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

Jialun Wu

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2017

Abstract

This paper presents a study of the factors that could describe the character of communities in the online platform Reddit (subreddits) and their relations to the popularity level of corresponding subreddits. The factors were identified with the count of comments and the outcomes of content analysis on the most scored comments in each subreddit. While the popularity was determined with centrality in the social network where subreddits were nodes and shared users were edges. 4 different network centrality metrics, namely betweenness, closeness, eigenvector and degree centrality, were adopted to provide an overall measurement of popularity. The relation between factors and popularity was examined with regression models, and the results of which were evaluated with statistical analysis. Both linear and stepwise regression models were built to compare the regression results. Our models adopted R-squared values and p-value to evaluate the fit of models and determine the significant predictors. The results revealed that comment count is a significant predictor in most cases, while it can provide a better explanation of variance in centralities if combined with features in outcomes of the content analysis.

Acknowledgements

Many thanks to my supervisor Professor Dragan Gašević who promoted such an interesting topic to study. This project familiarize me with the fundamental of data analysis and social network analysis, which maybe beneficial in my career.

Many thanks to Vitomir Kovanović and Srećko Joksimović for help and guidance. I really enjoyed your enthusiastic in the meeting and your timely reply on e-mail. You are so kind and patient to me.

Also big thanks to those who created *igraph* R package and built the online Reddit data dump, you make my life much easier.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Jialun Wu*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Internet has undergone a revolutionary development in recent years attributed to the astonishing growth of storage capacity, computation speed and data transmission technique. This advancement not only enables potentially enormous capacity to store information but also provides means to get access to the Internet both wired and wireless at anywhere, anytime, with almost no cost. The accelerated development and expanding accessibility of the Internet has altered the way people socialize and communicate on a fundamental level. Living in the era of ubiquitous Internet access, people are more and more accustomed to build their social image in the digital realm and participate in the online content generation. Therefore, the boundary between web content providers and audience has been blurred since the Internet users start to upload their own creations online. We are undergoing information explosion of user generated content. The rising of social network websites is a significant result of user participation into the Internet. Reddit, which was founded back in 2005, has now achieved the eighth place globally regarding the daily visitor and page view, estimated by the website Alexa[1]. So far, Reddit has become one of the largest online platforms where millions of registered users share information and discuss by means like comments. Being a leading community website, Reddit aggregate posts in the format of plain text or links to external websites, with the topic of which varying from general concerns like world news to minority interests such as the console game which just released. Registered users, also called 'redditors' by the website, always need to post content to certain 'subreddit'. The term 'subreddit' is the name for the community on the Reddit. Many subreddits features interesting topics such as the subreddit `AmA`, which stands for 'ask me anything', is a place where people get interviewed by answering questions

[1] http://www.alexa.com/siteinfo/reddit.com

other redditors posted in the discussion area. All submissions including posts and comments on the Reddit are graded with a voting mechanism, by which redditor can vote up and down on each submission to determine its ranking on the pages. Featuring the voting system and subreddit structure, Reddit has already become, as the slogan says, 'the front page of the Internet'. Unlike traditional web content provider, Reddit is a place where content is generated, shared, managed and ranked all by users. In other words, Reddit is a self-restraint website without any specific content generator.

After developed and expanded for more than ten years, Reddit has grown its own value and codes of conduct, known as 'Reddiquette'. According to the research of Smith (2016), Reddit already has 250 million registered unique users distributed in 217 different countries by April of 2017, with around 8 billion monthly pageviews. Considering the population that large and the time people have invested in contributing the website, it would be intriguing to look further into the pattern of communication within digital communities and investigate the underlying interaction between various subreddits. Despite the fact that the interaction among different social sub-groups has been widely studied in other disciplines (e.g sociology), the character of online interactions in a large social platform like Reddit remains largely unexplored. Fortunately, Reddit powers official application programming interface (API) for developers with easy access to its data warehouse. With this approach posts and comments can be retrieved to conduct online interaction analysis.

This project mainly focuses on examining the factors that describe prominent links between subreddits. In particular, we aim to utilize network structure to illustrate the latent relationship of the online community determined by shared users. Depending on social network analysis approach, the most important and popular subreddits can be identified under different selection criteria. To be more specific, 4 different measurements will be conducted to help define popularity of each subreddit that took into account. When exploring the aspects that bring about the social status of individual subreddit, we will pay major attention to the comments in the discussion area, which is the predominant component of subreddit content. The analysis of comments takes advantage of the voting mechanism with which to extract the most well-received contents with regards to scores. The analysis involves the count of total comments and linguistic examination, which will not only reflect the general traffic status but also provide an insight into the character of each subreddit with user preference revealed in content analysis.

Two particular research questions are being examined in the project. The first one

is **to what extent can the popularity of subreddits be represented by comment count**. After studying the correlation between count and popularity measurements, we then refer to explore **what factors of subreddits can describe its popularity**. To address the problem we will fit regression models to unveil the links between them.

In the following chapters, we first introduce the background of the research along with the previous work in the related area, followed by an elaborate description of the research question of the project. After that, the methods that involved during the research will be presented. With methods discussed thoroughly, results of the research will be illustrated with discussion and analysis. Lastly, we will have the conclusion summarizes the project followed by future work lists the scope the project could be extended.

Chapter 2

Background

This chapter mainly discuss the previous research and present related works mainly in four different domains. The first part introduces previous work on Reddit, then follows the study about mining online communities as well as computer-supported cooperative work. After that, we talk about the social network analysis research method by presenting the formation of social network and its application in a board research area. Lastly, content analysis method will be introduced associated with related works.

2.1 Study on Reddit

Publications about Reddit as a research object had been collected extensively at the start of our research. These studies gave us an insight of the fundamental state of Reddit and specific methods of data collection, which is beneficial to our project.

As figured out in the introduction section, Reddit has itself included in the list of top 10 globally most visited sites on the web with 8 million monthly traffic. As a data-rich website, it provides means to crawl the database of Reddit and extract considerable content that available in public domain. With fundamental structure inherited from the social new website *Digg* (Steinbauer, 2012), Reddit did not constrain itself in news voting, discussing and sharing but encouraged user participation, it now aggregates all sort of information online and distributes them with a massive amount of independent subreddits. Hence, substantial research has focused on the distinctive characteristics and comprehensive structure of Reddit along with the pattern of user behavior in the discussion area.

In their study Singer et al. (2016) demonstrated the development trajectory of Reddit as a social media platform. They took advantage of the Reddit official API and

crawled Reddit submissions from January 2008 to December 2012. The submission contains the title, author, up and down votes, number of comments, the link or text it contained and the submission time. They also A similar user's opinion on Reddit itself. With the obtained data, they illustrated the evolution of Reddit from its early age to the end of 2012, which includes not only the changes in the number of users and subreddits but also in the type of content users prefer through these years. Plots in the paper reflect the exponential growth in registered users, submissions and the percentage of images in the total number of submissions. Another interesting trend to notice is that self-posted content gradually take over, which means Reddit is not longer just a simple gather of links to other sites but also a large user community that have the ability to generate its own content.

While in the study, Weninger et al. (2013) evaluated the nature of hierarchical comment threads and studied the underlying features of top level comments. Due to large amounts of discussion in each post, he only captured submission with a four-month span of 25 most popular posts. Despite all this, he still collected more than 1 million users' id and karma (user score), 3 hundred thousand posts along with more than 16 million comments, both with votes. To ensure the vote score and comments stable enough, each post was collected 48 hours after creating. By observing the evolution of comment thread and implementing topical clustering algorithms, such as latent Dirichlet allocation (LDA) and hierarchical LDA, the author found a resemblance between hierarchical comment threads and topical hierarchy. Meanwhile, the determination of top level comments had been proved in the paper to have strong relationships with its appearance in the early stage of the post and whether it initiated a new subtopic.

This research, on the other hand, focuses on understanding the essence of subreddits with analysis of substantial comments, and in turn discovering the aspects of subreddit features that contribute to its degree of popularity. There exists research exploring the character of a subreddit with its comments. Chow and Hong (2016) trained a classifier with the purpose of identifying the corresponding subreddit of one particular comment. The training process utilized a neural network and adopted a corpus contains 1.65 billion posts from subreddits with specific themes, while largest subreddits are omitted due to the lack of a well-defined topic. Features to identify different subreddits were implicit in the trained neural model, nevertheless, these features were empirical and can not be visualized and extracted out of the model. Our research, however, removed the constraint on themes of subreddits.

2.2 Mining online communities with computer-supported cooperative work

To gain familiarity with methods on mining online communities with the help of computer system, we read papers to proper ways to analyze Reddit as a large online community.

Referring to communities, we may naturally considering a set of people physically live together. This knowledge has been challenged since the appearance of telecommunication and threatened by the rising of the Internet(Preece and Maloney-Krichmar, 2005). Coined by Hiltz (1985), the word 'online community' denotes a virtual community consists of members distant with each other. After then, the original definition of the community had faded while researchers were picturing the perspective of future communities (Rheingold, 2000; Boyd and Ellison, 2007). In the new definition of the term 'community', physical proximity is no longer a necessity, replaced with the emphasis of people aggregated with shared interest and minimum means of interaction provided with some basic rules(Porter, 2004).

Since online communities growing rapidly and have a potentially profound impact on the current human interaction pattern we have long been familiar with, researchers from disciplines such as social science, particularly sociology, anthropology and social psychology have conducted plenty of studies with methodology migrated from traditional community research. In the work of Kozinets (2002), he developed a technique named *netnography* to perform marketing research on online communities. The technique was adopted from the notion called ethnography, which is a documenting and study designed to examine cultural phenomena from observer's point of view (Hammersley and Atkinson, 2007). With netnography, thinking patterns of online consumer communities can be studied. In his work, he focused on the online coffee group called *alt.coffee* to learn coffee consumption patterns with netnography, which provided insights into the behavior of the future mainstream customer. The results in his work reveal how easily customers can be lured with new products in Starbucks, while small independent coffee shop has a potential risk of losing coffee lovers.

However, online communities have several characteristics that restrain traditional community analysis methods to fully take effect. With the absence of face-to-face interaction and eye contact, online community users have a possibility to preserve an inconsistent online identity, which is far from their genuine personality. As discussed in paper (Bergstrom, 2011), the existence of trolls, which infers users that have a will-

ingness to harm or at least discomfort others, had brought much trouble to maintain the Reddit online platform. Moreover, shortening of the distance between community members also brings in explosive growth on the dimension of online community data. For example, Reddit had already received 1,715,454,785 comments in total by the end of 2015 (Reddit, 2015). With the number this large we can not perform classic sociological analysis technology on comments by human labor, instead, we need to adopt computer-supported cooperative work (CSCW), as defined by Wilson (1991), involves human collaboration associated with computer networking skills. Beenen et al. (2004) utilized CSCW as a discipline to process data from 834 subjects on website *MovieLens*, to analyze the factors that motivate people to make contributions to the online community such as rating and writing comments. To take CSCW into effect, the behavioral theories should be included instead of empirical regularities, moreover, different outcomes need to be predicted by different designs of the model, with the aim of evaluating models with subsequent experiments.

To mine giant online communities like Reddit, our research requires quantitative analysis of Reddit comments, where we need CSCW discipline take place. Thus, we expect our model to produce variant results when feeding with different inputs. Furthermore, social science research method needs to be applied to analyzing the resources, here we consider social network analysis method, which will be introduced in detail in the following section.

2.3 Social network analysis

Since we adopted social network analysis research methods to evaluate Reddit data, we need further understanding of these methods.

The social network, which is a representation of social relations in the form of network structure, has proclaimed its ubiquity in both virtual society and the real one. The root of social network analysis can be found in the graph theory (Otte and Rousseau, 2002). Similar to the graph in the graph theory, the social network is characterized by the nodes (vertices) and the edges (links) between each associated pair of nodes. Nodes and edges in the network may have different attributes in different scenarios (Scott, 2000). In practice, the social network exists in a board range. For instance, the road map is an evident type of social network, with cities as nodes and roads as links between them. The same structure can be found in the map of airports. However, the social network for interaction and social relationship amidst humans is not that

obvious. Under different research circumstance, the network has different establishing methods. In the research of Vaquero and Cebrian (2013), the social network was built to study the relationship of students in the same class, featuring each individual students being nodes and interaction tie between them as a candidate for links. By analyzing the network, the researcher observed intense interactions within a small group of high-performance students. The members of this group received better academic improvement unless they produce reciprocity during the discussion. Those who failed to do so were selectively excluded to the group.

Once the network structure has been established, some key aspects of which regarding the analysis procedure can be examined. Those are, for example, density and centrality. Density here quantifies the degree of connections in a certain number of nodes. The density will reach 100% if all observed users in the group are connected with each other. On the other hand, Centrality, being another important aspect of the network structure, provides measurements to quantify the involvement of particular node in the whole network structure. High centrality is a strong evidence that the node has vast connections with the rest of nodes (Laat et al., 2007).

Centrality has many different measurements so as the word 'importance' has different meanings. Each variant of measurements follows a specific criterion. Degree centrality takes only the degree of the node into account, promoting nodes with the most direct links with the rest (Opsahl et al., 2010). Closeness centrality is raised to select nodes that have the shortest average path to the rest of nodes (Bavelas, 1950). On the other hand, betweenness centrality values the nodes that act as a bridge on the shortest paths for plenty of other node pairs (Freeman, 1977). Lastly, the eigenvector centrality, which provided insight for Google PageRank algorithm (Page et al., 1998), advocates nodes that have the most links associated with it (Ruhnau, 2000). These measurements are these most frequently implemented in research and proved stability compared with many other measuring mechanisms in the work of Costenbader and Valente (2003).

Centrality is qualified to be a norm to evaluate importance and popularity in the network structure. Xie et al. (1999) found in her research that for girls the centrality in the network of peer groups has a strong correlation with the level of popularity among others. This was the result of social network analysis of 506 students from 4 different schools with similar age. A similar analysis of centrality as the factor that describes popularity was adopted in the research of Cillessen and Rose (2005), which also focused on examining the underlying social interaction in the peer group. Zhang et al.

(2011) adopted centrality measurement as the standard for popularity when analyzing the People-Service-Workflow (PSW) network with the purpose of recommending services in a workflow composition process. Other norms of popularity include view count and the score generated by readers if the online community has this specific feature. Stoddard (2014) described in his work the discovering of connections between intrinsic article quality and popularity in two online communities. With methods to create an almost bias-free environment, he published articles in both Reddit and Hacker News and evaluated the quality of the article with user votes, while the popularity was represented by view count. The results demonstrated a strong correlation between the quality of the article and its popularity in the community.

For this project, centrality measurements function as indicators for popularity determination. As the first step, the social network will be constructed with subreddits and the shared users between subreddit pairs. After that, prominent subreddits will be identified by implementing social network analysis research methods.

2.4 Content analysis

Content analysis provides methods to quantify and arrange patterns in the content. The results of quantification can be analyzed with statistical means and methods, and thus help characterize the content itself with the statistical features generated from the analysis. In our research, content analysis is utilized to understand subreddit properties with top comments of different subreddits being extracted.

The development of computer technology enables advanced statistical software to analyze massive content in a short time. With this technique, researchers from all disciplines are able to analyze content in different media platforms, namely newspaper, books and comments in social communities. Nimrod (2010) examined 14 leading seniors' online communities to identify the possible benefits to elders. The examination included 686,283 messages from 13 major subjects in each selected communities over a 1-year span. The quantitative content analysis was implemented on those messages and revealed that more participation in these online communities can contribute to the happiness of seniors' daily life.

In our research, we will perform a basic level of content analysis with the implement of the Linguistic Inquiry and Word Count (LIWC) program[1]. It is a kind of dictionary-based content analysis, as discussed by Lowe (2012), which focus on word

[1] http://liwc.wpengine.com/

counting, sorting, and simple statistical tests of the raw content. The analysis can provide a fundamental character of the examined content. As in our case, comments can characterize the subreddits they associated with. Further detail about LIWC program is discussed in the method chapter.

Chapter 3

Method

In this chapter, we are going to introduce the design work along with the actual implementation of this project. The design work consists of the theoretical social network construction, the mathematical derivation of centrality measurements, the selection of features of subreddits and the configuration of regression models. The actual implementation part includes the data collection, introduction of open source packages chosen to do content analysis, social network analysis and regression separately.

3.1 Data collection

This project focused on analyzing the data from the Reddit social media platform. Reddit had made its official API and code open source and provided the documentation for API on their website[1].

For this project, we had two alternatives for data collection. The first is to download data via Reddit official API with scripts. Here we took advantage of a package called Python Reddit API Wrapper (PRAW), which provides simple access to Reddit's data. The package supports means to create a bot to automatically posting and collecting data from Reddit website and support to those non-Reddit websites to utilize contents on Reddit. However, for this project, we were required to analyze massive numbers of comments which would take too much time to download through API. Thus, PRAW was not suitable for our project. To find easy access to archived comments, instead of crawling data from Reddit directly, we decided to attach to an open source data dump for Reddit.

The second source of data is an online open source Reddit data dump stored at

[1] https://www.reddit.com/dev/api

Google BigQuery, which was what we were searching for. The dataset contains about 1.7 billion comments that compiled in JavaScript Object Notation(JSON). In this dataset, information about comments is classified into about 20 different categories. Each record of data is stored in JSON format and can be converted into Comma-Separated Values(CSV) format. An example of record is shown in listing 3.1.

```
{
    "body": "It sounds like you've given this some
       thought.",
    "score_hidden": null,
    "archived": null,
    "name": null,
    "author": "BoS_Knight3000",
    "author_flair_text": "",
    "downs": null,
    "created_utc": "1491350400",
    "subreddit_id": "t5_2rmt9",
    "link_id": "t3_63gf1m",
    "parent_id": "t1_dfu95z9",
    "score": "3",
    "retrieved_on": "1493802191",
    "controversiality": "0",
    "gilded": "0",
    "id": "dfu97bd",
    "subreddit": "BostonBruins",
    "ups": null,
    "distinguished": null,
    "author_flair_css_class": "PatriceBergeron"
  }
```

We do not need all these 20 different varieties of information. For this project, we only need 5 of them, including `body`, `author`, `created_utc`, `score` and `subreddit`. We provide in the Table 3.1 an detailed explanation of these five fields. We took data from `author` and `subreddit` to create social network and used data from `body` to extract content for later analysis, also we extracted timestamps in `created_utc` to generate daily statistics as properties of subreddits. The `score` was chosen to help ranking

comments to select the most appreciated ones by users in that subreddit.

Table 3.1: Description, format and example of selected 5 categories

Items	Description	Format	Example
body	The content of the comment	STRING	Very true.
author	The author of the comment	STRING	Nevermore416
created_utc	The time of the comment in UNIX timestamps	INTEGER	1498694621
score	The score of the comment	INTEGER	3
subreddit	The subreddit this comment posted in	STRING	Mariners

This dataset includes all comments since 2005 when the website just created. The comments data were initially stored by year since there were only 1,075 comments for the whole year at the very beginning. However, by the end of 2014, the total number of comments per year had reached 531,804,658 and the size of these data had soared to 115GB, which is too large to analyze as a whole and prone to damage and data loss. After then, the comments data was stored per month. The latest sub-dataset uploaded was comments for June 2017 and had 79,901,711 data entries with 21.7GB in total.

Given the challenges of analyzing the complete Reddit dataset, in this project we focused our efforts on a subset of Reddit data from May 2017. The statistics of the dataset is included in the table 3.2. We chose this subset because it was relatively up-to-date and large enough to cover almost all active subreddits. However, we still not able to process data up to 22GB on the personal computer. Due to this reason, the reduction was made to the size of data in the following process.

Table 3.2: Statistics of dataset from May 2017

Number of comments	Number of different users	Size of data
79,810,360	2,937,308	21.6 GB

3.2 Processing data

With the dataset available, the next step was to extract all data required for analysis. In the study, we first examined the network of subreddits to do popularity measurements (i.e., centralities) of individual subreddit and then evaluated properties of each subreddit with its statistics and top scored comments. After that, we built a model to find connections between subreddit properties and their popularity in the whole social network.

3.2.1 Social network construction

The social network for this project took subreddits as nodes, and links between subreddits as edges. Every time one user posted comments on two different subreddits, a link was built between them, assuming these two subreddits share some similarities. If another user visited the same two subreddits and posts comments, the link between them was emphasized by adding weight to existing link. The total weight of each link was the number of different users that visited and posted in these two subreddits it connected to, which means the more every two subreddits been visited the stronger their ties will be.

To generate such social network, we took all data in `author` and `subreddit` section. Since we have to run the program on the personal computer, only top 1,000 subreddits with highest comment count were included in our analysis. We tried to construct the graph with the top 10,000 and also top 4,000 subreddits; however, given the challenges of processing such large networks, the scope has been reduced to the top 1,000 subreddits. Regardless of the decrease in the size of subreddits, we still could generate a representative social network that could cover the majority of traffic on Reddit. For example, there were 79,810,360 comments in May and 63,111,585 of them posted in the top 1000 subreddits, which means the top 1,000 subreddits cover more than 79% of all traffic. Meanwhile, all of these top subreddits had at least 10,000 comments so they could be considered as active subreddits.

The average size of the original author-subreddit pair data is around 1.7GB, which is slightly large for data analysis. To reduce the size of data and speed up the processing, all different author and subreddits were indexed with a unique number. Then, a python script was written to select all subreddits that one user posts and build links between each other. Meanwhile, duplicate links were replaced with weights assigned to them. After that, a network was generated with nearly 500,000 weighted links, which means almost every two subreddits were connected but with different weights. To further prune the network and make it more representative, only the first 90% of links ranked by weights were kept and the rest links with weights below 2 were removed. With the above process, the size of data had been reduced down to only 5MB on average, which was totally possible to create a social network.

3.2.2 Subreddit properties processing

For subreddit properties, we were more interested in the daily statistics of each subreddit, which means the average post per day and the average unique user per day for each subreddit. These properties we suppose have a very direct connection with the popularity of subreddit based on an assumption that popular subreddit should generally have more people visiting and more commenting on a daily basis. To calculate these values, we first processed the data in `created_utc` and grouped comments of selected subreddits by different dates. After that, we counted the posts and different authors of each subreddit for every day, then calculated the average of them.

For content analysis, we took all data from `body` from May dataset, grouped by `subreddit`. These comments shall be analyzed with the Linguistic Inquiry and Word Count(LIWC) program which will be discussed in next section.

3.3 Analyzing the data

3.3.1 Linguistic Inquiry and Word Count

We analyzed the comments with the Linguistic Inquiry and Word Count(LIWC) program, which has been widely used in a large number of computerized text analysis (Coppersmith et al., 2014). The LIWC program contains a dictionary which categorizes words that belong to certain grammatical or psychological domain (e.g., common verbs and negative emotions) (Tausczik and Pennebaker, 2010). The dictionary was first developed at 1993 and kept updating itself in following years. The latest version of the dictionary is LIWC2015, which has about 6400 words, word stems and emoticons (Pennebaker et al., 2015), was selected to analyze the comments for this project. When doing analysis, the program simply read the text word by word and checked if the word has a match in the dictionary. Once matched, the corresponding category or scale obtained an appropriate increment. After analyzing the whole text file, the output was presented as a vector with approximately 90 different variables.

3.3.2 Centrality analysis

In social network analysis, network centrality metrics are widely used to identify the most popular or important nodes in the whole network. However, there are dozens of different centrality measurements that concern different aspects of the network.

For this project, we used four different centrality measurements, which were degree centrality, closeness centrality, betweenness centrality and eigenvector centrality. We measured them with the formulas provided in each following section, based on the subreddits network we generated in the previous section.

3.3.2.1 Degree

Degree centrality measurements are straightforward, the degree of a vertex means all other vertices that have distance one with it, in other words, the degree is the number of all other adjacent nodes of one specific node. The higher degree a node with the more central it should be. This is based on the social assumption that a person who has more direct connections with others in a social network shall be more important. In our network, we measured degree centrality by counting how many other subreddits one subreddit connected with. The measurement was based on the equation

$$C_D(V) = deg(V)$$

where $C_D(V)$ was the degree centrality of node V while $deg(V)$ was the number of nodes that have distance one with it. In practice, we normalised the value with the total number of all other nodes which is $n-1$.

As shown in the figure3.1, the node e is connected with f, h and a, so the degree of node e is 3. So as the degree for node a is 4 since it connected with e, b, c and d.

3.3.2.2 Closeness

Closeness centrality of a node is defined as the inverse of farness, which means the sum of the shortest distance to all other nodes. The higher the value is the more central a node will be. The focus of closeness centrality lies in the ability of nodes to spread information to the whole network. A node with high closeness centrality is considered to be capable of spreading information more conveniently through the whole network(Bavelas, 1950). A subreddit with high closeness centrality has a relatively short distance to all other subreddits. Closeness of one node x is calculated with the formula

$$C(x) = \frac{1}{\sum_y d(y,x)}$$

where $d(y,x)$ is the shortest distance between another node y and the node x.

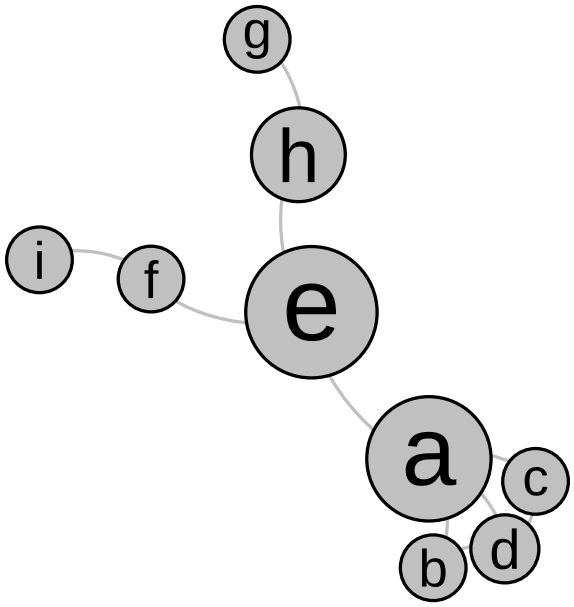


Figure 3.1: Example of social network

3.3.2.3 Betweenness

Betweenness centrality is a measurement that takes the number of times one node is on the shortest path between other two nodes. This was a concept raised by Freeman (1977), which qualifies the control of one particular node over the whole network. Take a power plant network as an example, a power station failure could cause more economic loss if it has high betweenness since a large number of nodes have to take a detour instead of taking the shortest path where the failed station is on. The formula to calculate betweenness is

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest distances from node s to node t while $\sigma_{st}(v)$ is the total number of shortest distances that cross node v.

As illustrated in the figure 3.1, though the node a has higher degree, node e is on the shortest path of 4 nodes, from node i, f, g and h to node b, c and d, while node a only have three.

3.3.2.4 Eigenvector

The measurement of the eigenvector centrality or so called eigencentrality involves the non-negative adjacency matrix $A(a_{v,t})$, where each element $a_{v,t}$ equals 1 if node v and node t are neighbors and 0 if otherwise (Press, 2017). Eigencentrality scores correspond to the first value of the eigenvector of the adjacency matrix. They are defined as

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

where $M(v)$ is a collection of the neighbour of the node v and λ is a constant. If rewrite the equation and consider it in a vector form, the equation will be:

$$\mathbf{Ax} = \lambda \mathbf{x}$$

Generally speaking, nodes with higher eigencentrality tend to have connections with those nodes that have many connections.

3.3.3 Regression

After the LIWC analysis results for each individual subreddit and the corresponding centrality scores were available, a model was fitted to find the connections between them. We decided to take a linear regression model, whose output is a linear combination of all input features plus a constant bias value. In this case, subreddit properties were chosen to be inputs and centralities were outputs.

For this project, two different kinds of regression models were generated. The first took all input features as predictor variables. And for the second one, predictors were examined and carefully selected in a two-step optimizing method.

For the first step, we evaluated their multicollinearity with the variance inflation factor (VIF) of each predictor, which indicates whether one input feature has a strong linear relationship with the other features (Fox and Monette, 1992). Then we removed the predictor that had the largest VIF value. After that, the procedure was repeated until all VIF values of predictors were below 10. This threshold was suggested by Myers (1990) who claimed that a value of 10 is a good threshold that indicates whether we should concern about the collinearity.

For the second step, we ran a stepwise regression. As in R, we were required to specify a direction, and we chose to take 'both' direction which means at each step, we not only add a new predictor that has the best correlation with the centrality so far

but also remove a redundant predictor. We evaluated the changes in predictors with the Akaike information criterion (AIC) (Akaike, 1973), which is a statistical measurement of the fitness of the model and penalize models with more variables. A drop in the value of AIC suggests a better fit of the model. In each step, we calculated AIC value to make sure new collection of the predictors provides a better fit. Once the AIC value can not be decreased by adding or deleting predictors, the step regression stopped.

3.3.4 Statistical Analysis

Many different criteria were raised in out research to evaluate the results of regression models. P-values of coefficients were computed to see how many predictors have reached the significant level. Multiple R-squared and adjusted R-squared value for each model were evaluated, which indicates the percentage of changes in the outcome variables that can be explained by input variables. We also had the results of an analysis of variance (ANOVA) with F-ratio that measures the level of improvement of the prediction of the model compared with the inaccuracy of the model (Field et al., 2012).

3.3.5 *Igraph* R package

Igraph is a R-based package designed for network analysis available on its website[2]. In our research, *igraph* was utilized to analyze the social network of subreddits, conduct centrality measurements and fit the regression models. Provided with nodes (subreddits) and links (shared user between subreddits) generated already, this package created a network object and provided tools to analyze and visualize this network. Centrality measurements were computed with built-in functions without too much effort. Lastly, both linear regression and stepwise regression fitting can be implemented on *igraph*, as well as VIF calculation.

[2]http://igraph.org/

Chapter 4

Results and analysis

In this chapter, we present the analyzing results of the data from the online media platform Reddit. The aim of this project was to determine the relationship between the characteristic of subreddits and their popularity in the whole social network of subreddits. The popularity was measured with 4 different centrality measurements, consisting betweenness, closeness, eigenvector and degree centrality, each from different aspects to assess the popularity.

This chapter starts with a presentation of basic statistics of the social network. This includes distribution of total comment count of subreddits, ranking and distribution of daily comment numbers and daily different users who post comments of subreddits, along with the distribution of weights of links in the social network, which quantify the frequency of users visiting the linked subreddits and post comments. After that, we present the ranking of top subreddits in term of 4 different centrality measurements separately, then, followed by the correlation analysis of centrality measurements with the number of comments per day. Lastly, a linear regression model and a stepwise regression model were fitted with the same data. The aim of presenting two regression models is to compare the results and choose the best model that could describe links between predictors and outcome variables.

4.1 Social network analysis

4.1.1 Subreddits properties and statistics

As the first part, we begin with some visualization of statistical results derived from the social network. The social network was generated based on the process discussed in the

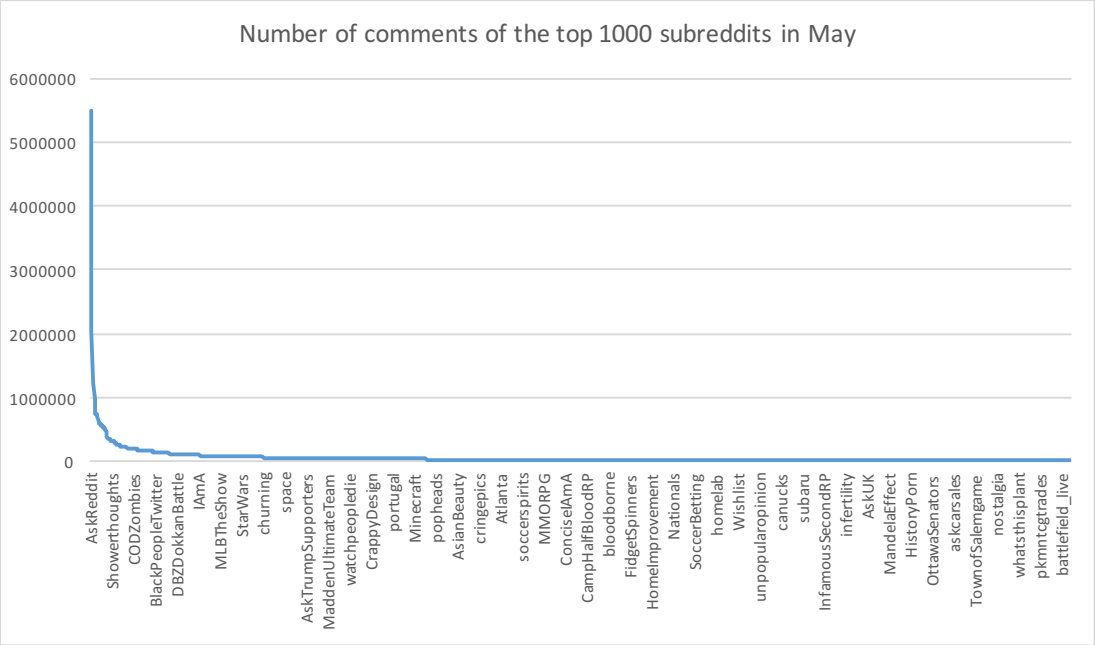


Figure 4.1: Distribution of comments among subreddits

methods chapter. Firstly, the distribution of total count for each subreddit is presented, along with the head and tail of the ranking of top 1,000 most frequently commented subreddits. With the distribution and ranking, we have the ability to identify those subreddits with highest comment numbers and examine the correlation with centrality measurements.

We present the distribution of the top 1,000 subreddits which has most comments along with the number of total comments. The distribution is presented in the figure 4.4. We can tell from the figure that this distribution seems to follow the power law, with the subreddit `AskReddit` gets first with more than 5,000,000 comments while the second one has a bit less than half of that number. Subreddits ranked 1,000 have about 10,000 comments a month, which are only 1 over 500 of the most visited ones. To make it clearer, the count of comments is displayed in table4.1, which not only including top subreddits but also some subreddits ranked 999 and 1,000.

In the following section, we present the results of centrality measurement distribution among subreddits and examine the correlation with the comment count we displayed here.

Table 4.1: The count of comments in top ranking and less famous subreddits

	name	count
1	AskReddit	5,477,134
2	Politics	2,030,218
3	nba	1,246,108
999	environment	11,865
1000	LiveFromNewYork	11,852

4.1.2 Centrality results

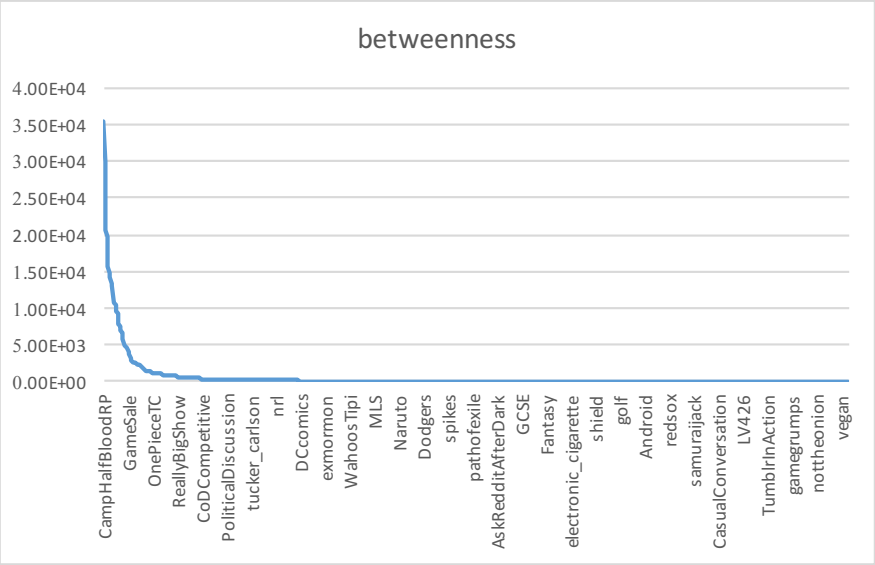
After the comment count derived, 4 different centrality measurements, namely betweenness, closeness, eigenvector and degree centrality, were calculated based on formulas we presented in the method chapter. Then we show the computation results of centrality measurements with the distribution of them. Meanwhile, the ranking of subreddits with highest centrality scores of 4 measurements are presented separately.

4.1.2.1 Betweenness

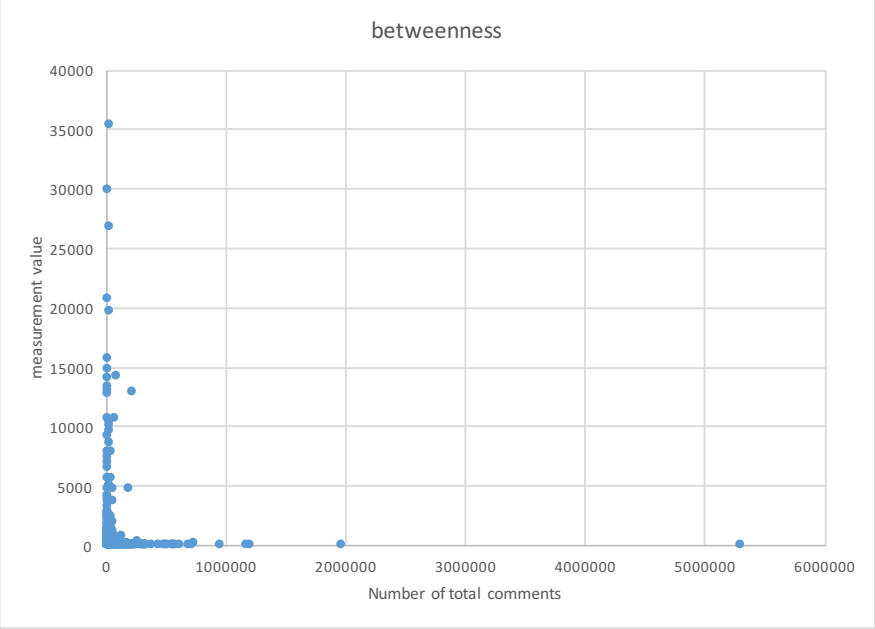
Table 4.2: Top 5 subreddits in betweenness centrality

	name	value	count
1	CampHalfBloodRP	35,325	39,635
2	ACTrade	29,846	19,424
3	SVExchange	26,834	41,415
4	InfamousSecondRP	20,670	16,896
5	soccerspirits	19,661	31,081

The distribution in the plot 4.2a shows a similar curvature with the count distribution, while the ranking is totally different. Moreover, the relation plot in figure 4.2b reveals that there are almost no direct links between comment count and the betweenness centrality. According to the ranking of the betweenness centrality, the top one got 39,635 comments in May which ranked about 500 in total comment count. Meanwhile, many of those subreddits who got the most comment count such as `AskReddit` scored 0 on betweenness centrality. We calculated the R-squared value between subreddit comment count and betweenness centrality and the result is 0.0008. Also the p-value



(a) Distribution of betweenness centrality scores of subreddits



(b) Relation between comment count and betweenness centrality

Figure 4.2: Analysis of betweenness centrality

for comment count is 0.356 which is not significant. With these result we can safely assume there are almost no correlation between them.

4.1.2.2 Closeness

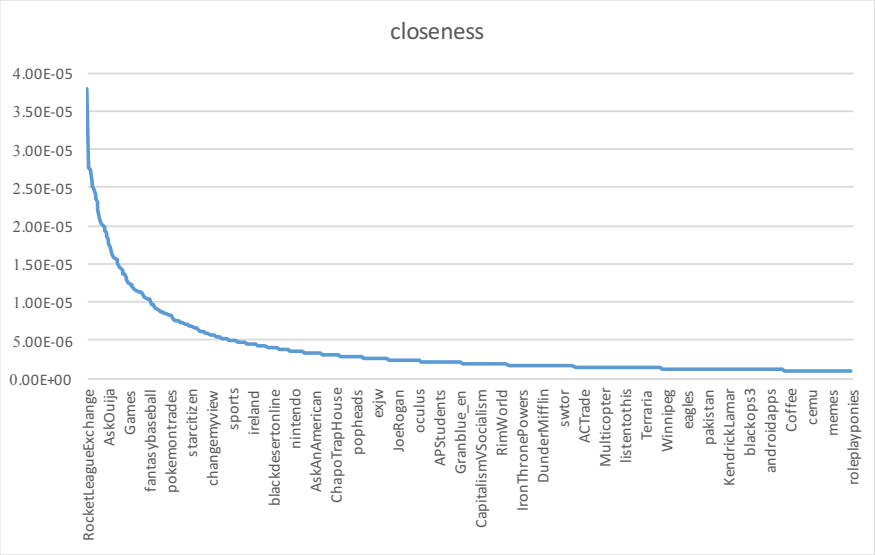
Table 4.3: Top 5 subreddits scored high at closeness centrality

	name	value	count
1	RocketLeagueExchange	3.78	754,433
2	conspiracy	2.75	316,987
3	nfl	2.74	357,950
4	anime	2.73	316,818
5	worldnews	2.73	988,527

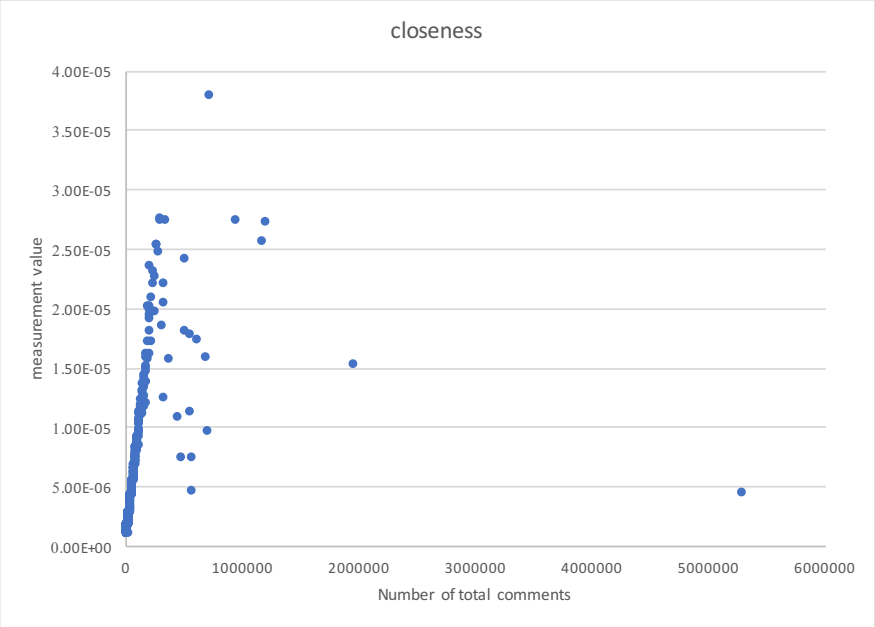
The curve in the figure 4.3a is smoother than that of betweenness. In the table, the subreddit `RocketLeagueExchange` gets first with a relatively big advance over the second. Subreddits with high closeness centrality are defined to have shorter average path to all other subreddits, which means users who post comments on subreddit like `RocketLeagueExchange` are more likely to have connections with other subreddits in a much broader range. Besides, these subreddits in the top ranking list for closeness centrality are all with relatively high comment count, even the subreddit `anime` whose score is the lowest count among all subreddits that reached top 5 in closeness ranking has 316,818 comments in total. The plot 4.3b shows a roughly linear correlation between closeness centrality and comment count. Meanwhile, the R-squared value between them is 0.1674, which is much higher than the one for betweenness centrality. Moreover, the p-value for comment count is lower than 0.0001 which proves it to be a very significant predictor.

4.1.2.3 Eigenvector

The distribution in figure 4.4a also has a similar shape compared with comment count. In the figure 4.4b we see a relatively clear linear relationship between eigenvector centrality and comment count, and the R-squared value is 0.4343 which is the highest so far, along with the lowest p-value that shows its significance. The idea behind eigenvector centrality is that for each subreddit, the link with a high score subreddit contributes more compared with the link with another relatively low score subreddit. Like Google's PageRank (Page et al., 1998), eigenvector centrality assigns a higher

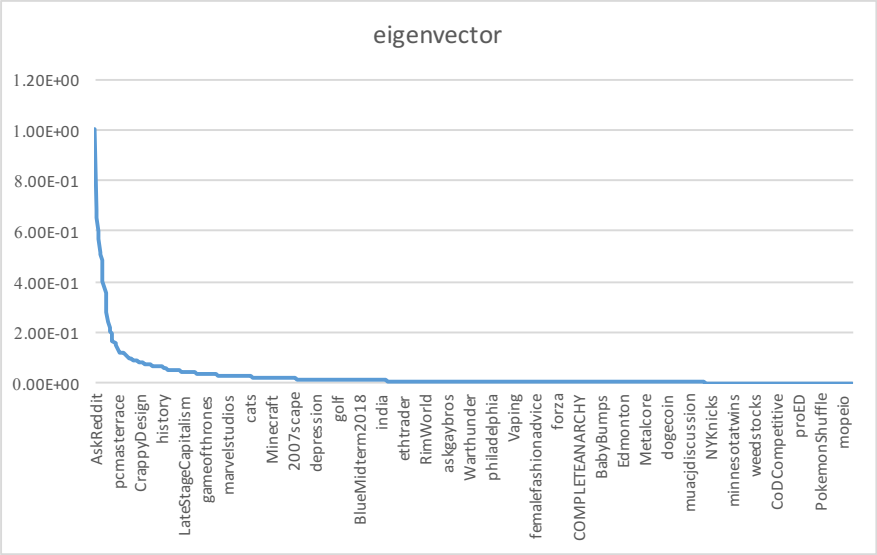


(a) Distribution of closeness centrality scores of subreddits

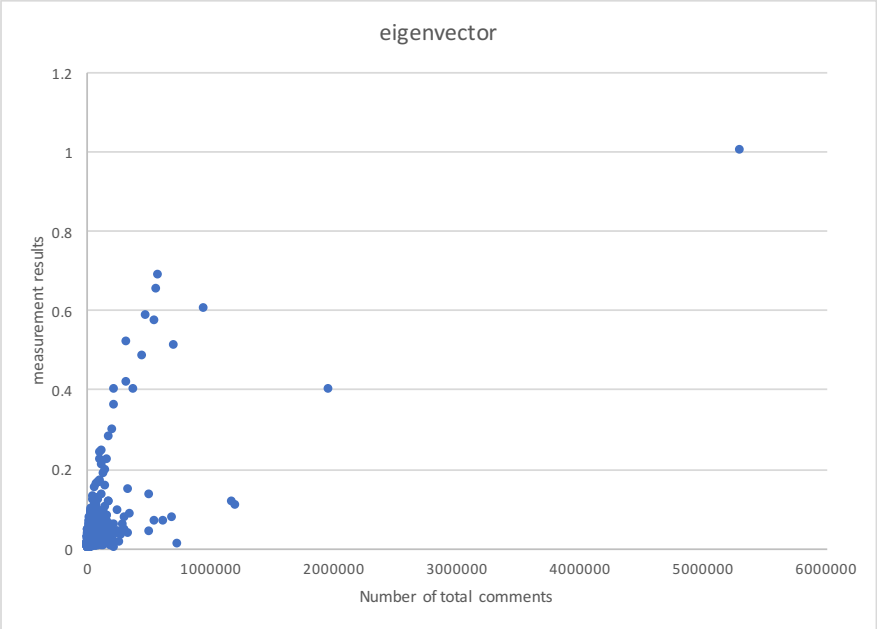


(b) Relation between comment count and closeness centrality

Figure 4.3: Analysis of closeness centrality



(a) Distribution of eigenvector centrality scores of subreddits



(b) Relation between comment count and eigenvector centrality

Figure 4.4: Analysis of eigenvector centrality

Table 4.4: Top 5 subreddits scored high at eigenvector centrality

	name	value	count
1	AskReddit	1.00	5,477,134
2	pics	0.69	600,409
3	funny	0.65	586,451
4	worldnews	0.60	988,527
5	todayilearned	0.59	496,050

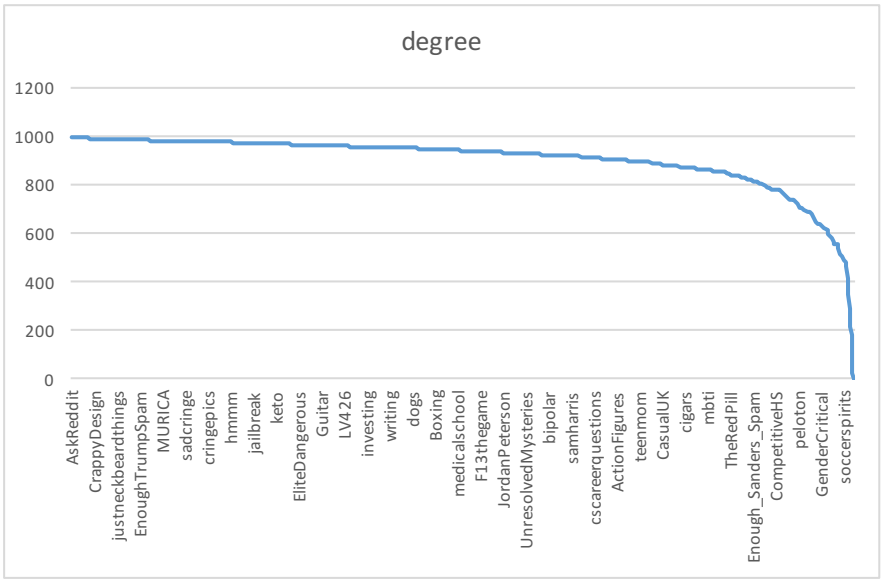
value to subreddit if it has been visited the most times in the network. This means a subreddit does not need to have top comments but its users are required to visit other popular subreddits to build links between them to rank to in terms of eigenvector centrality. The results identified subreddits with users visiting other top sites at the same time.

4.1.2.4 Degree

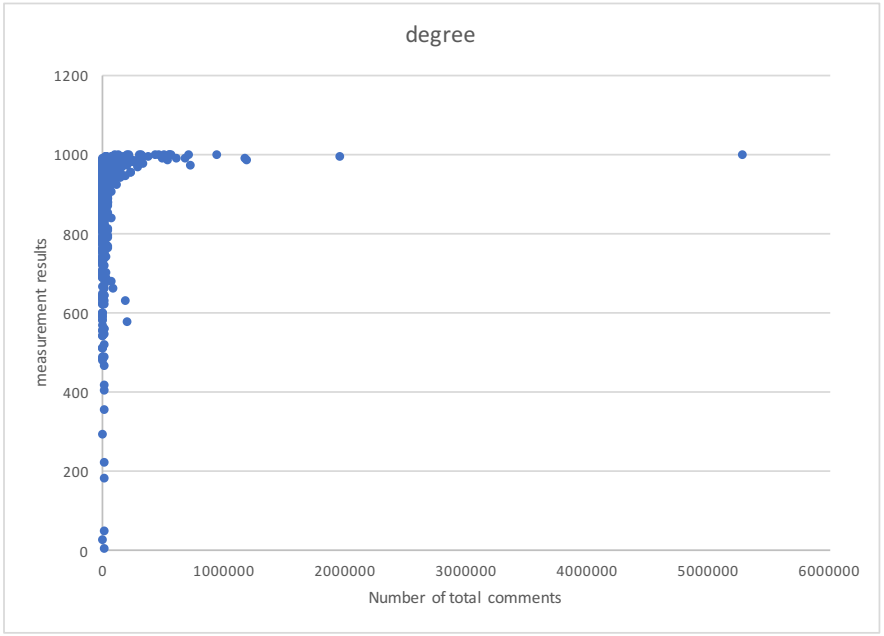
Table 4.5: Top 5 subreddits scored high at degree centrality

	name	value	count
1	AskReddit	996	5,477,134
2	pics	996	600,409
3	funny	995	586,451
4	todayilearned	994	496,050
5	videos	994	575,095

The results for degree centrality is relatively straightforward. Those subreddits whose degree centrality are high have direct connections with almost all other subreddits. The full ranking of degree centrality measurements show that about half of total subreddits have degree centrality larger than 900, so as we can see in the figure 4.5a, the distribution of degree centrality scores had a distinct shape compared with the distribution of count in comments. Considering we have 1,000 subreddits included in the network, and half of the nodes have connections with at least 90% of all the rest subreddits, this social network has a high density. The correlation from the figure 4.5b is not obvious, the R-squared value is 0.0165 while the p-value shows that the comment count is a significant predictor. These facts imply that though comment count has



(a) Distribution of degree centrality scores of subreddits



(b) Relation between comment count and degree centrality

Figure 4.5: Analysis of degree centrality

strong relationship with degree centrality, the variance in degree centrality can not be explained properly with comment count as the only predictor.

4.1.3 Regression with all features included

With the comment count, the variance of some centrality measurements can be explained properly such as eigenvector centrality, however, for the rest centrality measurements we still have not found the factors that have a high correlation with them. In order to find links between centralities and other subreddit properties such as LIWC results and daily comment and user count, a linear regression model was fitted, with all 95 different predictor variables included. These predictors were generated from two sources, with 93 of them came from the content of comments analysis based on LIWC program and the rest 2 of them were the average number of comments per day and the average number of different users per day for each subreddit. Since we got four different centrality measurements with totally different results, 4 models were fitted with respect to each centrality measurement separately. After fitting, some statistics were provided like residual, R-squared value and F-ratio of each model to evaluate their quality.

Table 4.6: Regression results with all predictors

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

The linear model for eigenvector centrality provides great results, with the smallest residual standard error and highest R-squared as well as F-ratio among all 4 models. Adjusted R-squared value reaches 0.80 which means around 80% of the variance in eigenvector centrality can be explained with the variance in 95 predictors of the model. Though the rest model was not as good as the one for eigenvector centrality, they all have significant improvement in terms of the R-squared value derived in the last section, when we were finding the correlation between comment count and the centrality. The last but not the least to mention is that all 4 different models were significant at

Table 4.7: Significant predictors for linear regression model for betweenness centrality

	Estimate	Std. Error	t value	Pr(>|t|)
Clout	0.7559	0.3396	2.23	0.0263*
Dic	-1.5110	0.7471	-2.02	0.0434*
prep	0.9341	0.3723	2.51	0.0123*
auxverb	0.6806	0.3581	1.90	0.0577.
adverb	0.4014	0.1935	2.07	0.0383*
negate	0.2931	0.1422	2.06	0.0396*
adj	-0.1776	0.1051	-1.69	0.0914.
interrog	-0.1818	0.0881	-2.06	0.0394*
number	0.1493	0.0610	2.45	0.0146*
social	-0.7276	0.4299	-1.69	0.0909.
family	-0.1030	0.0460	-2.24	0.0254*
friend	-0.1866	0.0578	-3.23	0.0013**
cause	0.1922	0.1115	1.72	0.0850.
discrep	0.1958	0.1018	1.92	0.0548.
differ	0.4420	0.2258	1.96	0.0506.
achieve	-0.1695	0.0722	-2.35	0.0190*
focuspresent	0.4843	0.2477	1.96	0.0509.
focusfuture	0.1071	0.0650	1.65	0.0997.
death	-0.0997	0.0373	-2.67	0.0077**
filler	0.0782	0.0327	2.39	0.0171*
.: $0.05 < p < 0.1$, *: $0.01 < p < 0.05$, **: $0.001 < p < 0.01$, ***: $0 < p < 0.001$				

$p < .001$, which tells us that all models were significantly better than if we use mean of input variables.

Also we provide down below all significant predictors with p-value lower than 0.1 in table 4.7, 4.8, 4.9, and 4.10.

Considering we have 95 different predictors, the number of predictors that considered significant is relatively low, which means more than half of the predictors are not pertinent to measurement of centralities. In the following section, the results of a regression model with selected predictors are examined.

4.1.4 Regression with selected features

The second regression model was fitted with predictors filtered in two steps. Firstly, variables with high multicollinearity were removed with the reason that high multi-

Table 4.8: Significant predictors for linear regression model for closeness centrality

	Estimate	Std. Error	t value	Pr(>|t|)
AvgAuthor	-0.5925	0.1346	-4.40	0.0000***
AvgPost	0.8864	0.1334	6.65	0.0000***
Authentic	-0.4475	0.2041	-2.19	0.0286*
interrog	0.1601	0.0930	1.72	0.0854.
friend	0.1457	0.0610	2.39	0.0172*
male	-0.2322	0.1388	-1.67	0.0946.
cogproc	-1.1268	0.5382	-2.09	0.0366*
insight	0.3276	0.1482	2.21	0.0273*
cause	0.2912	0.1177	2.47	0.0135*
certain	0.2771	0.1060	2.61	0.0091**
health	0.2118	0.0992	2.13	0.0331*
focusfuture	0.1742	0.0686	2.54	0.0112*
.: $0.05 < p < 0.1$, *: $0.01 < p < 0.05$, **: $0.001 < p < 0.01$, ***: $0 < p < 0.001$				

Table 4.9: Significant predictors for linear regression model for eigenvector centrality

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000	0.0140	0.00	1.0000
AvgAuthor	2.5331	0.0704	35.99	0.0000***
AvgPost	-1.8504	0.0697	-26.53	0.0000***
ipron	3.7439	2.1780	1.72	0.0860.
article	0.2101	0.1270	1.65	0.0985.
cogproc	-0.5321	0.2814	-1.89	0.0590.
cause	0.1467	0.0615	2.38	0.0173*
achieve	-0.0670	0.0398	-1.68	0.0925.
motion	0.1529	0.0894	1.71	0.0875.
.: $0.05 < p < 0.1$, *: $0.01 < p < 0.05$, **: $0.001 < p < 0.01$, ***: $0 < p < 0.001$				

Table 4.10: Significant predictors for linear regression model for degree centrality

	Estimate	Std. Error	t value	Pr(>|t|)
WC	-0.1024	0.0346	-2.96	0.0031**
Clout	-0.8534	0.2920	-2.92	0.0036**
Authentic	-0.6014	0.1663	-3.62	0.0003***
Dic	1.4370	0.6424	2.24	0.0255*
i	-7.6048	3.7244	-2.04	0.0415*
we	-2.0616	1.0734	-1.92	0.0551.
you	-5.7086	3.1421	-1.82	0.0696.
shehe	-6.9753	3.4063	-2.05	0.0409*
they	-2.3044	1.2253	-1.88	0.0603.
negate	-0.3068	0.1223	-2.51	0.0123*
adj	0.1589	0.0904	1.76	0.0790.
compare	0.1821	0.0759	2.40	0.0166*
quant	-0.2057	0.0691	-2.98	0.0030**
posemo	-1.9614	1.0165	-1.93	0.0540.
negemo	-1.7166	0.9067	-1.89	0.0587.
anger	0.3756	0.1343	2.80	0.0053**
sad	0.1113	0.0423	2.63	0.0086**
friend	0.0896	0.0497	1.80	0.0721.
discrep	-0.2369	0.0875	-2.71	0.0069**
certain	0.2009	0.0864	2.33	0.0203*
differ	-0.4251	0.1942	-2.19	0.0288*
affiliation	0.3344	0.1395	2.40	0.0167*
risk	0.0988	0.0442	2.24	0.0254*
motion	0.2608	0.1393	1.87	0.0614.
swear	-0.3818	0.1669	-2.29	0.0224*
netspeak	-0.6122	0.3350	-1.83	0.0679.
assent	-0.1131	0.0419	-2.70	0.0071**
nonflu	-0.1191	0.0549	-2.17	0.0305*

.: $0.05 < p < 0.1$, *: $0.01 < p < 0.05$, **: $0.001 < p < 0.01$, ***: $0 < p < 0.001$

collinearity will make variables largely depend on others and will lead to high resonance even with a small change in variables. The removal process was controlled by the calculation of VIF where predictor with the highest VIF was removed each turn. This process stopped until scoring of VIF for all predictors were lower than 10.

Since the removal did not involve the outcome variables, the 4 different model provide the same list of predictors that to be removed. The number of removed predictors is 29 and the full list is in the table 4.11

Table 4.11: All predictors that removed due to VIF higher than 10

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

To further reduce the degree of predictors, we ran a both-direction stepwise regression, which means for each step we computed the AIC score for adding or removing all candidate predictors and chose the movement that with the highest score. The process of regression stopped when no adding or removing predictors could improve the result of the current regression model. In the following table, we will present all remaining predictor variables that still exists and associated with their statistical significance.

We also provide the statistics of regression results in table4.12. After optimization with stepwise algorithm, we have about 30 different predictors left. From the results,

Table 4.12: Regression results with selected predictors

	Betweenness	Closeness	Eigenvector	Degree
Residual standard error	0.81	0.85	0.69	0.72
R-squared	0.36	0.29	0.54	0.50
Adjusted R-squared	0.34	0.27	0.52	0.48
F-ratio	18.81	15.02	38.91	23.47
Degree of freedom	970	972	970	957
P-value		$< 0.0001 * **$		

we can tell that all models except the one for eigenvector centrality have improved in terms of the f-ratio. The model for eigenvector centrality degenerates in all way round compared with the model with all predictors, with lower R-squared, adjusted R-squared and f-ratio and higher residual error. By contrast, we have considerable

growth in terms of f-ratio for other models, although the error remains and R-squared values decreased. To be more specific, the f-ratio for betweenness centrality model has a 12.08 increment which is almost 200% growth compared with the model with all 95 predictors. Meanwhile, we have also seen an improvement twice as the original figure in the model for closeness centrality, whose f-ratio increases from 5.08 to 15.02. Changes in f-ratio of the model for degree centrality was also significant, with an 88% improvement compared with the model with all predictors.

We provide as follows in the table 4.13, 4.14, 4.15, and 4.16 all the significant(p¡0.1) predictors as well as their coefficient, t-value and significant.

Table 4.13: Significant predictors for stepwise regression model for betweenness centrality

	Estimate	Std. Error	t value	Pr(>|t|)
WC	-0.0630	0.0327	-1.93	0.0544.
i	0.1433	0.0386	3.72	0.0002***
you	0.1398	0.0438	3.19	0.0015**
ipron	-0.2160	0.0507	-4.26	0.0000***
negate	0.1149	0.0431	2.67	0.0078**
compare	-0.2290	0.0528	-4.33	0.0000***
number	0.1311	0.0349	3.76	0.0002***
posemo	0.2174	0.0463	4.69	0.0000***
anger	0.1604	0.0518	3.10	0.0020**
family	-0.1785	0.0386	-4.62	0.0000***
friend	-0.2707	0.0425	-6.37	0.0000***
female	0.1789	0.0368	4.86	0.0000***
male	0.1301	0.0429	3.03	0.0025**
certain	-0.2451	0.0566	-4.33	0.0000***
affiliation	0.0716	0.0400	1.79	0.0733.
achieve	-0.1442	0.0470	-3.07	0.0022**
reward	-0.1638	0.0529	-3.10	0.0020**
focusfuture	0.1369	0.0413	3.32	0.0009***
space	0.1824	0.0364	5.02	0.0000***
death	-0.1039	0.0313	-3.32	0.0009***
swear	-0.1929	0.0490	-3.94	0.0001***
netspeak	0.0677	0.0360	1.88	0.0606.
filler	0.0815	0.0293	2.78	0.0055**
Comma	0.0797	0.0361	2.21	0.0274*
Colon	0.0809	0.0366	2.21	0.0271*
SemiC	0.1765	0.0273	6.46	0.0000***
Quote	0.1894	0.0356	5.32	0.0000***

.: $0.05 < p < 0.1$, *: $0.01 < p < 0.05$, **: $0.001 < p < 0.01$, ***: $0 < p < 0.001$

Table 4.14: Significant predictors for stepwise regression model for closeness centrality

	Estimate	Std. Error	t value	Pr(>|t|)
AvgPost	0.3374	0.0279	12.09	0.0000***
i	-0.0740	0.0384	-1.93	0.0539.
we	-0.1801	0.0473	-3.81	0.0002***
ipron	0.1693	0.0644	2.63	0.0087**
negate	-0.1172	0.0405	-2.89	0.0039**
interrog	0.1591	0.0635	2.51	0.0124*
anx	-0.0625	0.0346	-1.81	0.0712.
friend	0.1163	0.0420	2.77	0.0057**
female	-0.0789	0.0363	-2.17	0.0301*
male	-0.0719	0.0413	-1.74	0.0821.
tentat	-0.3133	0.0509	-6.15	0.0000***
see	0.0999	0.0365	2.74	0.0063**
health	0.0838	0.0337	2.49	0.0129*
affiliation	0.1587	0.0486	3.27	0.0011**
power	0.1316	0.0403	3.27	0.0011**
focusfuture	0.1731	0.0420	4.12	0.0000***
motion	0.0783	0.0432	1.81	0.0698.
space	-0.1967	0.0495	-3.97	0.0001***
money	0.0767	0.0324	2.36	0.0182*
assent	-0.0931	0.0383	-2.43	0.0152*
Colon	0.1367	0.0742	1.84	0.0657.
SemiC	0.1585	0.0292	5.43	0.0000***
QMark	-0.1306	0.0349	-3.75	0.0002***
Dash	-0.1866	0.0739	-2.52	0.0118*
Quote	0.1353	0.0361	3.75	0.0002***
.: $0.05 < p < 0.1$, *: $0.01 < p < 0.05$, **: $0.001 < p < 0.01$, ***: $0 < p < 0.001$				

Table 4.15: Significant predictors for stepwise regression model for eigenvector centrality

	Estimate	Std. Error	t value	Pr(>|t|)
AvgPost	0.6074	0.0226	26.82	0.0000***
WC	-0.0528	0.0278	-1.90	0.0581.
we	0.0883	0.0446	1.98	0.0481*
you	0.0949	0.0350	2.71	0.0068**
ipron	0.1888	0.0598	3.16	0.0016**
negate	-0.1699	0.0411	-4.13	0.0000***
interrog	-0.0842	0.0507	-1.66	0.0969.
family	0.0961	0.0303	3.17	0.0016**
friend	0.1013	0.0375	2.70	0.0070**
male	0.0715	0.0365	1.96	0.0501.
cause	0.0901	0.0434	2.08	0.0381*
tentat	-0.1567	0.0451	-3.47	0.0005***
see	0.0674	0.0300	2.25	0.0249*
health	0.0717	0.0263	2.73	0.0065**
affiliation	-0.1449	0.0543	-2.67	0.0078**
achieve	-0.0886	0.0370	-2.40	0.0168*
power	0.0708	0.0379	1.87	0.0617.
focuspast	0.0629	0.0337	1.87	0.0621.
focusfuture	-0.0823	0.0330	-2.50	0.0128*
work	0.1241	0.0303	4.09	0.0000***
leisure	0.1463	0.0351	4.17	0.0000***
relig	0.0794	0.0350	2.27	0.0237.
swear	-0.0962	0.0321	-2.99	0.0028**
Colon	0.1625	0.0533	3.05	0.0024**
Dash	-0.2015	0.0552	-3.65	0.0003***

.: $0.05 < p < 0.1$, *: $0.01 < p < 0.05$, **: $0.001 < p < 0.01$, ***: $0 < p < 0.001$

Table 4.16: Significant predictors for stepwise regression model for degree centrality

	Estimate	Std. Error	t value	Pr(>|t|)
AvgPost	0.0586	0.0235	2.49	0.0129*
WC	-0.1327	0.0300	-4.42	0.0000***
WPS	-0.0495	0.0277	-1.79	0.0744.
Sixltr	-0.0718	0.0383	-1.87	0.0612.
i	-0.1418	0.0463	-3.06	0.0022**
we	-0.1274	0.0427	-2.98	0.0030**
ipron	0.2674	0.0617	4.33	0.0000***
article	0.0888	0.0494	1.80	0.0727.
negate	-0.2473	0.0481	-5.14	0.0000***
compare	0.3000	0.0544	5.52	0.0000***
number	-0.1138	0.0302	-3.77	0.0002***
quant	-0.2078	0.0452	-4.60	0.0000***
posemo	-0.2390	0.0451	-5.30	0.0000***
family	0.1284	0.0353	3.63	0.0003***
friend	0.1425	0.0381	3.74	0.0002***
female	-0.1932	0.0352	-5.49	0.0000***
male	-0.0760	0.0392	-1.94	0.0529.
cause	0.0836	0.0464	1.80	0.0718.
discrep	-0.0998	0.0466	-2.14	0.0325*
tentat	-0.2445	0.0502	-4.87	0.0000***
certain	0.2874	0.0528	5.45	0.0000***
see	0.0686	0.0351	1.95	0.0512.
feel	-0.0747	0.0435	-1.72	0.0860.
ingest	0.0555	0.0272	2.04	0.0416*
affiliation	0.1250	0.0510	2.45	0.0145*
achieve	0.0747	0.0443	1.68	0.0924.
power	-0.1656	0.0423	-3.92	0.0001***
risk	0.0850	0.0313	2.72	0.0067**
focuspast	0.0962	0.0428	2.25	0.0249*
space	-0.1046	0.0509	-2.06	0.0400*
time	-0.1112	0.0454	-2.45	0.0146*
work	0.0747	0.0336	2.23	0.0261*
home	0.0520	0.0303	1.72	0.0861.
relig	-0.0805	0.0423	-1.90	0.0571.
assent	-0.0678	0.0322	-2.11	0.0355*
Period	-0.0789	0.0263	-3.00	0.0028**
QMark	-0.1292	0.0314	-4.12	0.0000***
Dash	-0.0744	0.0332	-2.24	0.0251*
Apostro	0.0697	0.0376	1.85	0.0642.
Parenth	-0.1377	0.0263	-5.24	0.0000***

.: $0.05 < p < 0.1$, *: $0.01 < p < 0.05$, **: $0.001 < p < 0.01$, ***: $0 < p < 0.001$

Chapter 5

Discussion

In this chapter, we attempt to answer the research questions on the basis of the analysis in the results part. These research questions are -(1) To what extent can the popularity of subreddits be represented by comment count? -(2) What factors of subreddits can describe its popularity? Answers are provided with results of correlation analysis as well as the comparison with previous work.

5.1 RQ1: To what extent can the popularity of subreddits be represented by comment count?

The analysis of the correlation between comment count and centrality measurements indicates that the subreddits with the most count are no longer among the most popular when the standard of popularity changes. In many types of research, comment count, page views, different user count and the number of a video being watched are all matrices to measure the popularity of websites (Szabo and Huberman, 2010; Chu et al., 2004). To sum up, these matrices are all favor objects with more clicks and viewers. However, when it comes to online social communities, these matrices are not always useful to reflect the popularity. In the network structure, we can not deem one subreddit popular if its users post heavily within and seldom visit outside.

When analyzing the correlation between comment count and popularity of subreddits, we adopted 4 different metrics, namely betweenness, closeness, eigenvector and degree centrality. These centrality measurements focused on different aspects on describing the popularity or importance of subreddits in terms of the whole network as discussed in the methods part. According to the results, the observed p-value was

0.356 which is not significant and the R-squared value was extremely low, which was down to 0.0008, we can assume that there is hardly any correlation between count and betweenness centrality. Comment count is significant in all other models with p-value lower than 0.0001, nevertheless, the R-squared value is relatively low especially for degree model. This fact provides an indication that more factors should be included to better explain the change in centrality models.

5.2 RQ2: What factors of subreddits can describe its popularity?

Though comment count proved to be a significant predictor in all models except that for betweenness centrality, it still can not provide sufficient explanation to the changes in centralities with such low R-squared value. Consequently, we introduced LIWC metrics as extra predictors with the purpose of increasing the explanatory power of the model. The analysis of the new model in results section implied the answer to this research question.

The linear regression model in the last chapter has the edge over stepwise regression model in terms of R-squared value, which indicates that model with all LIWC outputs and comment count as predictors can better contribute to the changes in the variance of response. However, with few significant variables and extreme high multicollinearity, the linear model can not specify all factors that have potential links with the formation of popularity. On the other hand, the stepwise regression model achieved a great improvement in terms of f-ratio and showed significant in more than 90% of remaining variables with slightly sacrifice of the R-squared value. The detailed lists of factors significant with corresponding centrality measurements are all available in the results chapter. Following we provide a brief interpretation of these results by sequence.

Firstly, the significant variables for betweenness centrality, as we expected, did not include comment count, which proved to have little contribution to explaining changes in betweenness centrality. Meanwhile, the model achieved 0.34 in adjusted R-squared value and have the most significant variables with the p-value lower than 0.0001. This result provides an indication that around 34% of changes in betweenness centrality can be explained by the variance in the result of content analysis alone.

Similarly, the significant variables for the degree centrality model were purely ex-

tracted from the outcome of LIWC analysis without comment count. Moreover, the adjusted R-squared value for degree model has reached 0.48 which is the second largest among 4 models.

The eigenvector centrality model was the only one whose adjusted R-squared value is higher than 0.5. Unlike previous models, the list of significant predictors for this model contains the comment count and the coefficient of which was the largest among others. Compared with the R-squared value of the eigenvector model in the last section whose with the only predictor was comment count, the extra content features provide minute improvements to the new model.

The closeness centrality model got the lowest adjusted R-squared score. With the smallest number of variables that significant, this model can hardly find set of appropriate factors to describe.

Chapter 6

Conclusion

This project began with the aim of examining factors that can describe the popularity of subreddits (communities on social online platform Reddit). As an initial step, we got access to an online data dump of archived Reddit comments. From the data dump, we collected information about comments in five different fields, namely content, author, time of creation, belonged subreddit and vote. Using information of author and belonged subreddits, we constructed a social network with subreddits as nodes and shared author as edges. The mechanism of weighted edges was adopted to represent the strength of ties between subreddits. Higher the number of the shared author was, stronger the edge connected these subreddits will be. Considering the processing power of a personal computer, only top 1,000 subreddits with most comments were included in our analysis. To remove noise and get rid of the long tail in the distribution of edge weights, all links whose weight were lower than 2 were deleted from the network.

In the generated social network, we implemented social network analysis method to compute the centrality measurements. For centrality, we mean the metric measuring how central one particular node is in the whole network. This project provides 4 different kinds of centrality measurements as the indicators of popularity, these are betweenness, closeness, eigenvector and degree centrality. In the case of factors that to be examined, comment count and content analysis results were selected as representations of subreddit features. Comment count was a plain count of comments of each subreddit while content analysis involved LIWC program to provide statistical results to the comment aggregation for each subreddit. Comments to be analyzed were chosen to have scores ranked in the top 1,000 in each subreddit. The outcomes of LIWC analysis to the comments selected were formed from 90 different variables, each with

unique statistical meaning.

With subreddit features as predictors and centrality to be the response, we fitted regression models to detect the relations between them. Firstly, we examined the correlation of comment count solely and 4 different centrality measures. After that, we combined comment count and LIWC outcomes as a new set of predictors and evaluated its links to the centralities. Moreover, we fitted a stepwise regression model with selected predictors and compared the results with the previous linear model.

The comparison provided insights to answer our research question. Regarding the first question, comment count proved to be a significant predictor except for analyzing betweenness centrality, while comment count can not explain much of the variance of the centrality measurements. Then it comes to the second question, where we can safely assume that with comment count and LIWC outcomes, the model provided a reasonable explanation for the popularity of subreddits in terms of both eigenvector and degree centrality, though they were not so good in explaining variances in betweenness and closeness centrality measurements.

6.1 Future work

This project has already identified some factors that make the contribution to describe the popularity of subreddits. However, further improvements can be made in several directions.

6.1.1 Data in a larger time span

The quantity of data could be extended in further analysis. This project analyzed the top 1000 subreddits in a 1-month span, which only provides a transient vision of the Reddit. To evaluate the long-term consistent status of subreddits, more data in a longer span should be included in our analysis. Meanwhile, those subreddits were ignored due to relatively low total count could have a potential influence on the analysis results, since they still account for around 20% of total traffic in the case of comments in May. However, these requirements call for much larger computation resources for the reason that any increment in the size of nodes can exponentially enlarge the time and memory required to conduct social network analysis.

6.1.2 Improvements in content analysis method

LIWC program does provide statistical results of comments in fine-grained categories. However, the analysis ignores the context of comments and treats sentence like the bag of words without sequence, hence it still superficial and lack insightful observation. Higher level content analysis softwares, namely *Profiler Plus* and *Visual Text* as discussed in the review (Lowe, 2012), not only extract statistics but also provide cognitive information inherited in the content. More comprehensive evaluation can be conducted to the comments in the subreddits with the help of advanced content analysis tools.

6.1.3 More features to be examined

In our analysis, we adopted comment count and features of content as the factors to describe the popularity of subreddits. These examined factors are one-sided which take comments into consideration and ignore other possible alternatives such as the number of different posts per day, the average number of comments for each post in the discussion area, the median absolute deviation of comments per day and the growth rate of posts and subscribers for individual subreddit. These factors are all worth exploring and could possibly provide a better explanation for the formation of the popularity of subreddits.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory*, pages 267–281.

Bavelas, A. (1950). Communication Patterns in TaskOriented Groups. *The Journal of the Acoustical Society of America*, 22(6):725–730.

Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., and Kraut, R. E. (2004). Using social psychology to motivate contributions to online communities. *Computer Supported Cooperative Work*, pages 212–221.

Bergstrom, K. (2011). "Don't feed the troll": Shutting down debate about community expectations on Reddit.com. *First Monday*, 16(8).

Boyd, D. M. and Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

Chow, A. and Hong, J. (2016). Topical Classification and Divergence on Reddit. pages 1–9.

Chu, K. K., Shen, T. C., and Hsia, Y. T. (2004). Measuring website popularity and raising designers' effort. In *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, volume 5, pages 4095–4099.

Cillessen, A. H. N. and Rose, A. J. (2005). Understanding popularity in the peer system. *Current Directions in Psychological Science*, 14(2):102–105.

Coppersmith, G., Dredze, M., and Harman, C. (2014). Quantifying Mental Health Signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

Costenbader, E. and Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283–307.

Field, A., Miles, J., and Field, Z. (2012). *Discovering Statistics Using R*, volume 58.

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

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness Author (s): Linton C . Freeman Published by : American Sociological Association Stable URL : http://www.jstor.org/stable/3033543 Accessed : 18-04-2016 12 : 00 UTC Your use of the JSTOR archive indicat. *Sociometry*, 40(1):35–41.

Hammersley, M. and Atkinson, P. (2007). *Ethnography: Principles in practice*. Routledge.

Hiltz, S. R. (1985). *Online communities: A case study of the office of the future*, volume 2. Intellect Books.

Kozinets, R. V. (2002). The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities. *Journal of Marketing Research*, 39(1):61–72.

Laat, M. D., 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.

Lowe, W. (2012). Software for Content Analysis A Review. *Journal of Conflict Resolution*, 3:1–18.

Myers, R. (1990). *Classical and Modern Regression with Application*, volume Second.

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

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251.

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.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web Internet And Web Information Systems*, 54(1999-66):1–17.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. *UT Faculty/Researcher Works*, (SEPTEMBER 2015):1–22.

Porter, C. E. (2004). A Typology of Virtual Communities: A Multi-Disciplinary Foundation for Future Research. *Journal of Computer-Mediated Communication*, 10(1):00–00.

Preece, J. and Maloney-Krichmar, D. (2005). Online Communities: Design, Theory, and Practice. *Journal of Computer-Mediated Communication*, 10(4):0000.

Press, C. (2017). Power and Centrality : A Family of Measures Author (s): Phillip Bonacich Source : American Journal of Sociology , Vol . 92 , No . 5 (Mar ., 1987), pp . 1170-1182 Published by : The University of Chicago Press Stable URL : http://www.jstor.org/stable/2. 92(5):1170–1182.

Reddit (2015). Happy 10th birthday to us! Celebrating the best of 10 years of Reddit. *Blog.Reddit*, pages 1–5.

Rheingold, H. (2000). *The virtual community: Homesteading on the electronic frontier*. MIT press.

Ruhnau, B. (2000). Eigenvector-centrality a node-centrality? *Social Networks*, 22:357–365.

Scott, J. (2000). *Social Network Analysis: a Handbook*, volume 22.

Singer, P., Ferrara, E., Kooti, F., Strohmaier, M., and Lerman, K. (2016). Evidence of online performance deterioration in user sessions on Reddit. *PLoS ONE*, 11(8):1–16.

Smith, C. (2016). 60 Amazing Reddit Statistics. *Dmr*.

Steinbauer, T. (2012). Information and Social Analysis of Reddit. *Retrieved from TROYSTEINBAUER@ CS. UCSB. EDU*.

Stoddard, G. (2014). Popularity Dynamics and Intrinsic Quality in Reddit and Hacker News. *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, pages 416–425.

Szabo, G. and Huberman, B. A. (2010). Predicting the Popularity of Online Content. *Commun. ACM*, 53(8):8088.

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.

Vaquero, L. M. and Cebrian, M. (2013). The rich club phenomenon in the classroom.

Weninger, T., Zhu, X., and Han, J. (2013). An exploration of discussion threads in social news sites: a case study of the Reddit community. *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference*, 579(2):579–583.

Wilson, P. (1991). *Computer supported cooperative work:: An introduction*. Springer Science & Business Media.

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):147155.

Zhang, J., Tan, W., John, A., Foster, I., and Madduri, R. (2011). Recommend-as-you-go: A novel approach supporting services-oriented scientific workflow reuse. In *Proceedings - 2011 IEEE International Conference on Services Computing, SCC 2011*, pages 48–55.