Popularity of content leads to its duplication on Reddit

Suren Nihalani **Revant Kumar** **Ashwini Khare** **Prajwal Prasad**

Abstract

On social network sites like Reddit, users submit links or share textual content. The community upvotes (for) and downvotes (against) and reddit uses community sourcing as a means to order content. The contributed are scored that is number of upvotes minus downvotes. The popularity of a user is represented by his/her karma. In this project, we have tried to judge how the popularity of contents leads to its duplication on Reddit. Users can re-post the popular posts on various other subreddits in order to receive more upvotes, and thus increasing their karma. Here, we are looking users in two ways, one being the initial person broadcasting the link trying to obtain the critical mass and second one being people reposting content. The results from this project will depict how the users satisfy their Reddit Karma and the design implications of this in online communities.

1 Objectives

In this project, we have looked at the following interesting phenomenon:

$$Karma\ in\ one\ reddit \quad \propto \quad repost$$

Thus, we have mainly focusing on to depict how the users satisfy their Reddit Karma and the design implications of this in online communities. We have correlated a post's score with its diffusion into other subreddits. We have tried to answer the questions such as – How often do links get reposted?, How many times does a user try to broadcast a link initially? We are trying to see whether there is a relationship between karma and repost frequency.

2 Background Study

As social networks and the user-generated content that populates them continue to grow in prevalence, size, and influence, understanding how users interact and produce this content becomes increasingly important. Insight into these community dynamics could prove valuable for measuring content trust, providing role-based group recommendations, or evaluating group stability and growth. For the related work, we have done readings of three papers, namely, Eric Gilbert's - "Widespread Under Provision on reddit", Cody Buntain's - "Identifying social roles in reddit using network structure" and Donn Morrison's - "Here, have an upvote: communication behaviour and karma on Reddit".

Online communities rely on their members to do work for the good of everyone on the site. The links with the most up-votes bubble up to the main page, pointing everyone toward the best content. However, when too many people rely on others to contribute without doing so themselves results in underprovision. Gilbert observed that Reddit overlooked 52% of the most popular links the first time they were submitted [1]. This suggests that many potentially popular links get ignored, thus jeopardizing the site's core purpose of showing the best voted content.

Cody Buntain has identified well-known behavioral patterns and social roles in the multi-community environments. In his paper, he has demonstrated the presence and identifiability of the answer-person role in reddit and showed that only a very small number of users participate across community boundaries [2].

Donn Morrison has tried to cluster the users of Reddit.com into behavioural roles using features derived from their egocentric reply-graphs. He has tried to explore the link between the distribution of karma (i.e. popularity measured by the number of up- and downvotes) and the behavioural role to which that user belongs. He observed that users with Contributor behaviour proved to be more popular, however the variance in karma in this role was high [3]. Thus, this indicates that the contributor role encompasses both popular and unpopular users. By predicting high-karma users, community owners or moderators, or even the users themselves, could more efficiently manage and maintain behaviour in the forums such as Reddit.com under analysis.

All three papers have tried to examine the users and the data posted by them on Reddit.com. They demonstrated that most of the popular links are overlooked in the first time, only a small number of users participate across community boundaries and the behavioural role of the user is directly linked to the distribution of karma.

In this project, we are also trying to examine the users and their activity. We are trying to demonstrate that popularity of content leads to its duplication on Reddit. In order to achieve this, we have used the Reddit API to collect the data across various subreddits at various times of the day.

3 Data Collection

3.1 Identification of Subreddits

So far, We chose subreddits related to photography. This was done as all reddit posts on such subreddits use image sharing services like imgur.com. These sharing services usually generate an unique url for each gallery that can be used to share on reddit. Hence, by searching for posts with same image urls, we were able to identify duplicate posts.

Reddit has an inbuilt subreddit recommendation system which lists down relevant subreddits for every search query. These subreddits have a varying frequency of user posts, thus giving a uniform result instead of results being biased to only few with the maximum activity. We choose 10 subreddits to get a preliminary idea about the posts and users in a subreddit.

The subreddits chosen for this study were:

- /pics
- /gifs
- /picrequests
- /photography
- /Images
- /itookapicture
- /postprocessing
- /photocritique
- /HDR
- /shittyHDR

For the project rest of the semester, we plan to re-choose subreddits after a meeting with the professor. Since search allows us to search throughout reddit, we don't have to pick subreddits close to each other.

3.2 Fetching and Storing

We then used Reddit API to retrieve the content (link submissions on hot page) from these subreddits by using an API wrapper (PRAW - python reddit api wrapper) provided for Python language. The

Number of Posts	42992
Number of Unique Authors	33044
Number of links reposted at least once	7541

Table 1: Data Analysis

script was executed for one hour each on four different times in the same day, to add the dimension of time in our content analysis. A MySQL database was used to store and update the tables of the content which was fetched. While data collection of top and news posts is running, in parallel, we also run a bot that looks at posts that has been collected more than 5 days in the past and we look up where have these links diffused in reddit and stores the tuple (link, sub_reddit, poster, time_created) into the db.

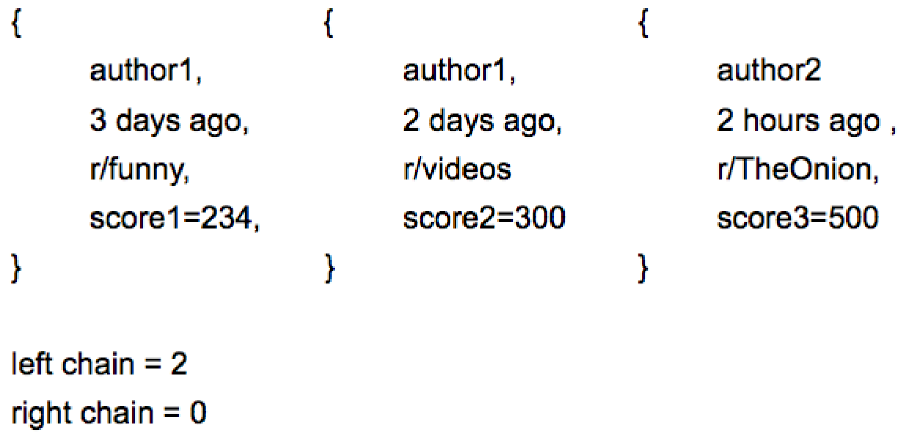
4 Data Analysis

The statistics of the data collected can be seen in Table1.

4.1 Chains

Once we have obtained all the posts, we search each of the posts on reddit in order to get their diffusion across the various sub-reddits since their inception. Thus, we thinking of the search results like a diffusion chain. We have sorted the results by time and we call the first person to post the link as the original author. How many times the original author posts initially is how much effort he puts into broadcasting the link. After this, we have found the search result with max score so that we can know when a particular reached its maximum popularity. All results to the left of the max post form the left chain and after wards form the right chain.

An example of chains can be seen in the following figure:



```
{                  {                  {
    author1,           author1,           author2
    3 days ago,        2 days ago,        2 hours ago ,
    r/funny,           r/videos           r/TheOnion,
    score1=234,        score2=300         score3=500
}                  }                  }

left chain = 2
right chain = 0
```

Figure 1: Description of Chain

In this example, we have three results, the last result has the max score, so our left chain length is 2 and right chain length is 0. Note that we don?t do normalization with respect to how many subscribers are in the subreddit.

4.2 Distribution of Chain Length

Figure 2 shows distribution of chain length of posts i.e. distribution of how far links go. It can be observed that most chain lengths are less than 50. There is a decreasing trend in distribution. Thus, posts with longer chains are less in number. However, there is one anomaly. We notice a tiny growth at the end near the 230-240 bucket with like nothing in between. This post has 235 shares and turned out to be a spam link where every-time a randomly generated unique author posted in /r/funny.

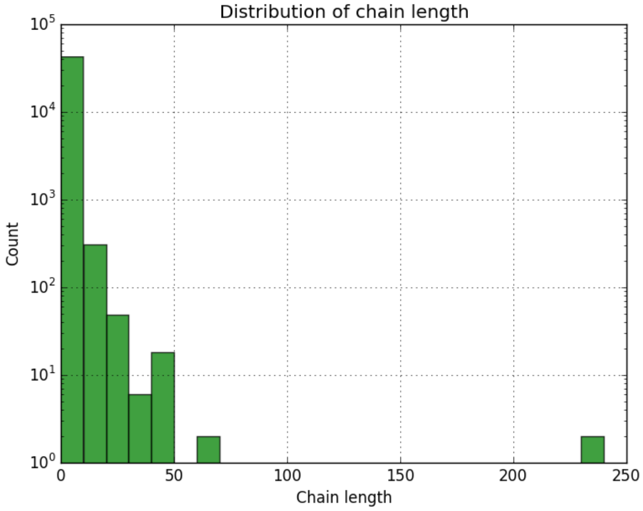


Figure 2: Distribution of Chain Length

4.3 Distribution of Original Author Re-posting

Figure 3 shows that distribution of posts by the original author. So, it basically shows the the number of times the original author is reposting his/her own link in order to broadcast it so that it becomes more popular which eventually leads to a higher karma for him/her. We observed that the maximum number of times author has reposted has gone to 12 in our dataset.

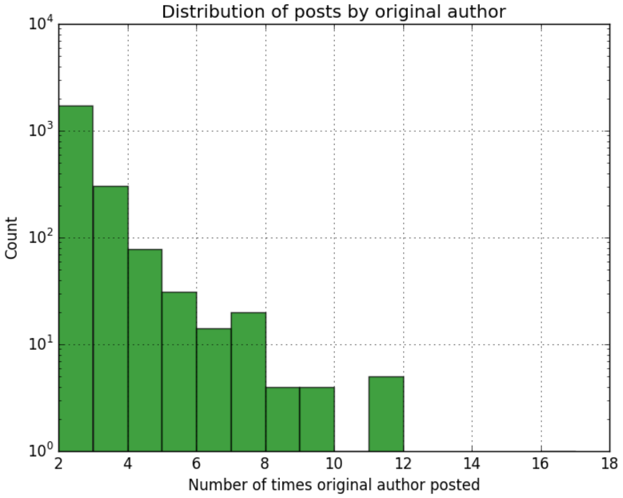


Figure 3: Distribution of Original Author Re-posting

4.4 Distribution of Re-posts by each Subreddit

Figure 4 shows the distribution of posts and re-posts by each subreddit. Thus, here we are looking how much repost happens in each subreddit. We were expecting a lot of reposts from subreddits adviceanimals or funny because most of them are meme based subreddits but the onion and radioreddit came out on the top.

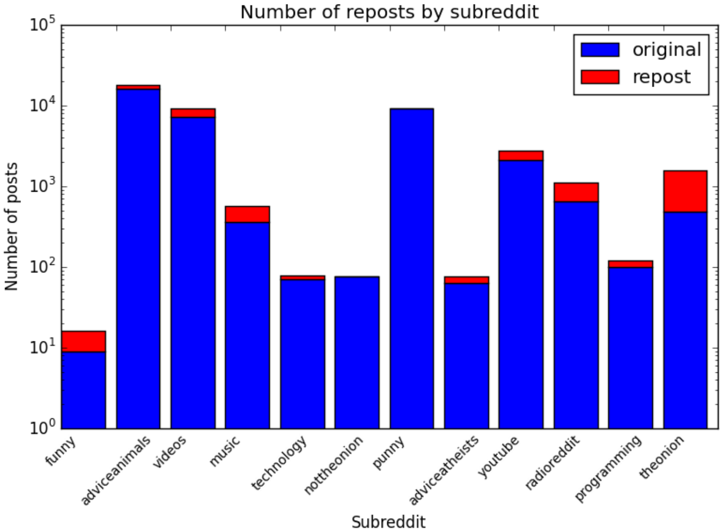


Figure 4: Distribution of Re-posts by each Subreddit

4.5 Distribution of Different Authors Re-posting

Figure 5 shows the distribution of different authors re-posting. So basically it calculates a ratio of unique authors by the chain length. If this ratio is 1, this means that all authors in post's chain are unique. We observed that that ratio is 1 in most of the cases. Thus, the chain of most of the posts on Reddit consist of unique authors i.e. most of the time it's only unique people who are reposting.

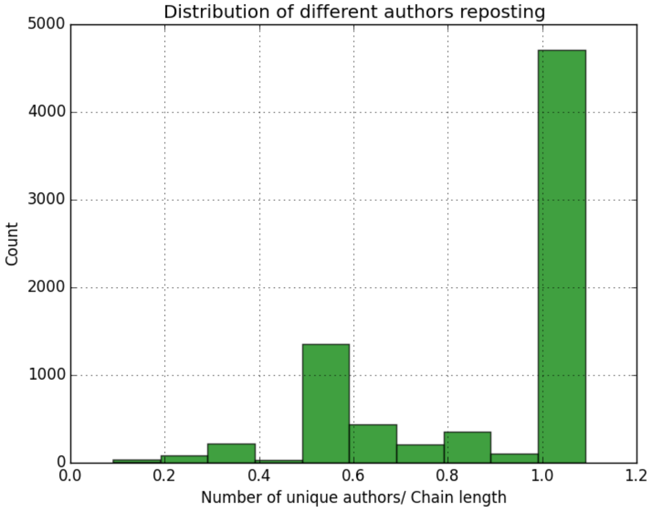


Figure 5: Distribution of Different Authors Re-posting

4.6 Distribution of Left Chain Length

Figure 6 shows distribution of left chain length. This distribution basically shows us how people try to make a particular post happening and popular. It tells the factors that led to the post's maximum popularity in the entire chain. It is observed that left chain for most posts is less than 30. Also, a decreasing behavior is observed as expected. However, again there is an anomaly in the graph near the 225-230 bucket. We looked at it, that post has a left chain length of 229 and found that it turned out to be a spam link where every-time a randomly generated unique author posted in /r/funny.

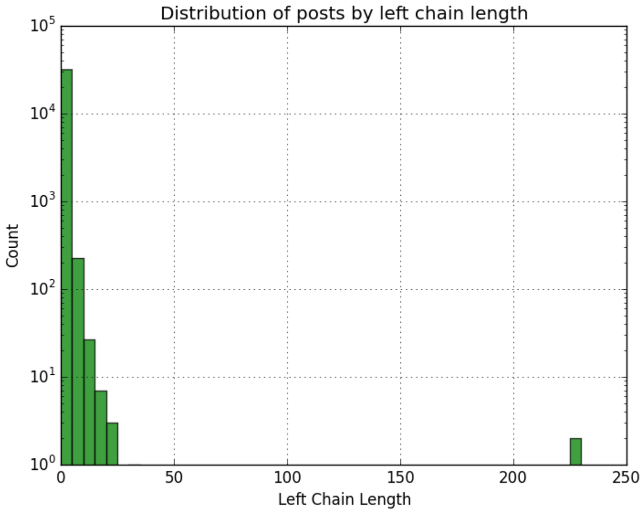


Figure 6: Distribution of Left Chain Length

4.7 Distribution of Right Chain Length

Figure 7 shows the distribution of the right chain length. It basically shows how far diffusion of a link can go. It was observed that no link made it more than 50 reposts after the point in the chain when the post was at its highest popularity. Also, a decreasing behavior was observed.

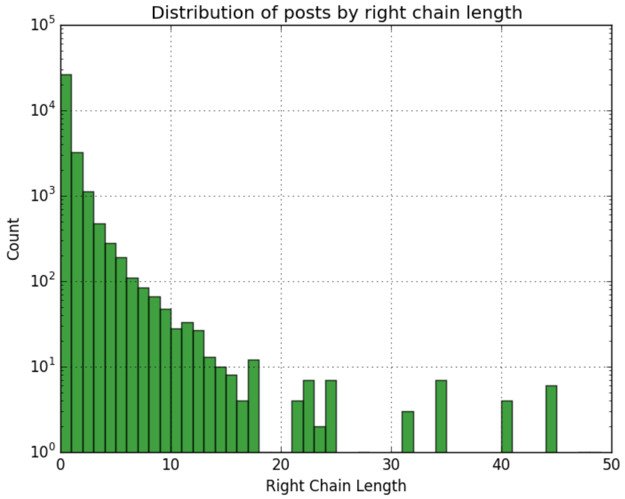


Figure 7: Distribution of Right Chain Length

4.8 Distribution of Max Score by the Right Chain Length

Figure 8 shows the scatter plot of the maximum score achieved and the right chain length of each post in the data-set. So, it was observed that for most of the posts, the right chain length is less than 20 even for very high maximum scores. Also, some unexpected results were observed such that some posts with even very low maximum score had very large right chain length and some posts with very high maximum score has very small right chain length.

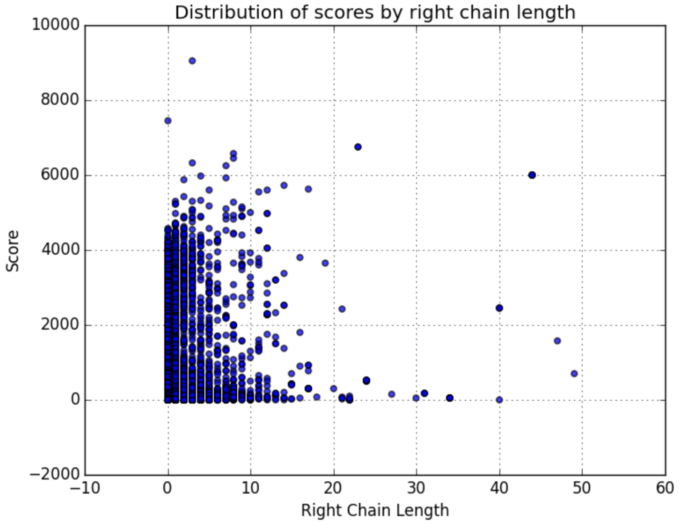


Figure 8: Distribution of Max Score by the Right Chain Length

4.9 Distribution of Maximum Score of the chains

Figure 9 shows the distribution of the maximum score of the chains in our dataset. We had expected the figure to be at highest around the low scores, and have a decreasing nature with a long tail. However, we noticed a surge in posts with scores around 3000.

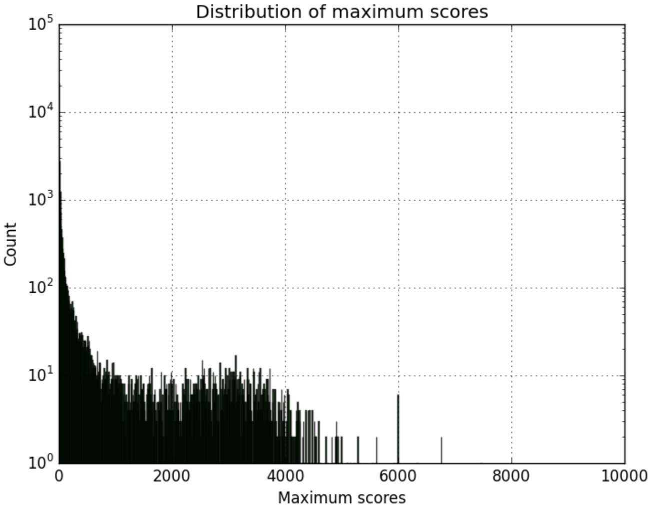


Figure 9: Distribution of Maximum Score of the chains

5 Comparison with Eric Gilbert's Paper on Widespread Under-provision on Reddit

We compared our results with a related study done by Eric Gilbert who found that only around half of the popular links were noticed in the initial submission. Gilbert?s dataset was popular links from /r/pics subreddit. We found similar results when we evaluated the same figures for our dataset of 12 sub-reddits.

- ?Widespread Under Provision on reddit? - Eric Gilbert
  - Analyzed 9,370 popular links
  - 48% links noticed on first submission
  - Refer figure 12

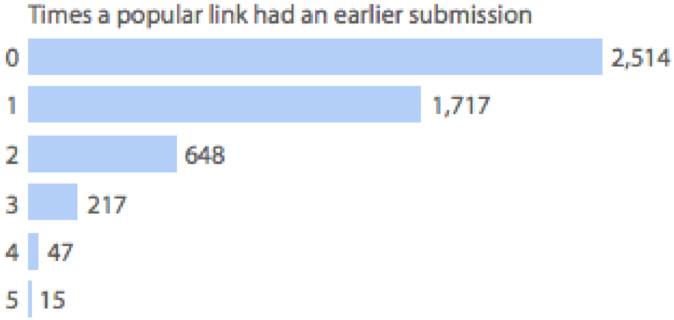


Figure 10: Times a popular link has an earlier submission

- Our Study
  - Analyzed 42,992 links
  - 12 subreddits
  - 78.28
  - Refer figure 13

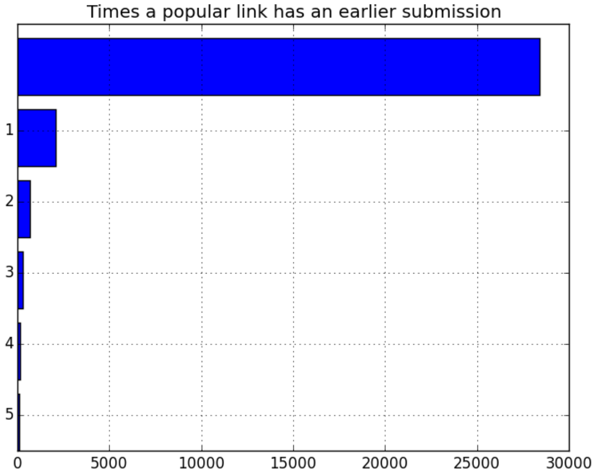
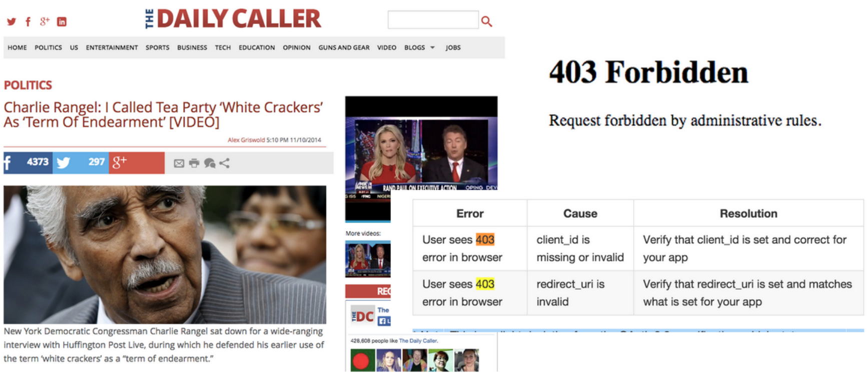


Figure 11: Times a popular link has an earlier submission

6 Anomalies with the Data

6.1 Forbidden Links

We found some posts that reddit has banned directly on API level. An online search of the given link redirected to a 403 Forbidden Error page. Interestingly, the link has a politically sensitive article. Moreover, it is no where mentioned in their API documentation to get an error of this nature. One of such pages can be seen in the following image:


THE DAILY CALLER
HOME POLITICS US ENTERTAINMENT SPORTS BUSINESS TECH EDUCATION OPINION GUNS AND GEAR VIDEO BLOGS JOBS
4373
297
More videos:
REC
THE DC
428,608 people like The Daily Caller.

POLITICS

Charlie Rangel: I Called Tea Party 'White Crackers' As 'Term Of Endearment' [VIDEO]

Alex Griswold 5:10 PM 11/10/2014

New York Democratic Congressman Charlie Rangel sat down for a wide-ranging interview with Huffington Post Live, during which he defended his earlier use of the term 'white crackers' as a "term of endearment."

403 Forbidden

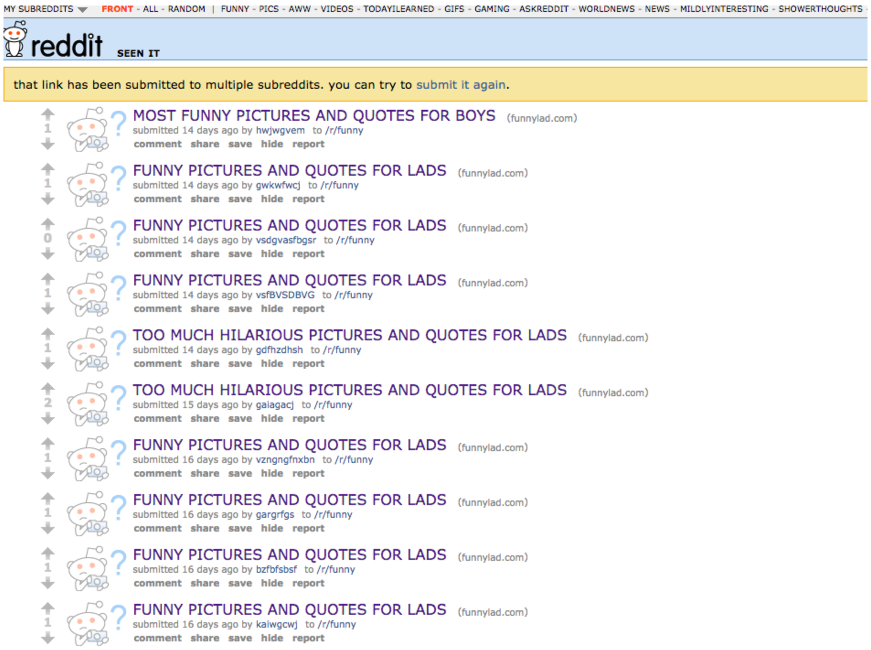
Request forbidden by administrative rules.

Error	Cause	Resolution
User sees 403 error in browser	client_id is missing or invalid	Verify that client_id is set and correct for your app
User sees 403 error in browser	redirect_uri is invalid	Verify that redirect_uri is set and matches what is set for your app

Figure 12: Link which is forbidden on Reddit

6.2 Posts having chain length of 235

Post having a chain length of 235 was too good to be true. So, we checked it manually and found that it turned out to be a spam link where every-time a randomly generated unique author posted in /r/funny. It?s pretty creepy to see that someone is using the API to growth hack their website. The domain name is expired. It doesn't make sense. A screenshot of that page is shown below:


MY SUBREDDITS
FRONT - ALL - RANDOM | FUNNY - PICS - AWW - VIDEOS - TODAYILEARNED - GIFS - GAMING - ASKREDDIT - WORLDNEWS - NEWS - MILDLYINTERESTING - SHOWERTHOUGHTS
reddit SEEN IT

that link has been submitted to multiple subreddits. you can try to submit it again.

1 MOST FUNNY PICTURES AND QUOTES FOR BOYS (funnylad.com)
submitted 14 days ago by hwjwgvem to /r/funny
comment share save hide report

1 FUNNY PICTURES AND QUOTES FOR LADS (funnylad.com)
submitted 14 days ago by gwkwfwcj to /r/funny
comment share save hide report

0 FUNNY PICTURES AND QUOTES FOR LADS (funnylad.com)
submitted 14 days ago by vsdgvasfbgsr to /r/funny
comment share save hide report

1 FUNNY PICTURES AND QUOTES FOR LADS (funnylad.com)
submitted 14 days ago by vsfBVSDBVG to /r/funny
comment share save hide report

1 TOO MUCH HILARIOUS PICTURES AND QUOTES FOR LADS (funnylad.com)
submitted 14 days ago by gdfhzdhsh to /r/funny
comment share save hide report

2 TOO MUCH HILARIOUS PICTURES AND QUOTES FOR LADS (funnylad.com)
submitted 15 days ago by gaiagacj to /r/funny
comment share save hide report

1 FUNNY PICTURES AND QUOTES FOR LADS (funnylad.com)
submitted 16 days ago by vzngngfnxbn to /r/funny
comment share save hide report

1 FUNNY PICTURES AND QUOTES FOR LADS (funnylad.com)
submitted 16 days ago by gargrfgs to /r/funny
comment share save hide report

1 FUNNY PICTURES AND QUOTES FOR LADS (funnylad.com)
submitted 16 days ago by bzfbfsbsf to /r/funny
comment share save hide report

1 FUNNY PICTURES AND QUOTES FOR LADS (funnylad.com)
submitted 16 days ago by kaiwgcwj to /r/funny
comment share save hide report

Figure 13: Post with a chain length of 235

Team member	Contributions
Ashwini	Data collection
Prajwal	Data visualization and scripting
Revant	Data analysis
Suren	Development of the bot

Table 2: Contributions

6.3 Exception Handling and Issues with PRAW

We just wanted to point out the difficulties that we faced in data collection. A lot of exception handling had to be done to take care of various issues to collect a dataset with a decent number of records. We had email alerts set up to notify whenever our bot found a new error. Also, the python wrapper we used to collect data had its own issues which forced us to recollect data in Phase 2.

7 Future Work

- Study user's reposting behavior
- Voting rings
- detect spam attempts

8 Contributions

We have been contributing about equally in this project. However we have distributed responsibilities based upon areas of interest as shown in Table 2.

9 References

- Eric Gilbert. 2013. Widespread underprovision on Reddit. In Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13). ACM, New York, NY, USA, 803-808
- Cody Buntain and Jennifer Golbeck. 2014 Identifying social roles in reddit using network structure. In Proceedings of the companion publication of the 23rd international conference on World wide web companion (WWW Companion '14) International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 615-620
- Morrison, Donn, and Conor Hayes. 2013. Here, have an upvote: communication behaviour and karma on Reddit. GI-Jahrestagung