Characterizing behavioral trends in a community driven discussion platform

Sachin Thukral[1], Arnab Chatterjee[1], Hardik Meisheri[1], Tushar Kataria[2],
Aman Agarwal[2], Ishan Verma[1], and Lipika Dey[1]

[1] TCS Research
sachi.2,hardik.meisheri,ishan.verma,arnab.chatterjee4,lipika.dey@tcs.com
[2] IIIT Delhi, New Delhi, India
tushar15184,aman15012@iiitd.ac.in

Abstract. This article discusses methods to systematically analyze the patterns of individual and group behavior observed in community driven discussion platforms like Reddit, where users exchange information and views on various topics of interest. We study the statistical behavior of posts and model user interactions around them. Reddit has grown exponentially from a small community to one of the largest social network platforms. Its large user base and popularity harbors a variety of user behavior in terms of their activity. Our study provides interesting insights about a huge number of inactive posts which fail to attract attention despite their authors exhibiting *Cyborg-like* behavior to garner attention. We also observe short-lived but extremely active posts emulating a phenomenon like *Mayfly Buzz*. A method is presented, to study the activity around highly active posts to determine the presence of *Limelight* hogging activity, if any. We also present a systematic analysis to study the presence of controversies in posts. We analyzed data from two periods of one-year duration but separated by few years in time, in order to understand how social media space has evolved over the years.

Keywords: Reddit · Social Network Analysis · Behavioral Analysis

1 Introduction

The availability of massive amounts of data from the electronic footprints of human social behavior on a variety of online social network platforms has triggered a lot of research and its applications. Tools from various disciplines have come together and is being developed into what is currently popular as computational social science [10]. The contemporary approach to social network analysis has much built upon the classical approaches [16] of sociologists, and the present interests span across operations research, market intelligence, survey science, and statistical computing. . Study of social network data not only reveals the structure of the connections between its individual components including strong and weak ties and their dynamics, but also the possible reasons as to why such structure and dynamics are prevalent.

A typical social network can be seen as a multi-dimensional graph where elements like posts, comments, users etc. are the nodes and their interaction define the links. Usual measurables like post lifespan, average number of posts per unit time shows an aggregate user behavior along the post dimension across the entire social media platform. User comments across posts render the interactivity flavor where user behavior can be segregated according to number of comments to time span of interaction. The layer of the number of distinct users involved opens the scope to differentiate user behavior in terms of the reachability and impact of the post.

The access to a huge amount of data allows a rigorous statistical analysis, which can be combined with behavioral studies that can bring out interesting spatial and temporal features, providing interesting insights. In this article, we study evolution patterns of posts over time based on user interactions with the posts and grouping them into different categories. We also categorized posts based on user interaction patterns emerging around them. We present methods to determine the focal points of interactions. Further, we presenting methods to identify behavioral trends exhibited by users in order to popularize their posts. Additionally, we also discuss methods to analyze the presence of controversial posts and comments, even before getting into the text content.

In the social news aggregation, web content rating and discussion website *Reddit.com*, members share content in the form of links, text posts and images, which are then voted up or down by other members, where from further discussions can emerge. Posts cover a variety of topics including news, science, movies, video games, music, books, fitness, food, and image-sharing. They are organized by subject into user-created *subreddits*, providing further opportunities for fostering discussion, raising attention and publicity for causes. While Reddit is known for its open nature and diverse user community across different demographics and subcultures that generate its content, posts are also moderated for various reasons.

In this article, we have tried to gather insights about where, when and by whom the content is being created in the community as a whole. The study of evolution patterns help us in understanding the characteristics of posts which garner huge number of responses. By studying characteristics from an author's perspective gives us an indication of which authors are more reliable in spreading information over the space. Similarly, identifying the focal points in a long discussion can lead to understanding of popular opinions. These markers and behavioral trends can be used as cues in various applications like advertisement placement, summarizing viral/popular topics from different perspectives, half life of information spread, etc.

With the increasing use of social media for collaboration and sharing of important information across individuals and even within enterprises, understanding human behavior and being able to characterize them as well as to understand the interaction dynamics within a group of users is itself turning out to be an important task. For instance, in the organization to which most of the authors of this work are affiliated to, more than $400,000$ employees engage in at least two organization specific, closed social networks serving completely different purposes. Analysis of temporal patterns and group dynamics presented in our work are important aspects which can not only aid in understanding the different categories of users, but also identify the information needs and push the right content or advertisement for the right group at the right time. The similarity of patterns observed over multiple data sources prove that user behaviors are fairly similar across social platforms in the same domain.

This contents of this article is essentially an extension of our recent paper [15] where we presented the primary analysis of behavioral trends observed in Reddit. The rest of the article is organized as follows: We present the earlier related work in Section 2. A brief description of the data used in our study is presented in Section 3. Section 4 presents the aggregate analysis of the data, which provides the basis for further analysis. Analysis of evolution patterns of posts is presented in Section 5. Section 6 shows the interaction dynamics, while Section 7 shows the behavior exhibited by authors over the space. Section 8 discusses methods to identify the presence of controversial content in posts and comments. Finally, the entire analysis is summarized along with the inferences in Section 9. .

2 Related Work

There have been several studies regarding social media dynamics from various perspectives. Researchers have examined the structure of the comment threads by analyzing the radial tree representation of thread hierarchies [8] in one such work, while another study presented the responses over a post using graph theoretic approach to infer the *for* and *against* communities for that particular post [1]. The basic assumption made was that every post contains at least one comment belonging to each community. Researchers have also studied the behavioral aspects of users by crowd-sourcing information from experiments on the platform. One such study focuses on how individuals consume information through social news websites and contribute to their ranking systems. A study on the browsing and rating pattern reported that most users do not read the article that they vote on, and in fact 73% of posts were rated before even viewing the content [5]. While user interactions (likes, votes, clicks, and views) serve as a proxy for the content's quality, popularity, or news-worthiness, predicting user behavior was found to be relatively easy [6]. The voting pattern in the Reddit [11] has been studied to analyse the upranking of posts from the new page to front page and behavior of users towards some posts which are getting positive or negative votings. They have studied the posts mentioning Wikileaks and Fox News and to see the impact of negative voting on them, although working on only one month of data. A study on rating effect on posts and comments [7] has revealed that random rating manipulations on posts and comments led to significant changes in downstream ratings leading to significantly different final outcomes – positive herding effects for positive treatments on posts, increasing final ratings on the average, but not for positive treatments on comments, while negative herding effects for negative treatments on posts and comments, decreasing the final ratings on average. An exploratory study [17] on the dynamics of discussion threads found topical hierarchy in discussion threads, and their possibility to be used to enhance Web search. A study on 'social roles' of users [2] found that the typical "answer person" role is quite prominent, while such individual users are not active beyond one particular subreddit. In another study, authors have used the volume of comments a blog post receives as a notion of popularity to model the relationship with the text [18]. Authors used several regression models to predict the volume of comments given in the text. This analysis is quite restricted in terms of the scale of dataset, limiting to political posts and only three websites which amount to four thousand posts. While the content analysis is most intuitive, it does not provide richer analysis. Text content shared over social media is usually noisy, full of non-standard grammar and spelling, often cryptic and uninformative to the outsider from the community. When one adds the scale of today's social media dataset, it is computationally non-viable to have content analysis over the whole corpus.

Most of the studies reported till date have performed analysis on a subset of data by restricting themselves to a limited number of posts, comments, top users, subreddits, etc., while we use two separate sets each of which are complete data for one year period. To the best of our knowledge, only, very few have used complete data for analysis. In Ref. [13], authors have presented evolution analysis over five years of subreddits with respect to text, images, and links though they have only considered posts and not analyzed comments. Ref. [4] has reported the effect of missing data and its implications over the Reddit corpus taken from 2005 to 2016.

3 Data description

3.1 Terminologies

Following are the terminologies that are frequently used throughout the article:

- A Reddit **Post** is text, link or an image submitted by a registered member. Posts are integral entities which allow users to express themselves and initiate a discussion.
- **Comment** is a response to the post that is active on Reddit. A comment can either be a direct response to the post or a response to any comment made on a post, thus creating a nested structure of a tree graph with possibily any number of children at any level.
- **Author** is a registered user on the platform who have at least one post or comment.
- **Score** is the difference between number of upvotes and downvotes.

3.2 Definitions

We define the following quantities, which will be used in the rest of the article:

- **Age** is the time difference between the last comment on the post and creation of the post, measured in seconds (unless otherwise mentioned).
- **Effective Comments** are the total number of comments on a post made by users other than the author of the post.
- **Automoderator** Official bot of Reddit.
- **Deleted Author** Authors who delete their post/comment.

3.3 Data

We use two separate data sets of Reddit [12], in order to see if the data exhibits any qualitative changes along with the quantitative changes, in a gap of few years –

- Period I: 1 January 2008 - 31 December 2008,
- Period II: 1 August 2014 - 31 July 2015.

The data contained posts and comments during those entire periods of one year each, and the associated variables like the title of the post, time of post/comment, subreddit, parent post/comment id, etc. BThe basic statistics of the data are presented in Table 1. After September 2015, there was a change in the number of fields that were being provided by Reddit API. So, in order to maintain consistency, we have used the data till July 2015.

In this study, we have considered only those comments which were made on the posts of Period I during the same period and similarly for Period II. We also neglected the comments made during Period I to posts created before Period I and also comments made on the posts of Period I beyond the time domain of Period I. Same procedure was carried out for Period II. We analyzed the data using parallel computation on a Hadoop setup.

Table 1. 2008 Data Table

	Period I	Period II
Number of Posts	2,523,761	63,118,764
Posts with deleted authors	425,770 (16.87%)	12,346,042 (19.56%)
Posts with zero comments	1,536,962	23,417,869
Posts with one comment	591,489	9,011,332
Number of Comments	7,242,871	613,385,507
Number of Comments on posts of the period	7,224,539	608,654,680
Number of Disconnected Posts	219 (0.009%)	1,380 (0.002%)
Number of Removed Comments	355 (0.004 %)	248,493 (0.04%)

Table 2 shows set of variables from the available metadata for comments and post that are used for our analysis. Available metadata contains 36 post variables and 21 comment variables. We have not considered score in our analyses, except for determining cyborgs.

Table 2. Used Data Variables for posts and comments

Posts	**Comments**
author	author
created utc	created utc
	link id
name	name
number of comments	parent id

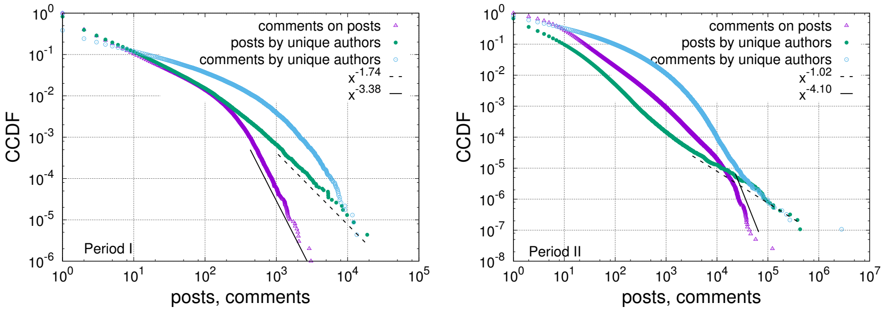

Fig. 1. The basic distributions of posts, comments and authors. Cumulative probability distribution (CCDF) that a post received at least c comments, CCDF that an author has posted at least p posts, and CCDF that an author commented at least c times. Plots for both Period I and Period II are shown, along with estimates (using MLE) of fits to asymptotic power law tails.

4 Analysis of aggregated data

For the analysis of the one-year aggregated data, first we computed the Complementary Cumulative Distribution Function (CCDF) which calculates the probability that a post has received at least c comments (Figure 1). For Period I, with an average of 7.3 comments per post, the CCDF shows a broad distribution with an asymptotic power law decay roughly beyond 500 comments: $Q(c) \sim c^{-\nu_c}$, with $\nu_c = 3.38(1)$. For Period II, with an average of 15.3 comments per post, the tail of the CCDF also has a similar broad distribution with an asymptotic power law decay roughly beyond $20,000$ comments: $Q(c) \sim c^{-\nu_c}$, with $\nu_c = 4.10(3)$. We can infer that while majority of the posts get small number of comments, there are also a significant yet diminishing number of posts with a large number of comments. Apart from difference in the size of the data and the average number of comments per post, the asymptotic power law region is observed much later in the Period II.

The probability (CCDF) that an author has posted at least p posts also show an asymptotic power law decay for Period I: $Q(p) \sim p^{-\nu_1}$ with $\nu_1 = 1.74(2)$, with around 9.2 posts per author on the average while for Period II, the asymptotic power law decay was $Q(p) \sim p^{-\nu_1}$ with $\nu_1 = 1.02(2)$, with around 6.7 posts per author on the average. It is interesting to note the qualitative

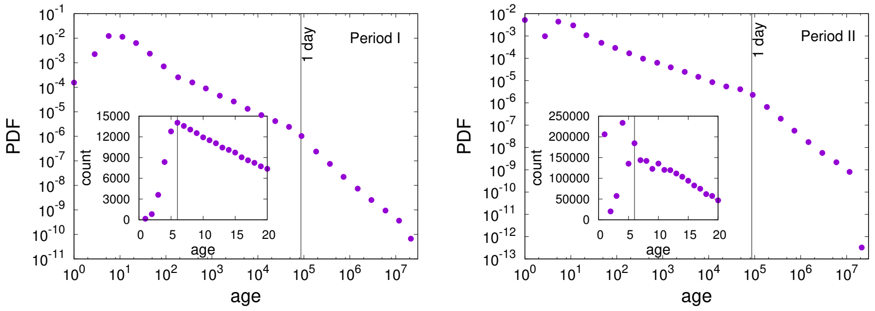

Fig. 2. The main plots shows the PDF of the age of a post (seconds) for the entire age range for Periods I and II. There is a marked departure in the nature of the probability distribution around ≈ 1 day, indicating that a large number of posts become inactive beyond that time. The insets show the histograms corresponding to the age distribution at small values of age. There is a prominent peak around 6 seconds for Period I and at values less than that for Period II.

difference between the distributions – while in Period I, the decay gets steeper after around 500 posts, in Period II, it gets slower beyond 1000 posts. However, the probability (CCDF) that an author commented at least c times also shows broad distribution, but with a faster decay, resembling a lognormal (Figure 1) distribution for very high values, with around 21.2 comments per author on the average for Period I, while for Period II the lognormal like behavior was followed by a slower decay with around 65.4 comments per author on the average. This is an indication of a higher tendency to comment than to create a post. One of the reasons behind this can be the less diversity of authors in contributing to posts compared to comments. Period II also shows the isolated data point for the *automoderator*. The power law fits in our analysis were performed using *maximum likelihood estimates* (MLE) [3].

While posting behavior is an intrinsic property of an user, and expected to be less correlated to comments, commenting is a part of the interaction with others and thus have a strong correlation with the behavior of others in the comment space, which justifies our observation that the former shows power law tail while the latter is lognormal.

5 Analysis of post evolution patterns

To analyze the evolution of the posts, we calculate the age and number of comments for each post.

5.1 Mayfly Buzz

The probability density function (PDF) of the ages of all posts (Figure 2) has a most probable value at 6 seconds for Period I, while the equivalent peak is smeared across values less than 6 seconds for Period II. Also, we observe that there is also a shift in slope around age of 1 day, following which, the PDF decays faster, suggesting that more posts tend to become inactive after a day. In fact, 88.6% in Period I and 71.1% in Period II of posts die before a day. We term this behavior of the posts as *Mayfly Buzz*, which resonates with the concept of creating a buzz for a day. Activity usually dies after a very short period of time, as seen in other social networking platforms. We observe the similar behavior on Reddit, where age is longer as we are dealing with a discussion platform as opposed to a microblogging site like Twitter [9], etc.

5.2 Cyborg-like behavior

Figure 3 shows the age distribution (frequency) of all the posts which have only a single comment. In Period I, there is a very prominent peak at 6 seconds, as found earlier (Figure 2). It can be seen that the ages of 72.78% of these posts do not exceed 600 seconds (= 10 minutes). Period II looks very similar except the peak is seen at 5 seconds with an additional peak at 1 second.

Table 3. Cyborg-like Posts Statistics

	Period I	Period II
Posts with first comment in less than 6 seconds	43138	1,804,374
Posts with same author of first comment	7,615	492,928
Cyborg-like Posts	6,389	387,845
Successful Cyborg-like Posts	3,446	70,237
Successful Non Cyborg-like Posts	866	28,892
Unsuccessful Cyborg-like Posts	2,943	317,608
Unsuccessful Non Cyborg-like Posts	360	76,191

Further, we analyzed posts whose first comment is posted within 6 seconds, which constitutes 43138 posts for Period I and $1,804,374$ for Period II. Out of these posts we found that there are $7,615$ and $492,928$ posts which have their first comment by the author of the post for Period I and Period II respectively. We observe an uncanny behavior from approximately 17% and 20% of people behaving in exactly the same manner respectively. To understand this uncanny behavior of posting comment by the same user, we checked the number of characters in the first comment of these posts. For instance, we find that 83.9% ($6,389$ of $7,615$) and 79% ($387,845$ of $492,928$) posts have number of characters more than 100 for Period I and II respectively. It is crucial to mention that we have left out posts which contain links to web-pages in this analysis, which may be copied from a certain source and pasted in the posts. Writing such long comments within 6 seconds is quite impossible for a genuine human. We categorize these posts to be exhibiting a *cyborg-like* behavior, where these posts may be just an advertisement or a message that these users intend to propagate.

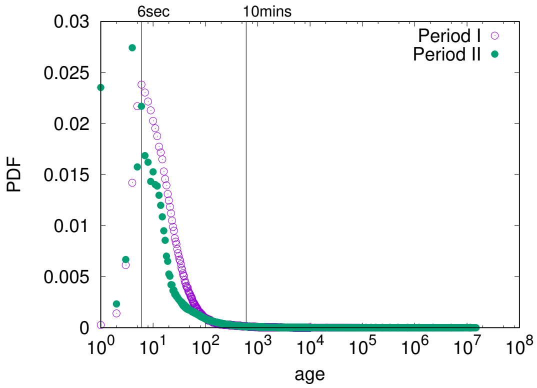

Fig. 3. PDF of ages for posts with one comment for Periods I and II.

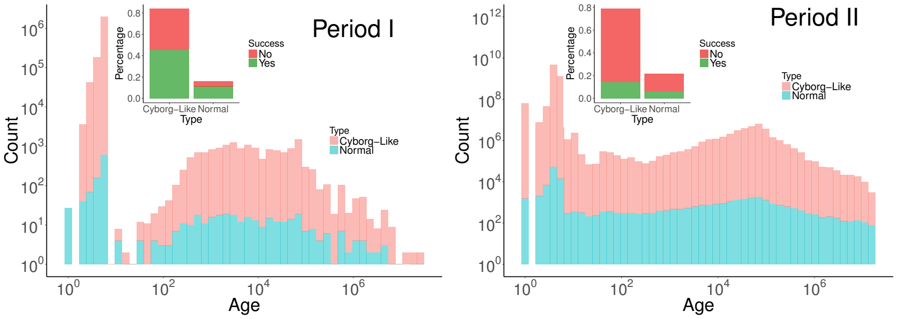

Fig. 4. The histograms of actual age of the posts whose comment is within 6 seconds. The insets show distribution of posts over success rate. Plots are shown for both Period I and Period II.

We further define a success criterion to check if this type of posts were successful in garnering attention or responses – if a post is getting any reaction (comment or vote) from other Reddit users, then they are considered successful in drawing attention. For instance, we find that 53.93% $(3,446$ of $6,389)$ and 18% $(70,237$ of $387,845)$ *cyborg-like* posts were successful in Periods I and II respectively. While 70.63% (866 of $1,226$) of normal posts (which have comments with less than 100 characters) of Period I are successful, which we assume can be possibly done by humans (Figure 4), which comes to about 27.5% $(28,892$ of $105,083)$ for Period II. Table 3 summarizes the data for this analysis. Hence, for Period I, we infer that machine generated content is less likely to garner interest as compared to human generated content. A possible reason behind the low success of the cyborg-like posts can be that lengthy comments and promotions/advertisements provide less room for any discussions. For Period II, however, we found a variety of behavior in the cyborg-like posts in the posts, which required much more granular analysis.

5.3 Analysis of depth and breadth of a post

Discussion happening on a post can be seen to have a tree-like structure. Depth of a post can be defined as the maximum length of nested replies on a post and breadth of a post can be defined as the maximum number of comments at a particular level. Figure 5 shows the variation of depth against the breadth of the posts for both the periods, with the heat map depicting the density. We observe that depth or breadth can grow independently, most prominent in Period II, where posts with simultaneously large values of depth and breadth are rare or absent. The plot also shows that breadth grows easily compared to the depth, which can be attributed to the larger effort to grow a nested discussion (increase depth) than to diversify a discussion (increase breadth).

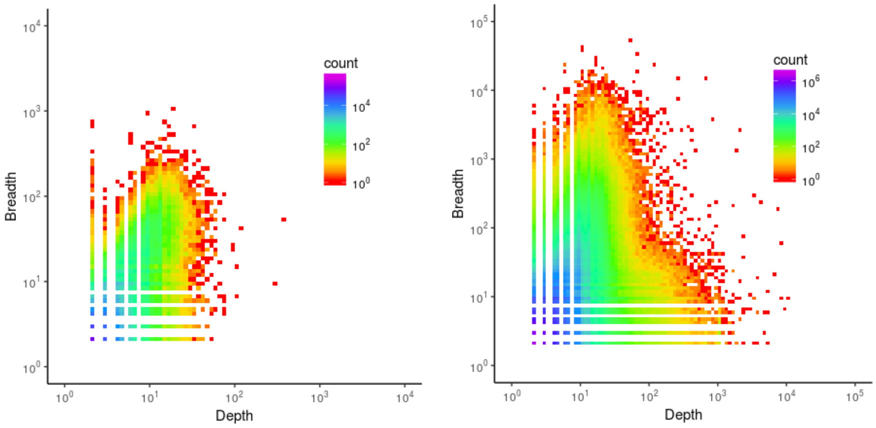

Fig. 5. Plot showing the depth of the posts against the breadth of the posts, with the density of posts shown through the heatmap.

5.4 Popular Post Dynamics

To understand the age dynamics of the popular posts and infer their behavior, plot the time evolution of the posts which have more than 500 comments. Three distinct categories are prominent (Figure 6):

- *Early bloomers* are rapidly growing posts, accumulating over 75% of their total comments within 1 day, creating the *Mayfly Buzz* as discussed earlier,
- *Steady posts* are characterized by ongoing activity throughout their lifespan.
- Slowly growing posts, which get suddenly very active at a late stage (after 30 days), can be termed as *late bloomers*.

We also study the behavior of the total number of comments with the age of each post, for all posts in our data. Figure 7 shows the heat map for all posts.

The overlaid binned average of all data indicates a marked departure in the gross behavior around 1 day which is also prominent from the density in the heat map.

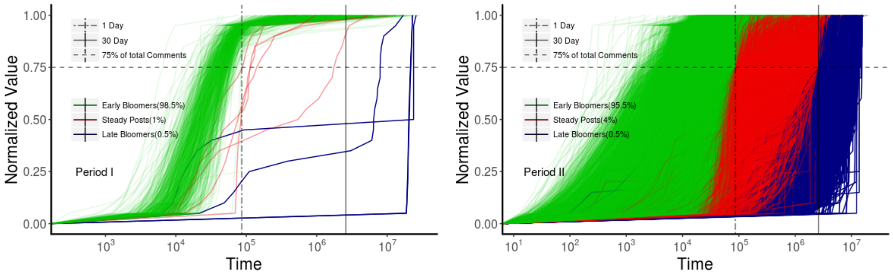

Fig. 6. Time evolution of the number of comments in a post (normalized by the final number of comments obtained in our defined time window of 1 year) for posts with more than 500 comments. The data has been coarse-grained to simplify visualization. The horizontal line corresponds to 75% of the total comments and the time to reach that fraction is used for characterizing the posts. The vertical lines are at 1 day and 30 days. Posts mostly active within 1 day (green) garner 75% of their comments during that period. Some posts grow throughout their active life span (red) taking time intermediate between 1 and 30 days to reach the 75% mark, while others grow slowly while becoming active at some later stage (blue), beyond the 30 days period. Plots are shown for Period I and Period II.

6 Analysis of interactions

The Reddit post-comment structure forms a tree graph, where posts can have its comments, and the comments can further garner replies. For this analysis, we have calculated a *limelight score* for each post based on the number of comments gathered as reply to a single first-level comment. In a way, this score computes the depth of discussion around a single comment for a post.

$$\text{Limelight Score} = \frac{\max(Comm_j)}{\sum_{k=1}^{N} Comm_j}$$

where $Comm_j$ is the total number of comments under j^{th} first level comment and N is the total number of first level comments for that post.

Figure 8 shows the histogram of the *Limelight scores* while the inset shows the CDF of the same. Here we have considered posts that have 500 or more comments only. We observe that in Period I, 56% of the total posts contain one comment with *Limelight score* of at least 0.25, which means that at least 25% of the discussion in this post is initiated and centered around a single comment.

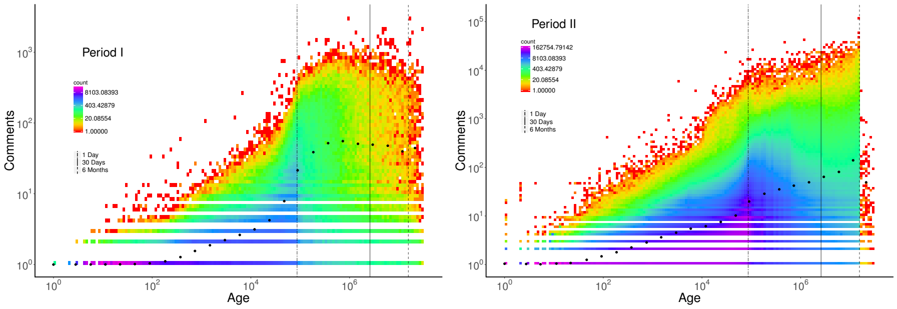

Fig. 7. The number of comments and the age of a post is shown as a density heat map for all posts, along with its binned average. A marked departure in the gross behavior around 1 day is prominent in both Period I and Period II.

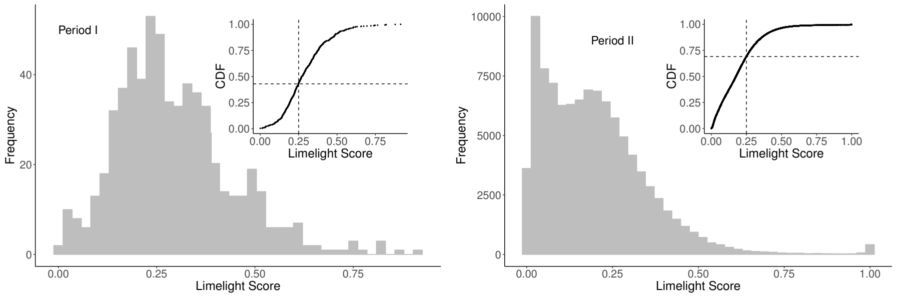

Fig. 8. Histograms of Limelight scores for the 628 posts for Period I and 100, 422 posts in Period II, which have more than 500 comments. The insets show the corresponding CDFs.

This behavior is also exhibited in 31% of the posts in Period II. Additionally in Period II, a finite number of posts actually have Limelight score close to unity, indicating absolute dominance of one branch of the comment tree.

We also observe that most of the time, the author of first level *Limelight* hogging comment is not the author of the post. For instance, this is true for about 97% of the posts during Period I.

This leads to an interesting insight that links virtual human behavior in social media to physical world social behavior. It is a rather common scenario that during any group discussion or meeting usually there are a few specific people, other than the presenter, who pro-actively initiates a conversation asking a question or making a comment, whereafter other people join the conversation. Interestingly, it is observed that lime-light hogging behavior is completely missing for posts whose authors exhibit *Cyborg-like* behavior. Thus, it may be inferred that posts automatically generated by bots have failed to garner garner human attention most of the times. However, we will conduct more rigorous studies in future to validate the inference.

To the best of our knowledge, characterizing content popularity by the depth of discussion around it has not been attempted before. Since it has been proved in earlier studies [5,14] that the number of upvotes-downvotes are not meaningful indicators for measuring interestingness or popularity of content, we claim that this can be a good way to measure them.

7 Analysis of Author Behavior

For analyzing author interactions, we define a network where, nodes represent unique authors and the edges represent the interaction between the authors through comments. We define the in-degree and out-degree for each node based on the number of interactions, where a self loop is ignored. Table 4 shows the statistics for the 3 categories – (i) authors who only put up posts are the pure *content producers*, (ii) authors who only comment are the pure *content consumers*, and (iii) rest of them indulge in both of the activities.

Table 4. Author Table

	Period I	Period II
Total Active Authors	229,488	9,369,708
Total Authors who only create posts	140,918	1,917,161
Total Authors who only comments	39,764	3,019,676
Total Authors who comment as well create posts	48,806	4,432,871

7.1 Quantifying author interactions to assess their influence

If A = total effective number of comments received and B = total number of comments on others' posts, then we define the **interaction score** of an author as $A/(A+B)$. Interaction score is zero for all authors who comment on others' posts but have not received any comments on their posts. Score is 1, if an author does not comment on others' posts but receives comments on one's own posts, though this is rarely observed. Figure 9 shows the histogram for the total count of authors along the whole range of interaction score.

There are some distinct authors who have the ability to consistently garner a large number of comments on each of their posts. To quantify this, we analyze the average number of effective comments received per post by authors. Figure 10 shows the normalized cumulative count across the effective number of comments per post. It is observed that 22% of the authors have fewer effective comments than the number of posts that they have put up which means no interaction for many posts for Period I, which happens to be around 6% in Period II. 11% have received an equal number of effective comments as their posts which can be attributed to an average of one comment per post for Period I, which is 13%

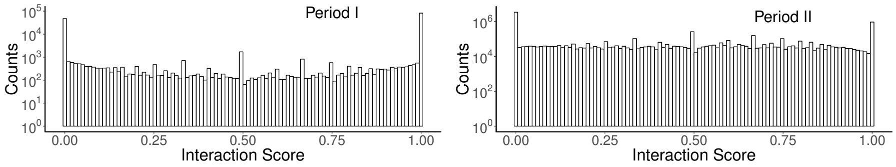

Fig. 9. Frequency counts for different interaction scores of authors for both periods. Equal reciprocative behavior is observed at score 0.5.

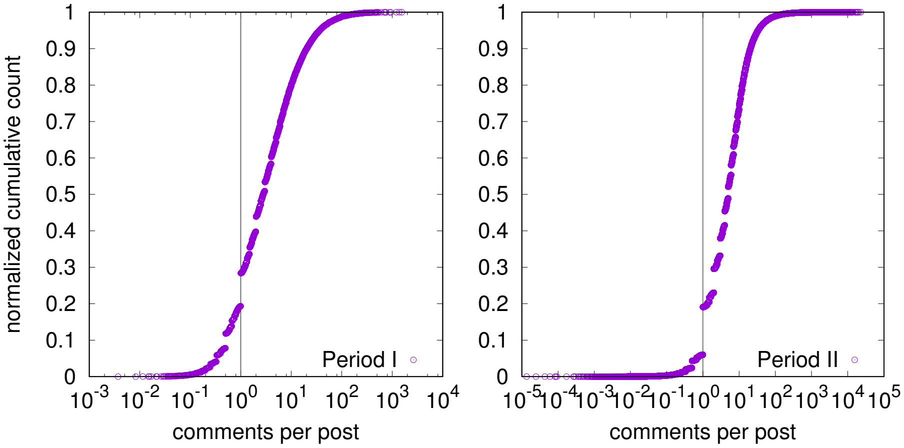

Fig. 10. Cumulative ratio of authors according to comments per posts. The vertical line corresponds to unity. For Period I, 22% fall below unity, 11% equals unity and the rest 67% are over unity. For Period II, 6% fall below unity, 13% equals unity and the rest 81% are over unity.

for Period II. The rest 67% received more comments than the number of posts put up for Period I, which comes to 81% for Period II.

The discussion above shows that authors who receive more attention on their posts are also the ones who are commenting on others posts. In other words, to gain attention on social media, authors have to be reciprocative. This is also highlighted by the peak at interaction score of 0.5 shown in Figure 9. It also emphasizes that the social media interactions are dominated by the phenomenon of mutual gratification.

8 On signatures of controversies

We observe that in the popular posts (with more than 500 comments), some comments have been deleted either by the author of the comment or by the moderators of the subreddits. In the latter case, deletion of a comment can happen only if the author made a comment that violates the rules of the subreddit set by the moderators, that can potentially lead to controversy in a social discussion platform. For further analysis, we have calculated the ratio of the number of deleted comments to the total number of comments, which can serve as a proxy for the measure of controversiality of a post and call it the *Controversiality Score*. In the top panel of Figure 11 we plot the Controversiality Score of a post against the number of unique authors for Period II. We observe that the plot branches roughly into two components for lower number of unique authors, one each for very high and very low level of controversiality score. This branching is absent beyond a certain number of unique authors, which is roughly 200 for our case. The colors map to the number of comments in the posts.

We observe that when there are fewer unique authors (less than 200 in our case) contributing in a discussion to a post, the outcomes can be quite extreme – it can either see a very high degree of controversy or a very low degree of controversy, as is seen by the left part of the graph. Beyond 200 unique authors, this extreme diversity vanishes – in fact, very high values of controversiality are absent. This indicates that more controversiality occurs within smaller groups over larger ones.

We further wanted to check if controversiality is related to the popularity of the subreddit in which the post is created. We define the popularity of a subreddit as the total number of posts that are being created in that subreddit during the period, and divide them into 5 categories, 1 being the least popular subreddit and 5 being the most popular subreddit (1 - $1-10$ posts, 2 - $11-100$, 3 - $101-1000$, 4 - $1001-2000$, 5 - above 2000). In the bottom panel of Figure 11, we plot the controversiality score of posts against the number of unique authors with colors indicating the popularity category of the subreddit. We can see that the popular subreddits (categories 4 and 5) have low controversiality score, around 0.25 or less, while high controversiality scores are prevalent in the less popular categories. Hence, we infer that controversiality is observed in smaller, closely knit groups (akin to *contempt breeds contempt*), and users stay clear of controversies in larger groups.

We have also checked the controversiality of the individual subreddits. We consider subreddits which have at least 100 posts. A post is considered *controversial* if its controversiality score is more than 0.2 i.e., more than 20% of comments of that posts are deleted. We calculated controversiality score of a subreddit as the fraction of posts in it that are controversial. The top panel of Figure 12 shows the controversial score of those subreddits in which user have posted at least 100 posts.

To check which are the users who are responsible for the above, we can see author-wise controversiality score. We can extend the above definition to *author controversiality score* which is the ratio of the number of controversial posts to the total number posts by the author where controversial posts are those with more than 20% percent of deleted comments in the posts. The bottom panel of Figure 12 shows the author controversiality score of authors who have more than 50 posts in our data.

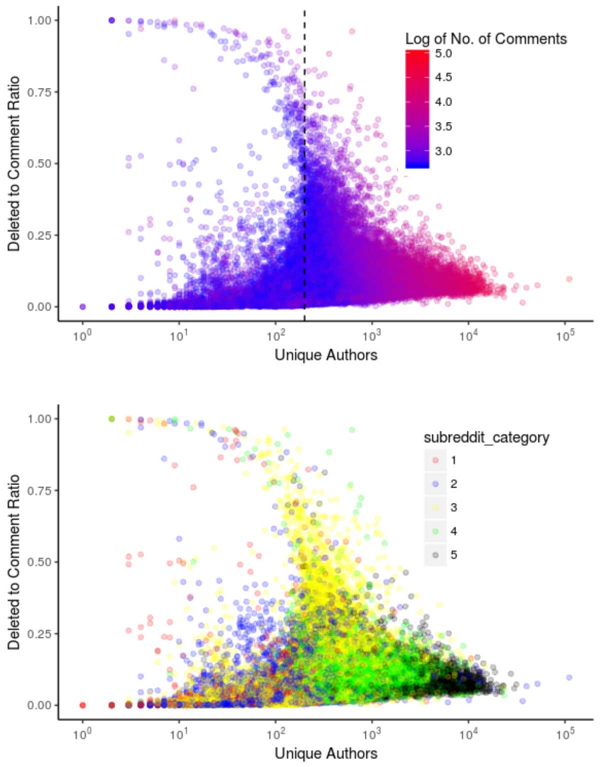

Fig. 11. (Top) The plot of controversiality score against the number of unique authors in the post, for all posts in Period II with at least 500 comments. Each post is coloured according to the number of comments it has. (Bottom) Same plot where the color of the point is according to the category of the subreddit in which they belong (discussion in text).

The above measures are the indicators that tell us, in which Reddit community (subreddit) controversial posts are put up and which user is responsible for initiating the controversy.

9 Conclusions and outlook

Reddit, the large, community-driven social network and discussion platform, harbors a plethora of behaviors as far as users are concerned. While a huge fraction of posts are left uncommented, the distribution of the number of comments on posts show correlation through the power law tail. Behavior of authors show a large variety – while many authors simultaneously post and comment, there are also a large fraction of purely *content producers* and *content consumers*, who restrict themselves only to posting and commenting respectively. The authors

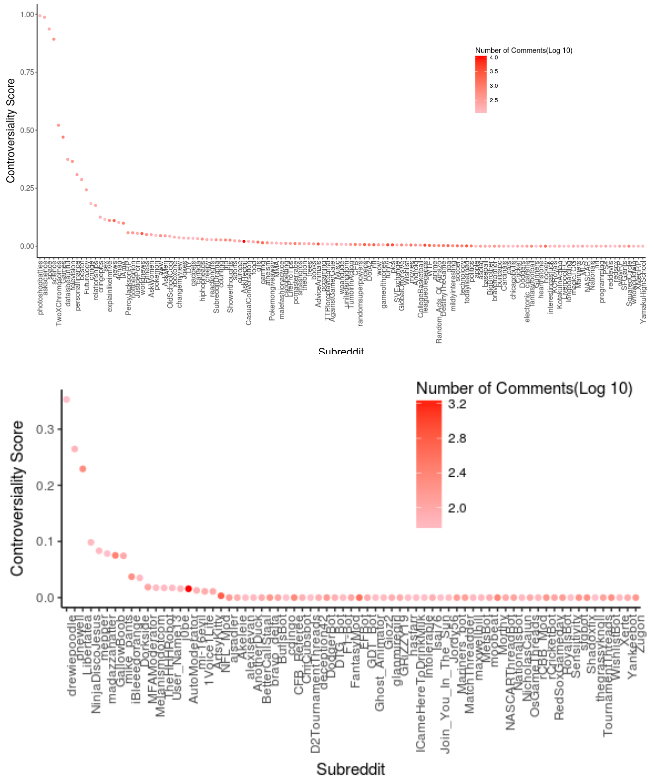

Fig. 12. (Top) Plot for the controversiality score for the subreddits which have more than 100 posts. (Bottom) Plot for the controversiality score for the authors which have more than 50 posts.

show a strong correlation between themselves and indication of the underlying multiplicative process in the form of lognormal distribution for the largest values for the distribution of number of comments by unique authors. Each post stay active as comments flow in and discussions are produced. However, a huge fraction of posts have only a single comment, and among them, a majority receive that only comment within 6 seconds, indicating a *Cyborg-like* behavior. A large fraction of posts seem to become inactive around the age of 1 day. This is consistent with the average active time of posts reported for micro-blogging site such as Twitter [9]. When we look at the time evolution of the top commented posts, we find three broad classes of posts – (i) *early bloomers* who gather more than 75% of their lifetime comments within a day, (ii) *steady posts* growing steadily throughout their lifespan, and (iii) *late bloomers* who show very little activity until the end of their lifespan. The early bloomers contribute to what we term as *Mayfly Buzz*, and constitute the majority of the posts. Posts also show *limelight hogging* behavior and upon appropriate characterization, we find that 56%

for Period I and 31% for Period II of posts have *limelight score* above 0.25, indicating that in such a large fraction of posts, at least one-fourth of the total weight of the discussions are contributed by one of the first level comments. In fact, this measure can be a more meaningful indicator of the interestingness or popularity of the content, compared to just votes or only number of comments. Social media discussion threads sometimes contain controversial content, and in Reddit this is moderated by deleting posts or comments. Our study tries to measure the controversiality from the fraction of such deletions, at the level of posts, authors and subreddits. We observe that controversiality is more prevalent in small, closely knit groups than large ones. Analysis of actual content can lead us to a better understanding, which we plan to carry out in future studies.

With the increasing use of social media even within closed groups as well as organizations, understanding human behaviors and able to characterize them is turning out to be an important task with potential impact and applications. One possible application of understanding temporal patterns of group behavior in such a scenario can be focused on injecting the right content or advertisement for the right group at the right time.

Our rigorous statistical analysis brings out a variety of behavioral elements from the authors and their interactions. There are few authors who are able to generate quite a lot of activity across a large number of posts. Going ahead, trend analysis of changing sentiment can be interesting. The insights gained from this analysis can be used to model different aspects from a large interactive population. In addition, predicting the recent trends can lead to better targeted reach e.g., innovative usage of *memes*.

References

1. Agrawal, R., Rajagopalan, S., Srikant, R., Xu, Y.: Mining newsgroups using networks arising from social behavior. In: Proceedings of the 12th international conference on World Wide Web. pp. 529–535. ACM (2003)
2. Buntain, C., Golbeck, J.: Identifying social roles in reddit using network structure. In: Proc. 23rd Int. Conf. World Wide Web. pp. 615–620. ACM (2014)
3. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Review **51**(4), 661–703 (2009)
4. Gaffney, D., Matias, J.N.: Caveat emptor, computational social science: Large-scale missing datain a widely-published reddit corpus. CoRR **abs/1803.05046** (2018)
5. Glenski, M., Pennycuff, C., Weninger, T.: Consumers and curators: Browsing and voting patterns on reddit. IEEE Transactions on Computational Social Systems **4**(4), 196–206 (2017)
6. Glenski, M., Weninger, T.: Predicting user-interactions on reddit. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. pp. 609–612. ACM (2017)
7. Glenski, M., Weninger, T.: Rating effects on social news posts and comments. ACM Transactions on Intelligent Systems and Technology (TIST) **8**(6), 78 (2017)
8. Gómez, V., Kaltenbrunner, A., López, V.: Statistical analysis of the social network and discussion threads in slashdot. In: Proc. 17th international conference on World Wide Web. pp. 645–654. ACM (2008)

9. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proc. 19th international conference on World wide web. pp. 591–600. ACM (2010)
10. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M.: Computational social science. Science **323**(5915), 721–723 (2009)
11. Mills, R.: Researching Social New – Is Reddit.com a mouthpiece for the 'Hive Mind', or a Collective Intelligence approach to Information Overload? In: Proceedings of the Twelfth International Conference, The Social Impact of Social Computing ETHICOMP 2011. pp. 300–310. Sheffield Hallam University (2011)
12. pushshift.io: Databases. https://pushshift.io/resources/databases/ (2017), accessed: 2017-10-01
13. Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., Strohmaier, M.: Evolution of reddit: from the front page of the internet to a self-referential community? In: Proceedings of the 23rd international conference on world wide web. pp. 517–522. ACM (2014)
14. Stoddard, G.: Popularity and quality in social news aggregators: A study of reddit and hacker news. In: Proceedings of the 24th international conference on world wide web. pp. 815–818. ACM (2015)
15. Thukral, S., Meisheri, H., Kataria, T., Agarwal, A., Verma, I., Chatterjee, A., Dey, L.: Analyzing behavioral trends in community driven discussion platforms like reddit. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 662–669. IEEE (2018)
16. Wasserman, S., Faust, K.: Social network analysis: Methods and applications, vol. 8. Cambridge Univ. Press (1994)
17. Weninger, T., Zhu, X.A., Han, J.: An exploration of discussion threads in social news sites: A case study of the Reddit community. In: Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. pp. 579–583. IEEE (2013)
18. Yano, T., Smith, N.A.: What's worthy of comment? content and comment volume in political blogs. In: ICWSM (2010)