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 its start in 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. 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 the cohort of users who join much later.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

Keywords

ACM proceedings, LaTeX, text tagging

1. INTRODUCTION

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 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 [?]. But the community evolving in itself changes the idea users outside have about it [?]. This internal evolution together with the novelty

WWW '16 Montereal, Canada

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?

Kooti et. al. [?] 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.

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?

2. PREVIOUS WORK

- The Taste for Privacy: An Analysis of College Student Privacy Settings in an Online Social Network [?]: 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 [?]: 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 [?]: 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 [?]: Initial experience matters!

- No Country for Old Members : User Lifecycle and Linguistic Change in Online Communities [?]: 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 [?]: Survival does depend on user initial activities.

- Wikipedians are Born, Not Made [?]: Users do have predispositions. Does that mean they do not change and we are simply sampling differently?

- Creating , Destroying , and Restoring Value in Wikipedia [?]: Not clear where it fits.

- The Impact of Membership Overlap on the Survival of Online Communities [?]: 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 an overview of the dataset and the community as a whole.

3.1 Raw size over time

Reddit has been growing in number of users and subreddits since its conception. The cumulative number of users and subreddits suggests that this growth is happening in an exponential fashion.

We have identified that about 16.2 million distinct registered users made comments in reddit since its conception until the end of 2014, while around 327 thousand subreddits received comments.

Since not all users post every month, the cumulative value might not be representative of how many users actually use reddit. If we consider only the users that authored a comment and the subreddits that had comments written at, we have the following graph.

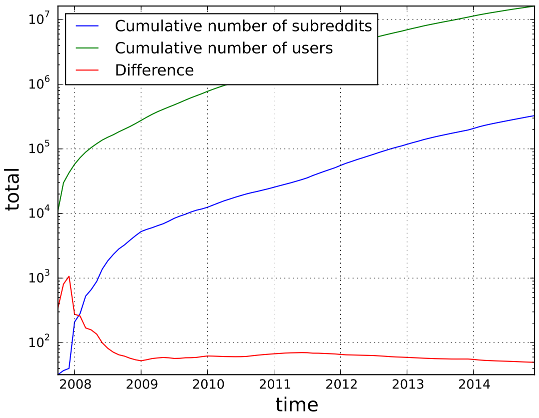

Figure 1: Caption

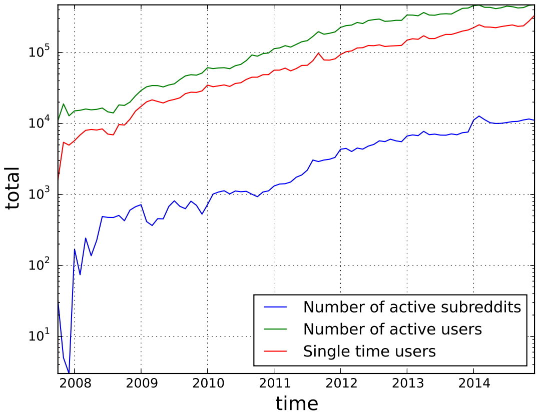

Figure 2: Caption

In this graph, we can see that reddit had around 470 thousand users that made comments in the last month of 2014, while around 11400 subreddits received comments in this same month. While the cumulative growth of reddit is impressive, the actual active community is significantly smaller than the total number of existing registered users and subreddits, as expected.

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.

3.2 Identifying cohorts

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

4.1 Users' Activity

One common metric of interest of a social network is the average number of posts that active users make in a determined time frame. For reddit, in particular, the overall evolution of monthly average posts per user is as follows.

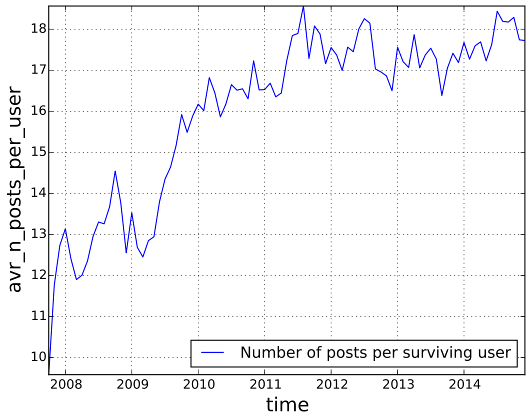

Figure 3: Caption

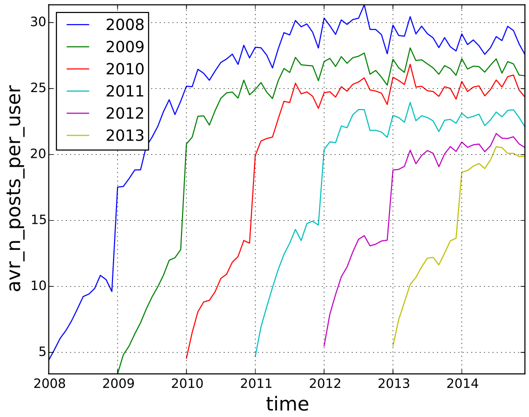

Figure 4: Caption

This is the first and most naive approach one can make to analyze the evolution of posting behavior for the users. This figure, however, averages over all active users in each evaluated time step. This hides cohort effects due to the user behavior evolution over the years of existence of the network. A first attempt to improve this would be to evaluate this posting average according to the users cohorts, as shown in Figure N.

Here we notice already a significant cohort effect: users from latter cohorts have a significantly lower posting average than users from earlier cohorts.

There are, however, two possible pitfalls in this analysis. The first is the fact that we are always considering the surviving users, that is, we count the users that presented some activity on the said month. Based on this, users from older cohorts could be posting more on average in 2014 because they are the ones that survived the longest and possibly are the ones that are most interested and get the most value out of the network, which would justify the increasing activity over.

To account for this, we change our time referential: each post time is considered as the time since the user creation (in our analysis, this referential is the first observed post for

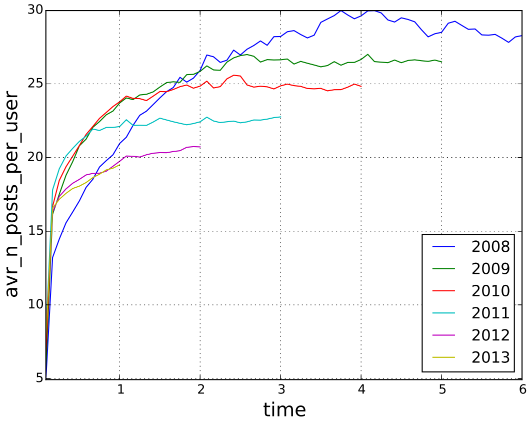

Figure 5: Caption

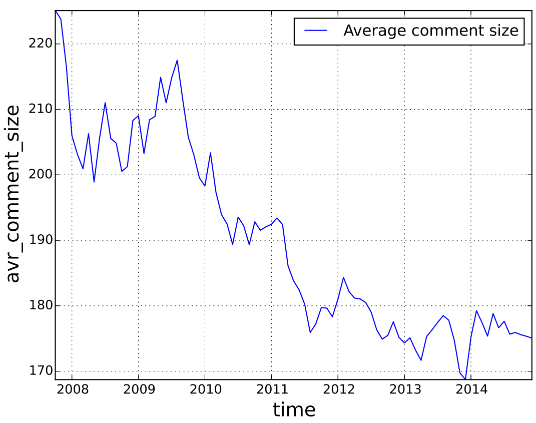

Figure 6: Caption

the user). This allows us to compare users based on how long they have existed in the network (but still treat them as cohorts), as seen in Figure N.

The second pitfall that one should be careful with is that users are being created during the cohort year, but not anymore afterwards. This accounts for the sharp increase in the average number of posts 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 anymore and the number of active users does not increase because of new users anymore. This fact, associated with the fact that young users tend to post less, drag the average down during the year, since young users are always coming in the network. This effect is more clear in Figure N, when we use the user-referential time frame, where it is clear that young users post less. This is particularly true for the first month, specially considering that there are many user accounts that are created and only show activity for a single day in reddit, which likely brings the average down.

It is unclear if in the initial year of existence of the users there are significant cohort effects in the average number of posts. However, we can observe that different cohorts apparently stabilize in different levels of behavior, with a general tendency for older cohorts leveling at lower values, eg. people from 2008 and 2009 level at higher posting frequencies than 2010. Some possible explanations for this would be that users that get most of the utility from the network are more likely to find it earlier, and this accounts for the higher activity for earlier cohorts. Also, earlier users could hold more status in the network, since they joined in an earlier stage of the network and therefore present a higher activity. Yet another explanation would be that earlier users demographics were different in terms of age and interests, for example, and these correlate to the fact that they present a higher activity.

4.2 Users' 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. 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 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 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.

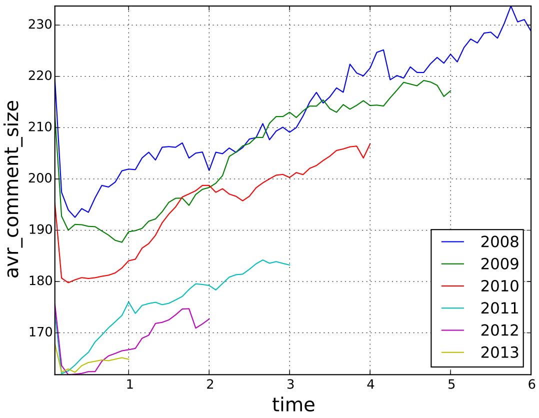

Figure 7: Caption

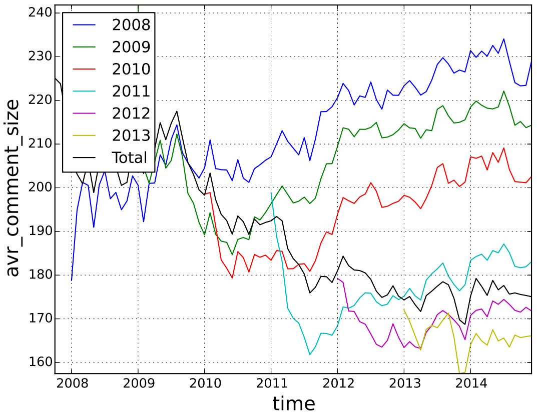

Figure 8: Caption

4.3 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 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 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.

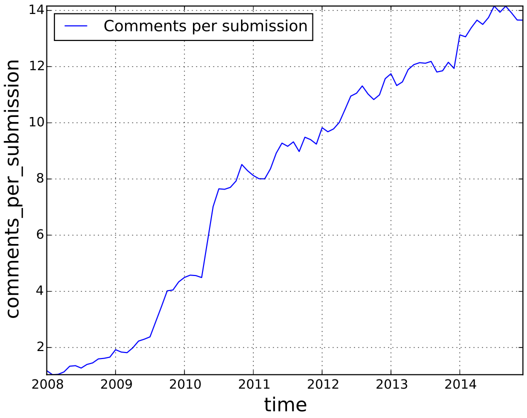

Figure 9: Caption

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.

4.4 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

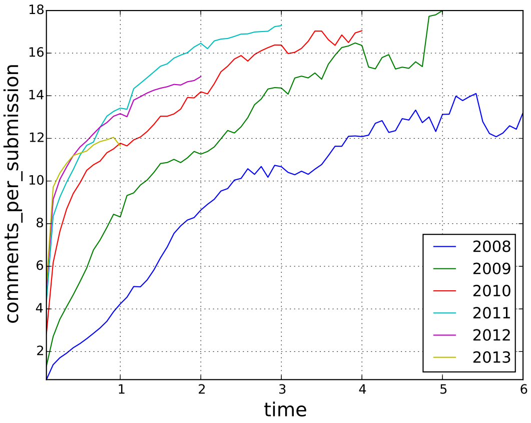

Figure 10: Caption

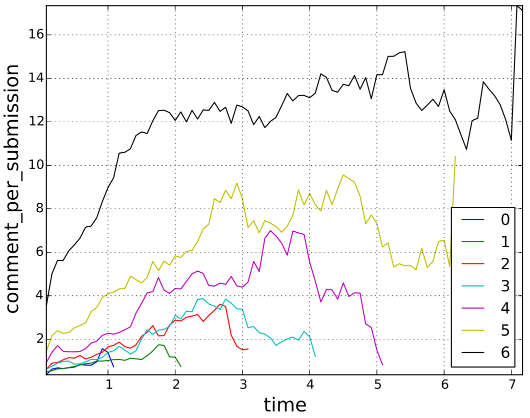

Figure 11: Caption

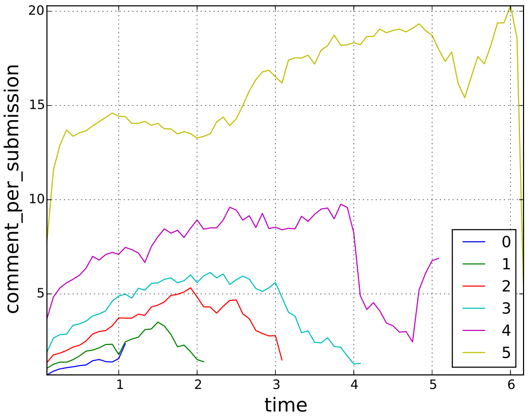

Figure 12: Caption

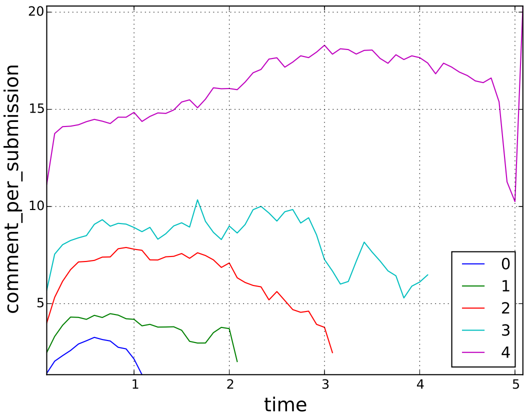

Figure 13: Caption

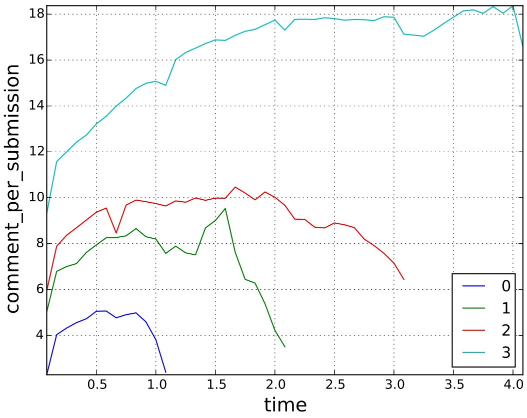

Figure 14: Caption

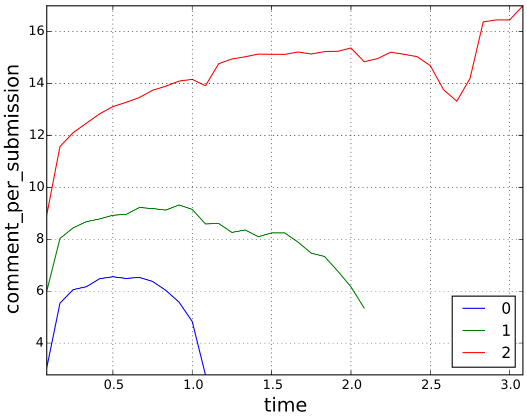

Figure 15: Caption

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.

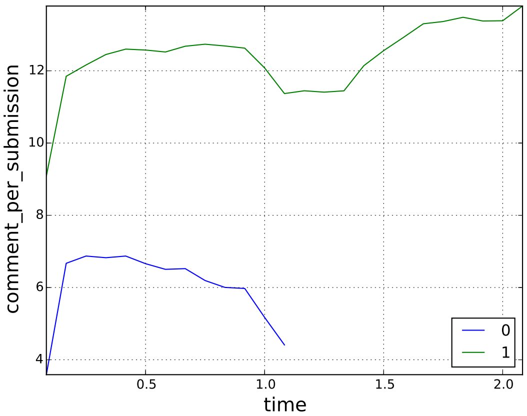

Figure 16: Caption

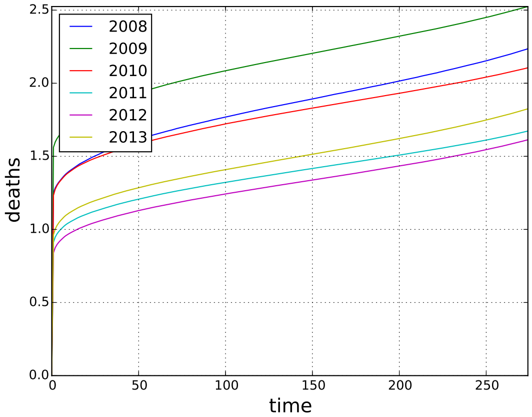

Figure 18: Caption

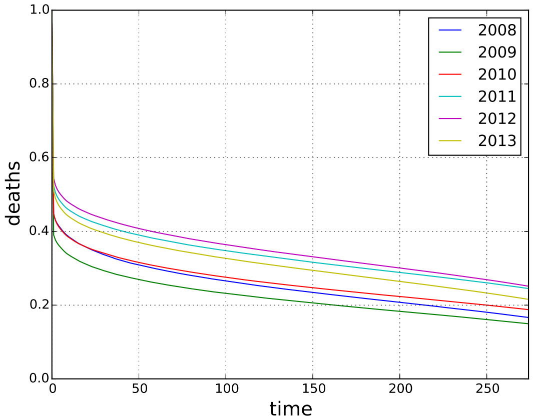

Figure 17: Kaplan-Meier estimator for one year of posting behavior for each user. Users for which the last posting day was in the first nine months of the one year window are considered "dead". This graph shows the percentage of surviving users per number of days since it first posted segmented by the cohort year the user joined the network.

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 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.

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

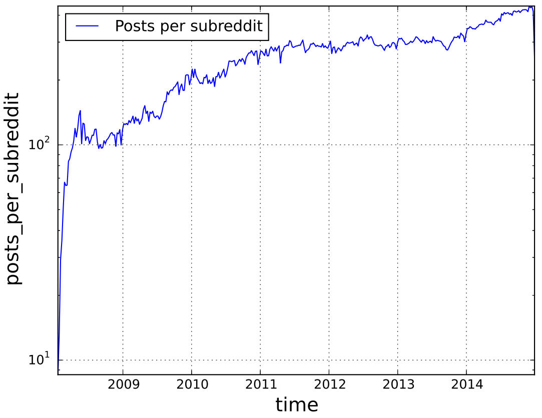

Figure 19: Caption

comments on average.

6. 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.

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.

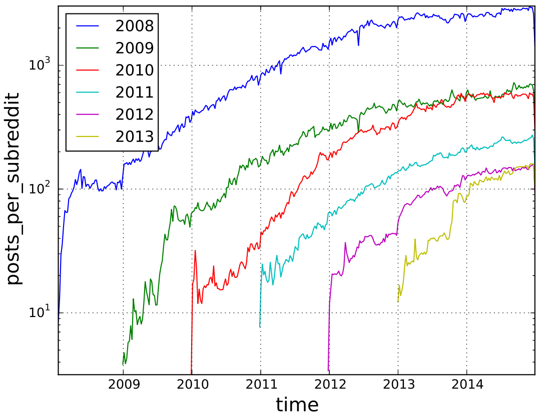

Figure 20: Caption

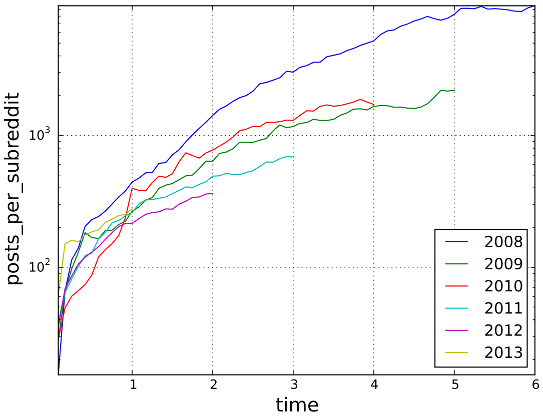

Figure 21: Caption

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, 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

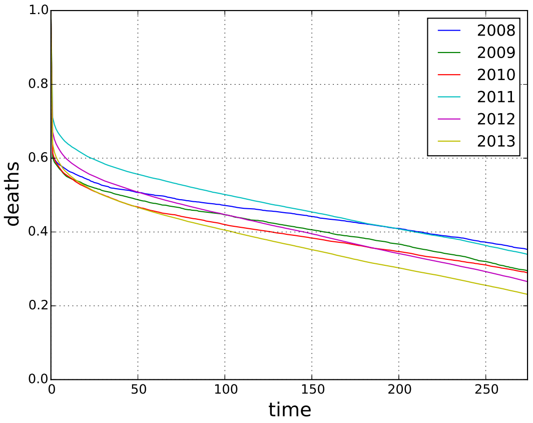

Figure 22: Caption

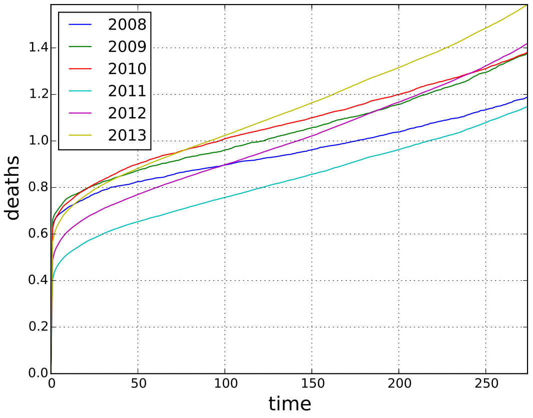

Figure 23: Caption

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. More than that, 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.

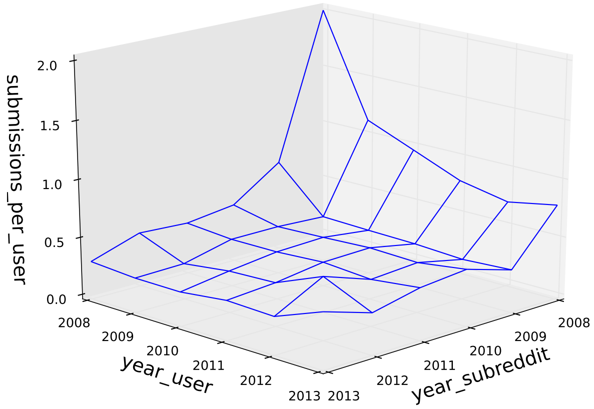

Figure 24: Caption

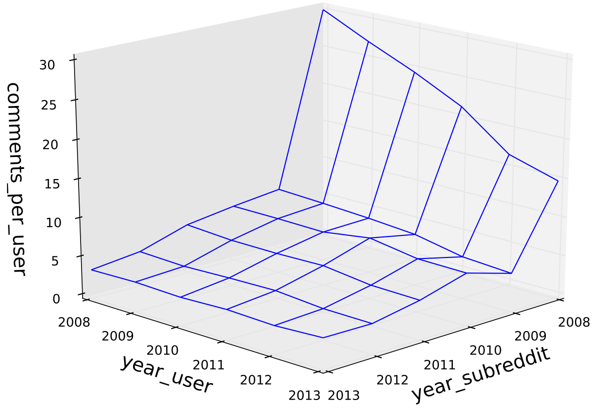

Figure 25: Caption

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.

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.