adult personality.

[de Laat et al., 2007] de Laat, M., Lally, V., Lipponen, L., and Simons, R.-J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for social network analysis. *International Journal of Computer-Supported Collaborative Learning*, 2(1):87–103.

[Digman, 1990] Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

[Dwolatzky, 2012] Dwolatzky, T. (2012). Public health for an aging society. *JAMA*, 308(23):2520–2520.

[Erdős and Gallai, 1960] Erdős, P. and Gallai, T. (1960). *Gráfok előírt fokszámú pontokkal*.

[Ethayarajh and Rudzicz, 2017] Ethayarajh, K. and Rudzicz, F. (2017). The effect of photoperiod on the mood of reddit users. *Cyberpsychology, Behavior, and Social Networking*, 20(4):238–245.

[Field, 2012] Field, A. P. (2012). *Discovering statistics using R*. Sage, London ; Thousand Oaks, Calif.

[Finlay, 2014] Finlay, S. C. (2014). Age and gender in reddit commenting and success. *Journal of Information Science Theory and Practice*, 2(3):18–28.

[Fox and Monette, 1992] Fox, J. and Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183.

[Freeman, 1977] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41.

[Grandjean, 2016] Grandjean, M. (2016). A social network analysis of twitter: Mapping the digital humanities community. *Cogent Arts & Humanities*, 3(1).

[Haralabopoulos and Simperl, 2017] Haralabopoulos, G. and Simperl, E. (2017). Crowdsourcing for beyond polarity sentiment analysis a pure emotion lexicon.

[Harman and Dredze, 2014] Harman, G. and Dredze, M. H. (2014). Measuring post traumatic stress disorder in twitter. *In ICWSM*.

[Herring, 2004] Herring, S. C. (2004). Slouching toward the ordinary: Current trends in computer-mediated communication.

[Hiltz, 1984] Hiltz, S. R. (1984). *Online communities : a case study of the office of the future*. Human/computer interaction. Ablex, Norwood (N.J.).

[hnerixh (reddit user), 2018] hnerixh (reddit user) (2018). Life and death of social networks. [Reddit user: hnerixh; Online; accessed 12-April-2018].

[Horne et al., 2017] Horne, B. D., Adali, S., and Sikdar, S. (2017). Identifying the social signals that drive online discussions: A case study of reddit communities.

[Indico Data Solutions, 2018] Indico Data Solutions (2018). Indico api documentation.

[Kaufmann, 1988] Kaufmann, M. (1988). *Computer-supported cooperative work : a book of readings*. San Mateo, Calif.

[Keneshloo et al., 2016] Keneshloo, Y., Wang, S., Han, E.-H., and Ramakrishnan, N. (2016). Predicting the popularity of news articles. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 441–449. SIAM.

[Langville, 2006] Langville, A. N. (2006). *Google's pagerank and beyond : the science of search engine rankings*. Princeton University Press, Princeton, N.J. ; Oxford.

[Lawyer, 2015] Lawyer, G. (2015). Understanding the influence of all nodes in a network. *Scientific reports*, 5:8665.

[Lerman, 2007] Lerman, K. (2007). User participation in social media: Digg study. pages 255–258. IEEE.

[Liu, 2011] Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Data-Centric Systems and Applications. Springer Berlin Heidelberg.

[Lowe, 2002] Lowe, W. (2002). Software for content analysis–a review. *Cambridge: Weatherhead Center for International Affairs and the Harvard Identity Project*.

[McCoy et al., 2017] McCoy, C. G., Nelson, M. L., and Weigle, M. C. (2017). University twitter engagement: Using twitter followers to rank universities.

[Mumford, 2017] Mumford, R. (2017). Itu releases 2017 global ict facts and figures. *Microwave Journal*, 60(9).

[Myers, 1990] Myers, R. H. R. H. (1990). *Classical and modern regression with applications*. Duxbury, Belmont, Calif., second edition.. edition.

[Newman, 2010] Newman, M. E. J. M. E. J. (2010). *Networks an introduction*. Oxford University Press, Oxford.

[Nimrod, 2010] Nimrod, G. (2010). Seniors online communities: A quantitative content analysis. *The Gerontologist*, 50(3):382–392.

[Oldenburg, 1999] Oldenburg, R. (1999). *The Great Good Place*. New York Berkeley, Calif.

[Otte and Rousseau, 2002] Otte, E. and Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6):441–453.

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques.

[Park et al., 2018] Park, A., Conway, M., and Chen, A. T. (2018). Examining thematic similarity, difference, and membership in three online mental health communities from reddit: A text mining and visualization approach. *Computers in Human Behavior*, 78:98–112.

[Pennebaker et al., 2015] Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

[Porter, 2004] Porter, C. E. (2004). A typology of virtual communities: A multi-disciplinary foundation for future research. *Journal of Computer-Mediated Communication*, 10(1):0–0.

[Ralston and Wilf, 1960] Ralston, A. and Wilf, H. S. (1960). *Mathematical Methods for Digital Computers, Vol.: 1*. Wiley & Sons, Incorporated.

[Rheingold, 1994] Rheingold, H. (1994). *The virtual community : homesteading on the electronic frontier*. HarperPerennial, New York.

[Rocchini, 2007] Rocchini, C. (2007). Hue scale representing node betweenness on a graph. [Online; accessed 12-April-2018].

[Selfhout et al., 2010] Selfhout, M., Burk, W., Branje, S., Denissen, J., Van Aken, M., and Meeus, W. (2010). Emerging late adolescent friendship networks and big five personality traits: A social network approach. *Journal of Personality*, 78(2):509–538.

[Sharma et al., 2017] Sharma, R., Wigginton, B., Meurk, C., Ford, P., Gartner, C. E., and Wolfson, M. (2017). Motivations and limitations associated with vaping among people with mental illness: A qualitative analysis of reddit discussions. *International Journal of Environmental Research and Public Health*, 14(1).

[Singer et al., 2014] Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., and Strohmaier, M. (2014). Evolution of reddit: From the front page of the internet to a self-referential community? *CoRR*, abs/1402.1386.

[Smith and Anderson, 2018] Smith, A. and Anderson, M. (2018). Social media use in 2018.

[Soukup, 2006] Soukup, C. (2006). Computer-mediated communication as a virtual third place: building oldenburg's great good places on the world wide web. *New Media & Society*, 8(3):421–440.

[Sowles et al., 2018a] Sowles, S. J., Mcleary, M., Optican, A., Cahn, E., Krauss, M. J., Fitzsimmons-Craft, E. E., Wilfley, D. E., and Cavazos-Rehg, P. A. (2018a). A content analysis of an online pro-eating disorder community on reddit. *Body Image*, 24:137–144.

[Sowles et al., 2018b] Sowles, S. J., McLeary, M., Optican, A., Cahn, E., Krauss, M. J., Fitzsimmons-Craft, E. E., Wilfley, D. E., and Cavazos-Rehg, P. A. (2018b). A content analysis of an online pro-eating disorder community on reddit. *Body Image*, 24:137 – 144.

[Statista, 2018] Statista (2018). Combined desktop and mobile visits to reddit.com from april 2017 to december 2017 (in millions).

[Steinbauer, 2012] Steinbauer, T. (2012). Information and social analysis of reddit. *TROYSTEINBAUER@ CS. UCSB. EDU*.

[Stoddard, 2015] Stoddard, G. (2015). Popularity and quality in social news aggregators: A study of reddit and hacker news.

[Tapiocozzo, 2015] Tapiocozzo (2015). These are six centrality measures on the same graph. [Online; accessed 12-April-2018].

[Tausczik and Pennebaker, 2010] Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

[Wasserman, 1994] Wasserman, S. (1994). *Social network analysis : methods and applications*. Structural analysis in the social sciences ; 8. Cambridge University Press, Cambridge.

[Wasserstein and Lazar, 2016] Wasserstein, R. L. and Lazar, N. A. (2016). The asa's statement on p-values: context, process, and purpose.

[Weninger, 2014a] Weninger, T. (2014a). An exploration of submissions and discussions in social news: mining collective intelligence of reddit. *Social Network Analysis and Mining*, 4(1):1–19.

[Weninger, 2014b] Weninger, T. (2014b). An exploration of submissions and discussions in social news: mining collective intelligence of reddit. *Social Network Analysis and Mining*, 4(1):173.

[Wilson, 1991] Wilson, P. (1991). Computer supported cooperative work (cscw): Origins, concepts and research initiatives. *Computer Networks & ISDN Systems*, 23(1-3).

[Wu, 2017] Wu, J. (2017). Mining reddit to identify factors that describe prominent links between different communities.

[Xie et al., 1999] Xie, H., Cairns, R. B., and Cairns, B. D. (1999). Social networks and configurations in inner-city schools: Aggression, popularity, and implications for students with ebd. *Journal of Emotional and Behavioral Disorders*, 7(3):147–155.