Averaging Gone Wrong: Using Time-Aware Analyses to Better Understand Behavior

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

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 (akin to Panciera et al.'s finding that Wikipedians are "born, not made" [48]), and the rise in average activity for surviving users over time is driven 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, and social norms and structure. One aspect of time often considered is the tenure of a user in the community, as in studies around modeling users' preferences [42] or analyzing the evolution of their language [12]. These analyses uncover insights about the lifecycle of a user in a community: users' preferences and behavior change with their age in a community [49], while their early experiences and activity shape future outcomes predictably [43, 48, 66, 70].

However, much 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 [37], develop new norms [33], and enact policies and rules guiding people's behavior [8]. 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 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

A common method to control for such confounds is cohort analysis, widely used in fields such as sociology [18, 41], economics [5, 47], and medicine [13, 15]. 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. People in a cohort are assumed to be exposed to the same state of the world and thus are more comparable to each other than to 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 [22, 30], and condition on the year in which

people were born to understand people's behavior, such as variations in financial decision-making [47] or opinions on issues [19, 28]. Similarly, medical studies interpret effects of drugs using cohorts of people within the same age group or amount of exposure to correlated conditions [13, 15].

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 [68]. Similarly, on review websites, users who join later tend to adopt different phrases than the older users who had joined earlier [12].

2.3 What might cause these differences?

These differences in activity between cohorts may be due to a number of reasons. One plausible explanation is 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 [40]. In this case, users who join earlier might be expected to be more active, committed users than those who join later.

Another possible explanation is that community norms may change over time. In many cases, it is a bottom-up process. 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, consciously or not [33]. Examples include adoption of 'RT', a retweeting norm by Twitter users and the subsequent introduction of the Retweet button on Twitter [33]; change in language use between new and old users on review websites [12]; and assumptions of clear roles and responsibilities on Wikipedia [32]. 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 [26, 35].

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 [7, 9]. 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 come online, their influx may lead to changes in activity patterns and communities (as with the yearly entry of college freshmen, and eventually all of AOL, gaining access to Usenet). The gradual penetration of technology also has age-related effects: people who did not grow up in a technological environment differ in their social media and search usage compared to younger generations [5, 10].
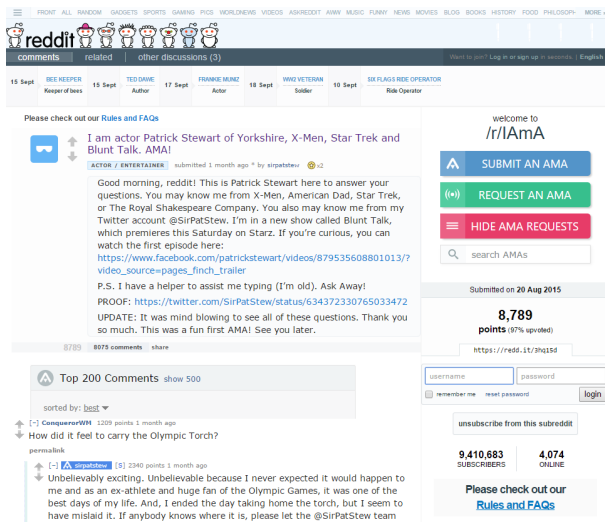
Figure 1: Reddit interface when visualizing a submission. This is Patrick Stewart's "AmA" (ask me anything) in "IAmA" (I am a), a submission where he answers users' questions in the comments. We can see the most upvoted comment and Patrick's answer right below.

2.4 Is Reddit getting "worse" over time?

All of the above reasons suggest that users from different cohorts are likely to be different, which has also been demonstrated in online and offline communities [10,12,51,60]. Further, they suggest a general hypothesis that communities "get worse" over time because newer users are likely to be less committed and knowledgeable about the community.

To address this hypothesis, we analyze both aggregate and cohort-based measures of user quality that are often raised about online communities: how active are users [25, 27, 38, 62], how much do they contribute [23,24,62], and what kinds of work do they engage in [9,48,68]?

We do this in the context of Reddit, a community that has been studied by many researchers [6, 17, 64, 66]. 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. DATA: REDDIT AS A COMMUNITY

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", and "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 submission can be imagined as the root of a threaded comment tree, in which Redditors can comment on submissions or each other's comments. Redditors can also vote on both submissions and comments; these votes affect the order in which submissions and comments are displayed and also form the basis of "karma", a reputation system that tracks how often people upvote a given Redditor's comments and submissions. We can observe these elements in Figure 1.

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, a number of Reddit users believe that it is, in fact, getting worse over time [1,14,20,46, 50,67]. Finally, 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 [65] 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[3]; 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 time-based analyses. In some analyses, we look only at comments; in some, we combine comments and submissions, calling them **"posts"**. We would also like to have looked at voting behavior as a measure of user activity[4], 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 [21], a big data processing tool. Redditor *fhoffa* imported the comments into BigQuery and made them publicly available [16]. We uploaded the submission data ourselves using Google's SDK.

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[5].

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.

---

[1]There is more to say about Reddit itself (see [55]).
[2] [57] provides more statistics about Reddit.

[3]A full description of the JSON objects is available at [56].
[4]This would also give us more insight than usual into lurkers' behavior; we'll return to this in the discussion.
[5]Ending with "_bot" or "Bot"; or containing "transcriber" or "automoderator".
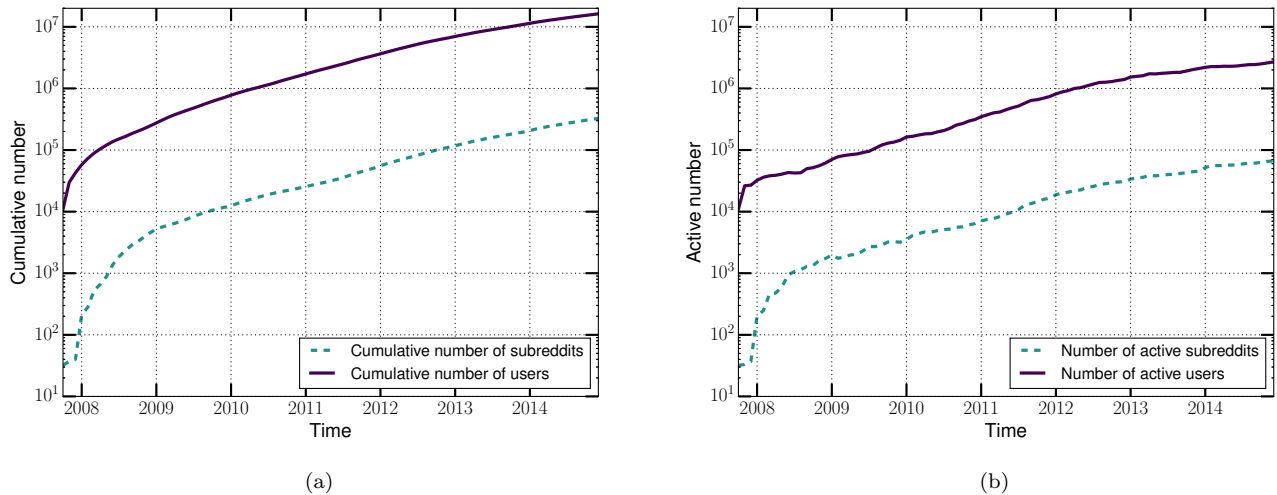
(a)   (b)

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.

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 [25, 27, 62] and communities [2] 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 around a fifth of 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 on the results; 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.
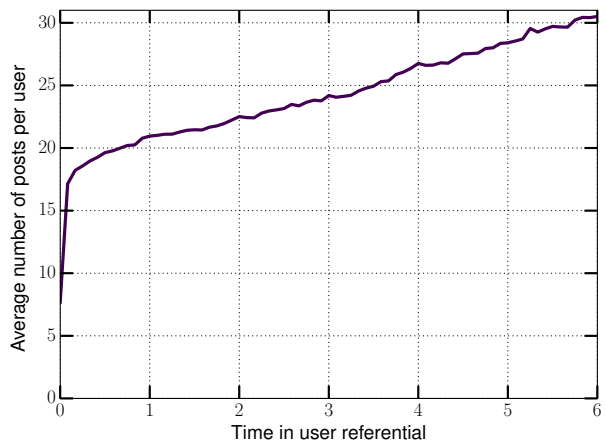
4. AVERAGE POSTS PER USER

One common way to represent user activity in online communities is quantity: the number of posts people make over time. Approaches that consider the total number of posts per user in a particular dataset [23] and that analyze the variation of the number of posts per user over time [24] have been applied to online social networks. In this section, we use this measure to address our first research question (**RQ1**): how does the amount of users' activity change over time?

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

Figure 3a shows that aggregate analysis, presenting the average number of posts per month by active users in that month. Taken at face 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

(a)   (b)

Figure 3: In Figure (a), monthly average posts per active user over clock time. In Figure (b), 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.

user creation follows an exponential distribution [36]. User lifetime, however, does not follow a exponential distribution and some types of user content generation follow a stretched exponential distribution [24]. Throw-away accounts are one example of very short-lived users in Reddit [6], for example.

To address these characteristics, Figure 3b shows a view that emphasizes the trajectory over a user's lifespan rather than the community's. To do this, 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 that the longer a user survives, the more posts they make (**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 of whether new users eventually follow in older users' footsteps (**RQ1a**).

Analyzing users' behavior by cohort is a reasonable way to address this question, and 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 the overall average (the black line in the figure) 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 [48], 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[6]. 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 is not **H1**: longer-lived users don't post more as they age. Instead, users who post more—right from the beginning—live longer, supporting (**H2**).

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 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

---

[6]We only show these figures for the sake of saving space, but the same trends are observed in the other cohorts.
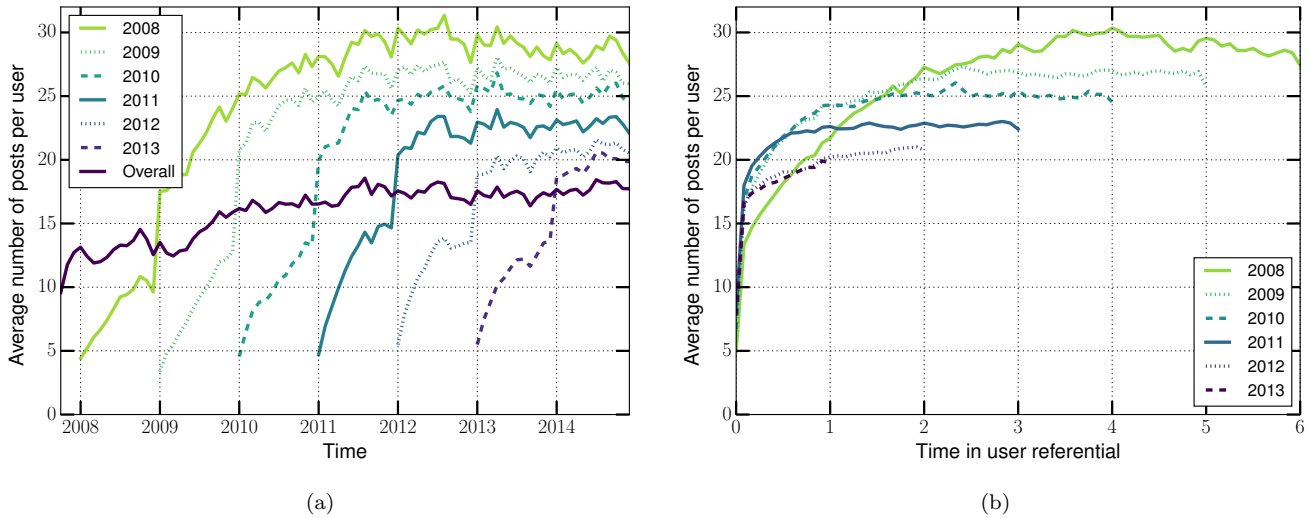
Figure 4: Figure (a) shows the average number of posts per active user over clock time and Figure (b) per active user in the user-time referential, both segmented by users' cohorts. The user cohort is defined by the year of the user's creation time. For comparison, the black line in Figure (a) represents the overall average.

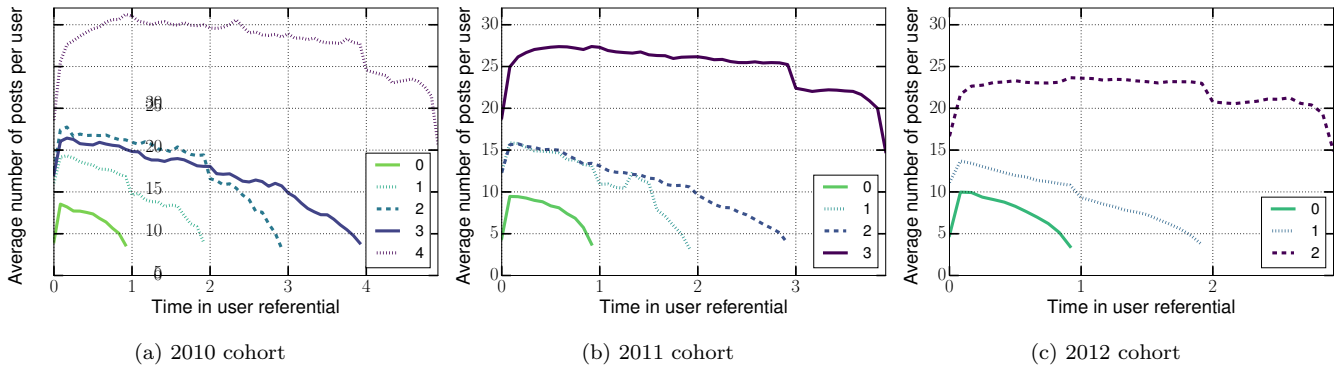| (a) 2010 cohort | (b) 2011 cohort | (c) 2012 cohort |

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.

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 change 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 Reddit has experienced exponential growth, the overall average for Figures 6a and 6b is heavily influenced by the ever-growing younger generations, who are more numerous than older survivors and who post shorter comments.

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 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
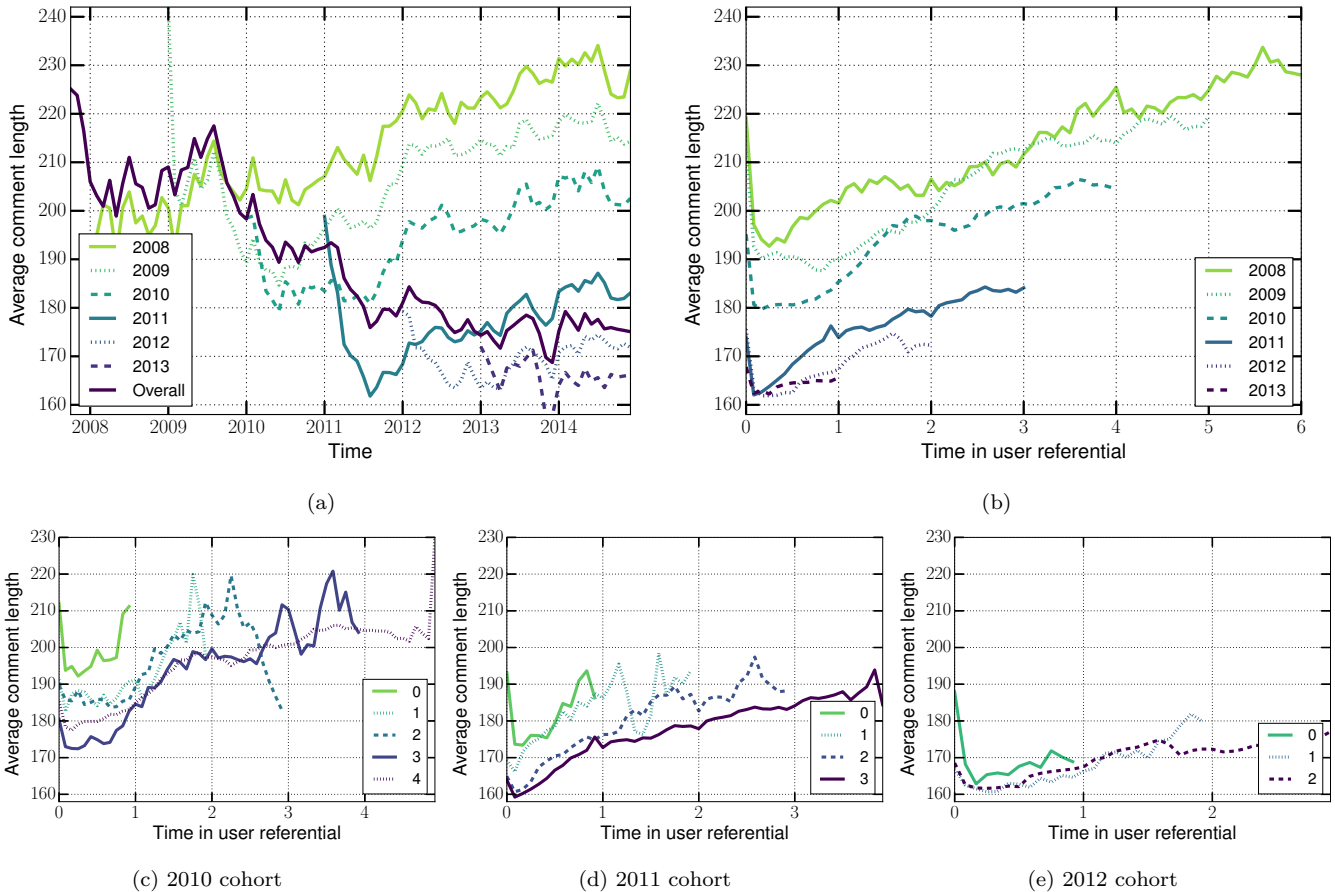
Figure 6: Figure (a) shows the average comment length over clock time and Figure (b) from the user-referential time. Both figures show the cohorted trends. The overall average length per comment decreases over time, although for any individual cohort, it increases after a sharp initial drop. Figures (c), (d) and (e), similar to Figure 5, show 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.

these numbers might think about ways to promote longer comments on Reddit.

However, Figure 6b shows that average comment length increases over time for every cohort. While later cohorts start at smaller comment length, after an initial drop, on average all cohorts write 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 [63], 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 column. What happens here is that the latter cohorts have many more users than earlier ones. Since their numbers

| | 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 start generating data in their cohort year, therefore the upper diagonal is blank. On the right column we see the overall average for all users.

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, contra (**H3**). Rather, those who tend to write less are joining the community in much larger numbers. Why later users write less is an open question we speculate about later in the discussion and future work section.

5.3 New users burn brighter

As with the number of posts per user, we cannot say if the increase in the curves seen in 6b is due to lower-effort users dying first or because users are writing more as they live longer. The sub-cohort analysis in 6c allows us to make two observations toward this question. First, *comment length does increase inside of each cohort*, no matter how long the user survives. Second, 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

In addition to questions of effort, the online community literature also often asks what sorts of activities users engage in, for instance, to categorize users into roles they play in the community [68]. As with comment length, we propose the following research quetion (**RQ3**): how do users' activities change 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 [69]. These might naturally map onto people who primarily comment and who primarily submit in Reddit, respectively. Submissions can be considered new content that an author generates, while comments can be considered as contributions toward existing content from another author.

Since the total number of comments always surpasses the number of submissions, we compute a user's ratio of comments per submission as a rough measure of the kinds of contributions they make. Figure 7a shows the overall and cohorted evolution of comments per submission from 2008 to 2013. Users who most prefer commenting to submitting come from 2009 to 2011, while 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, 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 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 useful early predictor of user survival.

Second, and unlike the case for average post length, surviving users' behavior changes over time. For post length, Figure 5 shows that even the most active users come in at a certain activity level and stay there, perhaps even slowly declining over time. Here, Figures 7c-f show that 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 AND FUTURE WORK

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 [40]. 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; several of our results support this explanation.

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.

(a)

(b)

(c) 2008 cohort
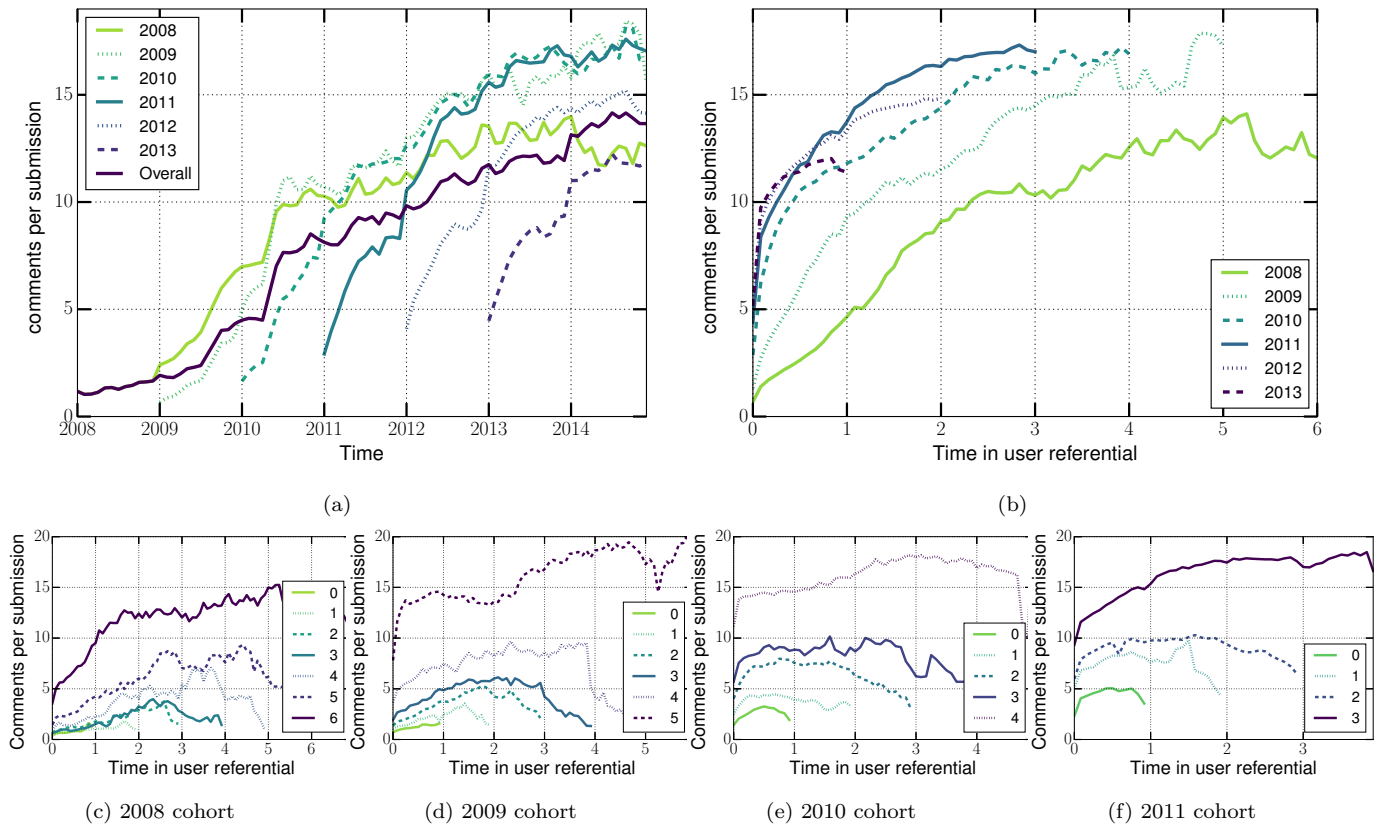
(d) 2009 cohort

(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).

7.2 Why are comments getting shorter?

We also observed that overall, comment lengths are getting shorter over time. One hypothesis is that users' behavior is being shaped by an "initial value problem"—that as users join the network, they tend to produce content according to the norms of what they see [12, 33]. Figure 6a presents some support for 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 [58] 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 around modeling 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, especially 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 to comment on; those comments in turn provide additional opportunities for commenting. In this reading, the value and ease of commenting has increased over time, making it a more common behavior.

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: seeking information and observing, rather than contributing content [44,54]. 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" [31,48]. We suspect such users are more likely to be ones who found Reddit earlier, when it was relatively small, and stuck with it.

7.4 Limitations and Future Work

In this paper we focused our attention on visible behavior attributable to specific users, which in this dataset meant submissions and comments. As with many analyses that fo-

cus on visible behavior, this means we miss important phe-
nomena. In particular, we discount lurkers despite their
known importance as audience members [45] and potential
future contributors [59]. Many lurkers likely vote, and thus
lurking may be even more important in a context like Red-
dit where votes affect content visibility and provide explicit
markers of attention and reputation.

However, the dataset does not have information on indi-
vidual voters or timestamps, just the aggregate number of
votes a post had received at the time of the crawl, making
it impossible to use them as activity measures for sure18(o)1(t)cx'1(c)1TJ0 -10.46 TD[(ku1(se)1(trs)-487(aT1(h)1(e)-3715(exi1(st)-)1(t)c

lreq1(u)1(aie)-3492conteietg anll-4441usotrl b e-342(a)1374(bmi1(st)-)1ki28(e)

CHI '06, pages 959–968, New York, NY, USA, 2006. ACM.

[3] 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.

[4] E. P. Baumer, P. Adams, V. D. Khovanskaya, T. C. Liao, M. E. Smith, V. Schwanda Sosik, and K. Williams. Limiting, leaving, and (re)lapsing: An exploration of facebook non-use practices and experiences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 3257–3266, New York, NY, USA, 2013. ACM.

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

[6] K. Bergstrom. "don't feed the troll": Shutting down debate about community expectations on reddit.com. *First Monday*, 16(8), 2011.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] G. L. Davis, M. J. Alter, H. El-Serag, 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.

[14] eleftheriaHthanatos. Is it just me or is reddit getting worse? `https://www.reddit.com/r/Anarchism/comments/2qmhgf/is_it_just_me_or_is_reddit_getting_worse/`.

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

[16] fhoffa. 1.7 billion reddit comments loaded on BigQuery. `https://www.Reddit.com/r/bigquery/comments/3cej2b/17_billion_Reddit_comments_loaded_on_bigquery/`.

[17] 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.

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

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

[20] GodOfAtheism. [Meta][Hypothetical] How would you make reddit worse? [Brackets]. `https://www.reddit.com/r/circlebroke/comments/19y889/metahypothetical_how_would_you_make_reddit_worse/`.

[21] Google. BigQuery. `https://cloud.google.com/bigquery/`.

[22] 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.

[23] 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.

[24] 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.

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

[26] M. Ingram. Digg Redesign Met With a Thumbs Down. `https://gigaom.com/2010/08/26/digg-redesign-met-with-a-thumbs-down/`, 2014.

[27] 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.

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

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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 & Social Computing*, CSCW '14, pages 499–504, New York, NY, USA, 2014. ACM.

[35] F. Lardinois. Digg redesign tanks: Traffic down 26%(updated with new reddit stats). `http://readwrite.com/2010/09/23/digg_redesign_tanks_traffic_down_26`, 2014.

[36] 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.

[37] 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.

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

[39] 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.

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

[41] 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.

[42] 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.

[43] 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 & Social Computing*, CSCW '15, pages 1728–1740, New York, NY, USA, 2015. ACM.

[44] 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.

[45] 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.

[46] NotaMethAddict. Reddit isn't getting worse at aggregating the news, everyone else has just gotten better. `https://www.reddit.com/r/TheoryOfReddit/comments/3o9wk1/reddit_isnt_getting_worse_at_aggregating_the_news/`.

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

[48] 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.

[49] 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.

[50] Phallic. Posts about "how reddit is getting worse" are getting worse. `https://www.reddit.com/r/reddit.com/comments/9y3yi/posts_about_how_reddit_is_getting_worse_are/`.

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

[52] 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.

[53] 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.

[54] 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.

[55] Reddit. About reddit.
`https://www.reddit.com/about/`.

[56] Reddit. JSON.
`https://github.com/Reddit/Reddit/wiki/JSON`.

[57] reddit admins. Blog.reddit.
`http://www.redditblog.com/2015/06/`
`happy-10th-birthday-to-us-celebrating.html`.

[58] rhiever. Retracing the evolution of Reddit through
post data. `https://www.reddit.com/r/`
`TheoryOfReddit/comments/1a7aoj/retracing_the_`
`evolution_of_reddit_through_post/`.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

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

[64] 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.

[65] Stuck_In_the_Matrix. I have every publicly available
Reddit comment for research. $\sim$ 1.7 billion comments
@ 250 GB compressed. Any interest in this? `https:`
`//www.Reddit.com/r/datasets/comments/3bxlg7/i_`
`have_every_publicly_available_Reddit_comment`.

[66] 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.

[67] voreSnake. Dear reddit, a warning.
`https://www.reddit.com/r/reddit.com/comments/`
`g5sdx/dear_reddit_a_warning/`.

[68] 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.

[69] 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.

[70] 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.

[71] 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.