UNDERSTANDING PROMINENCE ON MASSIVE AND ANONYMOUS SOCIAL NEWS SITES

By

Dan Kimball

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: COMPUTER SCIENCE

Approved:

_____

Sibel Adalı, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

March 2013
(For Graduation May 2013)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Many algorithms have been proposed to measure prominence of individuals in social networks. However, these algorithms are highly dependent on the underlying mechanisms for judging how individuals gain prominence in a given network. While there is a great deal of study in measuring prominence in networks of collaborations such as in academic networks and movie industry, there is relatively little work in understanding prominence in new and emerging social media sites. In particular, while there is some work in understanding the importance of friends and followers in Twitter, vote behavior in sites like Epinions, Digg and Slashdot, there is almost no study to this date on sites that encourage anonymous interactions like Reddit.

In this thesis, we explore the properties of interactions on Reddit. Reddit is different in that content is given either an "upvote", "downvote", or no vote from each user. The sum of the upvotes a user gets minus the downvotes is that user's Karma. Many users believe that Karma is an indicator of their prominence on Reddit, and others believe it is too noisy to be an effective measure. We evaluate how well a current algorithm based on collaborations, and several graph-based measures perform in finding prominent users, and how these results compare to various functions of Karma. We discuss what the performance of these algorithms reveal about the processes that lead to prominence on Reddit and how well Karma serves as a measure prominence.

CHAPTER 1
INTRODUCTION

In this thesis, I study the problem of prominence in open social news and entertainment sites. Prominence has been studied in many social networks [1] with the emphasis to understand what factors contribute to prominence in these networks and how prominence can be measured as a function of network structure [2] and activity [3]. However, most of these studies concentrate on networks with structured activities. For example, academic publishing has been studied in prominence research from many different angles [4]. Prominence in academic publishing is well-defined with many well-studied measures. Furthermore, activities in publishing such as writing papers together or citing each other are directly correlated with prominence. As a result, it is fairly easy to link these actions to prominence. The main challenge is finding the best method to link them. Similarly, for movie networks, being in many movies creates a network effect where more central individuals get better exposure to the network resources and hence can get better jobs in the future. In organizational settings, being central similarly is important as it provides one with access to crucial information in the organization. All of these are well-understood mechanisms that lead to one's prominence. Furthermore, these networks are often smaller in size and scope when compared to massive online new sites.

There is a great deal of study in understanding influence and prominence of users in Twitter [5]. For example, studies try to understand the nature of relationships between individuals and the value of having many friends or followers [6]. However, this type of study does not extend easily to open sites like Reddit which offers support for anonymous interactions and in fact encourages it. Twitter in contrast encourages users to be known to others and collect followers. As a result, Twitter evolved as a place for self-expression and self-promotion, as well as quick dissemination of information. Reddit on the other hand evolved as a site that is geared towards sharing and discussion of news without any specific emphasis on reputation outside of Reddit. This unique property of Reddit sets it apart from all

the other sites studied so far.

There is little work that helps us understand which mechanisms lead to prominence in anonymous social news site like Reddit and how to obtain outside measures of prominence. The only measures of prominence for this site rely on other users' level of engagement with the comments. But, comments are not propagated as in Twitter, they are discussed and elaborated on. Reddit is one of the most popular sites on the Internet but there is little study that explains how people use this site to discuss and disseminate information. This thesis is a first step towards addressing this problem.

We make the following contributions in this thesis.

- We develop a method to collect a representative subset of the Reddit messages for analysis. Without such a subset, it is not possible to arrive at valid conclusions about Reddit. We describe our method and argue why it is a valid method for sampling.

- We analyze network characteristics of Reddit in an effort to understand how it compares to many well-known and well-studied networks. We provide a large number of statistics to show that message exchange characteristics for the large part follow the exponential distribution common to many other networks.

- Finally, we introduce a study of prominence by computing many different measure of prominence introduced in the literature. We also develop various ground truth measures of prominence and compare which ones are the best predictors of prominence in Reddit.

1.1 Introduction to Reddit

Reddit is a social news website in which users submit any information that think the community will find interesting. Reddit messages are divided into Subreddits, which are topic specific. Within each Subreddit, users post content either in the form of a link to another website or a snippet of text, both with the intent of starting a conversation. Other users can then make comments on these posts, or in

reply to other comments already made. By commenting on comments, users build a Comment Tree for each post, with the root node of the tree being the post itself.

In addition to posting and commenting, users can vote on content. A user can choose to upvote or downvote a post or comment. Generally, users are encouraged to upvote relevant content and to downvote content that is not relevant to the discussion. An upvote counts as one point, and a downvote as negative one point. The sum of points for a post or comment are its score. Content is displayed in an order determined by a combination of age and its score (newer items and items with a higher score are shown first).

Users do not up and downvote only to ensure relevant content is displayed. Users are also rewarded for producing high-scoring content in the form of Karma. A user has two types of Karma: Link Karma and Comment Karma. A user's Link Karma is the sum of the upvotes on their posts that contain links minus the sum of the downvotes on their posts that contain links. A user's Comment Karma is the sum of the upvotes on all their comments minus the sum of the downvotes on their comments. Link Karma represents their ability to post useful stories, while their Comment Karma represents their ability to contribute to conversation. We are more interested in the latter.

1.2 Why Reddit

Reddit is a community built on reputation. Users with high karma generally are those that contribute to conversation in a meaningful way. It is these people who post often and meaningfully that are prominent within the website. But their reputation does not necessarily correspond to any specific rights within the site. For example, the posts of a user with high Karma are not necessarily displayed higher than other users' posts at first. However, as such users may be known to the community, they may get more attention in general.

Reddit is also known to support many communities who regularly discuss highly contentious topics like Atheism where users may end upvoting or downvoting comments that they do not agree with. In these cases, community standards are enforced by the other members' verbal reminders. The discussions on Reddit have a

strong community feel where the discussion is what matters, not necessarily a single post. This is a big departure from Twitter in which the original message matters the most as tweets are continuously propagated through retweets. As a result, the exchanges on Reddit is more geared towards discourse whereas on Twitter towards self-expression and self-promotion.

Sites like Digg [7, 8] are also built on reputation where reputation is gained by promoting a link that others' in the site will also like and give upvotes. But, there is no notion of community on Digg as in the case with Subreddits and there are no incentives for deep discussions for each promoted link similar to those in Reddit. As a result, it is often argued that Digg ratings are highly noisy as high ratings are achieved often for few sites that are well-known and at low risk of not getting good ratings. Reddit in contrast draws content from a much wider set of sites and audience with diverse interests.

Wikipedia is another well-studied site [9] where users gain prominence by their contributions. In this case, the prominence is explicit by a vote and it gives specific powers to its owner. In Wikipedia users work on a finished product, a Wikipedia page which is judged on its correctness. In contrast, Reddit discussions concentrate on passing fads and interests of its community with more ephemeral discussions. In fact, Subreddit activity displays strong preferential attachment. Users are drawn to a subreddit by interest in the topic of discussion. They are also drawn to where there is more activity. This results in more active subreddits having both more posts and more comments per post than less active subreddits.

Despite the interest on reputation, Reddit also strongly supports anonymity. In fact, it is often not allowed to post personally identifying information for self or others with the exception of specific posts by celebrities. This also means that Reddit supports multiple accounts and private discussions using what is called throw away accounts, one time accounts that are not linked to any specific user with a well-known reputation.

The closest site to Reddit is 4chan [10], which supports communication by posting of mainly images and short snippets of text. Posting anonymously is the default. This, in addition to the short time frame for which posts are active, results

in discussions that are further from the social norm than on typical websites.

Moderation plays an important role in Reddit. Slashdot [11, 12] has a distributed moderation system. Each user is allowed a limited amount of moderation power, with users who contribute more getting more power. 4chan [10] has some moderators, but they are to blend in like normal users and typically only delete content that violates rules. Reddit has moderators per subreddit, and they not only ensure that content follows the rules of the subreddit, but that content is relevant to the subreddit. Users can be moderators of multiple subreddits. When active in a subreddit that they are not a moderator of, they are no different from normal users.

As we can conclude from this discussion, while there are some similarities to other sites, Reddit is in many aspects a unique site. Furthermore, it is one of the most popular sites on the Internet often called the Internet's front page. Despite its popularity, it has never been analyzed like many other networks (Twitter, Wikipedia, Forums, etc.) have. My goal in this thesis is to perform preliminary analysis of Reddit and compare it to other well known social networks.

CHAPTER 2
RELATED WORK

There has been much work in analyzing various social networks. Reddit is similar to many of these, yet is different from them in various ways. We will try to use analysis methods that were successful on other networks on Reddit and see how they perform. Additionally, we will try to use insights from these studies to see if Reddit behavior matches behavior in other networks.

Prominence has always been an important part of social interaction. With the rise of online interaction, it is easier and easier to talk to many people, and many seek to separate themselves from the rest. Adali et. al. showed a method of finding prominence in [13] called iHypR. They found that when people worked together on objects, which were then grouped into collections, it was easier to rank users' prominence than without the collections. Additionally, the method works on a variety of datasets, and does not need to be tuned for analysis of new datasets.

Not all users seek reputation. Reddit allows for creation of new accounts without even an email, meaning a user may have many accounts. For discussion of sensitive topics, such as personal matters or the merits of their job, users tend to prefer a "throw-away" account which is only used for that topic of conversation. Slashdot has a similar setup. They allow users to post anonymously, or using an account. Gomez shows in [14] that only 18.6% of users posted anonymously, with the rest using real accounts in order to build reputation. Detecting which users are posting using a throw-away versus posting only once within our dataset is a problem for future study.

Anonymity can be important in a social setting. Bernstein et al. found in [15] that over 90% of posts in 4chan were submitted anonymously, and the vast majority of the remaining posts were made using obviously fake usernames. This, combined with the rapid cycling of content, leads to discussion much more willing to go outside of social norms. Reddit differs from this because threads are kept forever unless deleted by the poster, whereas 4chan automatically deleted threads

once enough new content is posted (median age is 3.9 minutes).

In some situations, reputation can mean everything. Zhang and Tran found in [16] that when evaluating reviews, reputation of the reviewer is of great importance when attempting to filter for useful reviews. Fake reviews are a great problem, and until detection of fake reviews is possible on a textual basis, user reputation will remain a key component of detecting the useful reviews.

Belk et al. examine the influence of overlapping communities on each other within forums in [17]. They first identify forums that strongly influence other forums through common posters, then examine how the influence changes over 9 years. However, they do not explore the influence of the individual posters, but their collective influence on a forum.

Bhatt and Barman examine the properties of Yahoo discussion groups in [18]. In particular, they observe that their users exhibit q-exponential behavior in response times, rather than power-law distributions. The combination of multiple users results in what appears to be a power-law distribution for the discussion as a whole. They also observe that threads become popular from their inception, rather than over time.

Morrison et al. look at the properties of forums over time in [19]. They group users into one of four roles within the forum. The "Supporter" receives many replies. The "Ignored" makes many posts but receives few replies. The "Grunt" Posts a moderate amount and receives a moderate amount of replies. Finally, the "Elitist" Posts a lot in a few threads. The authors examine how roles change over time for individual users. They found that grunts were extremely stable, while the other roles were not.

Gargi et al. build clusters of videos using similarities in viewers in [20]. They viewed two videos as being more similar if more users watched both of them. In addition, they used textual analysis of keywords for the videos to enhance their clustering.

Dave et al. in [21] find important people within networks by evaluating them on how their actions impact the actions of others. Applications of this are in marketing, specifically within viral marketing. The idea is that users whose actions

correlate with many other users taking the same action are influential within the network. Additionally, the users are evaluated on how far their actions reach. Users who influence other users who are then themselves influential are considered to be even more influential. The dataset was pulled from Flickr, using friend relationships and group memberships to determine if a user influences another.

CHAPTER 3
Data Collection

An important problem we address in this chapter is how to gather a representative sample of discussions on Reddit for analysis. Reddit activity is skewed heavily towards a small number of subreddits with very broad topics of conversation. However, there is also valuable conversation in the more focused but less active subreddits. Therefore, care was taken to ensure that all subreddits are represented proportionally to their activity. A random sample of subreddits creates an uneven distribution of activity by suppressing the activity on popular subreddits. A proportional sample of activity by subreddits creates the reverse problem as popular subreddits will dominate the collection in this method. To address these concerns, we collect posts in one solid time slice. We accomplish this by following post ids of posts. Given that posts are given incremental identifiers, this allows us to collect all the activity in a given time period. As a result, each subreddit will be represented proportional to its posting activity in this time slice.

We note that posts higher on the Reddit page for each subreddit are much more likely to receive commenting activity. Therefore, we do not want to collect comments off posts until they are old enough for all comments that are going to be made have already been made. As a result, we first collect a number of posts from a past time period and then collect all the comment data for these posts. In this thesis, we report on a collection of posts from mid December, since posts this old are unlikely to receive any further amount of comments or votes that are significant.

The collection itself is done with a Python library called "Python Reddit API Wrapper", or PRAW ([22]). This library directly translates subreddits, posts, comments, and users into Python objects. The collection code proceeds as follows:

Table 3.1: Data collected for a subreddit

Attribute	Description
name	The human readable name of the subreddit
id	The unique numerical identifier of the subreddit

Table 3.2: Data collected for a post

Attribute	Description
id	The unique numerical identifier of the post
sr-name	Human readable name of the subreddit containing the post
title	Title of the post
creation-date	Date and time the post was created
author	User who created the post

```
post-id = 150000 (A post id old enough such that
                  there is no additional comment activity)
while True:
    collect-subreddit-data(post-id)
    collect-post-data(post-id)
    comments = collect-comments(post-id)
    for comment-id in comments:
        collect-comment-data(comment-id)
    post-id += 1
```

For each subreddit, we collect the data shown in Table 3.1. For each post, we collect the data shown in Table 3.2. For each comment, we collect the data shown in Table 3.3. Note that all id's are numerical, but represented using base 36 numbers.

For our purposes, a post is really a comment with no parent (a comment in reply to nothing). Additionally, the comment object contains the same information as the post, but with more added. So, we create a comment object for every post with no parent-id, which allows us to look only at comments for some analysis.

Given that there is a rate limit on Reddit, we are able to collect about 8,120 posts per day. This collection has been running for about 16 days. In this thesis, we report on a data set of 904,802 comments, 127,857 posts from 8,819 subreddits and 292,197 users, ranging from December 17^{th}, 2012 to January 27^{th}, 2013.

Table 3.3: Data collected for a comment

Attribute	Description
id	The unique numerical identifier of the post
post-id	The id of the post the comment was made on
sr-name	Human readable name of the subreddit containing the post
parent-id	The id of the comment or post this comment was a reply to
up-vote	Total count of upvotes the comment received
down-vote	Total count of downvotes the comment received
comment-text	Text of the comment
creation-date	Date and time the comment was created
author	User who created the comment

Table 3.4: Collected data over time

Date	Comments	Posts	Active Users	Active Subreddits
17 Dec 2012	278153	61775	119715	6206
18 Dec 2012	559123	66082	205826	7220
19 Dec 2012	85186		50768	3511
20 Dec 2012	12612		9195	2293
21 Dec 2012	5060		3958	1645
22 Dec 2012	2911		2351	1225
23 Dec 2012	2166		1763	1009
24 Dec 2012	1671		1302	824
25 Dec 2012	1031		846	592
26 Dec 2012	1096		917	605
27 Dec 2012	940		771	547
28 Dec 2012	840		698	488
29 Dec 2012	547		466	351
30 Dec 2012	560		458	346
31 Dec 2012	429		366	295
01 Jan 2013	330		270	225
02 Jan 2013	444		361	271
03 Jan 2013	303		272	200
04 Jan 2013	241		211	158
05 Jan 2013	163		145	127
06 Jan 2013	147		137	120
07 Jan 2013	170		148	122
08 Jan 2013	139		123	101
09 Jan 2013	136		112	90
10 Jan 2013	108		80	62
11 Jan 2013	80		70	51
12 Jan 2013	69		60	47
13 Jan 2013	63		51	39
14 Jan 2013	56		51	51
15 Jan 2013	56		47	40
16 Jan 2013	59		50	43
17 Jan 2013	52		43	37
18 Jan 2013	58		49	38
19 Jan 2013	40		38	32
20 Jan 2013	27		25	23
21 Jan 2013	18		18	17
22 Jan 2013	31		30	25
23 Jan 2013	31		26	21
24 Jan 2013	17		15	14
25 Jan 2013	8		8	8
26 Jan 2013	11		11	8
27 Jan 2013	3		3	3

CHAPTER 4
NETWORK ANALYSIS OF REDDIT

In this chapter, we present an in-depth analysis of the Reddit network. We report on the top users and subreddits to show what type of activity is most common in the time period we have collected. We also look at the distribution of posts to users, reddits and time between posts, to assess the nature of the activity.

4.1 Top Users and Subreddits

We first identify the top subreddits and users to understand the most common topics of discussion and the users who participate in these.

4.1.1 Top Subreddits

In Table 4.1 we see the top subreddits by how many posts they have. In Table 4.2, we see the top subreddits by how many comments were placed within the subreddit. Looking at just the top two of each yields interesting results: "funny" receives over 30% more posts than the next most posted on, "AskReddit". However, "AskReddit" receives over 80% more comments than "funny". This makes sense when we look at what each subreddit is for. "funny" is a subreddit where humorous content is posted. "AskReddit" is a subreddit where users post questions and other users provide answers. It stands to reason that "AskReddit" will have vastly more comments than any other subreddit, since it is an area for questions of any type to be asked and they will receive many answers. Another subreddit like this is "IAmA", where users will start a thread about what they do for a job or something they have done, other users will post questions for this person as comments on their thread, and the original poster will comment with answers. Notably, President Obama and many famous tech figures have participated in IAmA posts, with resulting on Reddit being overwhelmed with too much traffic.

In general, the top subreddits by posts are topics that are post-oriented. These include subreddits for the posting of images. All of these image-oriented subreddits

Table 4.1: Top 50 subreddits by post count

Subreddit	Number of Posts	Subreddit	Number of Posts
funny	7187	tf2trade	498
AskReddit	5367	reportthespammers	488
AdviceAnimals	4317	business	471
pics	3686	Minecraft	463
POLITIC	2958	ModerationLog	459
gaming	2334	offbeat	456
aww	2181	nfl	382
videos	2142	ImGoingToHellForThis	374
trees	2031	IAmA	370
Music	1793	guns	365
politics	1705	askscience	361
WTF	1572	gifs	348
leagueoflegends	1459	starcraft	326
worldnews	1139	SteamGameSwap	316
technology	1111	mildlyinteresting	315
atheism	1002	reactiongifs	314
todayilearned	981	science	311
secretsanta	958	fantasyfootball	305
Dota2Trade	837	Android	295
circlejerk	811	bestof	275
gonewild	734	mylittlepony	275
news	732	cringe	274
movies	662	nsfw	268
buildapc	605	techsupport	268
fffffffuuuuuuuuuuuu	512	RisingThreads	267

rank lower when ranked by comment quantity. With this ranking, discussion and question oriented subreddits rank highly.

In Table 4.3, we see which subreddits have the most people talking within them. As one would expect, we see the results are similar to the top subreddits by other counts. We see a rapid dropoff from the top subreddits. The top, "funny", has nearly 24,000 users, while the 10^{th} has 5,600. The 50^{th} has only 1,190. While we see a sharp drop-off of distinct authors per subreddit, each author contributes to multiple subreddits. Each user need only contribute a small amount to make Reddit the strongly connected society that it is.

An interesting thing to note that the top subreddits are not dominated by top-

Table 4.2: Top 50 subreddits by comment count

Subreddit	Number of Comments	Subreddit	Number of Comments
AskReddit	77463	technology	4765
funny	42925	nba	4499
pics	27825	Android	4466
gaming	21974	Guildwars2	4463
leagueoflegends	19800	soccer	4275
WTF	19529	buildapc	4251
AdviceAnimals	18005	malefashionadvice	4106
politics	16663	AskWomen	3916
trees	13990	ffffffuuuuuuuuuuuu	3794
nfl	11785	blackops2	3759
videos	11755	cringe	3690
atheism	10371	halo	3604
gonewild	10252	AskMen	3593
todayilearned	9778	CFB	3561
IAmA	7894	POLITIC	3530
aww	7746	SquaredCircle	3478
worldnews	6842	Planetside	3476
guns	6793	sex	3453
MLPLounge	6195	relationships	3341
DotA2	6165	Dota2Trade	3233
Games	5500	gifs	3220
starcraft	5415	news	3054
movies	5229	conspiracy	3008
Minecraft	4841	fantasyfootball	3003
Music	4788	ImGoingToHellForThis	2937

ics related to technology, but to a wide range of topics from sports to relationships, politics, world events and many random topics that arise in AskReddit. In addition, there are some special topics special to Reddit that rise and fall in popularity over time (like discussions about a kids show called My Little Pony in our data set).

4.1.2 Top Users

In Tables 4.4 and 4.5 we see the top users ranked by commenting activity, the number of comments they have made, and the top (at most) 4 subreddits they have commented within.

The top users are obviously "bots" that use the same API used in data collec-

Table 4.3: Top 50 subreddits by author count

Subreddit	Number of Authors	Subreddit	Number of Authors
funny	23907	cringe	2065
AskReddit	20401	fffffffuuuuuuuuuuuu	2057
pics	16203	Android	2002
gaming	11566	guns	1866
WTF	10687	Guildwars2	1815
AdviceAnimals	10400	ImGoingToHellForThis	1764
leagueoflegends	8206	soccer	1732
videos	7158	nba	1595
trees	6401	malefashionadvice	1539
politics	5600	news	1511
todayilearned	5597	mildlyinteresting	1458
aww	5127	sex	1454
atheism	4570	pokemon	1417
nfl	3530	science	1410
worldnews	3480	halo	1349
gonewild	3373	secretsanta	1326
Music	3115	JusticePorn	1315
IAmA	3004	CFB	1304
movies	2805	buildapc	1293
technology	2776	circlejerk	1287
Games	2462	blackops2	1255
Minecraft	2414	wow	1223
starcraft	2352	Planetside	1220
DotA2	2352	LifeProTips	1200
gifs	2207	fantasyfootball	1190

tion to autonomously make comments. For example, the user "qkme_transcriber" is a bot that will autonomously read images from quickmeme.com and provide a comment with the text content of the image (This is possible because the url contains the text within the image).

However, for users that are not bots, we see an interesting pattern. Nearly all of the users have most of their activity within AskReddit, with many having it as the only subreddit they commented within. This makes sense, because AskReddit is by far the most active in terms of comments. When we look at the other subreddits commented within, we see a variety of subreddits, but not an overwhelming number. The top 50 users comment in 355 subreddits. The top 1000 comment in only 1,791.

Table 4.4: Top 50 users by comment count and subreddits most active within, part I

User	Comment Count	Subreddits in Order of Descending Activity
ModerationLog	918	ModerationLog, POLITIC
AutoModerator	713	relationships, aww, bestof, reactiongifs
qkme_transcriber	656	AdviceAnimals, funny, ImGoingToHellForThis, shittyadviceanimals
Purplebuzz	235	AskReddit, pics, funny, atheism
trevormatic	230	AskReddit, NotaMethAddict, notinteresting
savoytruffle	195	AskReddit
1368JM	190	AskReddit, AskModerators, britishproblems, SexyButNotPorn
LuckyD93	188	AskReddit, paracord, loseit
wingwalker	185	AskReddit
roger_bot	184	GuessTheMovie
ArchmageLudicrous	181	MLPLounge
GirPhralad	181	AskReddit
Thorse	165	gaming, reportthespammers, Patriots, Diablo
churro	161	MLPLounge
QualityEnforcer	157	pics, gentlemanboners, gaming, HistoryPorn
TweetPoster	156	nfl, funny, leagueoflegends, POLITIC
punkpixzsticks	155	AskReddit
original-finder	152	funny, AdviceAnimals, gaming, pics
MrAssKing	150	gonewild, aww
camopdude	149	AskReddit, books, PS3, psych
chocthunder4	147	gonewild, gonewildcurvy, dykesgonewild, grool
Beechet	144	MLPLounge
JonAudette	142	AskReddit
Late_Night_Grumbler	142	AskReddit, leagueoflegends, Maplestory
alwaystherenever	139	gonewild

Table 4.5: Top 50 users by comment count and subreddits most active within, part II

User	Comment Count	Subreddits in Order of Descending Activity
backpackwayne	139	RandomActsOfChristmas, Assistance, AskReddit, politics
doobie-scooo	139	AskReddit
heruskael	139	AskReddit, intj, zombies, aww
kubrick66	138	politics, guns, Libertarian, soccer
superolafboy	136	AskReddit
Val_Hallen	136	AskReddit, funny, politics
Sugisaki	134	MLPLounge, gaming, anime
YankeeQuebec	133	guns, Firearms, news, politics
DevaKitty	131	MLPLounge, mylittlefortress, Naruto, mylittlepony
syntaxxor	130	guns, loseit, Fitness, Glocks
gman524	127	AskReddit
PornOverlord	124	FoodPorn, EarthPorn, MapPorn, AbandonedPorn
IntoTheMystic1	119	AskReddit
ListenToTheMusic	118	Random_Acts_Of_Amazon, RAOWL, RandomActsOfPolish
ochtapas	118	AskReddit, pics, MensRights, funny
Tdaug	116	AskReddit, NoLimitations
dcmjim	110	AskReddit, funny
asharkey3	109	blackops2, gaming, gamegrumps, funny
tsingi	108	atheism, atheismplus
Bacon7	107	MLPLounge
ta1901	107	AskReddit, AskMen, sex, techsupport
zach2093	107	AskReddit, gameofthrones, asoiaf, television
FatGuyInALittleHouse	106	AskReddit
perogies	106	AskReddit, politics, worldnews, news
r_k_ologist	106	AskReddit, HITsWorthTurkingFor, PlayStationPlus, PS3

Table 4.6: Top 50 subreddits by unique commentors

Subreddit	Author Count	Subreddit	Author Count
funny	23907	cringe	2065
AskReddit	20401	fffffffuuuuuuuuuuuu	2057
pics	16203	Android	2002
gaming	11566	guns	1866
WTF	10687	Guildwars2	1815
AdviceAnimals	10400	ImGoingToHellForThis	1764
leagueoflegends	8206	soccer	1732
videos	7158	nba	1595
trees	6401	malefashionadvice	1539
politics	5600	news	1511
todayilearned	5597	mildlyinteresting	1458
aww	5127	sex	1454
atheism	4570	pokemon	1417
nfl	353	science	1410
worldnews	3480	halo	1349
gonewild	3373	secretsanta	1326
Music	3115	JusticePorn	1315
IAmA	3004	CFB	1304
movies	2805	buildapc	1293
technology	2776	circlejerk	1287
Games	2462	blackops2	1255
Minecraft	2414	wow	1223
starcraft	2352	Planetside	1220
DotA2	2352	LifeProTips	1200
gifs	2207	fantasyfootball	1190

Therefore, it appears that the connectivity of Reddit comes from not a few super-users, but from the majority rank and file users.

4.2 Distribution of Activity

In this section, we examine four different distributions to see how our data is distributed. The distributions of comments across posts, comments across subreddits, posts across subreddits, and comments across users.

We find that all these distributions fit a power-law regression (shown below). In particular, comments across users and posts across subreddits displayed the strongest fit. This fits with many other findings on social networks in which few

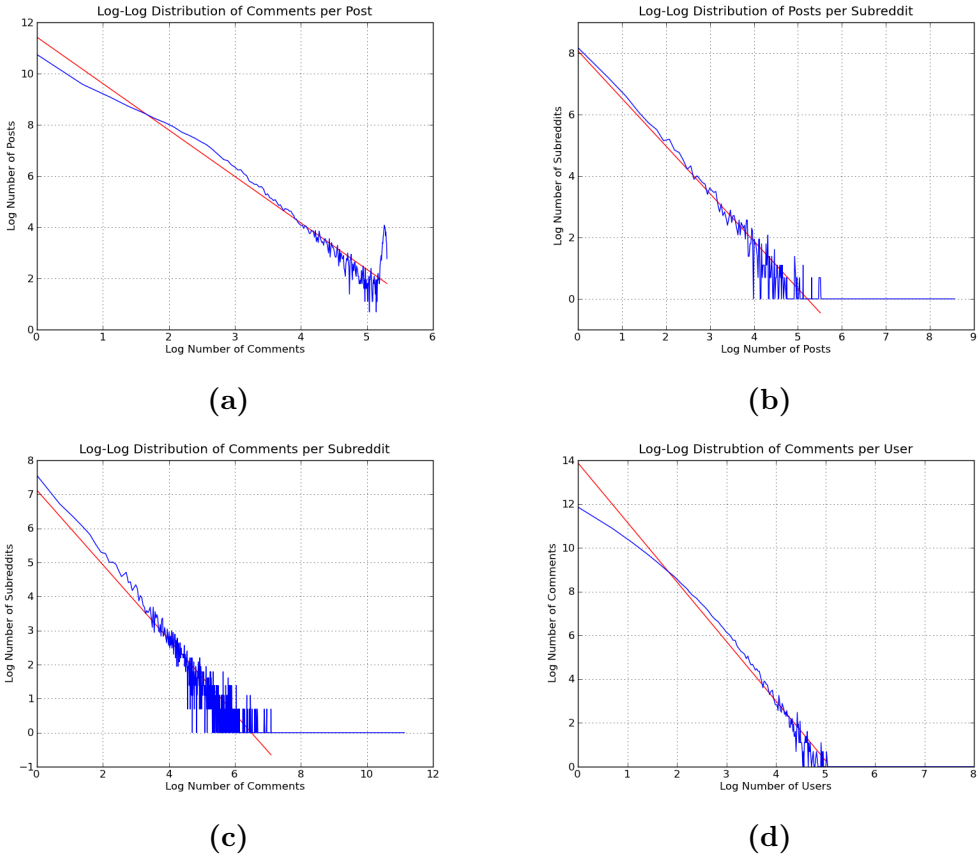


Figure 4.1: Graphs of the distribution of objects relative to each other

users and posts dominate most of the activity, followed by a long tail of many users and posts with very little activity such as Twitter [23, 24, 25].

4.2.1 Comments Across Posts

We examine the distribution of comments across posts. As might be expected, Figure 4.1a shows a clearly power-law relationship, with a best fit of $y = x^{-1.81} * 91491.376$, $r^2 = 0.877$. We also note that the post with the most comments had 201 comments. Additionally, there were 46,539 posts with zero comments.

4.2.2 Posts Across Subreddits

The distribution of posts across subreddits are shown in Figure 4.1b. The distribution follows a power-law distribution as well. The best fit is $y = x^{-1.55} * 3211.521$, $r^2 = 0.940$. The subreddit with the most posts had 5,239, and 3,564

subreddits had only one post.

4.2.3 Comments Across Subreddits

The distribution of comments across subreddits are shown in Figure 4.1c. This distribution also follows a power-law with a best fit of $y = x^{-1.10} * 1237.687$, $r^2 = 0.885$. The subreddit with the most comments had 68,443 comments, and 1,916 subreddits had just one post with no comments.

4.2.4 Comments Across Users

When we look at the distribution of comments across users in Figure 4.1d, we again see a power-law distribution with a best fit of $y = x^{-2.72} * 1072496.858$, $r^2 = 0.973$. One user had a rather astounding 2,954 comments, while 142,273 users had only one comment.

4.2.5 Time Before Replies

We take a brief look at the distribution of the time delta between when a post is made, and when the comments it receives are posted. In Figure 4.2a, we see what appears to be an power-law distribution. However, when looking at the log-log graph in Figure 4.2b, we see that there are some subtle but interesting variances. We see that there is a peak of comments around 3 seconds after the post, which then drops off and picks up at approximately one minute after the post was created. After this, the comment rate does drop off in a power-law type fashion.

When we look at the time for replies to be made to posts in addition to other comments, this property seems to hold even stronger. In Figure 4.2d, we see the same first peak at 3 seconds. However, the peak at one minute is even stronger (approximately double the size), and the same curved drop off. This implies that most replies to a post are made in a similar manner to replies for a particular item.

Other networks have exhibited the same drop-off behavior that is not quite power-law. However, none of them have shown the same peaking behavior. This could be because they do not exhibit it, but perhaps it is due to comments being posed with a longer delay than on Reddit, resulting in only one peak further out.

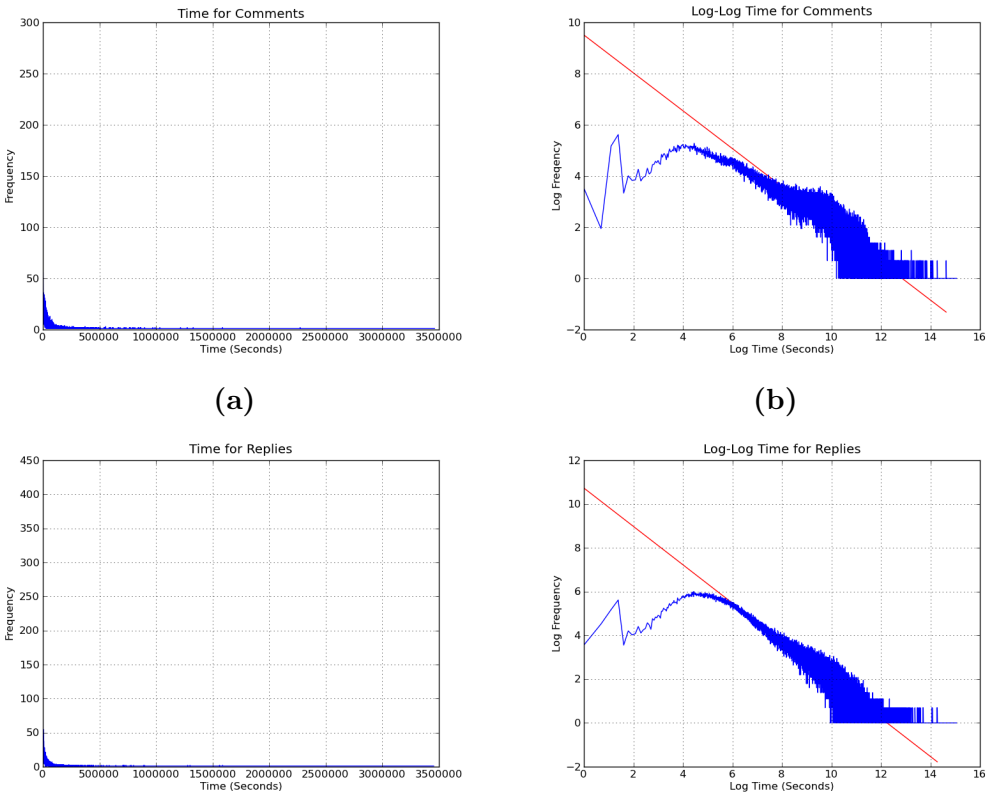


Figure 4.2: Temporal graphs

4.3 Analysis of Post Connectivity

Next, we analyze to which degree posts are related to each other in the whole Reddit network. This will show us to which degree the discussion on Reddit is truly connected. To give an example, consider an academic publishing network. In such a network, a paper is an equivalent of a post. Two papers are connected if they are written by the same authors. In fact, the more authors two papers have in common, the closer the posts are. Analysis of academic papers in this manner reveals that papers that are closely connected to each other correspond to the main topics in the network in the large scale. In the smaller scale, the connected papers represent research collaborations [13]. The overall connectivity of the network shows to which degree the collaborations and research topics are related to each other.

We conduct a similar analysis for the Reddit post network to understand what the local collaborations reveal. To accomplish this, we build a graph $G = (V, E)$

in which nodes V are posts. There exists an edge between $(e_1, e_2) \in V$ between two posts if there is a common author between any comments within the two posts. Edges are weighted by how similar the set of commenters for the two nodes are. If there are many common authors, the weight is small, representing a short distance. We use a number of different distance measures to compare posts.

4.3.1 Jaccard Distance

The Jaccard Distance ([26]) is a simple measure of distance, dependent only upon how many common authors two posts have relative to the set of all authors between two posts.

4.3.1.1 Unweighted Jaccard Distance

The simplest method for measuring distance is the Jaccard Distance which is based on the proportion of common authors to all authors for a pair of posts. To calculate this distance for two posts, we have two methods. The first is to ignore the number of times we see each author. Given two posts p_1, p_2, we have:

$$D_J(p_1, p_2) = \frac{|\Gamma(p_1) \cap \Gamma(p_2)|}{|\Gamma(p_1) \cup \Gamma(p_2)|} \tag{4.1}$$

where $\Gamma(p)$ is the set of distinct authors a post p has. Note that, the Jaccard distance does not depend on the number of authors two posts have. For example, two posts with total of 10 authors and 5 common authors will have the same Jaccard measure as two posts with 100 authors and 50 common authors. As a result, Jaccard is a more global measure and it disregards the differences in popularity of posts.

4.3.1.2 Weighted Jaccard Distance

The unweighted Jaccard does not account for users who comment multiple times on the same post. This is important to account for, since users who comment multiple times are much more likely to be contributing in a meaningful way to the conversation. To account for this, we introduce a new version of Jaccard called "Weighted Jaccard Distance". We will modify the above version to account for multiple instances of the same author. Instead of taking the intersections of the

author sets, we will take the minimum number of comments between the two posts for each author and sum them. Our equation becomes:

$$D_{wJ}(p_1, p_2) = \frac{\sum_{a \in (\Gamma(p_1) \cap \Gamma(p_2))} max(\kappa(a, p_1), \kappa(a, p_2))}{\sum_{a \in (\Gamma(p_1) \cup \Gamma(p_2))} min(\kappa(a, p_1), \kappa(a, p_2))} \tag{4.2}$$

where $\kappa(a, p)$ is the number of comments author a has in post p. Note that this measure, similar to the unweighted version, is not sensitive to the number of total authors a post has.

4.3.1.3 Example

In this section, we give a small example to show the difference between the two version of Jaccard distance. Suppose the number of comments for three authors is given as below.

Author	**# Comments in α**	**# Comments in β**
Alfred	1	2
Betty	3	2
Charles	0	4

In this example, $\Gamma(\alpha) \cup \Gamma(\beta) = 3$ and $\Gamma(\alpha) \cap \Gamma(\beta) = 2$. So, the unweighted distance between the two is $D_J(\alpha, \beta) = \frac{3}{2} = 1.5$.

For the weighted distance, we compute the max and mins as follows:

Author	Max	Min
Alfred	2	1
Betty	3	2
Charles	4	0

The sum of the max values is 9, and the sum of the mins is 3, giving us a weighted distance of $D_{wJ}(\alpha, \beta) = \frac{9}{3} = 3$. Note that the denominator is never 0 for both equations, since it will only be zero if the posts share no authors, and we only compute distances for posts that share authors. Also note that for the case of each author making the same number of comments, the unweighted and weighted distances are the same. The unweighted fails to capture the relative contributions made by different authors.

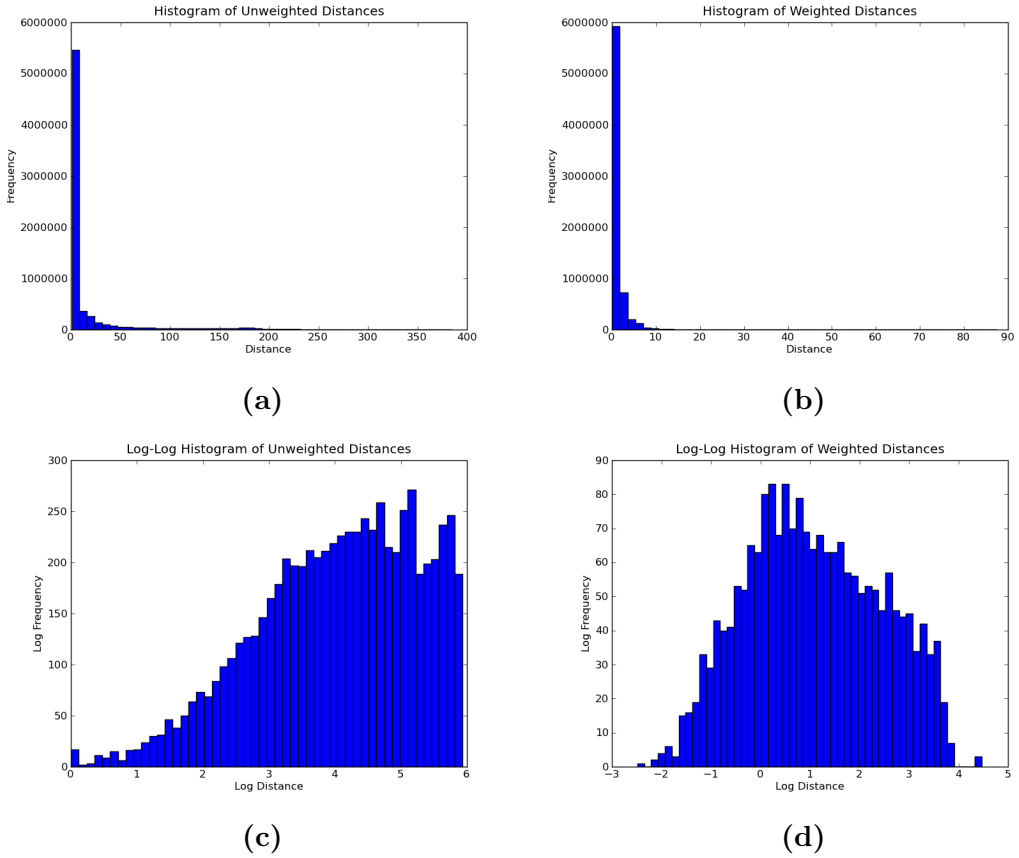


Figure 4.3: Jaccard distance distributions

4.3.1.4 Distribution of Distances

We show the distribution of both the weighted (Figure 4.3a) and unweighted (Figure 4.3b) distances. They show similar distributions, but the unweighted distances tend to be greater than the weighted. The maximal unweighted distance is 385, while the maximal weighted distance is 87.5. Additionally, the histogram drops off after 50 for the unweighted distances, but after 20 for weighted. Weighting seems to decrease the distances overall.

4.3.2 Adamic-Adar Distance

In addition to the Jaccard distance, we also consider a version of the Adamic-Adar Distance for posts. Unlike Jaccard, the Adamic-Adar distance ([27]) takes into account the number of authors a post has. The underlying logic in Adamic-Adar is that an author that has a lot of posts is unlikely to pay attention to each one.

As a result, she is unlikely to form relationships with those who comment on her posts. To compute this value, the distance function looks not only at the number of authors in common, but the number of other posts the authors have commented on. Just like with the Jaccard Distance, we have weighted and unweighted versions.

4.3.2.1 Unweighted

The unweighted version of the Adamic-Adar distance is dependent only on how many posts an author has commented, for the authors that are common to two posts. We care about how much of a user's attention was used to connect two posts. If an author only commented on the two posts, then all of their attention went to it. If a pair of posts corresponds to 2 out of the 100 posts the author has commented on, then not much attention went into the give pair. Given two posts, p_1, p_2, we have:

$$D_{AA}(p_1, p_2) = \frac{1}{\sum_{a \in \Gamma(p_1) \cap \Gamma(p_2)} ln\left(\frac{1}{\Theta(a)} + 1\right)} \tag{4.3}$$

where $\Theta(a)$ is the number of comments author a has made. We add one to $\frac{1}{Theta(a)}$ to make the distance positive in all cases.

4.3.2.2 Weighted

With the unweighted Adamic-Adar Distance, we account for an author commenting on many posts versus few posts. However, we do not account for multiple comments to a post. If a user comments on one post ten times, another 100 times, and made a total of 200 comments, we can connect the two posts more strongly than if they had commented on the first post only once. We weigh with the following formula:

$$D_{wAA}(p_1, p_2) = \frac{1}{\sum_{a \in \Gamma(p_1) \cap \Gamma(p_2)} ln\left(\frac{\kappa(a,p_1)+\kappa(a,p_2)}{\Theta(a)} + 1\right)} \tag{4.4}$$

with κ as defined for the Weighted Jaccard Distance.

4.3.2.3 Example

We revisit the same example we used for the Jaccard Distance, with a count of total comments made added.

Author	# Comments in α	# Comments in β	Total Comments
Alfred	1	2	100
Betty	3	2	5
Charles	0	4	10

We compute the unweighted distance to be

$$D_{AA}(\alpha, \beta) = \frac{1}{ln\left(\frac{1}{100} + 1\right) + ln\left(\frac{1}{5} + 1\right)} \approx 5.2 \tag{4.5}$$

We compute the weighted distance to be

$$D_{wAA}(\alpha, \beta) \frac{1}{ln\left(\frac{3}{100} + 1\right) + ln\left(\frac{5}{5} + 1\right)} \approx 1.38 \tag{4.6}$$

Note that the weighted distance is much less than the unweighted. Author Betty contributes much more strongly to a smaller distance in the weighted case, since she is devoting all of her attention to the two posts. Similar to the Jaccard Distance, the weighted and unweighted versions are the same in the case that the authors made at most one comment on any post. Also note that Charles does not contribute at all to our distance. We only examine the authors who commented on both, since only from them can we see what the relative attention to both was.

4.3.2.4 Distribution of Distances

The distance distribution shows a smooth decrease in frequency as distance increases only for low distances. Once we get higher, we see large spikes in frequency. These can be easily explained. Suppose we have a user who only made posts, and those posts never get comments (so they are the only author on the post). Suppose they made 1,000 posts. By both our weight and unweighted formula, the distance between any of these posts is:

$$D_{aa} = \frac{1}{log\left(\frac{1}{1000} + 1\right)} \approx 2303 \tag{4.7}$$

So, we would have 1,000 posts where the distance between any two is 2303. This gives us about 500,000 entries for a distance of 2303. The Jaccard distance

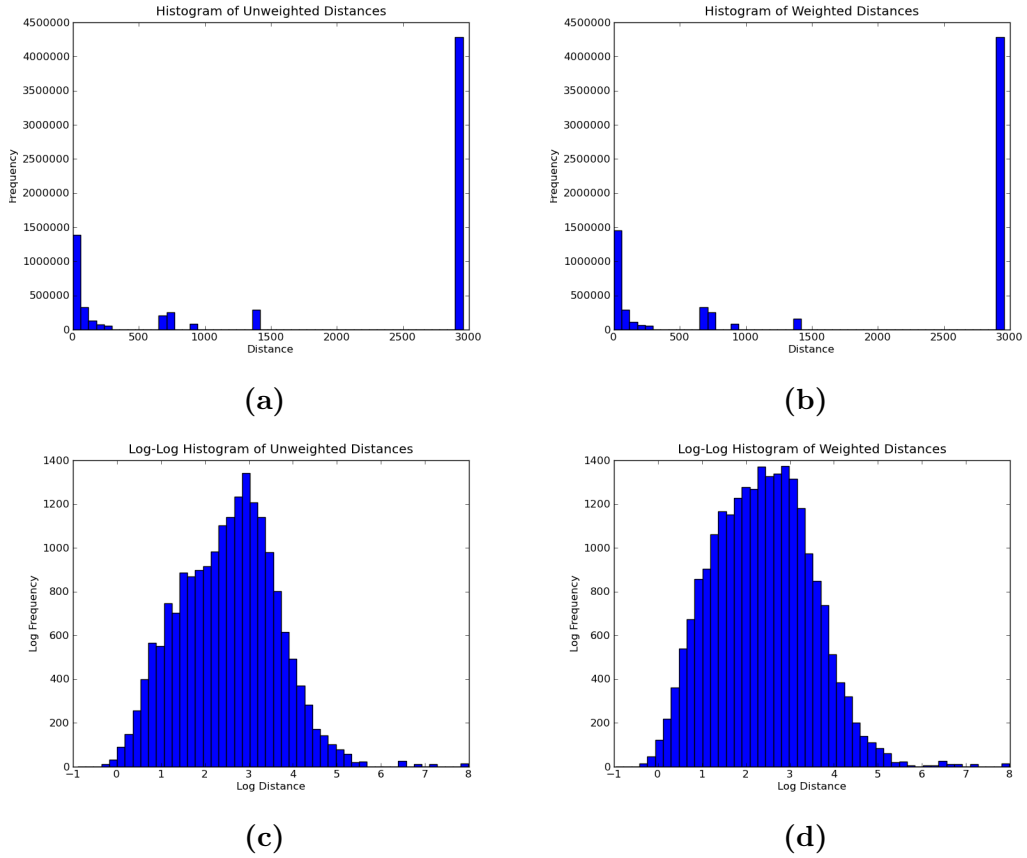


Figure 4.4: Adamic-Adar distance distributions

measure (both weighted and unweighted) will be 1, which is why these spikes do not appear in the Jaccard distance graphs.

Situations such as this one (and a slight variant of only one reply) create all large spikes for weights greater than 500. These posts are created by users, with the most prolific one by far being "PoliticBot", which is an automated poster that posts many articles, many of which get no or few comments. In total, 6722 users, approximately 2.3% of users, contribute to these spikes in 4766 posts. These posts account for 5,108,031 of the 7,087,883 links in the graph.

4.3.3 Degree Distribution

We examine the degree distribution for the graph. The set of edges in both graphs is the same, so the degree distribution is the same for both. In Figure 4.5, we see a great number of nodes with only one connect (9,369), tapering off quickly

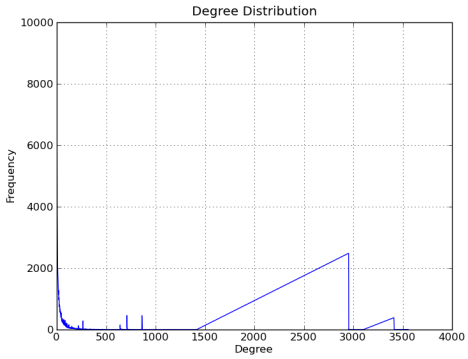


Figure 4.5: Degree distribution

Table 4.7: Cluster sizes

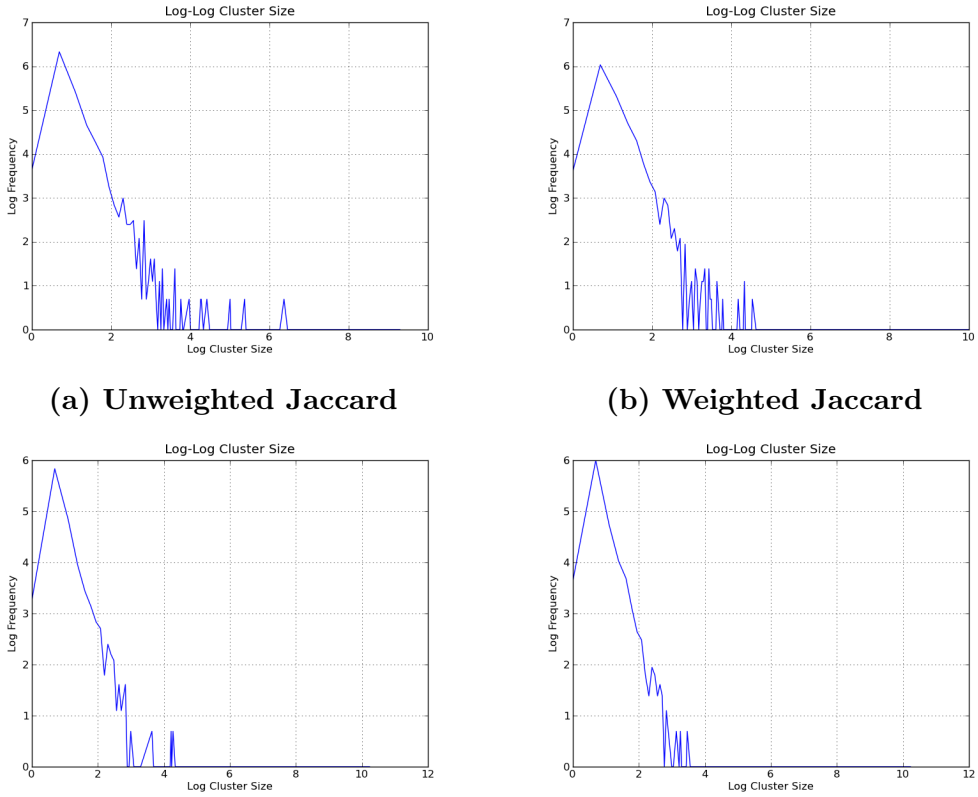
Distance Type	Count
Unweighted Jaccard	1307
Weighted Jaccard	1131
Unweighted Adamic-Adar	730
Weighted Adamic-Adar	792

to much more connected nodes. However, there are two interesting anomalies. At degree 2,953, there are 2,485 nodes with that degree. At degree 3,412, there are 388 nodes with that degree.

With a total of 7,087,883 edges in the graph and 127,857 nodes, there is an average degree of 110.9 for each node. This implies a very strongly connected graph. This is the case, as the clustering coefficient is .937. This follows from how easy it is to connect two nodes. If any user commented on both posts, the nodes for the posts are connected.

4.4 Cluster Analysis

We cluster the post to post graph using the distances described earlier with the Fast Community [28] algorithm. This algorithm uses the technique of modularity maximization to quickly find non-overlapping clusters. One issue found with this algorithm was the tendency to produce only a few very large clusters. To counteract this, the algorithm is run twice. After the first round, clusters of size greater than 100 have the algorithm run on them again, reducing them to smaller clusters. We



(a) Unweighted Jaccard **(b) Weighted Jaccard**

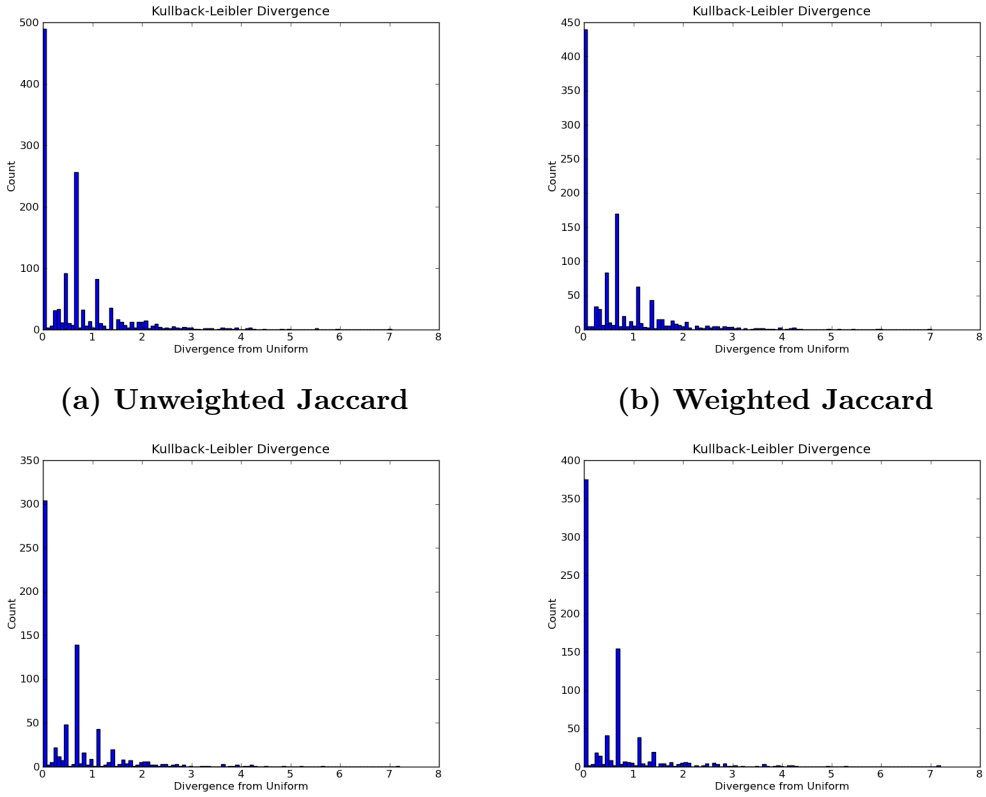
(c) Unweighted Adamic-Adar **(d) Weighted Adamic-Adar**

Figure 4.6: Distribution of cluster sizes

see the distribution of the size of clusters in 4.6. The number of clusters found for each distance type is shown in table 4.7. By taking a look at clusters by how much each subreddit is represented in them, we can learn more about the posting habits of users.

4.4.1 Difference from Uniform

Each cluster has a set of posts, and each post has a subreddit that it is part of. A uniform cluster would have a uniform distribution of subreddits represented. We measure each cluster's difference from uniform using the Kullback-Leibler divergence from a uniform distribution. We normalize each cluster so the sum of the components is 1, resulting in the formula for number of posts within the cluster being n and the number of posts within a subreddit being i:



(a) Unweighted Jaccard **(b) Weighted Jaccard**

(c) Unweighted Adamic-Adar **(d) Weighted Adamic-Adar**

Figure 4.7: Cluster deviation distribution

$$Div = \sum_{i} ln(i) \times \frac{i}{n}. \tag{4.8}$$

In Figure 4.7 we see the results of weighted and unweighted distances being used to generate the clusters. Overall, the distributions are quite similar. However, all four distance types have a large number of near-uniform clusters. If we look back at Figure 4.6 to the number of very small cluster, the number of nearly uniform clusters is around the same level. Most of the non-uniform clusters are the larger clusters. This implies that the majority of activity that encompasses multiple subreddits does so in a very non-uniform way.

CHAPTER 5
ANALYSIS OF PROMINENCE

5.1 Ground Truths

The ground truth of Reddit is Karma. Karma is a reward for contributing positively to a thread, and a punishment for contributing negatively. Each comment or post a user creates can be voted on by each other user. An upvote contributes +1 to the creator's Karma, and a downvote contributes -1. If a user in general makes positive contributions, they should have a high Karma. Likewise, if a user generally makes negative contributions, they should have a low Karma.

5.1.1 Total Karma

For this ranking, we take the user's total Karma across all of their posts and comments. If the user generally makes good posts, then this number should be high. However, most users with high total Karma have few posts, meaning they had one post that had receive a large number of upvotes.

5.1.2 Median Karma

The median Karma looks at the karma of the user's median comment. Since this requires half of the user's comments to have at least this much karma, it should provide a better measure of their overall quality. However, many users still had only one post that received many upvotes, which will rank high with this measure.

5.1.3 Average Karma

The average Karma is again a measure of overall quality. However, we again see users with few posts ranking highly. So, a measure that is immune to this is required.

5.1.4 Average Top 10

Average top 10 looks at users with at least 10 comments, and takes the average of their top ten comments. This eliminates users with less than 10 comments and favors users with many posts that have a high upvote count. Users with less than 10 comments receive a score of 0.

5.2 Degree Based Ranking

We can measure the amount of influence a user has by their general avitivy on Reddit. Using simple degree-basd measures involving their posting and commenting behavior (and the behavior of other users in response to them), we can get a measure of their prominence.

5.2.1 Number of Posts

For number of posts, we simply look at how many posts the user submitted. We might be able to see a user's prominence from how many posts they have made.

5.2.2 Number of Subreddits

For number of subreddits, we look at how many subreddits each user has commented in. A user who is active in many subreddits might be more prominent than a user who is only active in a few.

5.2.3 Number of Repliers

We look at the number of people who replied to the author's posts. A prominent user might have posts that many other users reply to.

5.3 iHypR

iHypR is an algorithm to determine prominence within a network with hyperedges. Hyperedges are groupings of objects within the network, which can be used to learn more about objects and the actors that created them within the network. It has been shown to perform well on DBLP (network of academic papers and authors in Computer Science), IMDB (network of movies and actors) and Enron Email

datasets (network of email exchange in the Enron organization), and is claimed to be a general algorithm with which to perform prominance evaluation within a network with hyperedges. Reddit has hyperedges in subreddits. If we consider users to be actors, posts to be objects, and subreddits to be hyperedges which link posts, then we arrive at a graph that can be used within iHypR. There are several varieties of iHypR, some of which ignore objects entirely, and some of which ignore the existing hyperedges and create new ones via clustering. We use the variant which uses objects, but forgoes the built in hyperedges for clusters. We instead use the clusters generated with FastCommunity. This is because the Reddit data is very strongly connected, and the distribution of posts across subreddits is strongly skewed. The set of natural hyperedges will have a few massive groups, and many small groups. Our clustering is more realistic in terms of typical user behavior. iHypR is based off the idea of social collaboration, and is useful as a social measure of prominence.

5.3.1 Modified iHypR

iHypR works in the following cycle:

```
Each User = 1/n, where n is the number of authors
while not converged:
    1. Each Post =  sum of the Users who commented on that Post
    2. Each Cluster = average of the top half of its Posts
    3. Each Post = the Cluster containing it
    4. Each User = sum of the Posts they commented on
```

In step 4, users are set equal to the sum of their posts. This favors users with a massive number of posts, even if they are not high quality. We provide a modified iHypR ranking where instead of summing the posts, we average across the top half. This removes the bias towards making many posts.

5.4 Rankings

5.4.1 Correlation

We examine the correlation between the rankings produced by all of the methods we have described.

Table 5.1: Measure abbreviations

Measure Name	Abbrev.
Total Karma	TK
Median Karma	MK
Average Karma	AK
Average of Top 10	$A10K$
Number of Posts	NP
Number of Subreddits	NS
Number of Repliers	NR
iHypR with Unweighted Jaccard	$iHypR_{UJ}$
iHypR with Weighted Jaccard	$iHypR_{WJ}$
iHypR with Unweighted Adamic-Adar	$iHypR_{UAA}$
iHypR with Weighted Adamic-Adar	$iHypR_{WAA}$
Modified iHypR with Unweighted Jaccard	$iHypR_{UJ}\_M$
Modified iHypR with Weighted Jaccard	$iHypR_{WJ}\_M$
Modified iHypR with Unweighted Adamic-Adar	$iHypR_{UAA}\_M$
Modified iHypR with Weighted Adamic-Adar	$iHypR_{WAA}\_M$

Table 5.2: Correlation table for karma based ranking

	TK	MK	AK	$A10K$
TK	**1.0**	0.394	0.658	0.985
MK	0.394	**1.0**	0.886	0.41
AK	0.658	0.886	**1.0**	0.687
$A10K$	0.985	0.41	0.687	**1.0**
NP	**0.0494**	0.00256	0.00661	0.0286
NS	**0.1**	-0.0155	-0.00274	0.0894
NR	**0.491**	0.192	0.328	0.488
$iHypR_{UJ}$	**0.0147**	-0.000117	-0.000171	0.000518
$iHypR_{WJ}$	**0.0144**	-0.000116	-0.000171	0.000507
$iHypR_{UAA}$	**0.0161**	-9.5e-05	-0.000138	0.000706
$iHypR_{WAA}$	**0.0161**	-9.5e-05	-0.000138	0.000706
$iHypR_{UJ}\_M$	0.0393	-0.0162	-0.00446	**0.0398**
$iHypR_{WJ}\_M$	0.0376	-0.0167	-0.00455	**0.0384**
$iHypR_{UAA}\_M$	0.0383	-0.0172	-0.0055	**0.0389**
$iHypR_{WAA}\_M$	0.0394	-0.0172	-0.00546	**0.0399**

Table 5.3: Correlation table for degree based ranking

	NP	NS	NR
TK	0.0494	0.1	**0.491**
MK	0.00256	-0.0155	**0.192**
AK	0.00661	-0.00274	**0.328**
$A10K$	0.0286	0.0894	**0.488**
NP	**1.0**	0.0502	0.574
NS	0.0502	**1.0**	0.123
NR	0.574	0.123	**1.0**
$iHypR_{UJ}$	**0.81**	-2.7e-05	0.417
$iHypR_{WJ}$	**0.778**	0.000853	0.4
$iHypR_{UAA}$	**0.974**	-0.000262	0.509
$iHypR_{WAA}$	**0.974**	-0.000262	0.509
$iHypR_{UJ}\_M$	0.0456	**0.296**	0.0856
$iHypR_{WJ}\_M$	0.0427	**0.301**	0.0834
$iHypR_{UAA}\_M$	0.0464	**0.308**	0.0893
$iHypR_{WAA}\_M$	0.0463	**0.312**	0.0902

Table 5.4: Correlation table for iHypR based ranking

	$iHypR_{UJ}$	$iHypR_{WJ}$	$iHypR_{UAA}$	$iHypR_{WAA}$
TK	0.0147	0.0144	**0.0161**	**0.0161**
MK	**-0.000117**	-0.000116	-9.5e-05	-9.5e-05
AK	**-0.000171**	-0.000171	-0.000138	-0.000138
$A10K$	0.000518	0.000507	**0.000706**	**0.000706**
NP	0.81	0.778	**0.974**	**0.974**
NS	-2.7e-05	**0.000853**	-0.000262	-0.000262
NR	0.417	0.4	**0.509**	**0.509**
$iHypR_{UJ}$	**1.0**	0.998	0.834	0.834
$iHypR_{WJ}$	0.998	**1.0**	0.802	0.802
$iHypR_{UAA}$	0.834	0.802	**1.0**	**1.0**
$iHypR_{WAA}$	0.834	0.802	**1.0**	**1.0**
$iHypR_{UJ}\_M$	-0.00126	**-0.00128**	-0.00107	-0.00107
$iHypR_{WJ}\_M$	-0.00145	**-0.00146**	-0.00119	-0.00119
$iHypR_{UAA}\_M$	-0.00144	**-0.00148**	-0.00115	-0.00115
$iHypR_{WAA}\_M$	-0.00147	**-0.00149**	-0.00117	-0.00117

Table 5.5: Correlation table for modified iHypR based ranking

	$iHypR_{UJ}\_M$	$iHypR_{WJ}\_M$	$iHypR_{UAA}\_M$	$iHypR_{WAA}\_M$
TK	0.0393	0.0376	0.0383	**0.0394**
MK	-0.0162	-0.0167	**-0.0172**	-0.0172
AK	-0.00446	-0.00455	**-0.0055**	-0.00546
$A10K$	0.0398	0.0384	0.0389	**0.0399**
NP	0.0456	0.0427	**0.0464**	0.0463
NS	0.296	0.301	0.308	**0.312**
NR	0.0856	0.0834	0.0893	**0.0902**
$iHypR_{UJ}$	-0.00126	-0.00145	-0.00144	**-0.00147**
$iHypR_{WJ}$	-0.00128	-0.00146	-0.00148	**-0.00149**
$iHypR_{UAA}$	-0.00107	**-0.00119**	-0.00115	-0.00117
$iHypR_{WAA}$	-0.00107	**-0.00119**	-0.00115	-0.00117
$iHypR_{UJ}\_M$	**1.0**	0.72	0.707	0.701
$iHypR_{WJ}\_M$	0.72	**1.0**	0.742	0.739
$iHypR_{UAA}\_M$	0.707	0.742	**1.0**	0.745
$iHypR_{WAA}\_M$	0.701	0.739	0.745	**1.0**

5.4.1.1 Degree Based Rankings

The degree based rankings (in Table 5.8 did not perform exceptionally well, except for the Number of Repliers. Number of Repliers correlated reasonably with the Karma based Rankings, particularly with the Total Karma.

5.4.1.2 iHypR Based Rankings

The iHypR based rankings in Table 5.9 show a strong correlation to each other (particularly between weighted and unweighted versions of the same distance), and strong correlation to the number of posts.

5.4.1.3 Modified iHypR Based Rankings

The modified iHypR rankings in Table 5.10 show very similar results as the unmodified iHypR, as well as showing an extremely strong correlation to the unmodified version of themselves.

5.4.2 Kendall-Tau

We use the Kendall-Tau measure of rank correlation [29] as another way to compare the rankings. The Kendall-Tau measure is defined as:

Table 5.6: Ties in rankings

Measure	Portion of Ties
TK	0.0684
MK	0.231
AK	0.0771
$A10K$	0.0684
NP	0.0453
NS	0.512
NR	0.00928
$iHypR_{UJ}$	0.000777
$iHypR_{WJ}$	0.00199
$iHypR_{UAA}$	0.00299
$iHypR_{WAA}$	0.00299
$iHypR_{UJ}\_M$	0.00124
$iHypR_{WJ}\_M$	0.00336
$iHypR_{UAA}\_M$	0.00471
$iHypR_{WAA}\_M$	0.00455

Table 5.7: Kendall-Tau table for karma based ranking

	TK	MK	AK	$A10K$
TK	**0.932**	0.505	0.663	0.927
MK	0.505	**0.769**	0.508	0.283
AK	0.663	0.508	**0.923**	0.656
$A10K$	0.927	0.283	0.656	**0.932**
NP	0.00872	**-0.334**	-0.0499	0.00868
NS	0.631	0.184	0.457	**0.633**
NR	-0.263	**-0.448**	-0.306	-0.263
$iHypR_{UJ}$	0.134	**-0.268**	-0.0207	0.135
$iHypR_{WJ}$	0.116	**-0.261**	-0.0222	0.117
$iHypR_{UAA}$	0.135	**-0.277**	-0.0264	0.136
$iHypR_{WAA}$	0.135	**-0.277**	-0.0264	0.136
$iHypR_{UJ}\_M$	0.103	**-0.285**	-0.0408	0.103
$iHypR_{WJ}\_M$	0.107	**-0.279**	-0.0344	0.108
$iHypR_{UAA}\_M$	0.11	**-0.275**	-0.0314	0.111
$iHypR_{WAA}\_M$	0.109	**-0.276**	-0.0333	0.109

Table 5.8: Kendall-Tau table for degree based ranking

	NP	NS	NR
TK	0.00872	**0.631**	-0.263
MK	-0.334	0.184	**-0.448**
AK	-0.0499	**0.457**	-0.306
$A10K$	0.00868	**0.633**	-0.263
NP	0.401	**0.433**	0.366
NS	0.433	**0.488**	-0.416
NR	0.366	-0.416	**0.437**
$iHypR_{UJ}$	**-0.555**	-0.262	-0.445
$iHypR_{WJ}$	**-0.551**	-0.283	-0.356
$iHypR_{UAA}$	**-0.477**	-0.258	-0.237
$iHypR_{WAA}$	**-0.477**	-0.258	-0.237
$iHypR_{UJ}\_M$	**-0.51**	-0.285	-0.371
$iHypR_{WJ}\_M$	**-0.46**	-0.282	-0.205
$iHypR_{UAA}\_M$	**-0.428**	-0.278	-0.123
$iHypR_{WAA}\_M$	**-0.433**	-0.278	-0.133

Table 5.9: Kendall-Tau table for iHypR based ranking

	$iHypR_{UJ}$	$iHypR_{WJ}$	$iHypR_{UAA}$	$iHypR_{WAA}$
TK	0.134	0.116	**0.135**	**0.135**
MK	-0.268	-0.261	**-0.277**	**-0.277**
AK	-0.0207	-0.0222	**-0.0264**	**-0.0264**
$A10K$	0.135	0.117	**0.136**	**0.136**
NP	**-0.555**	-0.551	-0.477	-0.477
NS	-0.262	**-0.283**	-0.258	-0.258
NR	**-0.445**	-0.356	-0.237	-0.237
$iHypR_{UJ}$	**0.99**	0.574	0.65	0.65
$iHypR_{WJ}$	0.574	**0.989**	0.578	0.578
$iHypR_{UAA}$	0.65	0.578	**0.988**	**0.988**
$iHypR_{WAA}$	0.65	0.578	**0.988**	**0.988**
$iHypR_{UJ}\_M$	0.698	0.519	**0.708**	**0.708**
$iHypR_{WJ}\_M$	0.623	0.579	**0.726**	**0.726**
$iHypR_{UAA}\_M$	0.63	0.548	**0.818**	**0.818**
$iHypR_{WAA}\_M$	0.632	0.55	**0.725**	**0.725**

Table 5.10: Kendall-Tau table for modified iHypR based ranking

	$iHypR_{UJ}\_M$	$iHypR_{WJ}\_M$	$iHypR_{UAA}\_M$	$iHypR_{WAA}\_M$
TK	0.103	0.107	**0.11**	0.109
MK	**-0.285**	-0.279	-0.275	-0.276
AK	**-0.0408**	-0.0344	-0.0314	-0.0333
$A10K$	0.103	0.108	**0.111**	0.109
NP	**-0.51**	-0.46	-0.428	-0.433
NS	**-0.285**	-0.282	-0.278	-0.278
NR	**-0.371**	-0.205	-0.123	-0.133
$iHypR_{UJ}$	**0.698**	0.623	0.63	0.632
$iHypR_{WJ}$	0.519	**0.579**	0.548	0.55
$iHypR_{UAA}$	0.708	0.726	**0.818**	0.725
$iHypR_{WAA}$	0.708	0.726	**0.818**	0.725
$iHypR_{UJ}\_M$	**0.989**	0.755	0.745	0.745
$iHypR_{WJ}\_M$	0.755	**0.988**	0.773	0.769
$iHypR_{UAA}\_M$	0.745	0.773	**0.987**	0.793
$iHypR_{WAA}\_M$	0.745	0.769	0.793	**0.987**

$$\tau(x, y) = \frac{|\text{Accordant Pairs}| - |\text{Discordant Pairs}|}{|\text{All Pairs}|}. \tag{5.1}$$

Unlike correlation, Kendall-Tau only considers the difference in the sorting of the rankings, not the magnitude of the difference. We modify the Kendall-Tau to penalize ties in the same way as [13]. However, we choose our penalties differently. Instead of fixing them as constants, we calculate the penalties for a comparison between rankings R_1 and R_2 as:

$$p_1 = \frac{t(R_1)}{t(R_1) + t(R_2)} \tag{5.2}$$

and

$$p_2 = \frac{t(R_2)}{t(R_1) + t(R_2)} \tag{5.3}$$

where $t(R)$ is the portion of pairs of objects within the ranking R that are ties. Table 5.6 shows us the tie proportions for the measures used. We see that all iHypR measures have a very low amount of ties, while Average of Top 10 and Number of Subreddits have very high tie counts. Average Top 10 assigns a score of zero to all users with less than 10 comments, resulting in many ties. Many users only post in

one subreddit, resulting in the high tie count for Number of Subreddits. The other Karma and Degree based measures have low tie proportions.

CHAPTER 6
DISCUSSION AND CONCLUSIONS

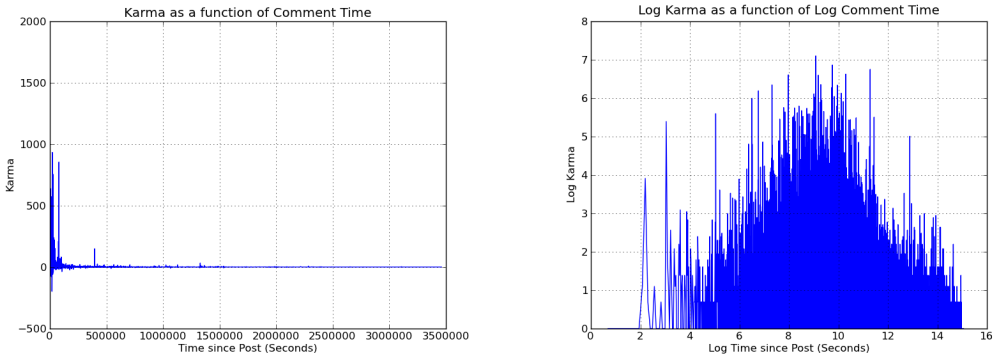
We find that Reddit portrays behaviors similar to other social networks. However, we see that both the correlation coefficient and the Kendall Tau comparison do not see any strong relationship between iHypR and the Karma based ground truths. iHypR, in this case, cannot predict how much Karma a user has. There are aspects of Reddit that set it apart from other networks which might account for this.

6.1 Karma is a Flawed Ground Truth

Karma is a highly volatile number. A single post can give a user more Karma than most other users will ever get. This does not even appear to be a matter of timing. In Figure 6.1, we see that the high-Karma comments are made at any time. There is a slight peak around two hours after the post, but if there was something to be gained by timing the system, the gap would not be nearly this large.

6.2 Reddit is not Collaborative in General

iHypR is focused on collaborations. One dataset where it performs well is author collaboration on academic papers. Here, authors are required to collaborate



(a) **(b)**

Figure 6.1: Karma over time

and must approve anyone who collaborates with them. Reddit, however, has no such limitations. Anyone who wants to comment can. In an academic paper, users must contribute meaninfully continually or risk no longer being allowed to collaborate. On Reddit, comments that are heavily downvoted are not shown by default, but there is no impact on a user's other comments. So, there is no global penalty to not collaborating meaningfully.

A way around this would be to look deeper into the nature of Reddit. Commonly, discussions will start in the comments, resulting in a "thread" of replies between two users, where few others will post. Also common is a similar structure of singular replies, but with many users posting. These structures show collaboration and actual discussion, rather than just replies. Using the presence and size of these "threads" could perhaps give a better ranking. Examining the impact of these factors is topic of future work.

CHAPTER 7
BIBLIOGRAPHY

[1] B. Zemljia and V. Hlebec, "Reliability of measures of centrality and prominence," *Social Networks*, vol. 27, no. 1, p. 7388, 2005.

[2] K. Faust and S. Wasserman, *Social Network Analysis: Methods and Applications*. Cambridge, MA: Cambridge University Press, 1994.

[3] J. Zheng, E. Veinott, N. Bos, J. S. Olson, and G. M. Olson, "Trust without touch: Jumpstarting long-distance trust with initial social activities," in *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, ser. CHI '02. New York, NY, USA: ACM, 2002, pp. 141–146. [Online]. Available: http://doi.acm.org/10.1145/503376.503402 [Accessed: Nov. 2012]

[4] M. Nascimento, J. Sander, and J. Pound, "Analysis of sigmod's co-authorship graph," *SIGMOD Rec.*, vol. 32, no. 3, pp. 8–10, Sep. 2003. [Online]. Available: http://doi.acm.org/10.1145/945721.945722 [Accessed: Dec. 2012]

[5] B. Suh, L. Hong, P. Pirolli, and E. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," in *2010 IEEE Int. Conf. on Privacy, Security, Risk, and Trust, and IEEE Int. Conf. on Social Computing*, Ottawa, Canada, 2010, pp. 177–184.

[6] C. Honey and S. Herring, "Beyond microblogging: Conversation and collaboration via twitter," in *HICSS'09. 42nd Hawaii Int. Conf. on System Sciences*, Hawaii, USA, 2009.

[7] K. Lerman and A. Galstyan, "Analysis of social voting patterns on digg," in *Proc. 1st Workshop on Online Social Networks*, ser. WOSN '08. New York, NY, USA: ACM, 2008, pp. 7–12. [Online]. Available: http://doi.acm.org/10.1145/1397735.1397738 [Accessed: Dec. 2012]

[8] K. Lerman. (2006) Social networks and social information filtering on digg. [Online]. Available: http://arxiv.org/abs/cs/0612046

[9] B. T. Adler, L. de Alfaro, I. Pye, and V. Raman, "Measuring author contributions to the wikipedia," in *Proc. 4th Int. Symp. on Wikis*, ser. WikiSym '08. New York, NY, USA: ACM, 2008, pp. 15:1–15:10. [Online]. Available: http://doi.acm.org/10.1145/1822258.1822279 [Accessed: Dec. 2012]

[10] L. Knuttila, "User unknown: 4chan, anonymity and contingency," *First Monday*, vol. 16, 2011.

[11] C. Lampe, E. Johnston, and P. Resnick, "Follow the reader: Filtering comments on slashdot," in *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, ser. CHI '07. New York, NY, USA: ACM, 2007, pp. 1253–1262. [Online]. Available: http://doi.acm.org/10.1145/1240624.1240815 [Accessed: Nov. 2012]

[12] N. Poor, "Mechanisms of an online public sphere: The website slashdot," *J. of Computer-Mediated Communication*, vol. 10, no. 2, pp. 00–00, 2005. [Online]. Available: http://dx.doi.org/10.1111/j.1083-6101.2005.tb00241.x [Accessed: Dec. 2012]

[13] S. Adalı, X. Lu, and M. Magdon-Ismail, "ihypr: Prominence ranking in networks of collaborations with hyperedges," *Trans. on Knowledge Discovery from Data (to appear)*, 2013.

[14] V. Gomez, A. Kaltenbrunner, and V. Lopez, "Statistical analysis of the social network and discussion threads in slashdot," in *World Wide Web 2008*, Beijing, China, 2008, pp. 645–654.

[15] M. S. Bernstein, A. Monroy-Hernndez, D. Harry, P. Andr, K. Panovich, and G. Vargas, "4chan and /b/: An analysis of anonymity and ephemerality in a large online community," in *Proc. 5th Int. Conf. on Weblogs and Social Media*, Barcelona, Spain, 2011, pp. 50–57.

[16] R. Zhang and T. Tran, "Helping e-commerce consumers make good purchase decisions: A user reviews-based approach," in *E-Technologies: Innovation in an Open World*, ser. Lecture Notes in Business Information Processing, G. Babin, P. Kropf, and M. Weiss, Eds. Springer Berlin Heidelberg, 2009, vol. 26, pp. 1–11. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-01187-0_1 [Accessed: Dec. 2012]

[17] V. Belk, S. Lam, and C. Hayes, "Cross-community influence in discussion fora," in *Proc. 6th Int. Conf. on Weblogs and Social Media*, Dublin, Ireland, 2012, pp. 34–41.

[18] R. Bhatt and K. Barman, "Global dynamics of online group conversations," in *Proc. 6th Int. Conf. on Weblogs and Social Media*, Dublin, Ireland, 2012, pp. 403–406.

[19] D. Morrison, I. McLoughlin, A. Hogan, and C. Hayes, "Evolutionary clustering and analysis of user behaviour in online forums," in *Proc. 6th Int. Conf. on Weblogs and Social Media*, Dublin, Ireland, 2012, pp. 519–522.

[20] U. Gargi, W. Lu, V. Mirrokni, and S. Yoon, "Large-scale community detection on youtube for topic discovery and exploration," in *Proc. 5th Int. Conf. on Weblogs and Social Media*, Barcelona, Spain, 2011, pp. 486–489.

[21] K. S. Dave, R. Bhatt, and V. Varma, "Modelling action cascades in social networks," in *Proc. 5th Int. Conf. on Weblogs and Social Media*, Barcelona, Spain, 2011, pp. 121–128.

[22] Python reddit api wrapper. [Online]. Available: https://github.com/praw-dev/praw [Accessed: Sept. 2012]

[23] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: Understanding microblogging usage and communities," in *Proc. 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, San Jose, California, USA, 2007, pp. 56–65.

[24] B. D. Alan Ritter, Colin Cherry, "Unsupervised modeling of twitter conversations," in *Proc. of HLT-NAACL*, Los Angeles, CA, USA, 2010, pp. 172–180.

[25] L. Hong, G. Convertino, and E. Chi, "Language matters in twitter: A large scale study," in *Proc. 5th Int. Conf. on Weblogs and Social Media*, Barcelona, Spain, 2011, pp. 518–521.

[26] P. Jaccard, "Etude comparative de la distribution florale dans une portion des alpes et des jura," in *Bulletin de la Socit Vaudoise des Sciences Naturelles 37*, 1901, p. 547579.

[27] L. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, pp. 211–230, 2003.

[28] A. Clauset, M. Newman, and C. Moore, "Finding community structure in very large networks." *Phys. Rev. E*, vol. 70, no. 6, p. 066111, 2004.

[29] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938. [Online]. Available: http://biomet.oxfordjournals.org/content/30/1-2/81.short [Accessed: Nov. 2012]

[30] Y. Yang, N. V. Chawla, X. Lu, and S. Adalı, "Prominence in networks: A co-evolving process," in *Proc. IEEE Network Science Workshop (to appear)*, West Point, NY USA, 2013.

[31] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proc. 20th Int. Conf. on World Wide Web*, Lyon, France, 2011, pp. 675–684.

[32] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proc. 19th Int. Conf. on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 851–860.

[33] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *HICSS-43. IEEE, January 6.*, Hawaii, USA, 2010.

[34] J. Huang, K. Thornton, and E. Efthimiadis, "Conversational tagging in twitter," in *Proc. 21st ACM Conf. on Hypertext and Hypermedia*, Toronto, Canada, 2010, pp. 173–178.

[35] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *CEAS 2010 - 7th Annu. Collaboration, Electronic Messaging, Anti- Abuse and Spam Conf.*, Redmond, Washington, USA, 2010.

[36] A. Wang, "Don't follow me: Spam detection in twitter," in *Int. Conf. on Security and Cryptography (SECRYPT)*, Athens, Greece, 2010, pp. 1–10.

[37] L. Graham and S. Gosling, "Can the ambiance of a place be determined by the user profiles of the people who visit it?" in *Proc. 5th Int. Conf. on Weblogs and Social Media*, Barcelona, Spain, 2011, pp. 145–152.

[38] A. Pal, S. Chang, and J. A. Konstan, "Evolution of experts in question answering communities," in *Proc. 6th Int. Conf. on Weblogs and Social Media*, Dublin, Ireland, 2012, pp. 274–281.

[39] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," in *Proc. 6th Int. Conf. on Weblogs and Social Media*, Dublin, Ireland, 2012, pp. 26–33.

[40] P. Agarwal, R. Vaithiyanathan, S. Sharma, and G. Shroff, "Catching the long-tail: Extracting local news events from twitter," in *Proc. 5th Int. Conf. on Weblogs and Social Media*, Dublin, Ireland, 2012, pp. 379–382.

[41] F. Shah and G. R. Sukthankar, "Using network structure to identify groups in virtual worlds," in *Proc. 5th Int. Conf. on Weblogs and Social Media*, Barcelona, Spain, 2011, pp. 614–617.

[42] L. A. Adamic, D. Lauterbach, C.-Y. Teng, and M. Ackerman, "Rating friends without making enemies," in *Proc. 5th Int. Conf. on Weblogs and Social Media*, Barcelona, Spain, 2011, pp. 2–9.

[43] S. Asur, B. A. Huberman, G. Szabo, and C. Wang, "Trends in social media: Persistence and decay," in *Proc. 5th Int. Conf. on Weblogs and Social Media*, Barcelona, Spain, 2011, pp. 434–437.

[44] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *Proc. 4th Int. Conf. on Weblogs and Social Media*, Washington, D.C., 2010, pp. 2–9.