Adoption and evolution of social networks from a cohort perspective

Samuel Barbosa
Institute of Mathematics and Statistics
University of São Paulo
São Paulo, Brazil
sam@ime.usp.br

Dan Cosley
Department of Information Science
Cornell University
Ithaca, NY 14853 USA
danco@cs.cornell.edu

Amit Sharma
Microsoft Research
New York, NY 10011 USA
amshar@microsoft.com

Roberto M. Cesar-Jr
Institute of Mathematics and Statistics
University of São Paulo
São Paulo, Brazil
cesar@ime.usp.br

ABSTRACT

Online communities provide a fertile ground for analyzing people's behavior and improving our understanding of social processes. However, like any complex social system, the key part is detail in identifying and accounting for underlying heterogeneity and selection effects among people in these communities. Using Reddit as an example community, we study the evolution of users based on comments and submissions data from 2007 to 2014, creating a cohort of users who join each year. Even with one of the simplest sources of differentiation between users—their age in the community—we find wide differences in people's behavior, including comment activity, effort and survival, both within cohorts and with the averages over the whole community. Not controlling for these variations may not only dilute the overall effects that we observe, but in some cases, it can lead us to the wrong conclusions (Simpson's paradox). These observations can be puzzling: for instance, we observe that average comment length decreases over any fixed period of time, but comment length in each cohort of users steadily increases during the same period after an abrupt initial drop. Finally, we analyze subcommunities on Reddit through the same lens of age and we find an enormous first-mover advantage: subreddits created early in the community's history are orders of magnitude more active than even successful subreddits created later, even among cohorts of users who join much later.

Keywords

user behavior;cohort;reddit

WWW '16 Montereal, Canada

1. INTRODUCTION

Understanding the evolution of users in a social network is essential for a variety of tasks: monitoring community health, predicting individual user trajectories, and supporting effective recommendations, among others. Many works aim at explaining these temporal aspects of evolution. Some adopt a point of view of the whole network and try to understand the patterns of behavior inside of the network [**?**, **?**], while others adopt a more user-centric point of view and try to model [**?**, **?**, **?**, **?**] or predict [**?**] individuals' behavior.

These analyses often combine all available data into aggregate analyses of the entire community over its entire history. This can be a natural response to limitations in the amount of available data: datasets may only capture a small part of the community's history; timestamps may not be available; snapshots may provide limited views of the community; or the community itself might be small [**?**]. Aggregate time-based analyses are also a natural first way to address these questions of community evolution.

In this paper, we argue it is likely that many of these aggregated views are misleading, because the conditions under which users join the community can vary greatly over time in ways that impact their behavior [**?**]. Among other things, the popularity, purpose, features, interface, and algorithms can change: Wikipedia circa 2005 and circa 2015 are very different, as are Facebook of 2005 and 2015. Analyses—including some of our own past work—that fail to account for this change may miss important details of what's really going on.

We support this argument through an analysis of user effort in Reddit, one of the most popular and long-running online communities, based on a very large, recently released dataset of posting behavior []. We address a number of questions commonly raised about users' effort in online communities: how often and how well do people post, what kinds of posts do they make, and where do they post them? In each case, we compare aggregate analyses of posting behavior to ones that treat users and subcommunities in Reddit as yearly cohorts, and views that look at calendar time to views that normalize behavior based on the creation of a user or subcommunity.

We find that even this simple accounting for time reveals additional insights about Reddit beyond what commonly performed aggreate analyses show. Users who join Reddit earlier post more and longer comments than those who join later. We also see evidence of larger behavior changes in these older users over time in terms of posting versus commenting, changes we think are driven in part by the growth of the community. Finally, we see that subcommunities created early in Reddit's history are likely to be much more active than those created later—and that the popularity of these communities spans across all user cohorts because of policy decisions Reddit makes in matching new users with popular subcommunities.

We see this paper as both making specific contributions to understanding behavior in Reddit and more general contributions in showing the importance of considering change over time in analyzing online communities.

2. TIME MATTERS

2.1 Why accounting for time is important

Why time is important: the incoming users and community both likely change over time. Differences in cohorts [**?**]

Users 1: The user adoption curve means that enthusiasts likely join earlier. Self-selection [**?**],

Users 2: Overall patterns of use might change because the overall population of internet users is still changing (this 'would be a place to critique the Mexican desensitization to violence paper).

Community 1: The nature or goals of the community can change. Flixter. Digg. Reddit iteself has a number of subcommunities that are the default set for new users—and they change over time.

Community 2: Successful communities often grow very rapidly, providing more content to react to (which is both good and bad – there's more to interact with, but getting responses might be a lot easier in a small than large community) and new challenges for management (Wikipedia policies).

Community 3: Behavior within the community changes as new users come. Conventions are created. Language changes. And, people likely model what they see.

Pithy summary of why time is important from the above.

2.2 Accounting for time and change

Using ordinal time when timestamps aren't available or are coarse. (Cosley et al. Suri paper from 2010 about ordinal time; k-exposure; must be others). We're more interested in the general case when time data is available.

Statistical tools: Using joining time as a predictor/control variable (e.g., in regressions). (Find examples). More sophisticated time series analysis. But, we're more interested in simpler analyses and effective ways to visualize behavior.

Adjusting time relative to phenomena or events of interest. Normalizing clock times to local time (Macy and Golder twitter paper) or internal clock time (Liz paper under submission to CHI). Normalizing to an event of interest (e.g., Crandall paper 'first interaction' in Wikipedia).

Using cohorts. (Find examples.) [**?**]

What can "different" be? Effort, activity, survival?

Are communities evolving in different ways based on when they are created in the network?

3. DATA: REDDIT AS A COMMUNITY

We start with a brief overview of both Reddit and the dataset that we use in this paper, focusing on aspects that directly impact our analyses[1].

3.1 What is Reddit, briefly

Reddit is one of the largest sharing and discussion communities on the Web. According to Alexa, as of late 2015 Reddit is in the top 15 sites in the U.S. and the top 35 in the world in terms of monthly unique visitors. It consists of a large number of subreddits (853 thousand as of June 21st, 2015[2]), each of which focuses on a particular purpose. Many subreddits are primarily about sharing web content from other sites: in "Pics", "News", "Funny", "Gaming', and many other communities, users ("Redditors") make "submissions" of links posted at other sites that they think are interesting. In other subreddits, Redditors primarily write text-based "self-posts": "AskReddit", "IamA", "ShowerThoughts" are places where people can ask questions and share stories of their own lives. Generically, we will refer to submissions and text posts as "submissions".

Each post can be imagined as the root of a threaded comment tree; in addition to posting, Redditors can make comments, and vote on both posts and comments. Votes are used both to sort comments within a post and posts within a subreddit, and also form the basis of "karma", a reputation system that essentially tracks how often people upvote a given Redditor's comments and submitted links. Redditors can also create and volunteer to moderate subreddits.

We choose Reddit as our target community for a number of reasons. It has existed since 2005, meaning that there has been ample time for the community to evolve and for differences in user cohorts to appear. Second, being composed of a number of diverse subreddits allows us to explore questions of how communities diverge over time. Third, Reddit data are publicly available through an API.

3.2 The dataset

Redditor *Stuck_In_The_Matrix* used reddit API to compile a dataset of the public available comments [3] from October 2007 until May 2015. Due to API call failures, he was not able to get about 350 thousand comments. The dataset is composed of 1.65 billion comments. He also compiled a submissions dataset for the period of October 2007 until December 2014, that was made available for us upon request. It contained a total of 114 million submissions.

These datasets contain the JSON data returned by the reddit API. The information in each of these objects [4] contained the UTC creation date, comment text, author username, subreddit, all relevant information for this work.

We focus on submissions and comments in the dataset because they have timestamps and can be tied to specific users and subreddits, allowing us to perform our time-based anal-

[1]There is much more to say about both Reddit itself (see `https://www.reddit.com/about/`).

[2]`http://www.redditblog.com/2015/06/happy-10th-birthday-to-us-celebrating.html` for more numbers on reddit size.

[3]Available in `https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment`.

[4]Full description of the JSON objects in `https://github.com/reddit/reddit/wiki/JSON`.

yses. In some analyses, we look only at comments; in some, we **combine comments and submissions, calling them "posts"**. We would also like to have looked at voting behavior as a measure of user activity[5], but individual votes with timestamps and voting user are not available through the API, only the aggregate number of votes that posts receive.

3.3 Our processing

To analyze the data, we used Google BigQuery [6], a big data processing tool that allowed us to efficiently navigate within the dataset. The work of importing the comments into BigQuery was done by the redditor *fhoffa*, and the data is publicly available [7]. For the submission data, we uploaded it using Gooogle's SDK [8].

As our initial pre-processing, we filtered comments and submissions presenting deleted users, data with no creation time and applied a very simple bot removal. The SQL constraints are shown below.

```
created_utc is not NULL
and author <> '[deleted]'
and not right(author, 4) = '_bot'
and not right(author, 3) = 'Bot'
and not lower(author) contains 'transcriber'
and not lower(author) contains 'automoderator'
```

We also considered only comment data from October 2007 until December 2014 in order to have a matching period for comments and submissions. After this process, we had a total of 1.17 billion comments and 114 million submissions.

3.4 An overview of the dataset

Here we present an overview of reddit and how it grew in the past few years. Figure 1a shows the cumulative number of user accounts and subreddits created over time as of the last day of every month. After an initial extremely rapid expansion from 2008–2009, both the number of users and subreddits have grown exponentially. As of the end of 2014, about 16.2 million distinct users have made at least one comment and 327 thousand subreddits received at least one comment since Reddit's inception.

However, as with many other online sites, most users [] and communities [**?**] do not stay active. Figure 1b shows the monthly number of user accounts and subreddits that made or received at least one post (our definition of post is either a comment or a submission). We define as an **active user the one that made at least one post in the month** in question. Similarly, an **active subreddit is the one that received at least one post in the month**. In December 2014, about 470,000 thousand users made posts and about 11,400 subreddits received posts; both are an order of magnitude less than the cumulative number of users or subreddits.

[5]Which would also give us more insight than usual into lurkers' behavior; we'll return to this in the discussion.
[6]`https://cloud.google.com/bigquery/`
[7]More information in `https://www.reddit.com/r/bigquery/comments/3cej2b/17_billion_reddit_comments_loaded_on_bigquery/`, users with a Google account can process up to 1TB of data for free in BigQuery by the time this paper was written.
[8]Part of the alpha code in the SDK, "gcloud alpha bigquery import".

The fact that such a significant amount of users stopped using the platform raises questions such as why users give up on their accounts, when they do so, and which users are more likely to stay active. In later sections we will take a closer look at how some of these things are happening in reddit.

3.5 Identifying cohorts

As we defined before, the account creation time is the time of the first post the user created. Similarly, the subreddit creation time is the time of the first post that it received. Throughout this paper, we will use the notion of user (subreddits) cohorts, which will consist of users (subreddits) with the same creation year.

In many cases, we will look at the time evolution of these cohorts. Since users (subreddits) can be created in the beggining of the referred year or in the end and our dataset has a date limit threshold, we are likely to have a variation on the data available for each user of up to one year, even though they are in the same cohort. To deal with this, many of our cohorted analysis will consider only the overlapping time window for which we collect data for all users. This means that we are going to cut one year at the end of some analysis to guarantee that we are not biased by users created early or later inside of the cohort.

Our data also starts in October 2007, but reddit has existed before it. That means that, not only we have incomplete data for the 2007 year (which compromises this cohort), but also there might be users and subreddits that show up in 2007 that were actually created in the previous years. Since we can not control for these, we will also omit 2007 cohort, for it is not representative of 2007. We will, however, include 2007 in the overall analysis over time (the non cohorted ones) for two reasons: first, it does not have any direct impact in the results, only extends the axis for 3 extra months and secondly, we often compare the cohorted approach with a simple, naive approach, and we would not expect a naive approach to do any kind of filtering.

4. POSTS PER USER

In this section, we will use a common metric of user activity in online communities, the number of posts per user over time.

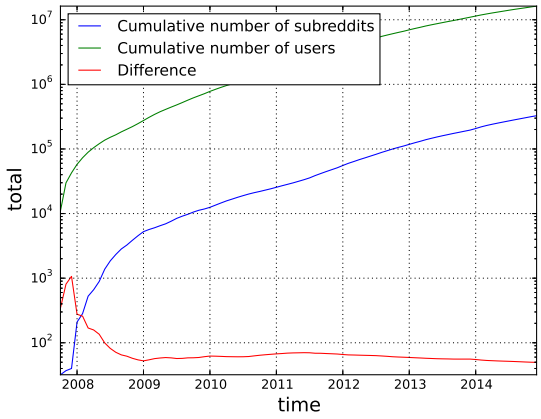
As we will see, both visualizing behavior relative to a user's join time rather than calendar time and using cohorts provide additional insight into how people's posting frequency in Reddit The first approach uses a notion of time relative to an event of interest (such as a user's first post); the second focuses on cohort effects.

4.1 The Aggregate View of Users' Activity

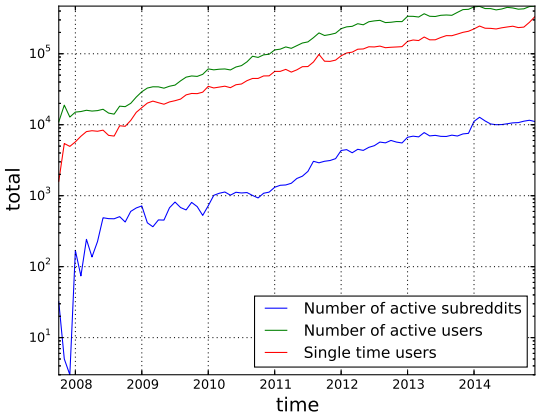
Figure 2a shows the average number of posts (submissions plus comments) per month by users who were active in that month. Taken at face value, this suggests that over the first few years of Reddit, users became more active in posting and that per-user activity has remained more or less steady since mid-2011.

4.2 Activity relative to a user's lifespan

This average view hides several important aspects of users' activity dynamics. In Figure 2b, we show a different view that emphasizes the trajectory over a user's lifespan. Here, we scale the x-axis not by clock time, as in the left figure,

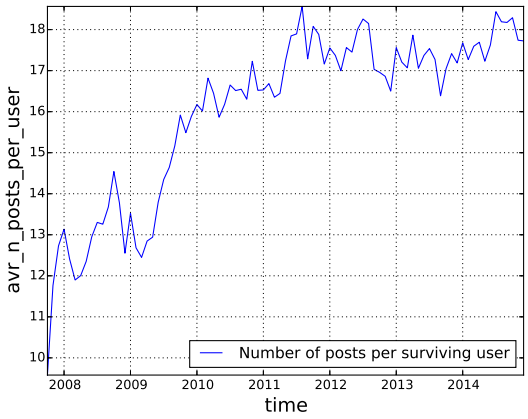


(a)

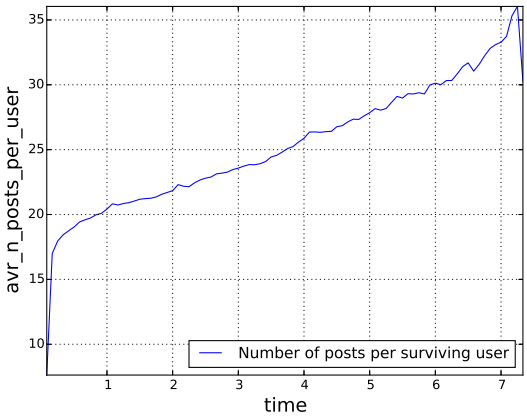


(b)

Figure 1: Figure a shows the cumulative growth of reddit for users and subreddits. Figure b shows the number of active users and subreddits in reddit over time. An active user or subreddit is the one that presented at least one post (comment or submission) in the time bin we used — time here is discretized by month.



(a)



(b)

Figure 2: On the left, monthly average of posts per active users over clock time. On the right, the monthly average of posts per active users in the user-time referential, i.e., each message creation time is measured in terms of when the user was created (defined by the first post). Each tick in the x-axis is one year. Since we are looking at the user time referential, the number of users at each month are the surviving users after x time.

but by time since the user's first post: "1" on the x-axis refers to one year since the user's account first post, and so on. We define this date as the **account creation time**. One caution about interpreting the graphs that are relative to the user's start time is that the amount of data available rapidly decreases over time, meaning that values toward the right side of an individual data series are more subject to individual variation.

A tempting conclusion at this point is: the longer a user survives, the more posts they make over time. This, however, is not exactly why we see this curve increasing from the user-time referential. The correct interpretation here is: given that the user survived, his/her posting average is going to be higher. As we will see, cohorting and segmenting the users will reveal more of the underlying process.

4.3 Cohort-based views of posting activity

Implicitly, Figure 2b suggests that older (in the sense of first account activity) users are more active than newer ones, raising the question of whether newer users are likely to eventually follow in older users' footsteps. Analyzing users' behavior by cohort, grouping them by account creation year, is one reasonable way to address this question. Figure 3a shows our first attempt at this analysis. This figure already shows a significant cohort effect: users from latter cohorts appear to level off at a significantly lower posting average than users from earlier cohorts.

However, the figure also has an awkward anomaly, the sharp increase in the average number of posts at the end of each cohort's first calendar year in Figure 3a. Since we segmented our users by cohorts, by the end of each year

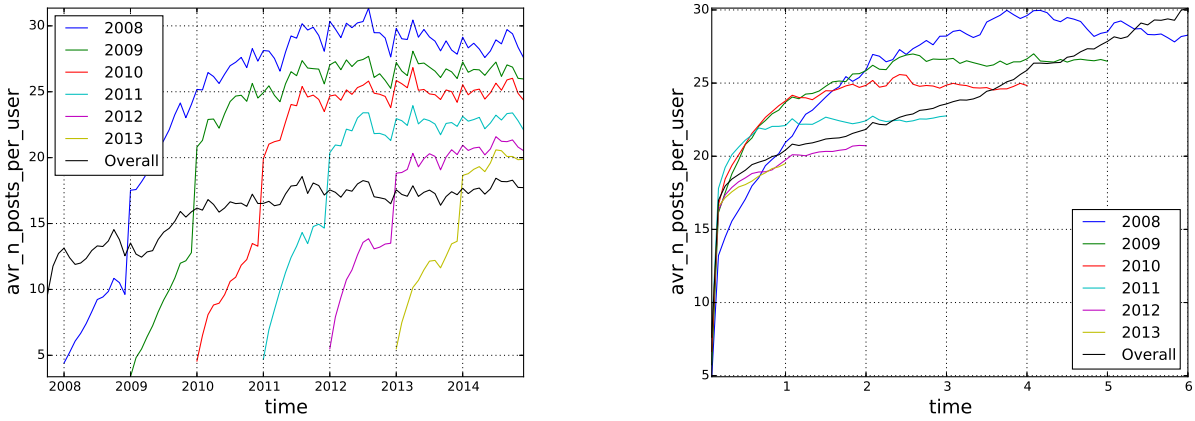


Figure 3: Average number of posts per active (left) users over clock time and surviving (right) users in the user-time referential, both segmented by users' cohorts. The user cohort is defined as the year the user made his/her first post.

onwards, we do not have users joining the network any more and the influx of low posting, low survival users stops. To support that it is indeed the case that these users are causing these effects, we need yet other representations of the users' evolution.

The fact that users in their early existence have a low posting average can be seen in the user-time referential Figure 3b. These account for the lower values in the cohort year for the in Figure 3a. We could, however, imagine that low posting users would keep around after the end of the cohort year and keep dragging the average down. Figure 4 show that this is not the case. It shows, for the years of 2010, 11 and 12[9] that, when segmenting users by the number of years they survived in the network, the low posting ones are the first to die.

Figure 3b actually reveals more than just why we see the anomaly on the Figure 3a. In the long run, a striking pattern emerges: different cohorts stabilize in different levels of behavior, and in particular, the steady state activity for surviving users goes down for every year from 2008 to 2012. Since Figure 4 tells us that the main reason why these curves increase is because the low posting users are dying sooner, it means that a higher number of low-posting users are surviving in the later cohorts. It is also important to notice that Figure 4 shows that users are more likely to post less as they live on.

5. COMMENT LENGTH AS A PROXY FOR EFFORT

In addition to the raw number of posts, comments length can also be considered as a proxy for user effort in the network. Users that type more put more of their time in the network, contribute with more content and might create stronger ties with the community. Similarly to what we did for the average posting per user, we analyze the overall average and cohorted trends.

Based on the downwards tendency of the overall comment length in Figure 5a, one could possibly imagine that the users' commitment to the network is decreasing over time. This, however, might not be the best way to interpret this information. Figure 5a also shows the comment length per cohort based on the user referential time and, unlike the average overall network comment length, surviving users increase the size of their contributions to the community over time. This puzzling observation is called Simpson's paradox, that we discuss in detail in the next section.

The important thing to notice here is that, while user comments get longer as they stay for longer in the network, younger users start from a lower baseline comment length than older users. Together with the fact that recent reddit has experienced exponential-like growth, the weight when evaluating the overall average for Figures 5a and b as the years go by is shifted towards the size of the ever growing younger generation, and this younger generation brings the average down since they start by writing less.

As with the posting per user, we can not say if the increase in the curves seen in 5b are due to the lower effort users dying first or because users are indeed writing more as they live on the network. To answer this, 5c allow us to make two important observations: first, *comment length do increase inside of each cohort*, no matter how long the user survives. Secondly, as a general trend, *users that make longer comments inside of each cohort die faster*. This is quite surprising, given that we would expect people to put less effort when they are more likely to stop using the network.

5.1 A Time-Based Example of Simpson's Paradox

Let us go back to Figure 5a, which shows the overall average comment size on Reddit over time. We see a clear trend towards declining sizes of comments. Across all users, we see that average size of a comment decreases as the community grows older. This could be a warning sign for reddit community managers, assuming longer comments are associated with more involved users and healthier discussions. A data analyst looking at these numbers might think about ways to

[9]We only show these figures for a matter of space, but the same trends are observed for the other cohorts.

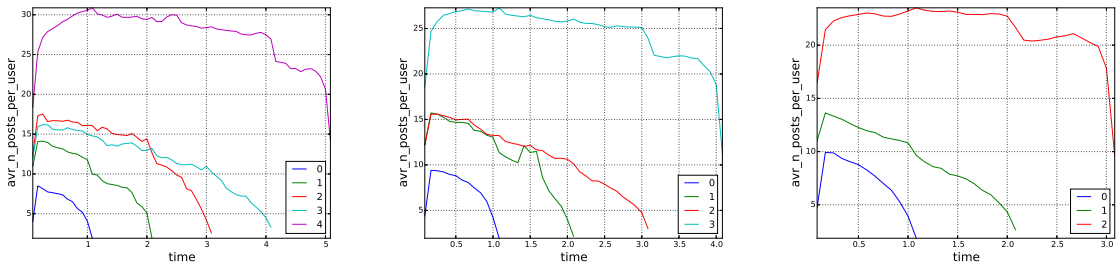


Figure 4: Each figure corresponds to one cohort, from 2010 to 2012, left to right. The users for each cohort are further divided in groups based on how long they survived: users that survived up to 1 year are labeled 0, from 1 to 2 years are labeled 1, and so on. We observe that, for all cohorts, users that have a lower posting average are the first to die.

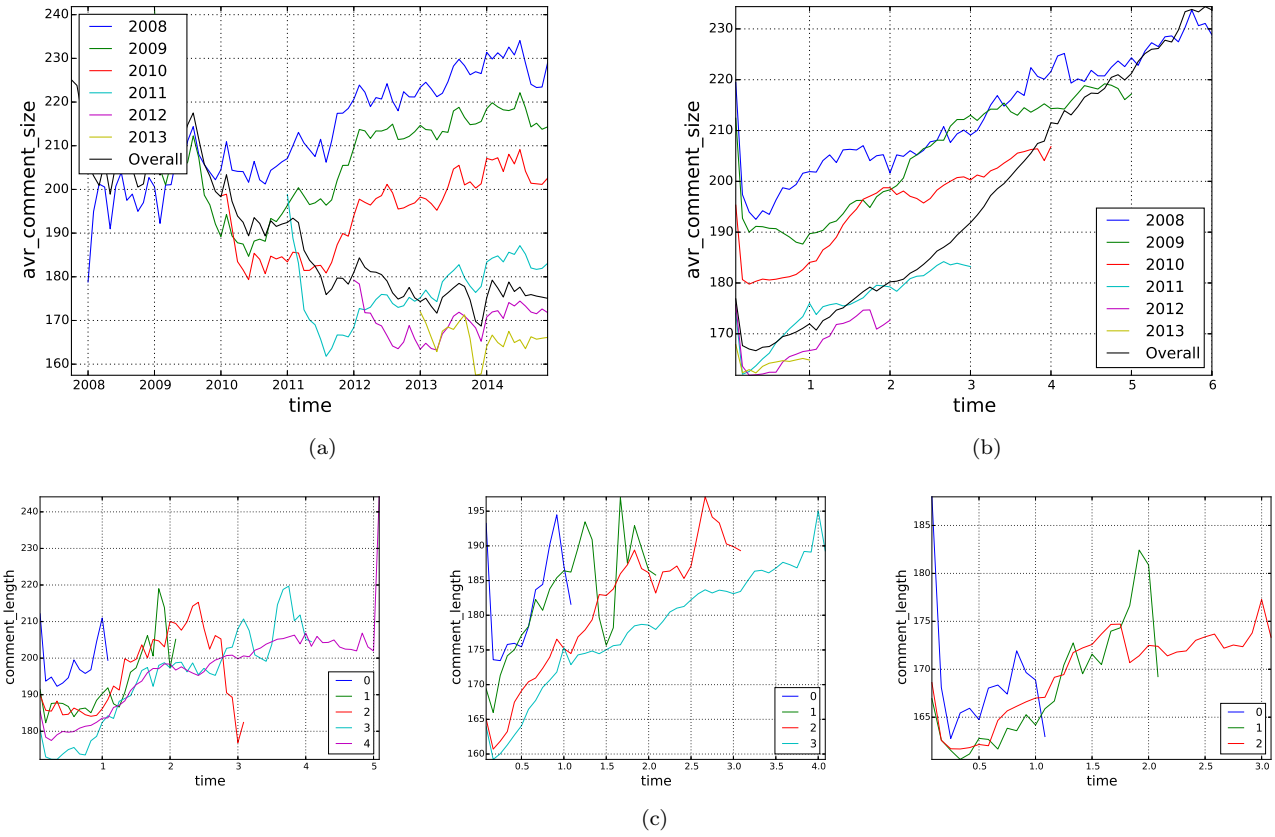


Figure 5: Figure a shows the average comment length over clock time and Figure b from the user-referential time. Both figures show the cohorted trends and the overall users trends. Figure c, just as in Figure 4, correspond to the cohorts of 2010, 11 and 12. They show the average comment length for users segmented by the number of years that the user survived in the network.

incentivize or promote longer comments on Reddit.

However, in the same Figure 5a, we saw that average comment size increases over time for every cohort. While later cohorts start at smaller comment sizes, after an initial drop, all of the cohorts show a positive trend towards writing bigger comments as time goes on. This is puzzling: when each of the cohorts exhibit a steady increase in their average comment size, how can the overall mean comment size decrease? This anomaly is an instance of the Simpson's paradox, and occurs because we fail to properly condition on different cohorts when computing mean comment length.

Table 1 provides some clues to what might be going on. When we move down the rows, we observe an increasing tendency in each column. It means that the average comment length increases for these users. However, when we move right through the columns, people in later cohorts tend to write less per comment. If we were to average each row, we would still get and overall increasing comment length per year, but that is not what we see in the overall column. What happens here is that the latter cohorts have

Cohort	2007	2008	2009	2010	2011	2012	2013	2014	Overall
2007	114	-	-	-	-	-	-	-	114
2008	106	99	-	-	-	-	-	-	103
2009	113	101	99	-	-	-	-	-	103
2010	114	103	96	91	-	-	-	-	96
2011	119	109	103	93	83	-	-	-	91
2012	125	114	110	101	87	81	-	-	89
2013	126	117	111	104	92	82	80	-	87
2014	128	119	113	106	95	87	83	82	88

Table 1: Evolution of the median throughout the years for each cohort. Each column here is one cohort and each line is one year in time. Cohorts only start having data on the cohort year, therefore the upper diagonal is blank. On the right column we see the overall median for all users.

many more users than earlier ones. Since their numbers increase year by year, we have a much larger contribution from them towards comments, compared to users of earlier cohorts. This uneven contribution leads to the paradox we observe in the Table 1.

Without the decision to condition on cohorts, one would have gathered an entirely wrong conclusion. People are not writing less as they survive, rather those who tend to write less are joining the community in much larger numbers. Knowing this, one may focus on better onboarding processes for newcomers, or evaluate why users in later cohorts tend to write smaller comments on average.

6. KINDS OF CONTRIBUTIONS PEOPLE MAKE

One common question from the literature is what sorts of activities users engage in; this can be used as a metric of community health (cites) or to categorize users into roles they play in the community (cite). In reddit, we do not have per-user voting behavior, but we do have the number of comments and submissions, and a naive view of this would look at the ratio of comments to submissions over time.

While submissions can be considered new content that an author generates, a comment can be considered as a contribution to an existing content from another author. Since the total number of comments always surpasses the number of submissions, Figure 6a shows the evolution of the overall and cohorted ratio of comments per submission over time for users created from 2008 until 2013. It is important to highlight here that we are not talking about the average number of comments a submission gets, but how many comments a user authors for each of his/her submissions. We observe an increasing trend in the overall and cohorted ratios.

Figure 6b is more informative of what is happening. It shows a clear pattern for users in older cohorts to have a lower comment per submission ration than the younger ones. Also, in Figure 6c we observe that, given the time a user will survive, the ratio seems to level at lower values for users that die faster. This suggests that the increasing tendency in Figure 6b is mainly due to the death of low-ratio users. Also, this indicates that this ratio, conditioned on the cohort, can be a good predictor of user survival.

7. PUTTING IT TOGETHER ON EFFORT

In the previous sections we observed that the average effort per post for older cohorts increases as the users survive in the network. We also observed that users from older cohorts present higher effort per post for the same survived time than users from earlier cohorts. This means that as you age in reddit, you write more per post, and the earlier you joined, the more you write. But we also observed that users from earlier cohorts are commenting more per submission than users from older cohorts for the same time survived in the network. Could users be actually putting the same effort in terms of number of written characters, but younger users do it writing more, shorter comments while older users write less, longer comments? To investigate this, we follow the same steps as in the previous sections, analyzing the overall and cohorted behaviors, over time and from the user time referential.

In Figure 7a, we observe that during most of the time, the overall average number of characters written per user per month stays between 3000 and 3500, peaking near 2010 and showing a slightly downwards tendency throughout the end of 2014. The cohorted curves show a different growth pattern in comparison with the overall trend, mainly increasing and then leveling at different values, with older cohorts higher than younger ones. The decreasing overall trend happens because the latter cohorts have a much more significant weight in the average due to the increased number of users that joined reddit in the later years. This highlights the differences of the overall trend for the cohorted trend: while the overall shows a slightly decrease towards 2014, the cohorts show an increasing and leveling behavior. This can lead to wrong conclusions if not treated properly.

To further investigate how users evolve in the network, we see in Figure 7b the number of written characters per month per user from the user time referential for the overall average and the user creation date cohorted curves. We observe a sharp increase in the beginning of all lines due to the fact that a significant number of users only survive a very short time and the total amount of characters they contribute is considerably lower in comparison with the ones that survive for longer. The effect these users have in the analysis is concentrated in the leftmost part of the graphic, which improves the analysis in this referential. We can see that users, as they survive, write more characters per month. This can be due to the fact that users write more as they age and/or because users that write less die first and the surviving ones are the ones that write the most. From the user perspective, we see that the evolution of overall trend and the cohorted ones are significantly different. The overall trend shows a positive second derivative and apparently keep increasing for older users, while the cohorted ones have a negative one and eventually level. The conclusions about how users behave based on this can be quite misleading, specially considering that it is not reasonable for users to be forever increasing the amount of written characters as they survive.

To understand the increasing amount of written characters as the users age in the network, Figure 7c shows the per cohort set of figures that segments the users in each cohort according to the number of years survived. We can see clear trends of users leveling in different values of written characters per month according to the number of years they are likely to survive. This means that most of the increasing behavior of the user-time referential is due to users that write few characters dying earlier. Also, this suggests that the number of characters written per month by users

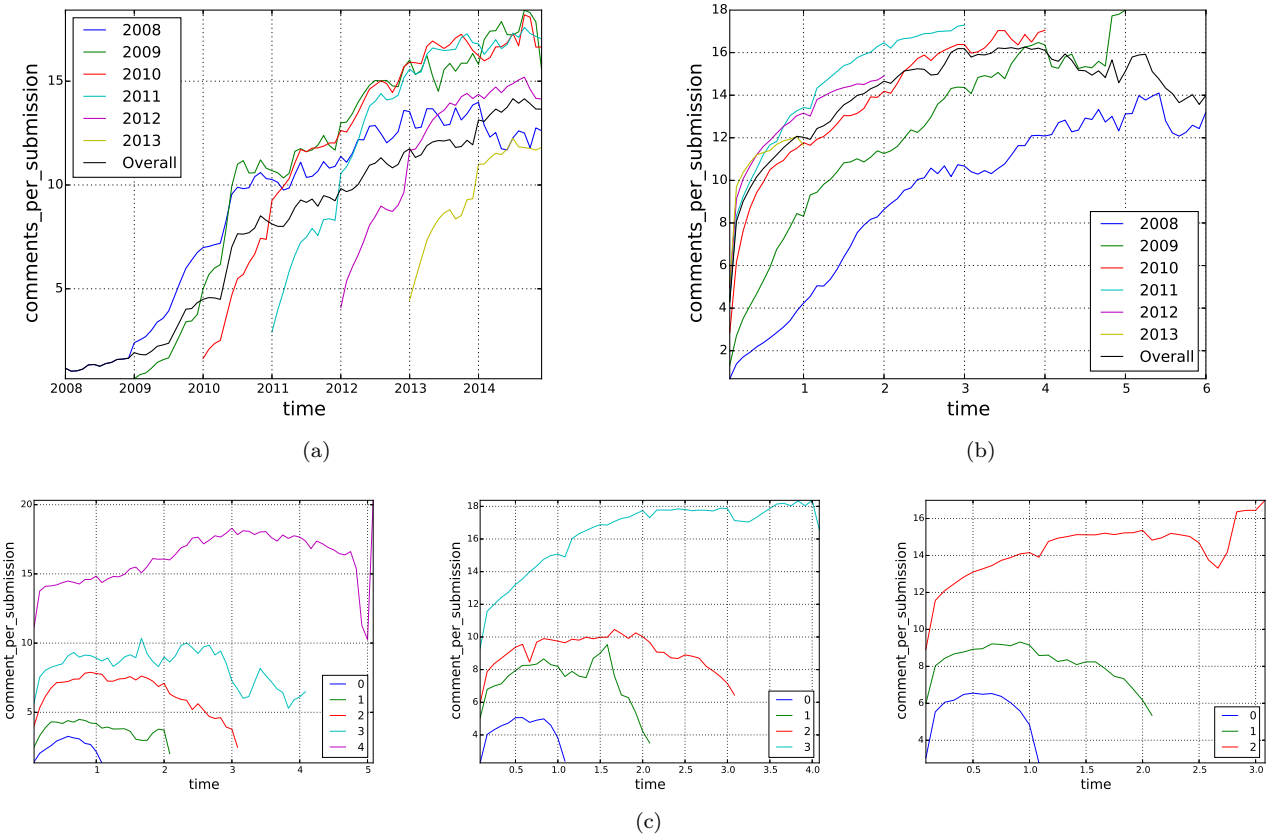


Figure 6: Figure a shows the average comment per submission ratio over clock time and Figure b from the user-referential time. Both figures show the cohorted trends and the overall users trends. Figure c, just as in Figure 4, correspond to the cohorts of 2010, 11 and 12. They show the average comment per submission ratio for users segmented by the number of years that the user survived in the network.

conditioned on the cohort year is a good predictor of user survival.

8. USER BIAS FOR EARLY CONTENT

The cohort a user belongs to has a significant impact to the user posting behavior, but that does not give us a picture of how these users coexist in the current community or the communities evolution on time. An interesting hypothesis that we could imagine is that users from a particular cohort are more interest in the communities from a particular cohort. We now look at the interplay between user and subreddit cohorts.

Figure 8a shows the number of posts in the subreddit-time referential for cohorts on the subreddit creation year. The most striking observation here is how 2008 subreddits are significantly more active than other subreddits for the same survived time. This led us to question what could be the process for this bias. One piece of evidence that we found that can significantly bias the user posting choices are the default subreddits a user is automatically subscribed upon creation [10]. These subreddits change over time and are said to be "a set of highly popular communities that the administrators of this site feel would give the average person an interesting first experience" [11]. Even if this set is defined only after the subreddits get popular, there might be a positive feedback that maintains the "core subreddits" as the most popular ones. We observe the bias for 2008 when we count each of the default subreddits set over time according to their creation year, seen in Table 2.

	2007	2008	2009	2010	2011	2012	2013	2014
December 31, 2009	5	6	1	-	-	-	-	-
October 18, 2011	3	14	2	2	-	-	-	-
October 19, 2012	2	16	3	2	2	-	-	-
July 17, 2013	2	15	2	1	2	-	-	-
January 1, 2014	3	14	2	2	3	-	-	-
April 19, 2014	3	13	2	2	2	-	-	-
May 7, 2014	4	23	6	5	4	7	1	-

Table 2: Count of subreddits per creation year for each default set of.

Figures 8b and 8c show a more current picture of red-

[10] https://www.reddit.com/r/defaults

[11] https://www.reddit.com/wiki/reddit_101

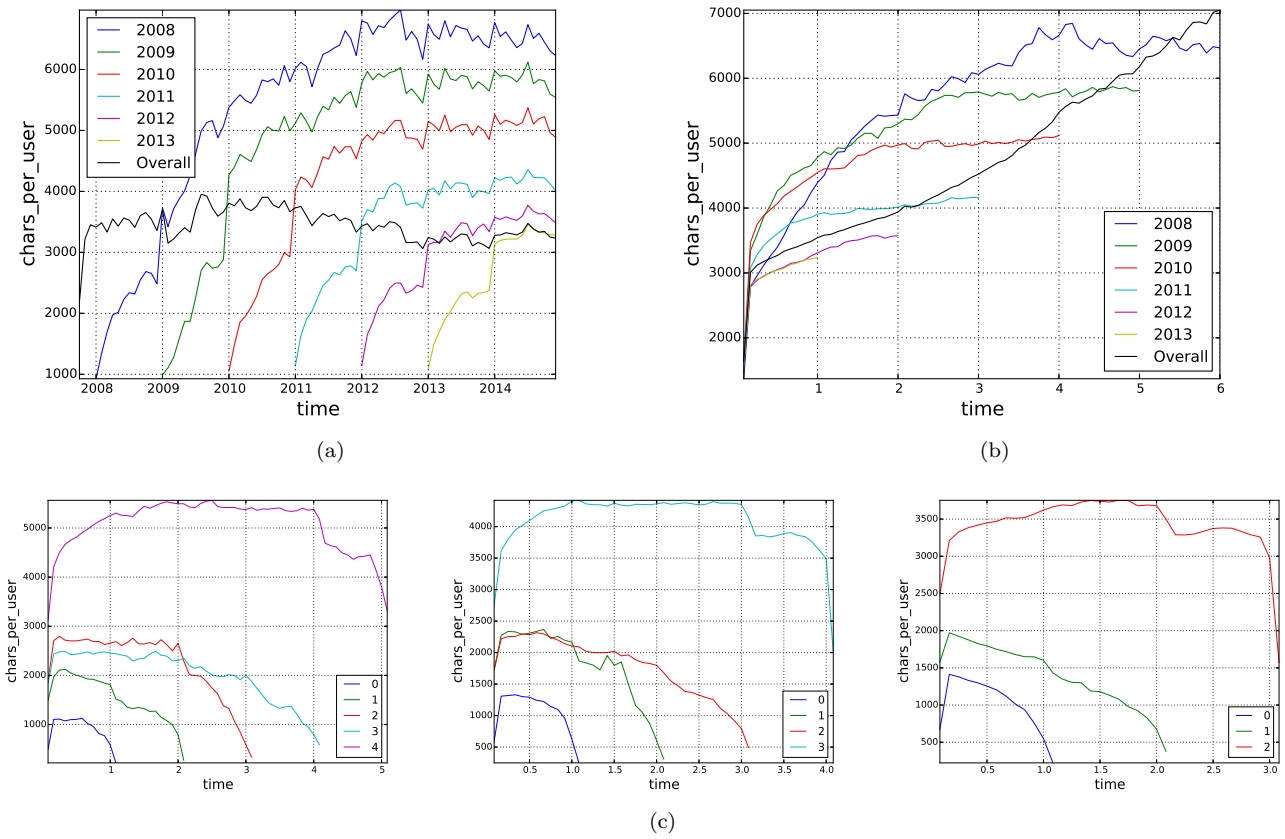


Figure 7: Figure a shows the average number of written characters per user over clock time and Figure b from the user-referential time. Both figures show the cohorted trends and the overall users trends. Figure c, just as in Figure 4, correspond to the cohorts of 2010, 11 and 12. They show the average number of written characters per user for users segmented by the number of years that the user survived in the network.

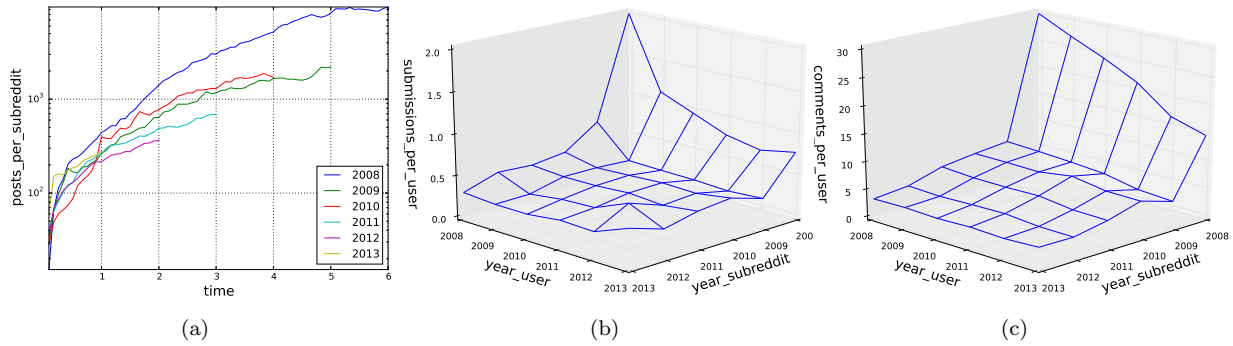


Figure 8: Figure a shows the monthly posts per surviving subreddit cohorted by the subreddit creation date (first post to the subreddit). Figure b and c shows, respectively, the number of submissions and comments per user in reddit for the last three months of 2014, segmented by user creation year and subreddit creation year. Each bin counts the number of comments or submissions based on the user and subreddit creation year and divide by the total number of active users created in that year.

dit. While submissions in Figures 8b are in general more skewed towards 2008, it is even more striking that users from 2008 submit even more. This might be due to the fact that these surviving users play a much more central role in these communities (moderators or key contributors) since they are more likely to be there from the start (relative to the total number of users from the cohort that are still active), as we see evidence in Table 3. Figures 8c show that comments are even more skewed than submissions towards 2008 subreddits. The tendency, however, is of decreasing commenting behavior for the earlier user cohorts (we have seen that users from earlier cohorts are posting less, which

	2007	2008	2009	2010	2011	2012	2013	2014
Moderators	757	1182	2108	4085	8059	9340	6868	4262
Percentage	8.52	6.02	5.2	3.91	2.54	1.5	0.82	0.18

Table 3: The first lines show the absolute number surviving users that posted as moderators in each cohort. The second line shows the relative percentage to the surviving users for the same cohorts. Were considered active the users in the last three months of 2014.

justifies this decrease).

9. DISCUSSION

In this section we discuss some of the processes that might explain our observations and refer to the associated literature. We're not taking a position on either of these as the mechanism that explains these results; both would be interesting avenues for future work. We do suggest that looking at Reddit from a cohort and user-based view rather than an aggregate community view helped us uncover interesting phenomena and questions that would have been invisible to more commonly-performed analyses of community behavior.

9.1 Why more low activity users are surviving?

We have seen that users from latter cohorts have a lower posting average than in earlier cohorts. One plausible explanation is that users who find a community suffer from a self-selection problem: users that find reddit early in its life are also more likely than average to be those who will be attracted to it. Previous work has shown that online book reviews have a self-selection problem, since early reviews tend to be positively biased [**?**]. This would mean that the mixture of users joining in the early stage of the community are more likely to be the most active ones and the latter ones are more likely to be less active. A higher number of less active users joining the network would also account for their longer survival, although other mechanisms might also be in play.

Another is an argument based on cumulative advantage, status, and attention-seeking: surviving users from earlier cohorts might be more capable of producing content that gets attention from other users. This would lead to them getting more comments and votes for their content, and people who get positive attention are more likely to return [**?**, **?**, **?**].

9.2 Why are comments getting shorter?

One hypothesis that we might consider to explain the decreasing comment length of the users is associated to an "initial value problem". We can imagine that users, as they join the network, tend to produce content according to the norms of what they see [**?**, **?**]. The observed behavior of the comments length for the users in reddit is a initial drop, followed by steady increase as the user survives. If the starting point for the initial drops are taken as the average of the network, that is what is observed by the user in the network, the initial drop would place each cohort starting at lower levels than the previous one.

9.3 Why comments per submissions are increasing?

The majority of users in social networks are known to be lurkers: users that only seek information and passively observe, not engaging and contributing to the network [**?**, **?**]. It is reasonable to expect the same from reddit. On the other hand, social networks often have a small number of "power contributors" [**?**, **?**].

When we consider the evolution of the number of comments per submission, we observer a decreasing trend for the older cohorts. Just as lurkers are the majority in the community and are attracted to it in search of information, we can imagine a set of users that are in the community in search of content and is whiling to engage in commenting, but hardly has the drive of a "power contributor" to bring new content into the network. Since in early reddit the amount of existing submissions was significant smaller than in the next few years, limited space existed for information seeking lurkers and "commenters". As the community grew, more content from an absolute point of view was present in the social network, therefore activities of information seek became more prominent and could explain the increase in commenting activity.

9.4 Why the bias to 2008?

Were we expecting different results? Defaults? Core consolidation? These observations allow us to conclude that, in the case of reddit, there are key subreddits that were created in 2008 that are the main focus of attention of all the users, although this is decreasing as new users join the network. Our hypothesis would not hold true in this case. This, however, might not hold true for other social networks, in which the communities or the content at the time at which users join the network might be their main focus of attention, highlighting again the importance of performing a cohort based analysis.

10. CONCLUSIONS

This work addresses some aspects of how to analyze the evolution of a social network and how to apply and avoid some pitfalls of cohort analysis. To do so, we analyse the reddit network and provide insights on the users posting behavior evolution, we identify a general tendency of newer users to write smaller comments and we discuss user survival from empirical standpoint.

We also analyse subreddits evolution, considering the volume of activity, how commenting and submitting change in function of cohorts and the matter of community survival as subreddits can be seen as independent communities.

11. FUTURE WORK

Although many observations were made regarding reddit, the main goal of this paper is not to study reddit in depth. Therefore, many observations still remain without an explanation. Being aware that users decrease the size of their comments and tend to submit more than comment allow us to question ourselves about the nature and motivation of these users. Are users commenting less because newer users they are less interested into engaging into existing submissions or are they simply "lazier" and writing is becoming less attractive for the newer generations? If it is a matter of writing being a less prefered method of communication, could reddit improve its users' interactions providing alternative ways to post media, such as pictures, sound and videos? Re-

garding subreddits, why are newer subreddits more likely to die? Is it because they can not compete with larger subreddits? Or is it because they are not in direct competition, but they cover a smaller and more niche like space of interests that make it less popular and more likely to die? Or it might just be that their creators from are from newer cohorts and, just as they are "lazier" to write, they can not keep up with maintaining their new communities?

These are some questions that we can ask and needs further investigation of the data, and might evolve into new models of how communities compete for spaces of interests or how user attention, effort and interaction in social networks are shifting through time.

One interesting question that recurrently arises when we were performing survival analysis was regarding the users' "breaks" from the network. As "death" in a social network is not well defined - users can delete their account, which is a clear sign of death, but they can also simple stop using the network - using the last posting date and setting thresholds for activity to be considered alive might not be enough. A next step would be to investigate better definitions of "death" and study different types of user behavior to characterize the burstiness of their behavior. Some users might only interact with the network in some specific occasions while some users might have a much more uniform pattern. Understanding how your network fare in terms of user burstiness is essential to understand how the users use the network and to set goals to improve or change the user experience. Also, a better definition of "death" would allow us to investigate the "rebirth" of users, that is, users that come back to the network. Whereas here we consider it as a right censorship problem, it might pose a much more central issue, as network survival might not only depend only in the ability to attract and retain users, but also in the ability to restore old users.

12. ACKNOWLEDGMENTS

Acknowledgments.