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.

We need categories, terms, and keywords
user behavior cohort reddit

1. INTRODUCTION

Understanding the users evolution in a social netowork is essential for a variety of tasks: information diffusion, behavior predicion, recommendation tools, among other possibilities. 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 **??**, some others adopt a more user focused point of view and try to assert more individuals characteristics **??** or predict their behavior **??**.

In dealing with these large temporal studies, it is quite common for researchers to have limited data in terms of time — just a snapshot of the network is known **??** — or limited in terms of scope — only a small community is observerd over the years **??**. It is also not unheard of cases in which datasets were made unavailable upon request from these social networks [1].

2. PREVIOUS WORK

In previous work, researchers have studied the relationship of different cohorts adopting new technologies and how users that did not grow in a technological environment show different characteristics when compared with the younger generations. This external variable to the social network might explain many different aspects of how adoption of a network happens. Just as users experience outside the network vary according to their age and influence their behavior, users' experience inside the network throughout time vary as the network evolves. Users in the early stages of a social network have a very different experience from latter users.

Are users evolving in different ways based on when they join the network? How is an early user different from a late user?

Evolution in this sense can be interpreted in many different ways. Researchers have looked into many aspects of how user behavior change, how frequent they post, how users adopt new language, how likely a user is to survive in the network (which is also related with the problem of predicting which users are going to depart from your network). Based

[1] `http://twitter.mpi-sws.org/`

WWW '16 Montereal, Canada

on this, we have to understand what we are looking for in the user behavior.

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

This evolving process of users changing inside of the network change the network itself. We know that the idea users have from a social network might change their willingness to try it, just as we know how the initial experience might impact in the user future behavior [7]. But the community evolving in itself changes the idea users outside have about it [3]. This internal evolution together with the novelty that the influx of users bring make reddit a very interesting evironment to understand, for sub-communities known as subreddits as being created all the time and in different contexts, which raises the following question.

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

December 31, 2009	announcements, AskReddit, blog, funny, gaming, pics, politics, programming, reddit.com, science, worldnews, WTF
October 18, 2011	AdviceAnimals, Announcements, AskReddit, AskScience, Atheism, Aww, BestOf, Blog, Funny, Gaming, IamA, Movies, Music, Pics, Politics, Science, Technology, TodayILearned, Videos, WorldNews, WTF
October 19, 2012	AdviceAnimals, Announcements, AskReddit, Aww, BestOf, Blog, Funny, Gaming, IamA, Movies, Music, News, Pics, Science, Technology, TodayILearned, Videos, WorldNews, WTF, Gifs, Television, Explainlikeimfive, Earthporn, books, AskScience
July 17, 2013	AdviceAnimals, AskReddit, Aww, BestOf, Books, EarthPorn, ExplainLikeImFive, Funny, Gaming, Gifs, IAmA, Movies, Music, News, Pics, Science, Technology, Television, TodayILearned, Videos, WorldNews, WTF
January 1, 2014	AdviceAnimals, AskReddit, AskScience, Aww, BestOf, Books, EarthPorn, ExplainLikeImFive, Funny, Futurology, Gaming, Gifs, IAmA, Movies, Music, News, Pics, Science, Sports, Technology, Television, TodayILearned, Videos, WorldNews
April 19, 2014	AdviceAnimals, AskReddit, AskScience, Aww, BestOf, Books, EarthPorn, ExplainLikeImFive, Funny, Gaming, Gifs, IAmA, Movies, Music, News, Pics, Science, Sports, Television, TodayILearned, Videos, WorldNews
May 7, 2014	announcements, Art, AskReddit, askscience, aww, blog, books, creepy, dataisbeautiful, DIY, Documentaries, EarthPorn, explainlikeimfive, Fitness, food, funny, Futurology, gadgets, gaming, GetMotivated, gifs, history, IAmA, InternetIsBeautiful, Jokes, LifeProTips, listentothis, mildlyinteresting, movies, Music, news, nosleep, nottheonion, oldschoolcool, personalfinance, philosophy, photoshopbattles, pics, science, Showerthoughts, space, sports, television, tifu, todayilearned, TwoXChromosomes, UpliftingNews, videos, worldnews, writingprompts

Table 1: Default subrredits over time.

Kooti et al. [4] showed that social conventions can define the evolution of a community and the early adopters play a major role in designing these conventions, even if at the time this is not known by them. Evidence for the need of a retweeting mechanism in Twitter was evident in the early stages of the community and, out of the many possibilities that coexisted, the "RT" tag survived. Early adopters of these conventions are core users, well connected and presenting high activity. Just as Twitter, reddit network evolved from a relatively small set of users and subreddits. Wheather or not these early adopter of reddit laid the foundations in terms of content and behavior is not necessarily clear. It is reasonable to imagine that users would always look for content in subreddits that were created around the time they joined the network, for they might refer to the current context they are inserted into. Therefore, we propose the following question.

	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.

Is there a consolidation point in a social network where the "core content" is established? Can this core change over time?

User-Network homophily? They connect because they are similar or do they become similar as the user evolves? Are the "dissimilar" leaving? Looking at how reddit looked like at a particular point in time is a different question from how users evolve, and much of the user evolution depends on the environment a user finds when they first join the network. In many ways, this is an initial value problem, but separating what is due to the evolution of the network and what comes from the different demographics outside the network is not always clear.

Are latter users intrinsically different from earlier users or are they having different initial experiences?

- The Taste for Privacy: An Analysis of College Student Privacy Settings in an Online Social Network [6]: Studies which characteristics are predictive of whether or not users are going to set their profile as public or private in Facebook. Raises questions about the limitations of the work because data collected came from a single cohort of users in a college.
- Social selection and peer influence in an online social network [5]: Yet another study based on a single cohort of Facebook data for college students. Discuss the relationship between homophily in creating connections and influence over the course of a connection.
- Who interacts on the Web?: The intersection of users' personality and social media use [2]: Studies how personality traits correlate with social media usage controlling for demographic variables age, gender, race, education and income. One of the research questions was whether user age cohorts influence social media usage. They found significant correlation of some personality traits with social media usage for the younger cohort (users from 18 to 29). They also acknowledge the lack of research on how age influences interaction on social media, pointing out that significant differences emerge from people that grew on a digital environment when compared to the ones that were introduced to the technology at a later time.
- "I LOVE THIS SITE!" vs. "It's a little girly": Perceptions of and Initial User Experience with Pinterest [7]:

Initial experience matters!

- No Country for Old Members : User Lifecycle and Linguistic Change in Online Communities [3]: User experience changes their behavior over time but they also come with some linguistic predispositions.
- All Who Wander : On the Prevalence and Characteristics of Multi-community Engagement [10]: Survival does depend on user initial activities.
- Wikipedians are Born, Not Made [8]: Users do have predispositions. Does that mean they do not change and we are simply sampling differently?
- Creating , Destroying , and Restoring Value in Wikipedia [9]: Not clear where it fits.
- The Impact of Membership Overlap on the Survival of Online Communities [11]: The survival of communities depends on the type of users that participate in it, and sharing certain types of users — core members from other communities that are not core members in the focal community — can be beneficial for community survival. Also, concepts of young and mature communities play a important role when analyzing community activity level, where young communities benefit from sharing members from matures communities.
- No Country for Old Members : User Lifecycle and Linguistic Change in Online Communities: Highlights the interplay on community language change and user adoption of new norms. As a general pattern, newcomers start learning the norms of the community and, as they age, they become more conservative in adopting new norms. Users that are more flexible in assimilating new norms have a higher survival rate.

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[2].

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 (X as of DATE), 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 2007, 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 the API to compile a more-or-less complete dataset of posts and comments since the inception of Reddit on XXXX. This dataset is publicly available via...

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 analyses. 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[3], 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

XXXX How did we process the dataset, both technically (using the existing BigQuery access, making our own tables), and in terms of filtering?

3.4 An overview of the dataset

Figures 1 and 2 show an overview of Reddit's growth over time.

Figure 1 shows the cumulative number of user accounts and subreddits created over time as of the last day of every month, along with the ratio between them. After an initial extremely rapid expansion from 2008–2009, both the number of users and subreddits have grwen exponentially. As of the end of 2014, about 16.2 million distinct users have made at least one comment and 327,000 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 2 shows the monthly number of user accounts and subreddits that received at least on comment. In December 2014, about 470,000 thousand users made comments and about 11,400 subreddits received comments; both are an order of magnitude less than the cumulative number of users or subreddits.

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.

3.5 Identifying cohorts

4. COHORTS MATTER: VARIATION IN ACTIVITY, EFFORT, SURVIVAL

[2]There is much more to say about both Reddit itself (see `https://www.reddit.com/about/`) and the dataset (see `https://news.ycombinator.com/item?id=9869871`

[3]Which would also give us more insight than usual into lurkers' behavior; we'll return to this in the discussion.

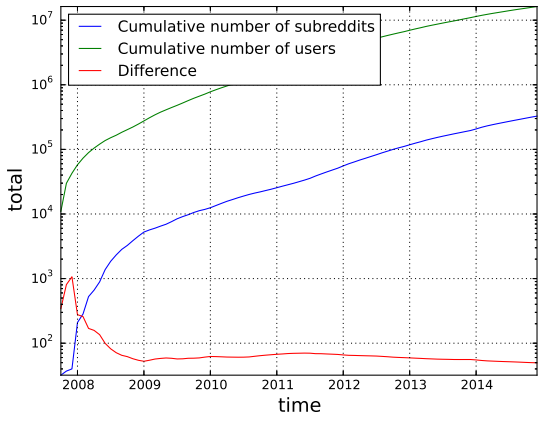

Figure 1: Cumulative growth of reddit for users and subreddits. The subtraction in log scale represents the average number of users per subreddits, that remains relatively constant for most of reddit history.

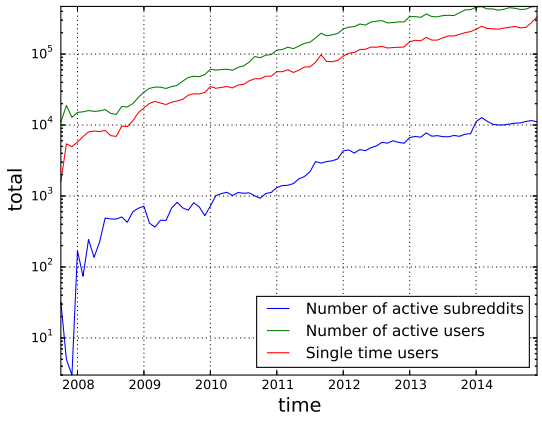

Figure 2: Number of active users and subreddits in reddit over time. An active user or subreddit is the one that presented at least one post in the time bin we used — time here is discretized by month.

In this section, we will use a common metric of user activity in online communities, the number of posts per user, to show how two main time-aware approaches to analyzing behavior provide additional insights beyond simple chronological aggregation. 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 3 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

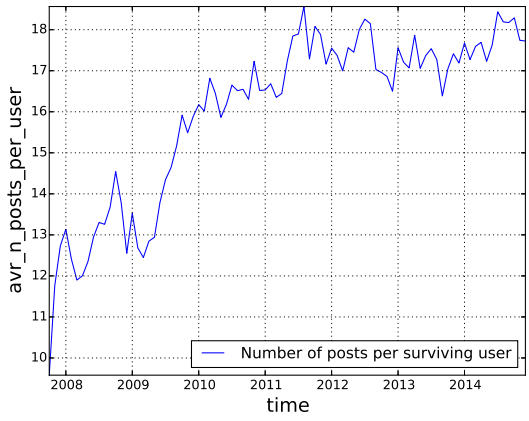

Figure 3: Monthly average of posts per active users over time. Posts as either comments or submissions, we count the total sum of comments and submissions for each month. Active users are the ones that made at least one post in the said month.

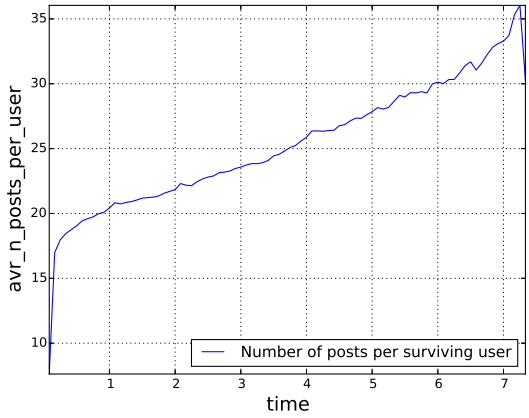

Figure 4: Evolution of all users in reddit. The x-axis is the time from the user creation referential, i.e., each message creation time is measured in terms of when the user was created. Each tick is one year and we discretized time by month — the n-th bin holds messages the user wrote in the n-th month. The user count for each month is the number of users that were active, that is, the users that authored at least one post in their n-th month. Since we are looking at the user time referential, you can understand this as the surviving users after x time. The y-axis is the number of posts per active users.

This average view hides several important aspects of users' activity dynamics. In Figure **??**, 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 prior figure, but by time since the user's first post: "1" on the x-axis refers to one year since the user's account creation, and so on. One caution about interpreting the graphs that are relative to the user's start time is that the amount of data available

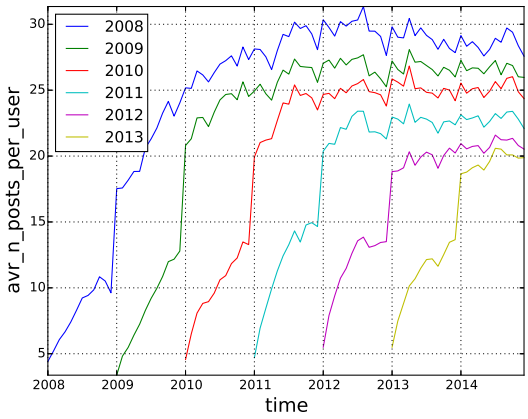

Figure 5: Average number of posts per user segmented by the user creation year. The user creation year is defined as the year the user made his/her first post. The averages over the cohort year is lower because users are constantly being created over the year. Once you move past the cohort year, the user creation halts (including the high number of "throw away accounts") and we only observe the "death" of users, that has a strong impact in the count of users for the average, therefore the significant discontinuity by the end of the cohort year. Here we observe that different years level in different values for posts per user, having older users posting more than younger users as a general trend.

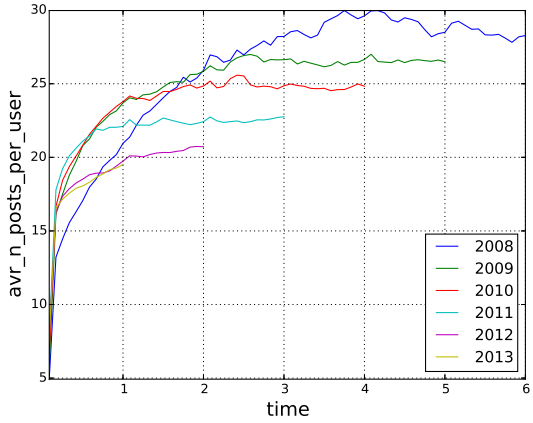

Figure 6: Number of posts per active (surviving) user for cohorts on the user creation date. An interpretation of this says "users that survived x time are posting on average y messages". Here we see that, although the 2008 cohort level at a higher value, the evolution of the number of posts took a longer time to increase. The other cohorts seem to follow a more regular pattern after the first year: older users that survived the first year post more on average than younger users that survived the first year. Since we are talking about surviving users, it is not clear from this figure whether these curves increase because the "low posting users" are dying earlier or because the users are actually increasing their activity as they live on. To differentiate these cases, Figure 7 shows, for each cohort, the average posting for users grouped by the number of years they survived in the cohort.

rapidly decreases over time, meaning that values toward the right side of an individual data series are more subject to individual variation.

This figure shows that a user's tenure matters: the longer a user survives, the more posts they make over time. Interestingly, we see that the curve rises much higher than it does in Figure **??**: users who survive over five years have almost 50% more posts per year than average.

4.3 Cohort-based views of posting activity

Implicitly, the prior figure 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 5 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 N. Since we segmented our users by cohorts, by the end of each year onwards, we do not have users joining the network any more and the number of active users does not increase because of new users anymore. This fact, along with the observation that young users tend to post less from Figure **??**, drags the average down during the first year, since young users are always coming into the network.

We can largely account for this by still treating users as cohorts based on the year they first posted to reddit, but adjusting the time reference to be relative to the user's first post rather than clock time as we did before. Figure 6 presents this view of the data.

This figure suggests that there may not be strong cohort effects early in a user's lifespan, although even after six months users in the most recent 2012 and 2013 cohorts appear to be less active than those in earlier groups. In the long run, however, 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.

This raises interesting questions of why we see this behavior. One plausible explanation is that users who find a community early in its life are also more likely than average to be those who will be attracted to it, in the same way that early ratings for a movie in a recommender system are likely to be higher than later ratings because the people who are most attuned to the movie are likely to see it earlier [**?**]. 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 [**?**, **?**, **?**].

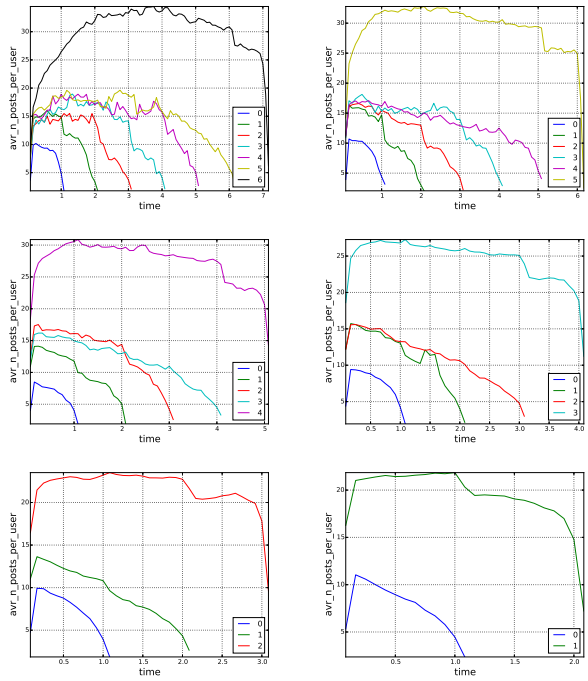

Figure 7: Each figure corresponds to one cohort, from 2008 to 2013, left to right, top to bottom. The users for each cohort are further divided in groups based on how long they survived: users that survived up to 1 year are labeled 0, from 1 to 2 years are labeled 1, and so on. We observe that, for all cohorts, users that have a lower posting average are the first to die. We also observe that the posting over time seem to decrease once we condition on the fact that users will survive a certain number of years. This suggests that the main reason for users to be posting less throughout the cohorts is because the mixture of users is different: there are more users that post less joining the network and the ones that post more are there to begin with, and they are the ones more likely to survive.

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.

4.4 Effort per Comment

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. The Figure N shows the evolution of the monthly average comment length in reddit.

Based on the downwards tendency of the comment length, one could possibly imagine that the user commitment with the network is lowering over time. This, however, might not be the best way to interpret this information. Figure

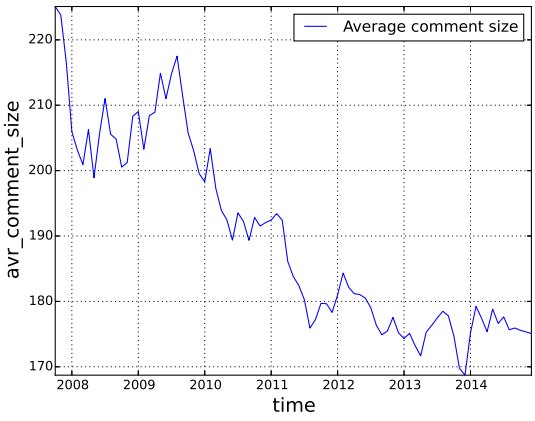

Figure 8: Average comment size (number of characters) over time for the reddit network. We observe that there is a decreasing trend for the average comment size. This means that users, on average, are making smaller comments in reddit as time passes. This, however, hides important aspects of user behavior over time and does not mean that users, as they survive in the network, write smaller comments.

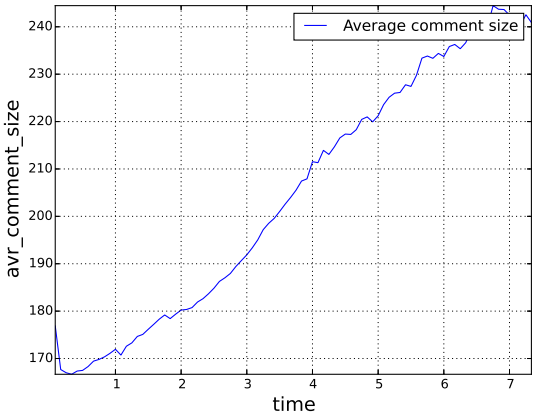

Figure 9: Average comment size from the users time referential. This shows that comment length made by users increase for users that lived longer in the network, save for a small decrease in the initial time. This, however, is also not enough to say that we should expect the average size of the comments in the network to increase since the users that survive write longer comments.

N shows the comment length per cohort based on the user referential time. This figure shows that, unlike the average overall network comment length, surviving users increase the size of their contributions to the community over time. This is true for all users cohorts. 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 than older users. Together with the fact

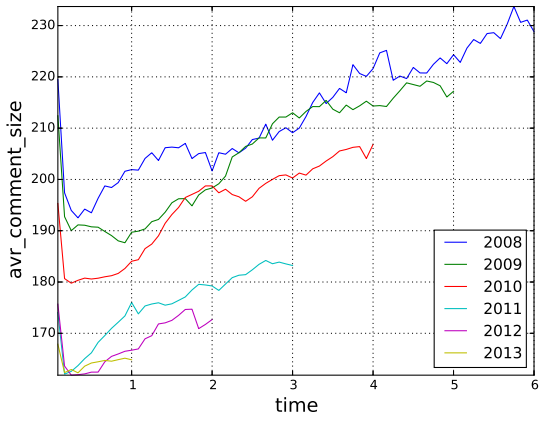

Figure 10: Average length of comments from the user time referential segmented by user creation cohorts. This figure start to explain why we see different trends in the overall network comment length and user referential overall comment length. It shows that, as users come from latter cohorts, they start from a lower commenting length average compared the the earlier cohorts. This, together with the fact that reddit is growing exponentially in terms of users means that *we have an influx of users that make smaller comments than the previous generations*, although even for them, as they survive, they make longer comments. Just as for the users' posting average, we can not distinguish based on this graphic alone whether users that make smaller comments leave the network earlier or they indeed write longer comments as they survive. Figure 11 sheds some light on this question.

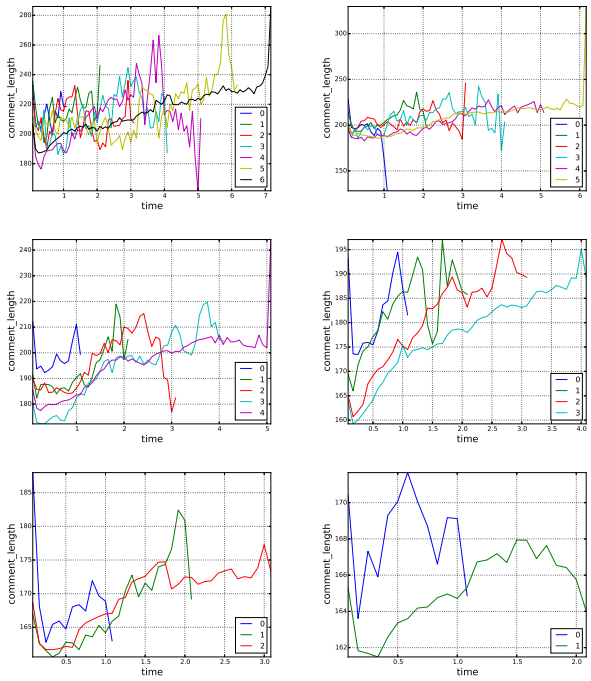

Figure 11: Just as in Figure 7, each figure corresponds to one cohort. These figures show the average comment length for users segmented by the number of years that he/she survived in the network, given a cohort. Here, we make two important observations: first, *comment length do increase inside of each 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.

that recent reddit has experienced exponential-like growth, the heavier weight when evaluating the averages for Figure N 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 writing less.

Some possible explanations for this difference in the starting points could be that older users are, again, sampled from a different demographics that is more committed and willing to spend more effort into developing their virtual identity. Also, it could be that it is a natural evolution of the community, as older users have taken most of the main space of interests when it comes to creating new subreddits and starting these communities, new users have it all already made and sometimes might feel intimidated or not motivated to create new topics or communities that already exist or that are less likely to compete with the existing ones. In a way, these new users could behave more as lurkers, while the older users are the ones that laid the foundation of reddit.

Yet another hypothesis that we might consider is that users are lowering their activity due to an "initial value problem". We can imagine that users, as they join the network, they tend to produce content according to the norms of what they see. If we look at the cohort posting size over time superimposed with the average size for the whole network, we can see that the starting point of each cohort seems to agree to a reasonable extent to the average over the total network. This way, users would be simply reproducing things as they see in their early months, but as we have seen in Figure N, users start their life posting longer content, but there is a strong decrease in size for the early months before the size increases for the surviving users.

4.5 Activity Nature

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 N shows the evolution of the 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

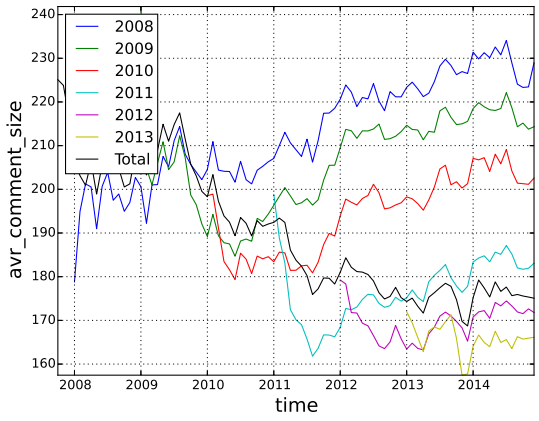

Figure 12: Average comment length over time for cohorts on user creation time superimposed over the trend for the total of users. Here we observe how the early existence for each cohort has a reasonable agreement with the overall trend. This might indicate how the new users might be just adopting the community norms regarding comment length.

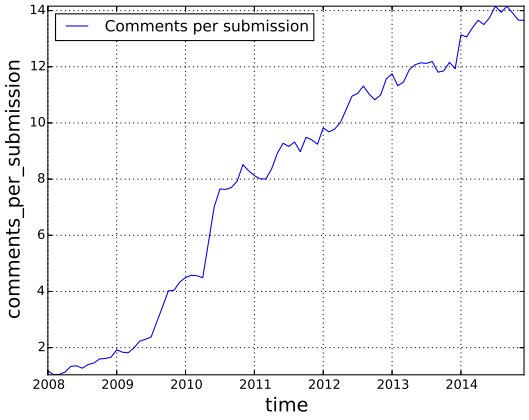

Figure 13: Comments per submission ratio for reddit over time. We observe an overall increasing trend of comments for each submission. Notice that these should not be interpreted as how many comments each submission gets, but as how many comments users author for each submission — submissions might get comments a long time after they are posted, by users from different years. This distinction becomes more important when we change the time referential and separate users by cohorts.

for each of his/her submissions.

We have found that segmenting users and subreddits by cohorts on the years that of the first comment highlights significant differences of behavior and help us to understand how reddit changed over these years.

Table 1: Number of distinct users that authored comments and submissions segmented by the year of the first post of the user. The Total numbers are based on posting

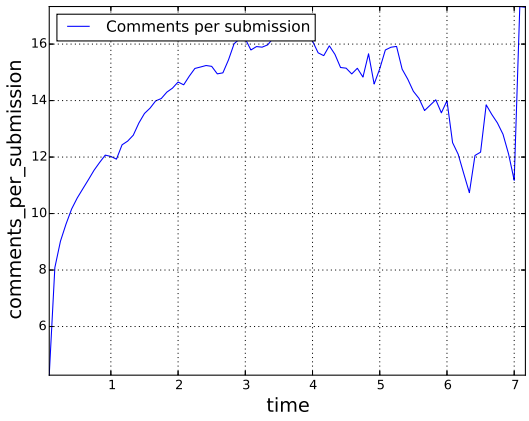

Figure 14: Comments per submissions ratio from the user referential. This should be interpreted as how many comments a user makes for each submission in their x-th year of existence. We observe an interesting overall trend that peaks between 3 and 4 years of existence. This, however, does not mean that users will necessarily decrease their behavior as they live longer, but that given that a user has survived for x years, what is his/her comment per submission ratio likely to be.

data from 2007 until 2014, corresponding to our full dataset. The Oct 1st, 2014 onwards numbers are based on the last 3 months of data we have, and we consider this as the current, active reddit.

Table 2: Number of distinct subreddits segmented by the year of the first post of the user. The Total numbers are based on posting data from 2007 until 2014, corresponding to our full dataset. The Oct 1st, 2014 onwards numbers are based on the last 3 months of data we have, and we consider this as the current, active reddit.

Table indicates that reddit grew significantly from 2007 until 2012, practically doubling the number of new users per year for each of these years, with similarly significant growth in subreddits. Although the most expansive growth happened in the first years, more than half of the registered users are from the last 2 years, and their behavior is significantly different than previous users, impacting in the overall behavior of the community. For instance, users from the 2014 cohort have a higher tendency to make submissions instead of comments, in contrast with all the previous cohorts.

Looking at the user time referential, the evolution of the number of comments per submission shows a decreasing trend for the older cohorts. One explanation for this is that, as the community grew, more content from an absolute point of view was present in the social network, and therefore users had more reason to make contributions commenting instead of submitting new content that was likely to already exist.

4.6 Users' 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

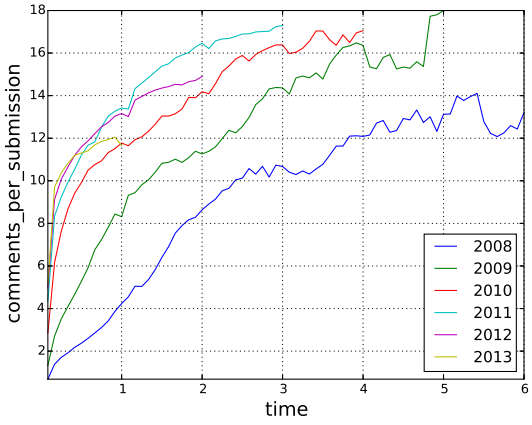

Figure 15: Comments per submissions cohorted by user creation year. Here we observe that, unlike the total aggregated graphic, all users are are increasing the number of comments per submissions, but latter cohorts show a much higher level of comments per submissions than earlier cohorts. This brings the initial part of the aggregated user-referential curve up, while the end of the curve consists only of users from the latter cohorts that preset an lower ratio. It is also important to notice that, as these curves move to the right, less comments and submissions exist in the bins, for there are less users that survived for such long periods. This results in some spiky behavior in the rightmost end of some curves due to the reduced amount of data. Just as with the previous user-referential cohort curves, we can not distinguish solely based on this graphic if users are increasing their commenting behavior or if the users that do not comment die earlier. Figure 16 help us to answer this question.

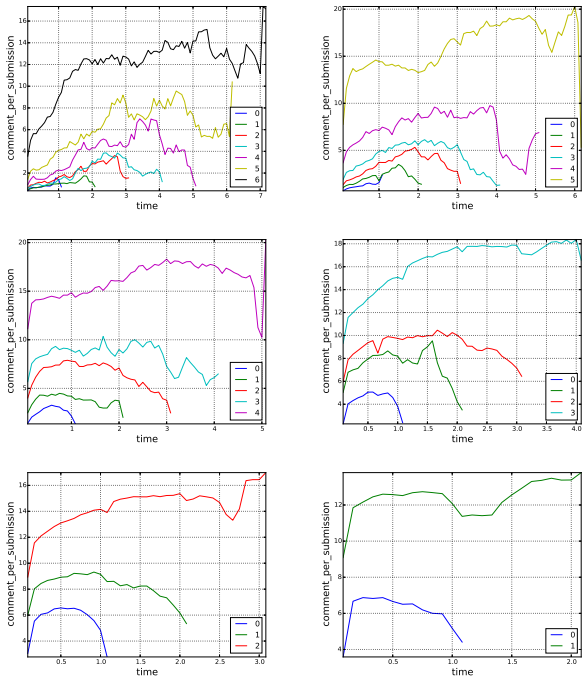

Figure 16: Just as in Figures 7, each figure corresponds to one cohort. Each cohort was segmented based on users according to how many years they survived. Each figure show the comment per submission ratio for these groups of users per cohort. We can observe a clear pattern of users that die earlier having a lower comment per submission ratio. Also, these curves do not show clear signs of increase over time, which suggests that the main reason for the cohorted user-referential ratio in Figure 15 to increase is due to the fact that the low performers drop out earlier.

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 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 posts while older users write less, longer posts? 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 17, 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 18 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,

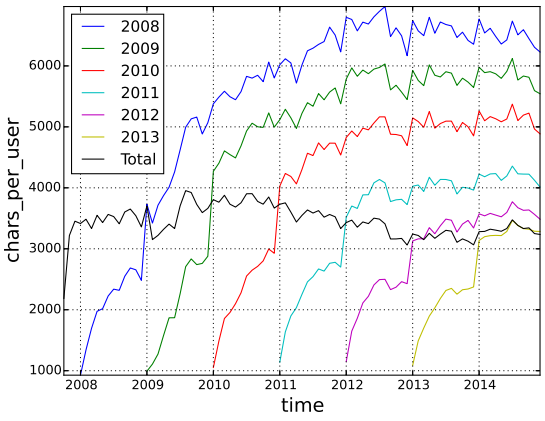

Figure 17: Average number of written characters per user over time for cohorts on user creation time superimposed over the trend for the total of users. The overall average number of characters a user writes per month in reddit for the total of users had it maximum at around 2010. After that, we observe a slightly decreasing pattern for this proxy for user effort on average in the network. We also observe in the cohorted curves that, at any point in time, the average number of characters that older cohort users write per month is always higher than the earlier cohorts. Also, as with other per user averages, these curves are significantly different in the cohort year, since users are being created in the period.

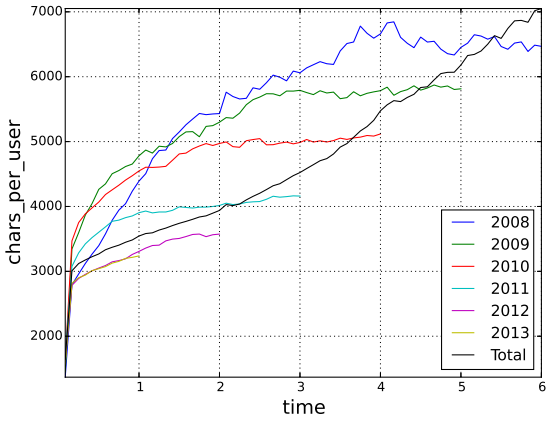

Figure 18: Average number of written characters per user from the user time referential segmented by user creation cohorts superimposed with the average number of written characters per user from the users time referential for the total of users in the network. The overall line shows that, as users survive in the network, the number of characters they are likely to write every month increases. We also observe for the cohorted curves that the surviving users level at different values, with a clear general pattern for users in older cohorts to level at higher values than younger ones, given the same time lived in the network. The partial exception is 2008, that presents a lower evolution in the first one and a half year. Since, as in the previous cases, we can not know if this is because users took a longer time to write more of the low-effort users stayed around for a longer time, we have to further segment these cohorts by the time each user existed in the network. It is also important to notice how the shapes of the cohorted curves are different from the total line. This happens because the cohorts have different weights in the total curve because of larger number of users in them.

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 19 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.

4.7 Users' Survival

The simplest definition of an active user in reddit is to set a threshold date and define that every user that posted after that date is an active user and users that do not show any kind of behavior are "dead". This, however, is a limited interpretation of how users decide to stay or leave the network, specially if we want to analyse how this behavior changed over time. Also, since our users might always come back to the network at a later time, they might be "reborn", that means we have right censored data.

To account for these, we look at a one year window of time for each user. This way, we avoid the right censored data and the possibility that a user might have come back to the network at a later time. Given this, we segment users by their cohort and define that users active in the last 3 months of this one year window are active users. Based on this data manipulation, we present the Kaplan-Meier (cite) survival curve in Figure N.

As previously mentioned, reddit shows a significant number of "single time users" that only post once in their existence. This can be seen in the initial drop in the first day. An interesting thing to see is that, although different cohorts level in different survival values, the "user decay" is similar throughout all of them. Not only that, but there is a general trend for older cohorts to die faster than younger ones. One possible explanation for that is that early reddit still lacked in content, with few subreddits to submit and few submissions to comment. This could lead to a higher number of users that did not stayed around after their initial impressions.

5. IMPORTANCE OF COHORTS: PUZZLING OUTCOMES

From the prior analysis, cohorts emerge as an important factor when analyzing activity on a community like Reddit. Because of this heterogeneity, any analysis that speaks of

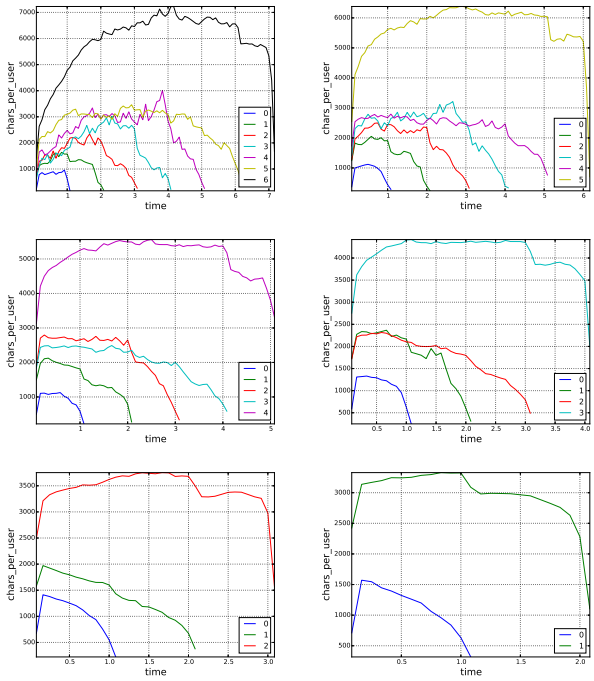

Figure 19: Just as in Figure 7, each figure corresponds to one cohort. These figures show the average number of written characters per user for users segmented by the number of years that he/she survived in the network, given a cohort. Here we observe that users average number of written characters per month tend to level at higher values the longer one individual survive in the network. For the 2008 cohort from Figure 18, this shows evidence that the initial lower behavior of the 2008 curve is because the low-effort users did not die as fast as in the latter cohorts. Also, it shows that the main reason for the average user effort increases as users survive is due to the fact that the lower-effort users die faster.

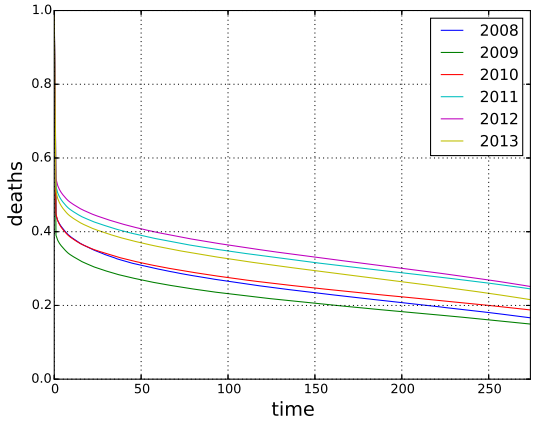

Figure 20: Kaplan-Meier estimator for one year of posting behavior for each user. Users for which the last posting day was in the first nine months of the one year window are considered "dead". This graph shows the percentage of surviving users per number of days since it first posted segmented by the cohort year the user joined the network.

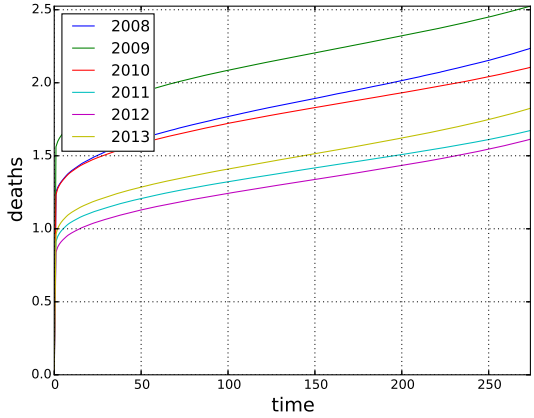

Figure 21: Nelson-Aalen empirical hazard estimation for the users survival. This curves show the pointwise probability of a user to die in time.

users' activity should account for cohorts and perhaps other kinds of heteregeneity in a community's participants. In this section, we show that accounting for cohorts is not only a desirable property, but a vital one: not doing so can lead to analysis with absolutely wrong conclusions.

Let us consider Figure x, which shows the 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 incentivize or promote longer comments on Reddit.

However, in Figure 7, we saw that average comment size increases over time for different cohorts. While later cohorts start at smaller comment sizes, 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 x provides some clues to what might be going on. For illustration, we consider the change in average comment length from the year 2011 to 2012. Overall, comment length is increasing. If all users had similar average comment lengths, then we would also see that average length across cohorts is decreasing with time. However, people in later cohorts tend to write less per comment. 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.

Year	Median
2007	114
2008	103
2009	103
2010	96
2011	91
2012	89
2013	87
2014	88

Table 3: Evolution of the median throughout the years for the whole reddit dataset.

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

Table 4: 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.

With the decision to condition on cohorts, one may have gathered an entirely wrong conclusion. People are not starting to write less, rather those who tend to write less are joining the added to the community. 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. COHORTS OF SUBREDDITS

6.1 Subreddits Activity

One way to look at reddit is as a multi-community social network. Each subreddit can be considered as a semi-independent community, and as such, we can study the evolution of these communities based on time, cohorts and survivability. A number of other online communities have similar properties, with tighter (Wikiprojects, enterprise social network discussion groups) or looser (Wikia, StackOverflow) interdependence between the individual sub-communities.

One of the initial question we can ask is how is the number of posts in these surviving communities evolving? One thing to be aware here though is that this variable is likely to be sensitive to the ration of active users per active subreddits. If we imagine that users don't change their posting patterns rapidly, and increased number of users per subreddits implies in an increased number of posts per subreddit. The overall number of users per subreddits, however, seems fairly stable throughout the years.

Figure N shows the evolution of number of posts per active subreddits in time for subreddits created between 2008 and 2013. We can observe that the average number of posts per subreddit increases over time. Since the ratio of active users per subreddit remains relatively stable, one could imagine that subreddits are receiving more posts throughout the time.

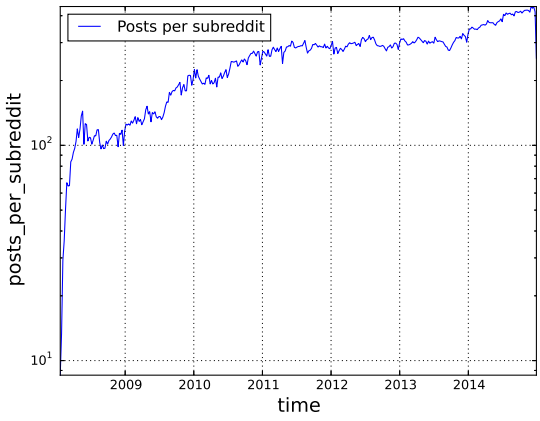

Figure 22: Caption

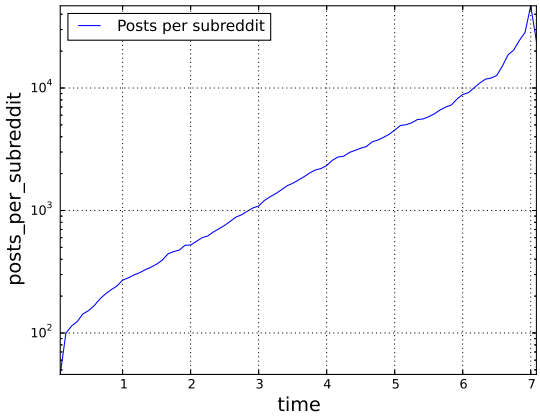

Figure 23: Caption

To better understand how correct is this conclusion, we cohort the subreddits in time. We can observe that the majority of posts in reddit are made in 2008 subreddits, and the posting averages for this cohort dominates the total posting average for the whole social network. Also, we notice that for most points in time, the number of posts per subreddits increases as we move to older cohorts. This, however, is not sufficient for us to conclude that these communities are evolving in different ways, since subreddits from older cohorts had more time to consolidate their reach and popularity and most of the "likely to die" subreddits that bring the average number of posts down in newer cohorts are still alive.

To properly compare these communities starting from the same baseline, we evaluate every posting time according to the subreddit creation time (fists post ever made in the subreddit). The x-axis then becomes the time the subreddit has lived, grouped by cohort. This approach reveals a general trend of subreddits from newer cohorts stabilizing in a lower posting average than older cohorts. This, however, does not hold true for the 2009 and 2010 cohorts, although they stabilize in very similar levels, and for the 2013 cohort,

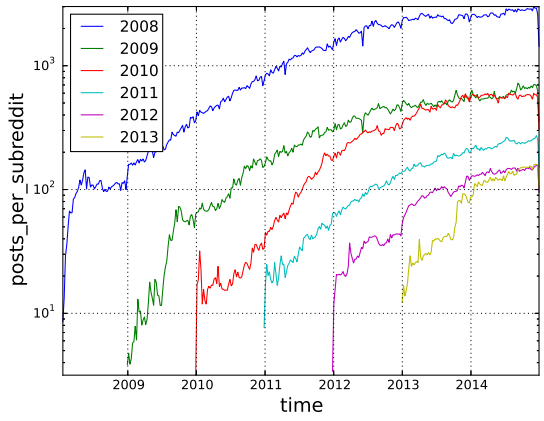

Figure 24: Caption

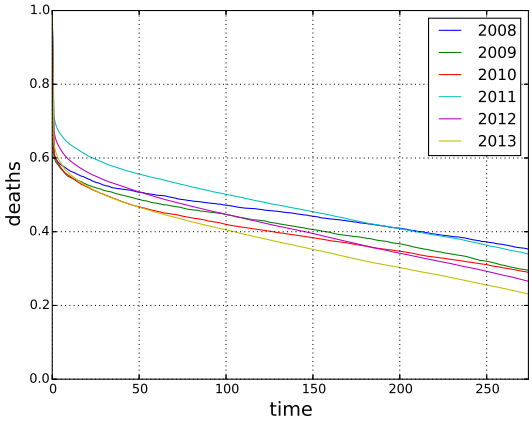

Figure 26: Caption

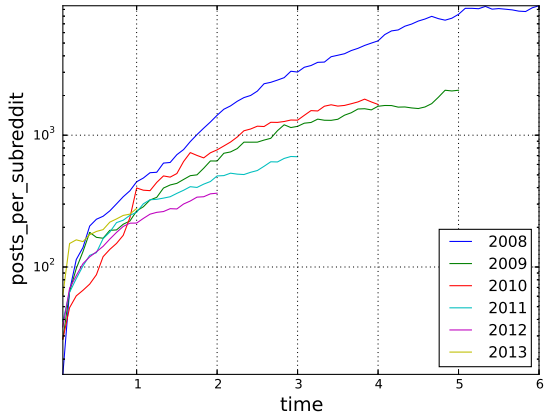

Figure 25: Caption

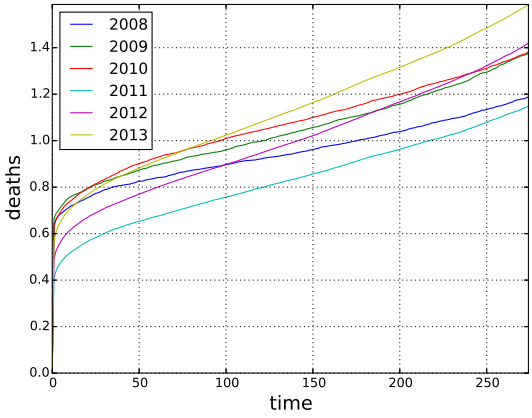

Figure 27: Caption

for which we have only one year of data in the overlap for all subreddits.

Assuming that subreddits that survive have, on average, a higher number of posts than the ones that do not survive, part of the higher levels of posting for the older cohorts could also be explained by a faster "death rate" of the low posting subreddits. Therefore, the faster the number of posts per subreddit grows, the faster the non-fit subreddits are being eliminated.

6.2 Subreddits' Survival

Similarly to what we did for users, we look in a one year time window for the last post that was created for each subreddit and define as subreddits that died as the ones that the last post happened in the first nine months of the one year window. The Kaplan-Meier curve is shown in Figure N.

In this survival curve, we also observe that there is a significant number of subreddits that survive only through the first day, just as seen with the users, although the proportion in this case is not as high as the users . Also, unlike the users, there are significantly differences in the "decay of subreddits".

7. 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. 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.

An initial hypothesis would be that users would be interested in the communities that were being created at the same time they joined the network. To test for that, Figures N and N show the number of submissions and comments per user, respectively, based on the user and subreddit cohorts.

It is possible to see that users' behavior, independently of the cohort, are biased to subreddits created on 2008. 2008 users' submissions are particularly more biased to 2008 subreddits. 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.

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

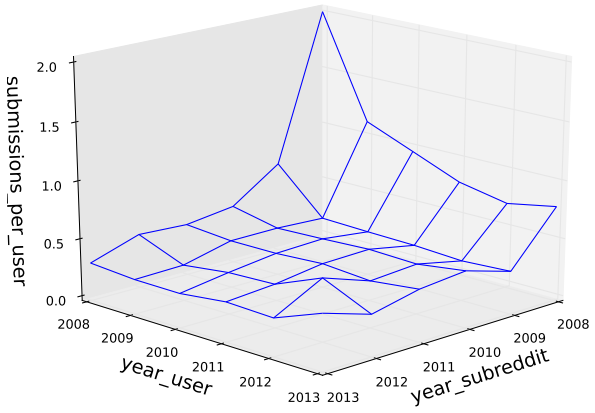

Figure 28: Caption

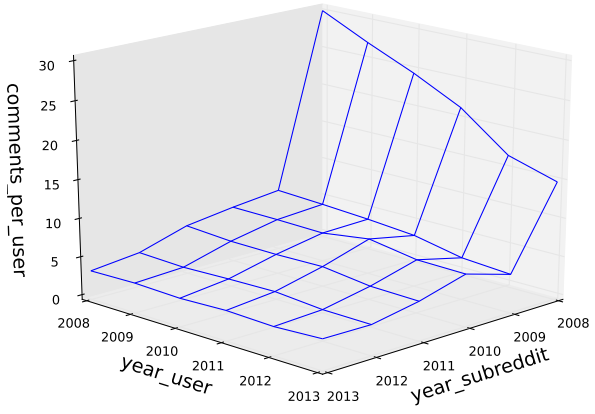

Figure 29: Caption

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.

8. DISCUSSION

9. CONCLUSIONS

This work addresses some aspects of how to analyse 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.

10. 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? Regarding 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.

11. ACKNOWLEDGMENTS

Acknowledgments.

12. REFERENCES

[1] A B8DFEHGPIRQTSTU @ QWVYX ‘ B a bPEcSWVed f d @ g7hRi6QTf QWB VeSTp pFQTSTU @ QWVYX ‘ B ÂĎÂd’ XrEcSWVed fYd Âę g7hW ÂĚÂĘ STd $ X ÂĞ p ÂÍ VeEcfYhYdF ÂĹÂÍ pFQWV6bPQTQWU bPEcSWVed f d @ g7hRi6QTf QWB VeSTp XrQTQWUF ÂĎÂd’ XrEcSWVed f d @ g7h8 ÂĚÂĽ STd $ X ÂŘ V ÂŚ QT ÂŠ d $ Ve .

[2] T. Correa, A. W. Hinsley, and H. G. de Zúñiga. Who interacts on the Web?: The intersection of usersâĂŹ personality and social media use. *Computers in Human Behavior*, 26(2):247–253, 2010.

[3] C. Danescu-niculescu mizil, R. West, D. Jurafsky, and C. Potts. No Country for Old Members : User Lifecycle and Linguistic Change in Online Communities. 2013.

[4] F. Kooti, K. P. Gummadi, and W. a. Mason. The Emergence of Conventions in Online Social Networks. *Artificial Intelligence*, pages 194–201, 2010.

[5] K. Lewis, M. Gonzalez, and J. Kaufman. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109(1):68–72, 2012.

[6] K. Lewis, J. Kaufman, and N. Christakis. The Taste for Privacy: An Analysis of College Student Privacy Settings in an Online Social Network. *Journal of Computer-Mediated Communication*, 14(1):79–100, 2008.

[7] H. J. Miller, S. Chang, and L. G. Terveen. "I LOVE THIS SITE!" vs. "It's a little girly": Perceptions of and Initial User Experience with Pinterest. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, pages 1728–1740, 2015.

[8] K. Panciera, A. Halfaker, and L. Terveen. Wikipedians are born, not made: a study of power editors on Wikipedia. *Human Factors*, pages 51–60, 2009.

[9] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating , Destroying , and Restoring Value in Wikipedia. 2007.

[10] C. Tan and L. Lee. All Who Wander : On the Prevalence and Characteristics of Multi-community Engagement. 2015.

[11] H. Zhu, R. Kraut, and A. Kittur. The Impact of Membership Overlap on the Survival of Online Communities. 2014.