Evolution of effort and activity in online communities from a cohort perspective

Samuel Barbosa
University of São Paulo
São Paulo, Brazil
sam@ime.usp.br

Dan Cosley
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
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. Because both people and communities change over time, we argue that analyses of these communities that take time into account will lead to deeper and more accurate results. Using Reddit as an example, we study the evolution of users based on comment and submission data from 2007 to 2014. Even using one of the simplest temporal differences between users—yearly cohorts—we find wide differences in people's behavior, including comment activity, effort and survival. Further, not accounting for time can lead us to both misinterpret and miss important phenomena. 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, an example of Simpson's Paradox. Dividing cohorts into sub-cohorts based on the survival time in the community provides further insights; in particular, longer-lived users start at a higher activity level and make more and shorter comments than those who leave earlier. These findings both give more insight into user evolution in Reddit in particular, and raise a number of interesting questions around studying online behavior going forward.

Keywords

user behavior; cohorts; computational social science; Reddit; temporal

Categories and Subject Descriptors

H.0 [**Information Systems**]: GENERAL; J.4 [**SOCIAL AND BEHAVIORAL SCIENCES**]: Sociology

WWW '16 Montréal, Canada

1. INTRODUCTION

Understanding the evolution of users in a social network is essential for a variety of tasks: monitoring community health, predicting individual user trajectories, and supporting effective recommendations, among others. Many works aim at explaining these temporal aspects of evolution. Some adopt a point of view of the whole network and try to understand more general patterns of behavior [27, 57], while others adopt a more user-centric point of view and try to model [8, 41, 44, 54] or predict [10] individuals' behavior.

These approaches often combine all available data into aggregate analyses of the whole community over its entire history. This can be a natural response to limitations in the amount of available data: many datasets capture a small part of the community's history [2]; timestamps may not be available [44, 45]; snapshots may provide limited views of the community [9]; or the community itself might be small [33]. Aggregate time-based analyses are also a natural first way to address questions of community evolution.

However, we argue it is likely that many of these aggregated views are misleading. The conditions under which users join the community may vary greatly over time, and this might impact their behavior [37]. Among other things, popularity, purpose, features, interface, and algorithms can change: Wikipedia circa 2005 and circa 2015 are very different, as are Facebook of 2005 and 2015. Analysis—including some of our own past work—that fail to account for this change may miss important details of what is really going on.

We support this argument through an analysis of user effort in Reddit, one of the most popular and long-running online communities, based on a very large, recently released dataset of posting behavior. We address a number of questions commonly raised about users' effort in online communities: how active are users, how hard do they work, and what kinds of things do they do? In each case, we compare aggregate analyses of posting behavior to ones that treat users in Reddit as yearly cohorts, and views that focus on calendar time versus user-referential views that normalize behavior based on the creation date of a user. We also look at differences within yearly cohorts, seeking differences between shorter and longer-lived users.

We find that even simple accountings for time reveal additional insights about Reddit beyond what commonly performed aggregate analyses can provide. Users who join Reddit earlier post more and longer comments than those who join later, while users who survive longer start out both more active and more likely to comment than submit compared to users who leave Reddit early; none of these findings are obvious from commonly used analysis of user behavior.

Further, we find that aggregate analysis can be downright misleading. For instance, although average comment length decreases over time in an aggregate view, the comment length for surviving users increases over time in every cohort. Likewise, an aggregate analysis suggests that longer-lived users post more over time; this is not the case. Instead, users come into Reddit as active as they will ever be (somewhat akin to Panciera et al.'s finding that Wikipedians are "born, not made" [41]), and the rise in average activity for surviving users over time is driven fully by lower-activity users leaving early.

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

2. TIME MATTERS

2.1 Why accounting for time is important

Communities grow and, with time, die. For any community, its users play a role in its evolution, but they are also simultaneously affected by the evolution of the community. Untangling this interplay can help make sense of patterns of activity in a community.

One useful way to understand the evolution of a community and its users is through time, as it provides a linear account of the growth (or decay) of overall activity, types of content, social norms, and structure of communities. To account for time, users on online communities are differentiated based on their age, such as when modeling their preferences [36] or analyzing the evolution of their language [10]. These analyses uncover insights about the lifecycle of a user in a community: users' preferences and behavior change with their age in a community [42], and their early experiences and activity shape future outcomes predictably [37, 41, 53, 56].

However, much of past work on online communities ignores the time at which a user joins the community and analyzes all users together. This might be a mistake: communities may grow denser or sparser with time [31], develop new norms [27] and/or enact policies and rules guiding people's behavior [6]. These changes mean that people experience different versions of a community at different times, which can, in turn, affect their observed behavior. This interaction with the state of a community can confound conclusions about people's behavior, because the differences one observes may simply be due to changes in the community, rather than any significant change in the outcome variable of interest or the user population.

2.2 Cohorts are analytically useful

To prevent such confounding, a common unit of analysis to control for such biases is cohort analysis, widely used in fields such as sociology [14, 35], economics [3, 40], and medicine [11, 12]. A cohort is defined as a group of people who share a common characteristic, generally with respect to time. For example, people born in the same year, or those who joined a school at the same time, or got exposed to an intervention at similar times can be considered as cohorts. Such people in a cohort can be assumed to be exposed to the same state of the world and thus are more comparable to each other than people in other cohorts.

For example, sociological studies often use students who join a school in the same year to understand the effect of interventions [16, 24], and condition on the year in which people were born to understand people's behavior, such as variations in financial decisions-making [40] or opinions on issues [15, 22]. Similarly, medical studies interpret effects of drugs using cohorts of people with the same age group or lifelong exposure to correlated conditions [11, 12].

Recent work shows that cohorts' importance transfers to online communities as well. Just as people's behavior varies according to their biological age, their experience in an online community may vary with their age in the community and their year of joining. In Wikipedia, we find substantial differences in the activities of cohorts of users who joined earlier versus those who joined later [54]. Similarly, on review websites, users who join later tend to adopt different phrases than the older users who had joined earlier [10].

2.3 What might cause these differences?

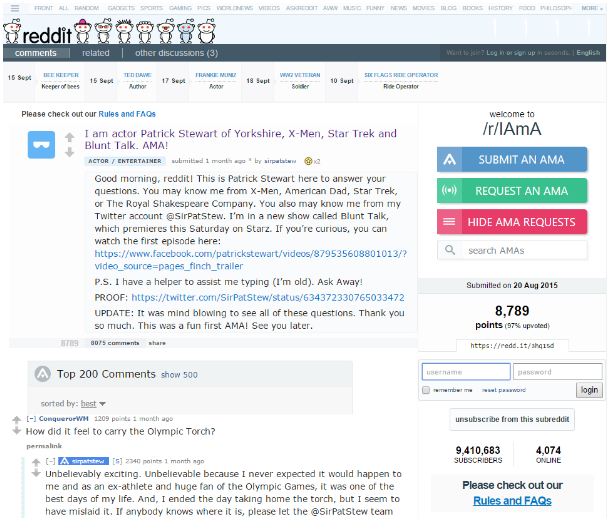
These differences in activity between cohorts may be due to a number of reasons. It could be due to selection effects: people who are enthusiastic about a community or its goals are more likely to self-select as early members of a community, while others may be more likely to join later [34].

The norms in community may change over time, which could explain why users in later cohorts may behave differently. In many cases, it is a bottom-up process. Kooti et al. [27] 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. Examples include adoption of 'RT', a retweeting norm by Twitter users and the subsequent introduction of the Retweet button on Twitter [27]; change in language use by new and old users on review websites [10]; and assumptions of clear roles and responsibilities on Wikipedia [26]. In other cases, it may be directed by the community managers. For instance, the makers of Digg unilaterally changed the nature of the community by introducing a new version of the website, leading to a sudden change in norms and behavior in the community [20, 29].

The growth of a community may also affect people's behavior. Successful communities often grow very rapidly, which can be both good and bad for users' experience. On one hand, growth would imply availability of a larger chunk of content to choose from. On the other, it might be harder to connect to others and get responses in a bigger community. A community may also need to adopt new rules and policies to manage growth and newcomers, as in the evolution of Wikipedia [5, 7]. In those cases, the experience of later cohorts of users may be vastly different from the initial ones who joined before formal rules were in place.

Finally, patterns of use may change because the overall population of Internet users is still changing. As more and different people become connected with the web, their influx may lead to observed change in activity patterns. This also affects technology use: people who did not grow up in a technological environment differ in their social media and search usage compared to younger generations [3, 8].

All of the above reasons suggest that users from different cohorts are likely to be different, which has also been


FRONT ALL RANDOM GADGETS SPORTS GAMING PICS WORLDNEWS VIDEOS ASKREDDIT AWW MUSIC FUNNY NEWS MOVIES BLOG BOOKS HISTORY FOOD PHILOSOPH MORE »
reddit
comments
related
other discussions (3)
Want to join? Log in or sign up in seconds. | English
15 Sept BEE KEEPER Keeper of bees
15 Sept TED DAWE Author
17 Sept FRANKIE MUNIZ Actor
18 Sept WW2 VETERAN Soldier
10 Sept SIX FLAGS RIDE OPERATOR Ride Operator
welcome to
/r/IAmA
SUBMIT AN AMA
REQUEST AN AMA
HIDE AMA REQUESTS
search AMAs
Submitted on 20 Aug 2015
8,789
points (97% upvoted)
https://redd.it/3hq15d
username
password
remember me
reset password
login
unsubscribe from this subreddit
9,410,683
SUBSCRIBERS
4,074
ONLINE
Please check out our Rules and FAQs

Please check out our Rules and FAQs

I am actor Patrick Stewart of Yorkshire, X-Men, Star Trek and Blunt Talk. AMA!

ACTOR / ENTERTAINER submitted 1 month ago * by sirpatstew x2

Good morning, reddit! This is Patrick Stewart here to answer your questions. You may know me from X-Men, American Dad, Star Trek, or The Royal Shakespeare Company. You also may know me from my Twitter account @SirPatStew. I'm in a new show called Blunt Talk, which premieres this Saturday on Starz. If you're curious, you can watch the first episode here: https://www.facebook.com/patrickstewart/videos/879535608801013/?video_source=pages_finch_trailer

P.S. I have a helper to assist me typing (I'm old). Ask Away!

PROOF: https://twitter.com/SirPatStew/status/634372330765033472

UPDATE: It was mind blowing to see all of these questions. Thank you so much. This was a fun first AMA! See you later.

8789 8075 comments share

Top 200 Comments show 500

sorted by: best

[–] ConquerorWM 1209 points 1 month ago

How did it feel to carry the Olympic Torch?

permalink

[–] sirpatstew [S] 2340 points 1 month ago

Unbelievably exciting. Unbelievable because I never expected it would happen to me and as an ex-athlete and huge fan of the Olympic Games, it was one of the best days of my life. And, I ended the day taking home the torch, but I seem to have mislaid it. If anybody knows where it is, please let the @SirPatStew team

Figure 1: Reddit interface when visualizing a submission. This is Patrick Stewart's "AmA" (ask me anything) in "IAmA" (I am a), a submission where he answers users' questions in the comments. We can see the most upvoted comment and Patrick's answer right below.

demonstrated in online and offline communities [8,10,43,48]. In this paper, our motivating research question was "How users' behavior evolve when accounting for cohorts?". To do so, we considered as behavior their activity levels, effort when commenting and kinds of contributions in the network. As we will see, users from different cohorts exhibit significant differences in behavior. More than that, we find that not accounting for underlying time-evolving processes of the network might lead to wrong conclusions about their behavior.

3. DATA: REDDIT AS A COMMUNITY

In this paper, we explore how cohort based analysis can give us deeper insights into three common questions about online communities: how active are users [19,21,32,50], how much do they contribute [17,18,50], and what kinds of work do they engage in [7,41,54]? We do this in the context of Reddit, a community that has been studied by many researchers [4, 13, 52, 53]. We begin with a brief overview of both Reddit and the dataset that we use in this paper, focusing on aspects that directly impact our analyses[1].

3.1 What is Reddit, briefly

Reddit is one of the largest sharing and discussion communities on the Web. According to Alexa, as of late 2015 Reddit is in the top 15 sites in the U.S. and the top 35 in the world in terms of monthly unique visitors. It consists of a large number of subreddits (853,000 as of June 21st, 2015[2]), each of which focuses on a particular purpose. Many subreddits are primarily about sharing web content from other sites: in "Pics", "News", "Funny", "Gaming', and many other communities, users ("Redditors") make "submissions" of links posted at other sites that they think are interesting. In other subreddits, Redditors primarily write text-based "self-posts": "AskReddit", "IAmA", "ShowerThoughts" are places where people can ask questions and share stories of their own lives. Generically, we will refer to submissions and text posts as "submissions".

Each post can be imagined as the root of a threaded comment tree; in addition to submitting, Redditors can make comments, and vote on both submissions and comments. Votes are used to sort comments within a submission and submissions 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 submissions. We can observe these elements in Figure 1. Redditors can also create subreddits and volunteer to moderate them.

We choose Reddit as our target community for a number of reasons. It has existed since 2005, meaning that there has been ample time for the community to evolve and for differences in user cohorts to appear. Second, it is one of the most popular online communities, allowing different types of contributions—comments and original submissions—across many different subreddits. Third, Reddit data are publicly available through an API.

3.2 The dataset

Redditor *Stuck_In_The_Matrix* used Reddit's API to compile a dataset of almost every publicly available comment[3] from October 2007 until May 2015. The dataset is composed of 1.65 billion comments, although due to API call failures, about 350,000 comments are unavailable. He also compiled a submissions dataset for the period of October 2007 until December 2014 (made available for us upon request) containing a total of 114 million submissions. These datasets contain the JSON data objects returned by Reddit's API for comments and submissions[4]; for our purposes, the main items of interest were the UTC creation date, the username, the subreddit, and for comments, the comment text.

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

3.3 Preprocessing the dataset

To analyze the data, we used Google BigQuery[6], a big data processing tool. Redditor *fhoffa* imported the comments into BigQuery and made them publicly available[7]. We uploaded the submission data ourselves using Google's SDK.

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

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

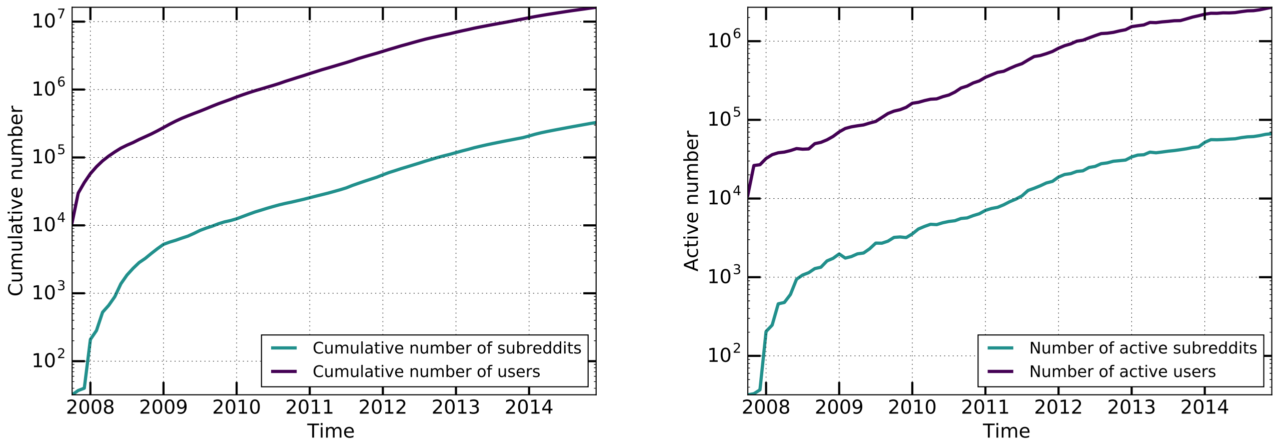
[3] Available in https://www.Reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_Reddit_comment.

[4] A full description of the JSON objects is available at https://github.com/Reddit/Reddit/wiki/JSON.

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

[6] https://cloud.google.com/bigquery/.

[7] See https://www.Reddit.com/r/bigquery/comments/3cej2b/17_billion_Reddit_comments_loaded_on_bigquery/.



(a)

(b)

Figure 2: Figure (a) shows the cumulative growth of Reddit for users and subreddits. Figure (b) shows the number of active users and subreddits in Reddit over time. An active user or subreddit is one that had at least one post (comment or submission) in the time bin we used—here, discretized by month.

For the analysis in the paper, we did light preprocessing to filter out posts by deleted users, posts with no creation time, and posts by authors with bot-like names[8].

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

3.4 An overview of the dataset

Here we present an overview of the dataset that shows Reddit's overall growth. Figure 2a presents the cumulative number of user accounts and subreddits created as of the last day of every month. After an initial extremely rapid expansion from 2008–2009, the number of users and subreddits have grown exponentially. As of the end of 2014, about 16.2 million distinct users and 327 thousand subreddits made/received at least one post based on our data.

However, as with many other online sites, most users [19,21,50] and communities [1] do not stay active. We define as an "**active user**" one that made at least one post in the month in question. Similarly, an "**active subreddit**" is one that received at least one post in the month. In December 2014, about 2.7 million users and 66 thousand subreddits were active, both an order of magnitude less than the cumulative numbers. Figure 2b shows the monthly number of active users and subreddits.

Our interest in this paper is not so much whether users survive as it is about the behavior of active users. Thus, in general our analysis will look only at active users and subreddits in each month; those that are temporarily or permanently gone from Reddit are not included.

3.5 Identifying cohorts

We define the "**user's creation time**" as the time of the first post made by that user. Throughout this paper, we will use the notion of user cohorts, which will consist of users created in the same calendar year.

In many cases, we will look at the evolution of these cohorts. Since users can be created at any time during their cohort year, and our dataset ends in 2014, we are likely to have a variation on the data available for each user of up to one year, even though they are in the same cohort. To deal with this, some of our cohorted analyses will consider only the overlapping time window for which we collect data for all users in a cohort. This means that we are normally not going to include the 2014 cohort in our analyses.

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

4. AVERAGE POSTS PER USER

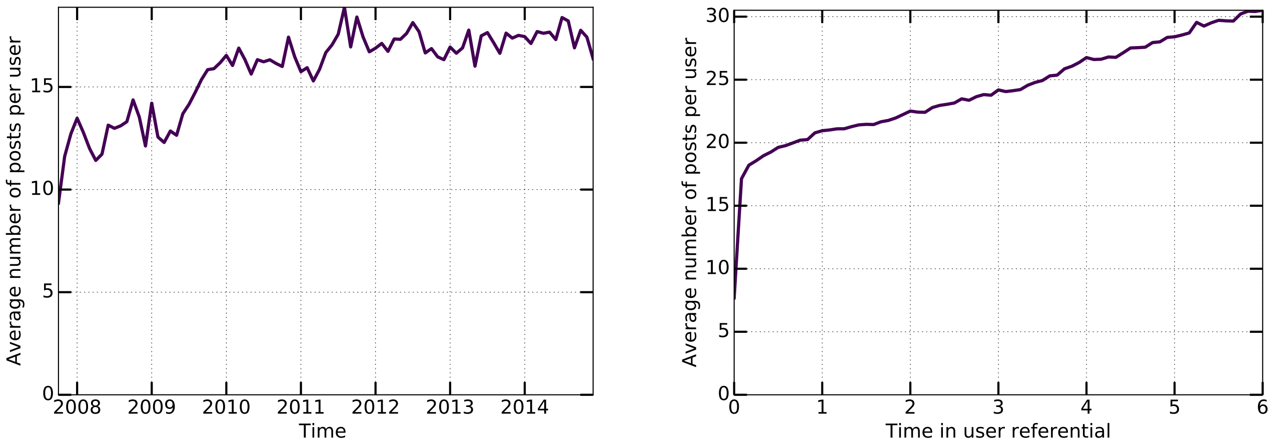
In this section, we will use a common metric of user activity in online communities, the number of posts per user over time. Approaches that consider the total number of posts per user in a particular dataset [17] and that analyzes the variation on the number of posts per user over the days [18] have been applied to online social networks.

As we will see, both visualizing behavior relative to a user's creation time and using cohorts provide additional insight into posting activity in Reddit compared to a straightforward aggregate analysis based on calendar time .

4.1 Calendar versus user-relative time

We start with a common analysis used in this kind of work: aggregating behavior in the community based on calendar time. Figure 3a shows the average number of posts per month by active users in that month. Taken at face

[8] Ending with "_bot" or "Bot"; or containing "transcriber" or "automoderator".



(a)

(b)

Figure 3: In Figure (a), monthly average posts per active user over clock time. In Figure (b), the monthly average posts per active users in the user-time referential, i.e., message creation time is measured relative to the user's first post. Each tick in the x-axis is one year. In both figures (and all later figures), we consider only active users during each month; users that are either temporarily or permanently away from Reddit are not included.

value, this suggests that over the first few years of Reddit, users became more active in posting, with per-user activity remaining more or less steady since mid-2011.

This average view hides several important aspects of users' activity dynamics. Previous work has looked into behavior relative to the user creation time. It has been shown that edge creation time in a social network relative to the user creation follows an exponential distribution [30]. User lifetime, however, does not follow a exponential distribution and some types of user content generation follow a stretched exponential distribution [18]. Throw-away accounts are one example of very short-lived users in Reddit [4], for example.

To address these characteristics, Figure 3b shows a different view that emphasizes the trajectory over a user's lifespan. We scale the x-axis not by clock time, as in Figure 3a, but by time since the user's first post: "1" on the x-axis refers to one year since the user's account first post, and so on. We call this the **time in the user referential**. One caution about interpreting graphs with time in the user referential is that the amount of data available rapidly decreases over time as users leave the community, meaning that values toward the right side of an individual data series are more subject to individual variation.

The evidence at this point supports the tempting hypothesis is that the longer a user survives, the more posts they make over time (**H1**). This hypothesis, however, is incorrect; we will present a more nuanced description of what is happening informed by cohort-based analyses.

4.2 New cohorts do not catch up

Figure 3b suggests that older users are more active than newer ones, raising the question (**RQ1**): will new users eventually follow in older users' footsteps? Analyzing users' behavior by cohort is a reasonable way to address it.

Figure 4a shows a first attempt at this analysis. We can already observe a significant cohort effect: users from later cohorts appear to level off at significantly lower posting averages than users from earlier ones. It suggests that newer users likely will never be as active as older ones on average. It also shows that surviving users are significantly more active than a naive average would suggest.

However, Figure 4a also has an awkward anomaly: a rapid rise in the average number of posts during each cohort's first calendar year, especially in December. Combining cohort segmentation with user-referential analysis, as in Figure 4b, helps smooth out this anomaly and aligns cohorts with each other. Doing this alignment makes clear that differences between earlier and later cohorts are apparent early on.

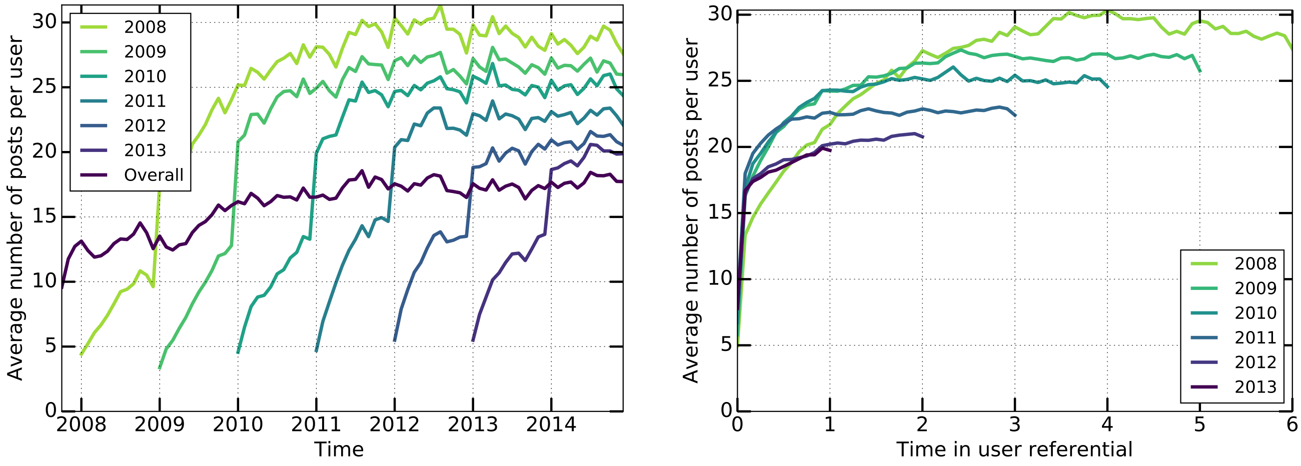
4.3 Does tenure predict activity, or vice versa?

These graphs still support our initial hypothesis **H1** and they do not explain the rapid increase in posting activity in the first few months. An alternative hypothesis, inspired by the "Wikipedians are Born, not Made" paper [41], is that individual users come in with different posting propensities, and the rise over time is not that individual users become more active but that low-activity users leave the system (**H2**). To examine this, we further segment each cohort by the number of years they were active in the system, as defined by the difference between their first and last post times.

Figure 5 shows this analysis for the 2010, 2011 and 2012 cohorts[9]. Across all cohorts and yearly survival sub-cohorts, users who leave earlier come in with a lower initial posting rate. Thus, the rise in average posts per active user is driven by the fact that users who have high posting averages throughout their lifespan are the ones who are more likely to survive. As the less active users leave the system, the average per active user increases. In other words, the correct interpretation of Figure 3b isn't that longer-lived users post more. It actually is that users who post more—right from the beginning—live longer.

Combining Figure 5's insight that the main reason why these curves increase is because the low posting users are dying sooner with the earlier observation that the stable

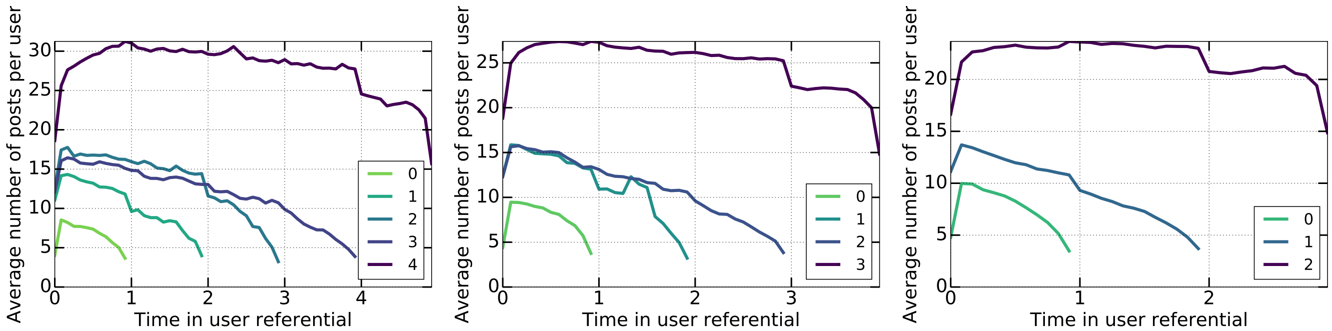
[9] We only show these figures for the sake of saving space, but the same trends are observed in the other cohorts.



(a)

(b)

Figure 4: Figure (a) shows the average number of posts per active users over clock time and Figure (b) the active users in the user-time referential, both segmented by users' cohorts. The user cohort is defined by the year of the user's creation time. For comparison, the black line in Figure (a) represent the overall average.



(a) 2010 cohort

(b) 2011 cohort

(c) 2012 cohort

Figure 5: Each Figure corresponds to one cohort, from 2010 to 2012, left to right. The users for each cohort are further divided in groups based on how long they survived: users that survived up to 1 year are labeled 0, from 1 to 2 years are labeled 1, and so on. For all cohorts, longer-tenured users started at higher activity levels than shorter-tenured ones.

activity level is lower for newer cohorts suggests that low-activity users from later cohorts tend to survive longer than those from earlier cohorts. That is, people joining later in the community's life are less likely to be either committed users or leave than those from earlier on: they are more likely to be "casual" users that stick around.

5. COMMENT LENGTH

Activity as measured by the average number of posts per user is one proxy for user effort. Comment 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. Thus, we put forward the following question (**RQ2**): how does comment length changed in the community over time, both overall and by cohort?

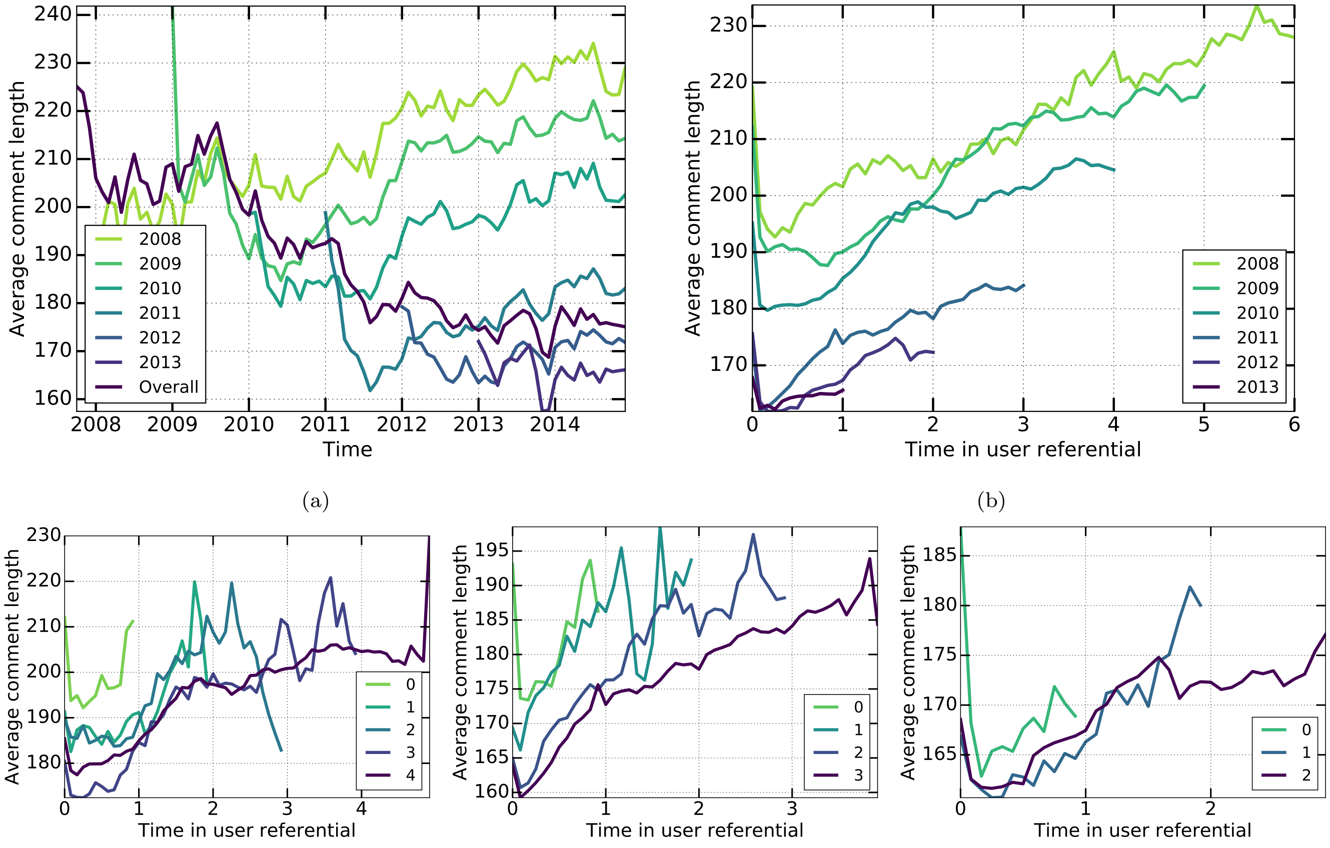
5.1 Comment length drops over time

Figure 6a shows the overall comment length in Reddit over time (the darker line) and the overall length per cohort. Based on the downwards tendency of the overall comment length in Figure 6a, one might hypothesize that users' commitment to the network is decreasing over time (**H3**), or that there is some community-wide norm toward shorter commenting (**H4**).

However, this might not be the best way to interpret this information. Figure 6b shows the comment length per cohort in the user referential time. An important observation here is that younger users start from a lower baseline comment length than older ones. Considering the fact that recent Reddit has experienced exponential growth, the weight when evaluating the overall average for Figures 6a and 6b is shifted towards the comment length for the ever-growing younger generation as the years go by; this younger generation brings down the average since their average is lower.

5.2 Simpson's Paradox: the length also rises

Let us go back to Figure 6a, which shows the overall average comment length on Reddit over time. We see a clear



(a)

(b)

(c) 2010 cohort

(d) 2011 cohort

(e) 2012 cohort

Figure 6: Figure (a) shows the average comment length over clock time and Figure (b) from the user-referential time. Both figures show the cohorted trends. The overall average length per comment decreases over time, although for any individual cohort, it increases after a sharp initial drop. Figures (c), (d) and (e), similar to Figure 5, shows the monthly average comment length for active users in the cohorts of 2010, 2011 and 2012, segmented by the number of years that the user survived in the network. Opposite the analysis for average posts, which showed that low-activity users were the first to leave Reddit, here, people who start out as longer commenters are *more* likely to leave.

trend towards declining length of comments in the overall line (the black line that averages across all users). 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 promote longer comments on Reddit.

However, in Figure 6b, we saw that average comment length increases over time for every cohort. While later cohorts start at smaller comment length, after an initial drop, all cohorts show positive trends towards writing longer comments over time. This is puzzling: when each of the cohorts exhibits a steady increase in their average comment length, how can the overall mean comment length decrease? This anomaly is an instance of the Simpson's paradox [51], and occurs because we fail to properly condition on different cohorts when computing mean comment length.

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

	Cohorts								
Year	2007	2008	2009	2010	2011	2012	2013	2014	Overall
2007	220	-	-	-	-	-	-	-	220
2008	208	198	-	-	-	-	-	-	204
2009	224	204	201	-	-	-	-	-	208
2010	223	204	189	184	-	-	-	-	193
2011	233	211	199	184	167	-	-	-	182
2012	241	221	212	197	173	167	-	-	178
2013	244	225	214	199	177	167	164	-	174
2014	246	229	217	204	183	172	165	176	176

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

umn. What happens here is that the latter cohorts have many more users than earlier ones. Since their numbers increase year by year, we have a much larger contribution from them towards comments, compared to users of earlier cohorts. This uneven contribution leads to the paradox we observed in Figure 6a.

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

5.3 New users burn brighter

As with the posting per user, we can not say if the increase in the curves seen in 6b are due to the lower effort users dying first or because users are writing more as they live on the network. To answer this, 6c allow us to make two important observations: first, *comment length does 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.

6. KINDS OF CONTRIBUTIONS

One common question from the literature is what sorts of activities users engage in, for instance, to categorize users into roles they play in the community [54]. As with comments length, we propose the following research quetion (**RQ3**): how does users' activities changed in the community over time, both overall and by cohort?

6.1 Over time, responsiveness increases

Consider the case of Usenet: people who never start threads and only respond play the role of answerer, while there are other roles that include fostering discussion [55]. These might naturally map onto people who primarily comment and who primarily submit in Reddit, respectively. 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 7a shows the overall and cohorted evolution of the number of comments for each submission users made from 2008 to 2013. Here we see that users who most prefer commenting to submitting come from 2009 to 2011, and we observe that, over time, the average ratio of comments to submissions increases both overall and per-cohort for active users.

Again, we analyze our data from the user-time referential, as seen in Figure 7b. It shows a clear pattern for users in earlier cohorts to have a lower comment per submission ratio than users in later cohorts ones, given that they both survived the same amount of time. Surviving users from later cohorts also exhibit a more rapid increase in comments per submission than those from earlier cohorts. In particular, the 2008 and 2009 cohorts increase much more slowly over time than those from 2010 onwards; later cohorts are more similar (although the 2012 and 2013 cohorts may level off lower than 2011 based on the limited data we have).

6.2 Comment early, comment often

Figures 7c-f shows the cohorts from 2008 to 2011 segmented by surviving year. Three interesting observations arise from these data. First, we see that just as in the analysis of average posts per user, the users who survive the longest in each cohort are also the ones who hit the ground running. They start out with a high comment-to-submission ratio relative to users in their cohort who abandon Reddit more quickly. This suggests that both the count of posts and the propensity to comment might be a strong predictor of user survival.

Second, and unlike the case for average post length, surviving users' behavior changes over time. Figure 5 shows that even for the most active users, they come in at a certain activity level and stay there, perhaps even slowly declining over time. Here, in Figures 7c-f, the ratio of comments to submissions increases over time; combined with the observation that overall activity stays steady, this suggests that the ratio is changing because people *substitute* making their own submissions for commenting on others' posts.

Finally, this increase is most pronounced in the earlier cohorts of 2008 and 2009, with ratios more than doubling over their first year, much more than for later cohorts.

7. DISCUSSION

In this section we discuss some of the processes that might explain our observations, and how they connect to other literature. We're not arguing here that we know the answers; instead, we see these as interesting avenues for future work.

7.1 Why are newer "active" users less so?

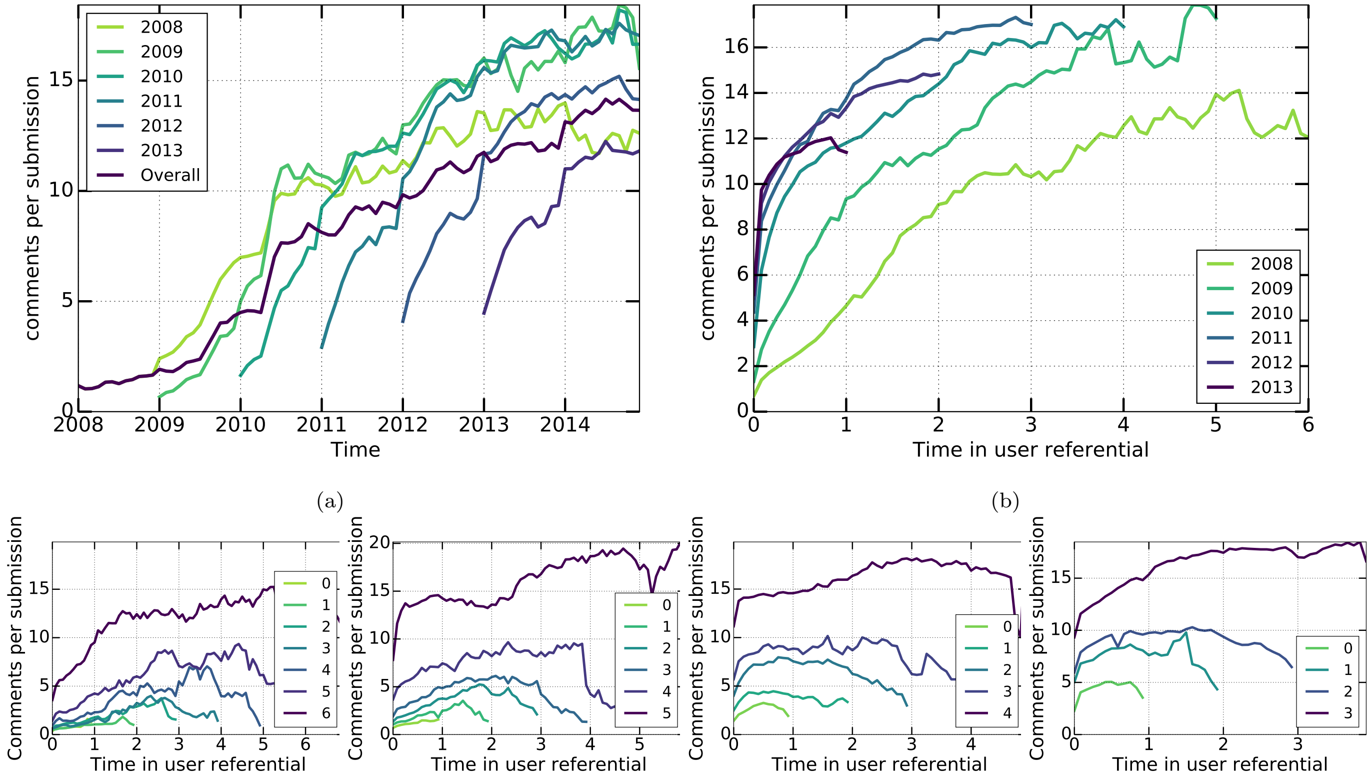
We have seen that users from later cohorts have a lower posting average than in earlier cohorts. One plausible explanation is that users self-select: users that find Reddit early in its life are also more likely than average to be those who will be attracted to it. Previous work has shown that online book reviews have a self-selection bias, where people who are more likely to like (or promote) the book review it earlier, leading to a positive early bias in an item's life [34]. In Reddit's case, this would mean that the mixture of users joining in the early stage of the community would be disproportionately likely to be the most active ones and the latter ones are more likely to be less active.

Another plausible hypothesis for later cohorts having a higher number of less active users could be that, over time, Reddit has accumulated an increasing number of valuable-but-small/niche communities. The increased diversity might support a wider set of users in getting value, explaining the increased survival percentage. The niche/smaller nature of newer communities might provide fewer opportunities to both submit and comment, explaining the lower average activity for surviving users.

A third hypothesis is that Reddit overall is becoming more about consumption and voting on content rather than producing it. Older users with contribution norms continue to contribute; newer users tend to provide audiences and feedback. High-resolution voting data could be a real boon in understanding if this is true.

7.2 Why are comments getting shorter?

We also observed that overall, comment lengths are getting shorter over time.



(a)

(b)

(c) 2008 cohort

(d) 2009 cohort

(e) 2010 cohort

(f) 2011 cohort

Figure 7: Figure (a) shows the average comment per submission ratio over clock time for the cohorts and the overall average. Figure (b) shows the average comment per submission from the user-referential time for the cohorts. Figures (c), (d), (e) and (f), similarly to Figure 5, shows the 2008, 2009, 2010, and 2011 cohorts, segmented by the number of years a user in the cohort survived. As with average posts per month, users who stay active longer appear to start their careers with a relatively higher comments per submission ratio than users who abandon Reddit sooner. Unlike that analysis, however, the early 2008 cohort ends up below the later cohorts in Figure (b).

One hypothesis is that users are being shaped by an "initial value problem". We can imagine that users as they join the network, tend to produce content according to the norms of what they see [10,27]. The observed behavior of the comments length for the users in Reddit is a initial drop, followed by steady increase as the user survives. If the starting point for the initial drops are taken as the average of the network, that is what is observed by the user in the network, the initial drop would place each cohort starting at lower levels than the previous one. Figure 6a presents some support this hypothesis: the initial month of each cohort year, which consists of data only from users who joined in that month, is quite close to the overall line from the prior month.

Another hypothesis advanced by community members[10] is that Reddit's karma system favors shorter comments. That is, people can get more upvotes for a given amount of effort by writing more, shorter comments. This could be directly measured even with the available data, and might be the start of a very interesting line of future work that tries to model strategic posting and attention distribution behavior in Reddit.

7.3 Why do comments per submission increase?

We also saw that comments per submission increase over time for surviving users, and that this is most dramatic for users who join earlier.

One process hypothesis is that this is because early in Reddit's life, there simply weren't as many submissions to comment on, meaning that people who wanted to be active contributors more or less had to submit in order to do so. As the community grew, more content became available, making information seeking more valuable—and perhaps increasing the value and ease of commenting.

This question of ease and value might be more general, and tie to our observations about self-selection and karma accumulation. Most users in social networks are known to be lurkers: only seeking information and passively observing, not engaging and contributing with content [38, 46]. Consumption in Reddit is valuable and easy, and some contributions are easier than others: reading is easier than voting; voting is easier than commenting; commenting is easier than submitting. Only users for whom finding and submitting comments is relatively easy or relatively valuable are likely to be frequent submitters or "power users" [25, 41]. We suspect such users are more likely to be ones who found Reddit earlier, when it was relatively small, and stuck with it.

[10] See https://www.reddit.com/r/TheoryOfReddit/comments/1a7aoj/retracing_the_evolution_of_reddit_through_post/.

7.4 Limitations and Future Work

In this paper we focused our attention on behavior attributable to specific users, which in this dataset meant submissions and comments. As with many analysis that focus on visible behavior, this means we miss important phenomena. In particular, we discount lurkers despite their known importance as audience members [39] and potential future contributors [47]. Many lurkers likely vote, and thus lurking may be even more important in a context like Reddit where votes affect content visibility and provide explicit markers of attention and reputation.

However, the dataset does not have information on individual voters or timestamps, just the aggregate number of votes a post had received at the time of the crawl, making it impossible to effectively treat them as activity measures and ways to understand the behavior of those who voted. The existing voting data might be much more useful, however, in addressing questions that involve predicting a given user's future behavior based on whether and how other users respond to a user's early contributions [23, 49].

Another blind spot that focusing on visible behavior can induce is our emphasis on active users. This is a reasonable view of the community that focuses on what is happening, but our results should all be interpreted in the context of "given the set of active users at any given time". Applying these results to questions that require considering all users would be a mistake.

One of our assumptions was that one account is associated with one user. This might not be the case, as more than one user can share the same account [28] or one user can have multiple accounts that are userd regularly or simply thrown away [4]. While it is not clear what is the impact of such behavior in our analyses, it raises the question of whether these users-accounts' associations can be de-anonymized.

We did, implicitly, consider survival in the analyses that broke cohort down by survival time; we see careful thinking about what it means to "survive" in a community as an interesting problem in its own right. Potentially, users' "breaks" from the network can influence both our results and other analyses that assume users depart on their last visible day of activity. Focusing on activity also fails to account for actual deletion in many contexts. In Reddit, activity from users is marked with a username of "[deleted]" (which we were able to ignore after realizing that one author had millions of comments!), but in some contexts, such as Wikipedia articles that are deleted, edit behavior on those articles do not show up in many data dumps.

These questions of how to define active users and dead users and distinguishing patterns of behavior seems an interesting venue to pursue. Better definitions of "active" and "dead" users might allow us 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; in Wikipedia, the practice of leaving temporarily is so common it is called a "wikibreak". Understanding how your network fares in terms of user burstiness is essential to understand how the users use the network and to shape the user experience. A better definition of "death" would allow us to investigate the "rebirth" of users, that is, users that come back to the community. Rather than an annoying right censorship statistical problem, it might pose a much more central issue, as a community's survival might not depend only in its ability to attract and retain users, but also in the ability to "resurrect" old users.

8. CONCLUSIONS

This work highlights the importance of taking time into consideration when analyzing users' evolution in social networks. We do so by cohorting the users based on their creation year. Although simple, this approach provides evidence of significant differences between methods that account for time with methods that do simple overall analyzes. We also analyze the evolution of users and communities from a shifted time referential: considering the time of an action in relation to the user creation date. This also reveals unexpected phenomena that we would otherwise not notice.

While investigating our research question **RQ1**, we found that user posting activity for surviving Reddit users is actually significantly higher than a naive average would suggest, that older users who survive are considerably more active than younger survivors, and that these newer users are unlikely to catch up. Controlling for survival provided evidence for our hypothesis **H2**, supporting that users have a stable level of posting activity over time (with slightly decreasing patterns), the percentage of surviving but low-activity users is increasing in the younger cohorts and that low activity users dying faster is the main reason for the drop in the overall average curve.

Similarly, when dealing with our research question **RQ2**, we analyzed user effort based on average comment length. We found that, while the overall average in Reddit seems to decrease, users actually write longer comments as they survive, no matter when they joined. Still, later cohorts of users that joined the network are writing smaller comments; their greater number leads to this version of Simpson's paradox, where where the overall average decreases while the series for each individual cohort increases.

Finally, we analyze the type of activities users engage in our research question **RQ3**, differentiating comments and submissions. We found that users with a higher comments per submission ratio are more likely to survive longer in the network. Even more, this behavior changes as the users survive—particularly for earlier cohorts. Users' comments per submissions patterns change, and their main mechanism to do so seems to be replacing their submitting by commenting behavior, as their posting activity remains stable.

An important observation that we made is that the overall evolution of users' behavior in a network is driven by three factors: actual changes in users' behavior over time, users joining the network and users leaving the network. Failing to account for different demographics joining and leaving the network might limit our interpretation of the data (**H1**, **H3** or **H4**) and lead to wrong conclusions.

Both our and work and its limitations suggest fruitful directions for better understanding of users' evolution in both Reddit and online communities in general, directions we hope inspire other work in this area.

9. ACKNOWLEDGMENTS

The authors are grateful to FAPESP grant #2011/50761-2, CAPES grant #99999.009323/2014-07, NAP eScience - PRP - USP, Google and RepNerv.

10. REFERENCES

[1] J. Arguello, B. S. Butler, E. Joyce, R. Kraut, K. S. Ling, C. Rosé, and X. Wang. Talk to me: Foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 959–968, New York, NY, USA, 2006. ACM.

[2] Y. Artzi, P. Pantel, and M. Gamon. Predicting responses to microblog posts. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 602–606, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[3] S. Beldona. Cohort Analysis of Online Travel Information Search Behavior: 1995-2000. *Journal of Travel Research*, 44(2):135–142, Nov. 2005.

[4] K. Bergstrom. âĂIJdonâĂŹt feed the trollâĂİ: Shutting down debate about community expectations on reddit.com. *First Monday*, 16(8), 2011.

[5] S. L. Bryant, A. Forte, and A. Bruckman. Becoming wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '05, pages 1–10, New York, NY, USA, 2005. ACM.

[6] B. Butler, E. Joyce, and J. Pike. Don't look now, but we've created a bureaucracy: The nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1101–1110, New York, NY, USA, 2008. ACM.

[7] B. Choi, K. Alexander, R. E. Kraut, and J. M. Levine. Socialization tactics in wikipedia and their effects. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 107–116, New York, NY, USA, 2010. ACM.

[8] T. Correa, A. W. Hinsley, and H. G. de Zúñiga. Who interacts on the web?: The intersection of users' personality and social media use. *Comput. Hum. Behav.*, 26(2):247–253, Mar. 2010.

[9] D. Cosley, D. Huttenlocher, J. Kleinberg, X. Lan, and S. Suri. Sequential influence models in social networks. In *Proceedings of the Fourth International AAAI Conference on Web and Social Media*, 2010.

[10] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 307–318, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[11] G. L. Davis, M. J. Alter, H. ElâĂŞSerag, T. Poynard, and L. W. Jennings. Aging of hepatitis c virus (hcv)-infected persons in the united states: A multiple cohort model of {HCV} prevalence and disease progression. *Gastroenterology*, 138(2):513 – 521.e6, 2010.

[12] L. EM, V. CM, and H. RI. The effect of acute renal failure on mortality: A cohort analysis. *JAMA*, 275(19):1489–1494, 1996.

[13] E. Gilbert. Widespread underprovision on reddit. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 803–808, New York, NY, USA, 2013. ACM.

[14] N. D. Glenn. *Cohort analysis*, volume 5. Sage Publications, 2005.

[15] K. E. D. Glenn Firebaugh. Trends in antiblack prejudice, 1972-1984: Region and cohort effects. *American Journal of Sociology*, 94(2):251–272, 1988.

[16] K. A. Goyette. College for some to college for all: Social background, occupational expectations, and educational expectations over time. *Social Science Research*, 37(2):461 – 484, 2008.

[17] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 491–501, New York, NY, USA, 2004. ACM.

[18] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. E. Zhao. Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 369–378, New York, NY, USA, 2009. ACM.

[19] A. L. Hughes and L. Palen. Twitter Adoption and Use in Mass Convergence and Emergency Events. In *ISCRAM Conference*, May 2009.

[20] M. Ingram. Digg Redesign Met With a Thumbs Down, 2014.

[21] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[22] M. K. JENNINGS. Political knowledge over time and across generations. *Public Opinion Quarterly*, 60(2):228–252, 1996.

[23] E. Joyce and R. E. Kraut. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11(3):723–747, 2006.

[24] A. M. P. Karl L. Alexander, Scott Holupka. Social background and academic determinants of two-year versus four-year college attendance: Evidence from two cohorts a decade apart. *American Journal of Education*, 96(1):56–80, 1987.

[25] A. Kittur, E. H. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Alt.CHI at 25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007)*, 2007.

[26] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: Conflict and coordination in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 453–462, New York, NY, USA, 2007. ACM.

[27] F. Kooti, H. Yang, M. Cha, K. Gummadi, and W. Mason. The emergence of conventions in online social networks. In *Proceedings of the Sixth International AAAI Conference on Web and Social Media*, 2012.

[28] A. M. I. Lampinen. Account sharing in the context of networked hospitality exchange. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, CSCW '14, pages 499–504, New York, NY, USA, 2014. ACM.

[29] F. Lardinois. Digg redesign tanks: Traffic down 26%(updated with new reddit stats), 2014.

[30] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 462–470, New York, NY, USA, 2008. ACM.

[31] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 177–187, New York, NY, USA, 2005. ACM.

[32] M. R. LEVY and S. WINDAHL. Audience activity and gratifications: A conceptual clarification and exploration. *Communication Research*, 11(1):51–78, 1984.

[33] 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.

[34] X. Li and L. M. Hitt. Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474, 2008.

[35] W. Mason and S. Fienberg. Introduction: Beyond the identification problem. In W. Mason and S. Fienberg, editors, *Cohort Analysis in Social Research*, pages 1–8. Springer New York, 1985.

[36] J. McAuley and J. Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 165–172, New York, NY, USA, 2013. ACM.

[37] 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. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, CSCW '15, pages 1728–1740, New York, NY, USA, 2015. ACM.

[38] B. Nonnecke and J. Preece. Lurker demographics: Counting the silent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 73–80, New York, NY, USA, 2000. ACM.

[39] B. Nonnecke and J. Preece. Silent participants: Getting to know lurkers better. In C. Lueg and D. Fisher, editors, *From Usenet to CoWebs*, Computer Supported Cooperative Work, pages 110–132. Springer London, 2003.

[40] G. W. Orazio P. Attanasio. Consumption growth, the interest rate and aggregation. *The Review of Economic Studies*, 60(3):631–649, 1993.

[41] K. Panciera, A. Halfaker, and L. Terveen. Wikipedians are born, not made: A study of power editors on wikipedia. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, GROUP '09, pages 51–60, New York, NY, USA, 2009. ACM.

[42] K. Panciera, R. Priedhorsky, T. Erickson, and L. Terveen. Lurking? cyclopaths?: A quantitative lifecycle analysis of user behavior in a geowiki. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1917–1926, New York, NY, USA, 2010. ACM.

[43] M. Prensky. Digital natives, digital immigrants part 1. *On the Horizon*, 9(5):1–6, 2001.

[44] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, GROUP '07, pages 259–268, New York, NY, USA, 2007. ACM.

[45] J. M. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, and P. Rodriguez. The little engine(s) that could: Scaling online social networks. In *Proceedings of the ACM SIGCOMM 2010 Conference*, SIGCOMM '10, pages 375–386, New York, NY, USA, 2010. ACM.

[46] S. Rafaeli, G. Ravid, and V. Soroka. De-lurking in virtual communities: a social communication network approach to measuring the effects of social and cultural capital. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, pages 10 pp.–, Jan 2004.

[47] C. Ridings, D. Gefen, and B. Arinze. Psychological Barrier: lurker and poster motivation and behaviour in online communities . *Communications of the Association for Information Systems*, 18(16), 2006.

[48] N. B. Ryder. The cohort as a concept in the study of social change. In W. Mason and S. Fienberg, editors, *Cohort Analysis in Social Research*, pages 9–44. Springer New York, 1985.

[49] C. Sarkar, D. Y. Wohn, and C. Lampe. Predicting length of membership in online community "everything2" using feedback. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion*, CSCW '12, pages 207–210, New York, NY, USA, 2012. ACM.

[50] S. Scellato and C. Mascolo. Measuring user activity on an online location-based social network. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 918–923, April 2011.

[51] E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241, 1951.

[52] G. Stoddard. Popularity and quality in social news aggregators: A study of reddit and hacker news. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 815–818, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[53] C. Tan and L. Lee. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International*

Conference on World Wide Web, WWW '15, pages 1056–1066, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[54] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, iConference '11, pages 122–129, New York, NY, USA, 2011. ACM.

[55] H. T. Welser, E. Gleave, D. Fisher, and M. Smith. Visualizing the Signatures of Social Roles in Online Discussion Groups. *The Journal of Social Structure*, 8(2), 2007.

[56] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *Proceedings of the Fourth International AAAI Conference on Web and Social Media*, 2010.

[57] H. Zhu, R. E. Kraut, and A. Kittur. The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 281–290, New York, NY, USA, 2014. ACM.