INFORMATION AND SOCIAL ANALYSIS OF REDDIT

TROY STEINBAUER – TROYSTEINBAUER@CS.UCSB.EDU

ABSTRACT

Social news websites have the ability to crowd-source the identification and promotion of news articles with the ability for users to discuss topics. Like Online Social Networks (OSNs), social news sites are information rich and contain a new underlying structured network. Reddit is a rising social news website which shares many characteristics to other popular social news websites include StumbleUpon, Digg, and Slashdot, yet has never been analyzed at a high or low level. This paper will provide a broad analysis of Reddit, starting with the initial data crawling, as well as in depth analysis of underling social structures between the article subsections and the users of the site, focusing primarily on the comment section where the bulk of the data exists.

1 INTRODUCTION

Reddit is a social news website driven by user content. The content is either a link to another webpage or a text blob accompanied by a title. Each submission is voted upon by users of the site, either up or down, to rank the submission. Users can also discuss the submission in a threaded comment section attached to each submission.

Reddit is a data-rich website with a complex network structure. Each subreddit is functionally independent of each other, yet sometimes share common submissions and even cross-posting happens. The users of the site drive the content and the rating of the content, along with the large discussions of every article.

Reddit's structure is most similar to the social news site Digg, sharing almost all features except for the visibility of a user's actions. StumbleUpon has the comments and voting aspects, but you do not browse StumbleUpon, you are presented with a random submission from sections you subscribe to. Slashdot, one of the oldest social news sites, relies on moderators to filter and select user submissions to display online. These four social new sites all contain a threaded comment section.

The first goal of this paper is to collect a large dataset from Reddit due to no dataset is currently available. This paper will focus on the popular sections of Reddit in order to find more user activity, and thus more interesting data. The second step in the paper is to analyze Reddit at a high level. This will start with the subreddits and drill down into the submissions and comments.

1.1 About Reddit

Reddit was founded in June 2005 by Steve Huffman and Alexis Ohanian and is funded though both advertising as well as paid subscriptions to Reddit Gold. Reddit became an open source project, hosted on GitHub, on June 18th, 2008. Currently Reddit employs 10 full time workers. The site is written in Python using the Pylons web framework (Reddit, Frequently Asked Questions). Reddit's traffic continues to rise in recent years, approaching the 100th most visited site on the web globally and in the top 50 in the US (Alexa).

Reddit's front page is where a user's favorite, or default selection, of subredddits are displayed. A subreddit is a community portal where individual submissions are placed in. Subreddits themselves are also user created and cover a wide range of topics. Viewing a subreddit displays a list of submissions ordered either by submission date, popularity, controversially, or overall highest scoring in a time period. Each submission has a text title, a score based off of user up and down votes, either a link to a webpage or a text description, and a threaded comment section for users to discuss the reddit. The comment section is where users can discuss the submission and vote upon other comments, allowing popular comments to be filtered to the top.

2 Prior Work

The comment section of social news sites is not new to researchers. In "Statistical Analysis of the Socially Network and Discussion Threads in Slashdot" (Vicenc Gomez, 2008), researchers dove into that properties of the discussion section. The paper analyzes the user social graph derived out of the comment section of posts in a very similar manor to what this paper aims to do. According to the paper, "Slashdot presents common features of traditional social networks, namely, a giant cluster of connected users, small average path length and high clustering." Due to Slashdot and Reddit's similarities, it can be expected for Reddit's user graph generated from the comment section to also follow traditional social network structure. This paper identified connected users in the comment section in three ways,

In Jamali and Rangwala's paper "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis", the comment section is investigated and related to the popularity of a submission (Salman Jamali, 2009). This paper uses a slightly different metric than the Slashdot paper for identifying edges between users. This paper identifies two users are connected via an undirected edge when both users comment on the same submission four unique times or more. This threshold of four was chosen for the analysis sections of the paper, but they did mention that other thresholds could be used. This user graph was used to compare the eight different subsections of Digg and how user relationships look. It found that the popular categories, like sports and technology, not only had more activity but also more social connections. The World Business section was the least connected of all sections, even lower than the Off Topic, or Other, section in Digg.

3 DATA COLLECTION

3.1 OVERVIEW

Currently no large Reddit datasets exist. Thus a key part of this project is to crawl Reddit and retrieve the content. There are three parts to crawling Reddit:

1. Get list of subreddits
2. Get list of submissions for each subreddit
3. Get content and comments for each submission

The first step is to obtain a listing of all the different subreddits. Since any user can start a subreddit, there are close to 25,000 subreddits with many of them empty or non-active. Each subreddit contains information about itself, often including links to related subreddits.

The second step is to obtain a list of submissions for each subreddit. This listing includes most of the submission's metadata except for the comments attached to it. Due to time constraints and the scope of this project, only the top 1% most popular and active subreddits will be crawled. This will turn out to be around 250 subreddits, ordered in popularity by Reddit itself.

The final step is to obtain the submission and its comments. Reddit's threaded comment system provides the most interesting data. Each comment contains the user, a timestamp, a list of replies in the form of comments, and the score (up votes minus down votes) for that individual comment. It is important to note that while Reddit exposes the number of up votes and down votes for a comment and submission, these numbers have been fuzzed. The difference, or score, is the actual value though.

3.2 CRAWLER DESIGN AND IMPLEMENTATION

A distributed, URL queue based web crawler will be implemented for crawling Reddit. The individual crawler's logic follows a few simple steps similar to most crawlers:

1. Get next URL from queue
2. Download and save content at URL
3. Put new URLs from content onto queue
4. Repeat

Utilizing Reddit's API avoids HTML based page scraping. Most all of Reddit's publicly facing pages exposes its content in multiple ways. Appending a simple ".json" to most any page on Reddit returns the main part of the page in a JSON format, easily parsed and stored. The seed URL for the first part of the crawl, getting the list of subreddits, is reddit.com/subreddits/. The data obtained in part one become the seed for part two, and similarly part two for part three. A minor optimization utilized to minimize the number of requests required for crawling was Reddit's ability to increase the default

return of 25 items in a list to a maximum of 100. This proved to be most useful for part one and two where the results we paginated.

The crawler is written in Python utilizing the Pyro library. Pyro allows the sharing of python objects across the network. This will allow a synchronized Queue to be used to provide a simple URL server to the distributed crawlers. Python's Queue class will allow the storing of arbitrary objects, making the development very flexible and rapid.

The majority of the crawl was executed on UCSB's Engineering Instructional Computing (ECI) lab computers. They all have access to a shared storage where each page's data is stored individually.

3.3 Reddit Fights Back

The crawler had to evolve over time due to Reddit's aggressive rate limit and anti-crawling techniques implemented. If requested at a rate to fast, the IP address of an individual crawler would be blacklisted and then be denied any further requests for a designated time period. This was very common during the development of the crawler as no code was in place to limit its requests. In the end, the final crawler makes once request per minute in order to not impact Reddit in a negative way.

Reddit naturally prevents crawling in a number of ways. For example, a non-authenticated session (no user logged in) cannot browse the subreddit ordered by date. An authenticated user can browse a subreddit by date, but the distance you can browse back is capped. Thus it is inherently not possible to get a complete crawl of the site. This single factor caused a shift in focus of the paper when in the experimental phase due to the inability of obtaining ordered submission data.

3.4 Difficulties Along The Way

Crawling Reddit proved to be a must larger task then initially assumed. One thing that was unavoidable and was not planned for was reddit.com actually being down. This killed more than one attempt at crawling since the crawler was unable to discern the difference between reddit.com being down and a bad URL. Thus large chunks of data were missing in the middle of the crawl.

A second circumstance that was uncontrollable was the reliability of the machines used for crawling. Since they were a shared resource, their average uptime was very short and machines would periodically get restarted. This led to the termination of crawlers, slowing down the overall rate. More than once overnight half of the machines would be restarted and would cause a setback in crawl completion. Because of this, the URL serving machine, the central part of the distributed crawler, was run on a machine not part of the cluster and not a shared resource.

3.5 Crawling Results

Table 1 below displays some statistics on the pure data crawled.

	Subreddits	Listings	Submissions
Disk Size	21 MB	640 MB	4.3 GB
Requests	243	2,348	257,624
Entries	24,298	227,792	3,851,223

Table 1 – Crawled data statistics.
Note: Listings only shows the 1% of subreddits crawled.

Due to the compact and dense format that JSON provides, the results take up much less disk space than an HTML based crawl would. A simple comparison of the HTML version versus the JSON version shows a factor of four times the saving in size for the same raw data. Not only does it save more space, the JSON data is naturally easier to work with when moving to the Analysis section.

4 Subreddits Analysis

The listing of subreddits provides the highest level view of Reddit. Table 2 below shows some general statistics of the crawl. Note the large difference between the largest amounts of subscribers to a single subreddit versus the average subreddit. This is a clear indication that an overwhelming majority of subreddits are not followed.

Count Subreddits	24,298
Max Subscribers	1,140,436
Avg Subscribers	1077
S.D. Subscribers	23,514

Table 2 - Subreddits

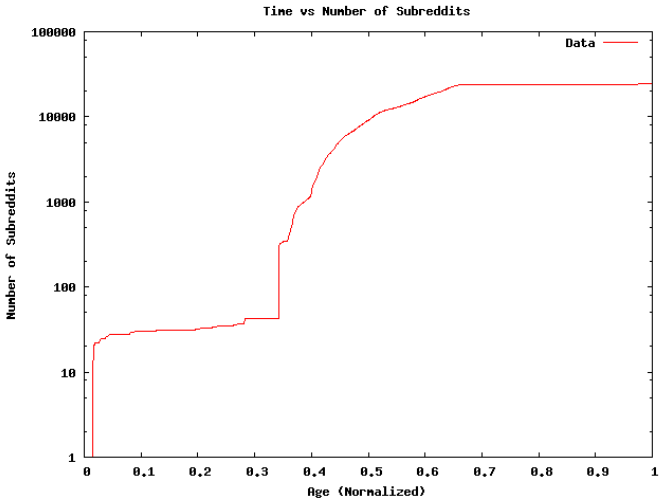

Figure 1 – Subreddits versus Age

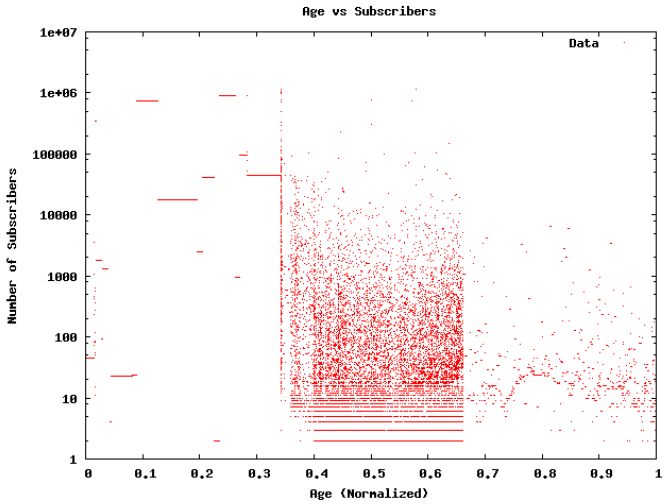

Figure 2 – Subscriptions versus Age

The two figures on the previous page relate the subreddits to their creation time. Figure 1 shows the growth of subreddits over time. Figure 2 shows the number of subscriptions versus its age. The x-axis is time normalized between the oldest and newest subreddit creations.

These two graphs show very interesting properties about the distribution of subreddits. From Figure 1 it can be seen that there was an initial group of subreddits. Then about 1/3 into the timeline there was a huge surge of subreddits created. This is due to the early 2008 change to Reddit, opening the creation of subreddits to the public (Reddit, Reddit Blog, 2008). The number of new subreddits created decreases over time until a sharp cutoff most likely due to a new rule limiting the rate in which subreddits are created.

Subreddits which have the largest subscriber base tend to be older. Figure 2 shows that the newer a subreddit is the less likely it will have a large amount of subscribers. This is very common and shows that it is difficult to compete with existing giants.

4.1 Related Subreddit Graph

Many subreddits include links to related subreddits in their description. For example, /r/comics references /r/webcomics. These links are placed there by the subreddit's monitors and are usually picked based on the content of the subreddit as well as the subreddit's user base. Using these links to form the edges of a graph with the subreddits themselves as nodes, a simple graph can be formed. It is assumed that all edges are undirected, and that if subreddit A states it is related to subreddit B, then B must also be related to A.

Number of Nodes	27,091
Number of Edges	9,050
LCC Nodes	4,582
LCC Edges	8,778
LCC Diameter	14
Largest Degree	108
Average CC	0.038566
Triangles	5,270

Table 3 – Related subreddit graph statistics

Table 3 above shows the graph properties the undirected graph formed out of related subreddit links. The graph itself is broken up into many components, with the largest component only having about 25% of the nodes. This is signs that the overall graph is not tightly clustered like most social networking graphs. The largest component, however, does show properties of a social network with its greater clustering coefficient and an overall larger degree distribution following powerlaw properties.

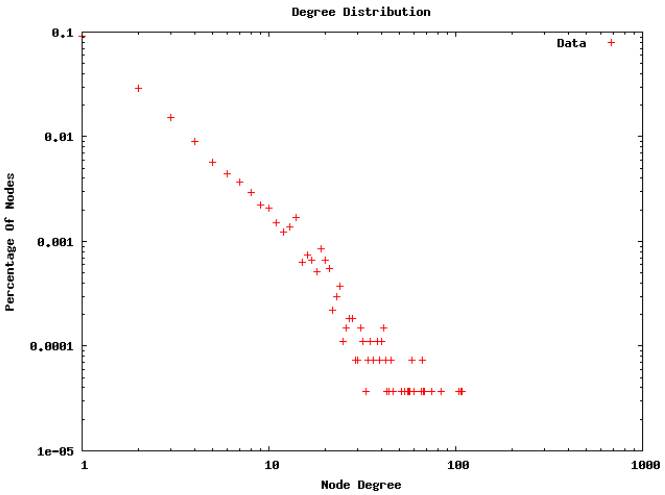

Figure 3 – Related Subreddit Distribution

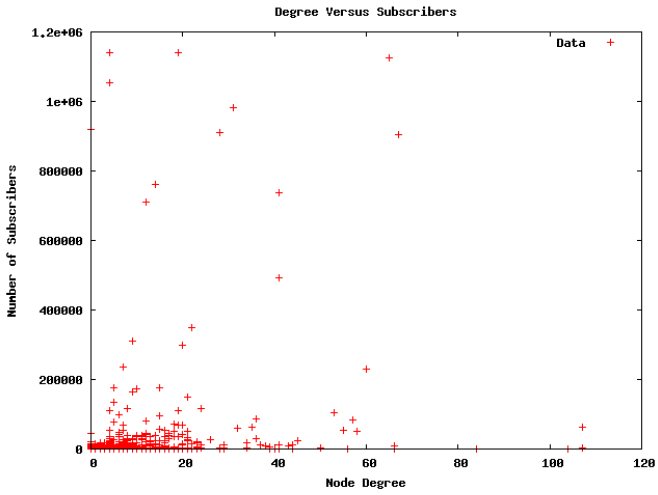

Figure 4 – Degree versus Subscribers

below shows the distribution of node degree on the related subreddit graph. As you can see, the vast majority of subreddits do not contain links to other subreddits inside their description. Of those who do have links, they tend to have many.

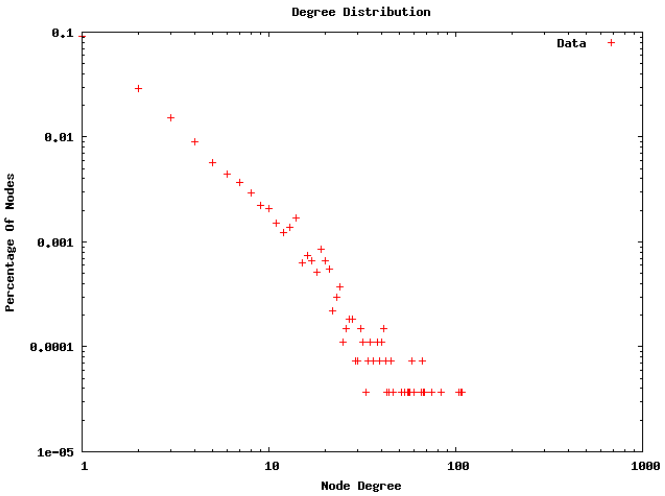

Figure 3 – Related Subreddit Distribution

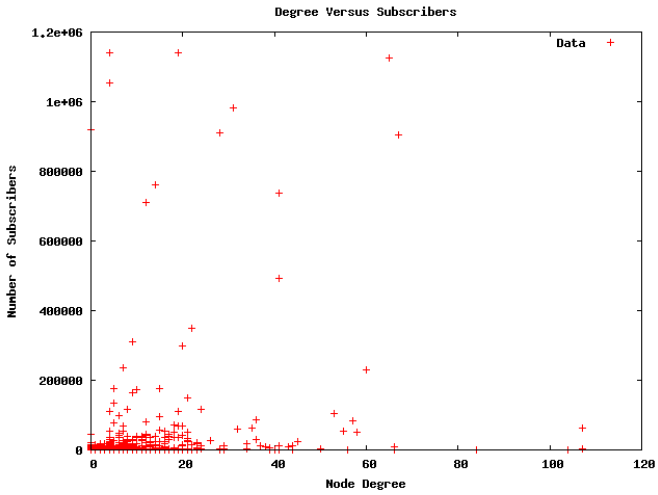

Figure 4 – Degree versus Subscribers

Figure 4 above shows the average number of edges a subreddit has relative the number of subscribers. This figure was not presented on a log-log scare in order to empisize the distribution of number of subscribers. On average, the larger subscribed subreddits are not well connected in the graph and have lower related links.

Next, the subreddits with the largest degrees and lower subscription counts were inspected by hand in order to identify any trends. Many of the well-connected subreddits turned out to be either states, including /r/california and /r/ohio, as well as cities, like /r/bosten/ and /r/georgia/. This turns out because most of the subreddits devoted to a specific city also reference the state that they are in and other nearby and major citys. All of these city and state subreddits however have less than ten thousand subscribers. A second category of highly connected subreddits with fewer subscribers are subreddits devoted to broader, non-popular categories. For example, the top two most well connected subreddits /r/listentothis and /r/misc both are highly linked from other categories, but are not specific enough to draw subscribers.

5 Comment Tree Analysis

The comment section in Reddit can be viewed as a directed acyclic graph, or more specifically, a tree with the root node the submission itself. While each reddit's comment tree is unique, many common properties appear to reoccur. For example, the higher score a comment has, the more likely it is to have many replies. Many factors can be analyzed including the height of the tree, the number of leaf nodes, average span, and many more. These statistics can be then compared to other social news sites' comment systems which also follow a simple tree structure.

5.1 Digging Into The Comments

The average submission on Reddit had fewer than 10 comments, with 95% of all submissions having less than 100. Not clearly visible in Figure 5 below is that the most comments on a single submission were approaching to 1600. From past personal experience though, the author has seen greater than 5000 comments on a single submission. Other properties including maximum width and depth were found to have similar distributions to the overall tree size. These values had a direct relationship with the distribution of the tree size.

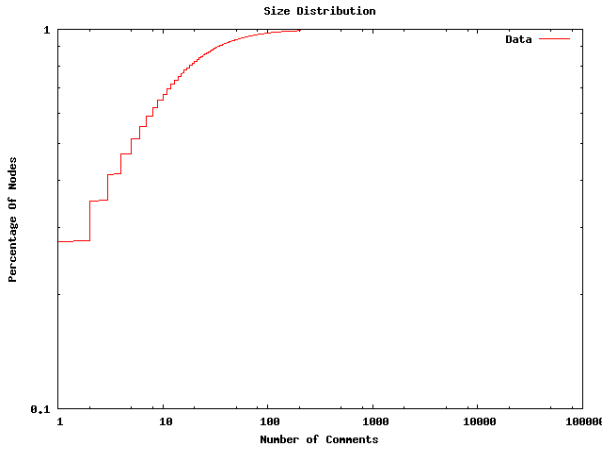

Figure 5 – Tree Size Distribution

5.2 Temporal Properties

One interesting property is the rate comments are added to a submission. Due to time stamped data, it is possible to calculate the distribution of comments since the original post occurred. Figure 6 below shows the time elapsed since submission to comment post versus the percentage of comments at that same time.

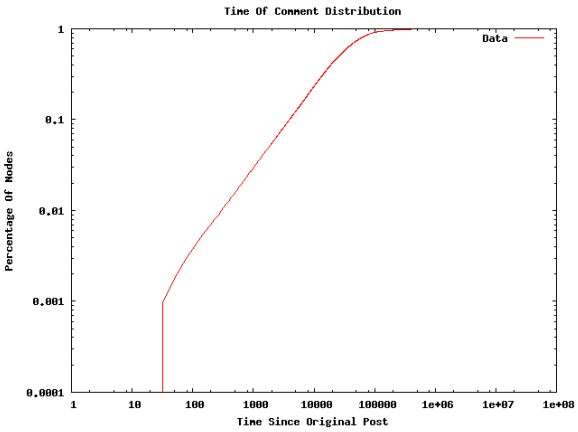

Figure 6 – Comment time distribution

Figure 6 shows when a comment was posted relative to the original submission time. There is an initial wall before the first post can occur. It can be seen the 90% of comments occur within the first 24 hours of a submission, with the rate of additional posts dropping off rapidly after. This is most likely due to the constant flow of new articles and the average daily active users. Actively participating users will view Reddit more frequently, thus seeing newer articles. It can be assumed that the majority of users will not view the same post multiple times unless actively engaged in the comment section.

6 User Comment Graph Analysis

While users in Reddit have the ability to 'friend' one another, it is more structured like Twitter's follow rather than a classical bidirectional online social network friending. When you friend another user, their names will be highlighted in posts to help you identify comments and submissions by them. Information about who a user's friends are, sadly, is not public information. A user's up and down vote history can be publicly available, but it is private by default.

A social graph will be constructed out of the comment section of each submission. In this paper, three different rules will be used to construct three different graphs from the comment section which relate closely to the Slashdot and Digg methods for extracting graphs. Each graph will be of increasing strictness in defining the relationship requirements. Note: All graphs ignore the root, or original poster, of the comment tree and only operate with bidirectional relationships.

- Loose Graph – User A commenting on User B is undirected edge between A and B
- Tight Graph – User A commenting on User B and user B commenting on User A is an undirected edge between A and B
- Strict Graph – User A commenting on User B four times and user B commenting on User A four times is an undirected edge between A and B

The action of user A commenting on user B's submission or comment will be interpreted as a directed edge from A to B. This in turn is essentially the directed tree structure with all link directions inverted. Properties of the Reddit user graph are seen in Table 4 below.

	Loose Graph	Tight Graph	Strict Graph
Number of Nodes	317316	176385	13802
Number of Edges	1804974	435413	13251
Components	603	5694	2876
LCC Nodes	316219 (99.65%)	166730 (94.53%)	7142 (51.75%)
LCC Edges	1804327 (99.96%)	428952 (98.52%)	9003 (67.94%)
Largest Degree	25736	8665	525
Average CC	0.049533	0.019251	0.014971

Table 4 – User Graphs Statistics

The distribution of user degrees can be seen in Figure 7 below.

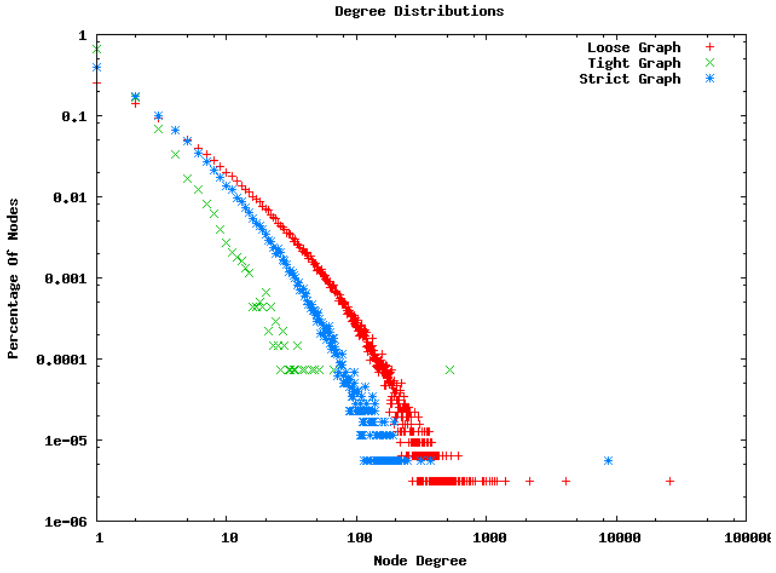

Figure 7 – Degree Distribution of the three user graphs constructed from comment interactions

7 ADDITIONAL WORK

There is a large amount of data located in the Reddit crawl, and was not fully utilized in this paper due to time constraints. Starting at the high level, the content of the submissions as well as the distribution of active users can be used to identify related subreddits. This could also be used to generate recommendation algorithms based on which subreddits you currently participate in.

While temporal properties were touched upon, relating these to the creation of links in the user graph would provide insight into the creation of the graph. The content of the posts and comments were also largely ignored.

8 CONCLUSION

This paper has provided the first dataset as well as an overview into Reddit's structure. Starting with the subreddits, we identified standard social network properties and interesting relationships between subreddits and their popularity. The comment section was analyzed and three distinct social graphs were created out of it.

9 WORKS CITED

Alexa. (n.d.). *Reddit.com Site Info*. Retrieved December 6, 2011, from Alexa - The Web Information Company: http://www.alexa.com/siteinfo/reddit.com

Reddit. (2008, March 12). *Reddit Blog*. Retrieved December 8, 2012, from Reddit - : http://blog.reddit.com/2008/03/make-your-own-reddit.html

Reddit. (n.d.). *Frequently Asked Questions*. Retrieved December 1, 2011, from Reddit - The Front Page of the Internet: http://www.reddit.com/help/faq#NerdTalk

Salman Jamali, H. R. (2009). Digging Digg : Comment Mining, Popularity Prediction, and Social Network Analysis. *International Conference on Web Information Systems and Mining.* IEEE.

Vicenc Gomez, A. K. (2008). Statistical Analysis of the Social Network and Discussion Threads in Slashdot. *WWW'08: Proceeding of the 17th International Conference on World Wide Web* (pp. 645-654). New York: ACM.