Adoption and evolution of social networks from a cohort perspective

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

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

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

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

ABSTRACT

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

Keywords

user behavior;cohort;Reddit

1. INTRODUCTION

Understanding the evolution of users in a social network is essential for a variety of tasks: monitoring community health, predicting individual user trajectories, and supporting effective recommendations, among others. Many works aim at explaining these temporal aspects of evolution. Some adopt a point of view of the whole network and try to understand the patterns of behavior inside of the network [12,30], while others adopt a more user-centric point of view and try to model [5, 21, 23, 28] or predict [6] individuals' behavior.

These analyses often combine all available data into aggregate analyses of the entire community over its entire history. This can be a natural response to limitations in the amount of available data: datasets may only capture a small part of the community's history []; timestamps may not be available []; snapshots may provide limited views of the community []; or the community itself might be small [15]. Aggregate time-based analyses are also a natural first way to address these questions of community evolution.

In this paper, we argue it is likely that many of these aggregated views are misleading, because the conditions under which users join the community can vary greatly over time in ways that impact their behavior [19]. Among other things, the popularity, purpose, features, interface, and algorithms can change: Wikipedia circa 2005 and circa 2015 are very different, as are Facebook of 2005 and 2015. Analyses—including some of our own past work—that fail to account for this change may miss important details of what's really going on.

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

We find that even this simple accounting for time reveals additional insights about Reddit beyond what commonly performed aggregate analyses show. Users who join Reddit earlier post more and longer comments than those who join later. We also see evidence of larger behavior changes in

WWW '16 Montereal, Canada

these older users over time in terms of posting versus commenting, changes we think are driven in part by the growth of the community. Finally, we see that subcommunities created early in Reddit's history are likely to be much more active than those created later—and that the popularity of these communities spans across all user cohorts because of policy decisions Reddit makes in matching new users with popular subcommunities.

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

2. TIME MATTERS

Communities grow, and die, with time. 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.

2.1 Why accounting for time is important

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. Communities may grow denser or sparser with time [13], develop new norms [**?**] and/or enact policies and rules guiding people's behavior [3].

These changes mean that people experience a different version 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.

To prevent such confounding, a common unit of analysis to control for such biases is the cohort, widely used in fields such as sociology [7, 17], economics [], and medicine []. 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 [], and condition on the year in which people were born to understand people's behavior, such as variations in financial decisions-making [1] or opinions on issues [] . Similarly, medical studies interpret effects of drugs using cohorts of people with the same age group or lifelong exposure to correlated conditions [**?**, 14]. These results are also hold for behavior on the web. For instance, people who did not grow up in a technological environment differ in their social media and search usage compared to younger generations [2, 5].

Recent work shows that the importance of cohorts 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, for example, we find substantial differences in the activities of cohorts of users who joined earlier versus those who joined later [28]. Similarly, on review websites, users who join later tend to adopt different phrases than the older users who had joined earlier [6].

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 [16].

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. [12] 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, change in language use by new and old users on review websites [6], and assumptions of clear roles and responsibities on Wikipedia [**?**] . 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.

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 people's experience with the community. 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 grow big enough to have its own rules and policies, such as Wikipedia [], and 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.

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 [**?**, **?**]. Accounting for these differences can be helpful for making conclusions about outcomes of interest, such as user's activity levels, their survival, and other outcomes of interest.

2.2 Accounting for time and change

' To account for time, users on online communities are differentiated based on their age, such as when modeling their preferences [18] or analyzing the evolution of their language [**?**]. These analyses uncover insights about the lifecycle of a user in a community: users' preferences and behavior change with their age in a community [22], are and their early experiences and activity shape future outcomes predictably [19, 21, 26, 29].

However, much of past work on online communities ignores the time at which a user joins the community and analyzes all users together, irrespective of when they joined a community. These analyses normalize clock time to a user-relative time, such as measuring time since a user's first post in a community. In this paper,we argue that gnoring the actual wall-clock time at which people joined can be a mistake.

Changes in activity we might see between users might get washed out because of the internal differences due to their joining date, or may appear exaggerated when they are, in fact, no effects. As we will show, through a careful analysis of effect of a user's join cohort, not accounting for cohorts can lead us to absolutely wrong conclusions in some cases.

We use Reddit as an example community to understand people's behavior, effort and survival in a community, through the lens of cohorts. We focus on the following research questions:

Evolution of activity. How does the activity of users change as the community evolves?

Activity levels of users are a good, first proxy for understanding individual users' behavior [], their survival rate [], and gauging the general health of a community as a whole []. We look at two ways of studying activity: overall number of contributions, and the type of contributions that people make on Reddit.

Evolution of effort. How does the effort that people put in, measured in terms of their comment length, change as the community evolves?

We choose effort as our second question because work in communities such as Wikipedia shows strong associations of user effort with engagement and survival []. We look at the length of posts by people as a proxy for the effort they put in the community.

3. DATA: REDDIT AS A COMMUNITY

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

3.1 What is Reddit, briefly

Reddit is one of the largest sharing and discussion communities on the Web. According to Alexa, as of late 2015 Reddit is in the top 15 sites in the U.S. and the top 35 in the world in terms of monthly unique visitors. It consists of a large number of subReddits (853,000 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 both 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 submitted links. 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, being composed of a number of diverse subReddits allows us to explore questions of how communities diverge over time. Third, Reddit data are publicly available through an API.

3.2 The dataset

Redditor *Stuck_In_The_Matrix* used that 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 were not available. He also compiled a submissions dataset for the period of October 2007 until December 2014 that was made available for us upon request, containing a total of 114 million submissions. These datasets contain the JSON data returned by the Reddit 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 analyses, we look only at comments; in some, we combine comments and submissions, calling them "posts". We would also like to have looked at voting behavior as a measure of user activity[5], but individual votes with timestamps and usernames are not available through the API, only the aggregate number of votes that posts receive.

3.3 Pre-processing the dataset

To analyze the data, we used Google BigQuery [6], a big data processing tool that at the time of writing this paper provided free processing for up to 1TB of data for users with a Google account. Redditor *fhoffa* imported the comments into BigQuery and made them publicly available [7]. We uploaded the submission data ourselves using Google's SDK [8].

For the analyses in the paper, we did light preprocessing to filter out posts by deleted users, posts with no creation time, and posts by authors with names that follow Reddit's conventions for auto-posting bots.

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

[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.html`for more numbers on Reddit size.

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

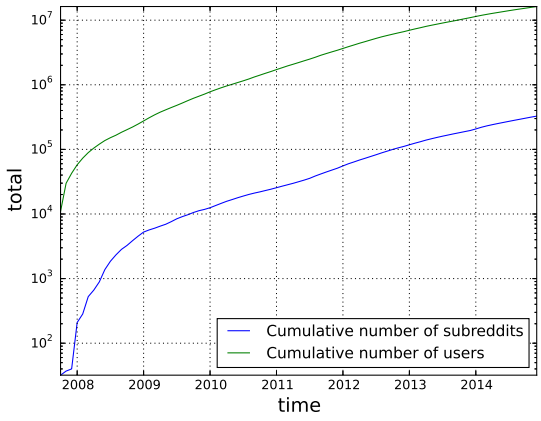
[4]A full description of the JSON objects is available at `https://github.com/Reddit/Reddit/wiki/JSON`.

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

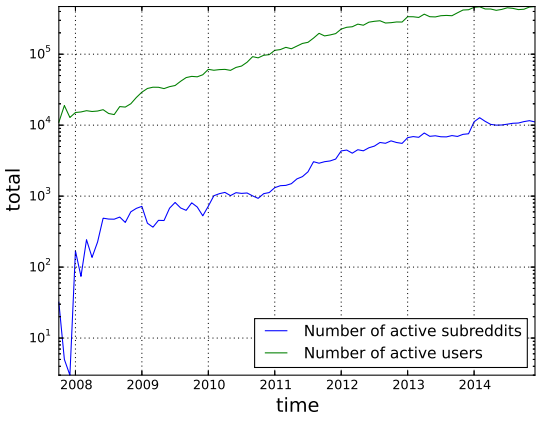
[6]`https://cloud.google.com/bigquery/`.

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

[8]Part of the alpha code in the SDK, "gcloud alpha bigquery import".



(a)



(b)

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

Here we present an overview of the dataset that shows Reddit's overall growth. Figure 1a 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, both the number of users and subReddits have grown exponentially. As of the end of 2014, about 16.2 million distinct users have made at least posts and 327,000 subReddits received at least post since Reddit's inception.

However, as with many other online sites, most users [] and communities [**?**] do not stay active. Figure 1b shows the monthly number of active users and subReddits.

We define as an *active user* as 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 470,000 thousand users and 11,400 subReddits were active, both an order of magnitude less than the cumulative numbers.

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 analyses 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 a user's creation time as the time of the first post by that user or in that subReddit. 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. ACTIVITY: AVERAGE POSTS PER USER

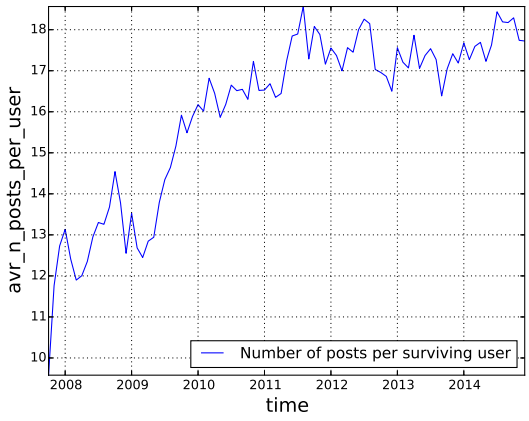
In this section, we will use a common metric of user activity in online communities, the number of posts per user over time. This metric has been used for characterizing user behavior in blogspace [8],

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

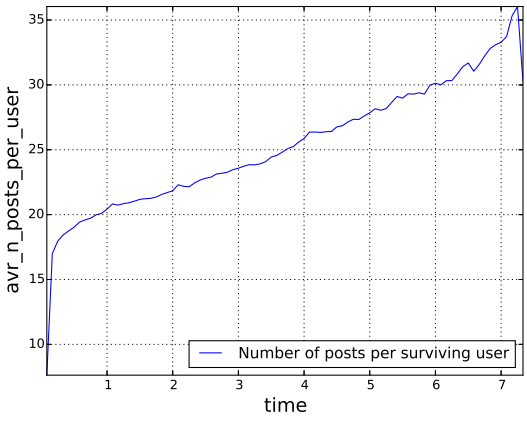
4.1 Calendar versus user-relative time

We start with a common analysis used in this kind of work: aggregating behavior in the community based on calendar time. Figure 2a shows 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 user creation follows an exponential distribution [27]. User lifetime, however, does not follow a exponential distribution and some



(a)



(b)

Figure 2: 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.

types of user content generation follow a stretched exponential distribution [9]. In Figure 2b, we show a different view that emphasizes the trajectory over a user's lifespan. Here, we scale the x-axis not by clock time, as in the left figure, 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.

One caution about interpreting the graphs that are relative to the user's start time is that the amount of data available rapidly decreases over time as users leave the community, meaning that values toward the right side of an individual data series are more subject to individual variation. A tempting conclusion at this point is that the longer a user survives, the more posts they make over time. This conclusion, however, is incorrect; we will present a more nuanced description of what is happening informed by cohort-based analyses.

4.2 New cohorts don't catch up

Figure 2b suggests that users who started earlier are more active than newer ones, raising the question of whether newer users will eventually follow in older users' footsteps. Analyzing users' behavior by cohort is a reasonable way to address this question.

Figure 3a shows our first attempt at this analysis. This Figure already shows a significant cohort effect: users from later cohorts appear to level off at a significantly lower posting average than users from earlier cohorts. That is, it suggests that newer users probably won't ever be as active on average as older ones.

However, Figure 3a also has an awkward anomaly, the rapid rise in the average number of posts during each cohort's first calendar year, especially in December. Combing cohort segmentation with user-referential analysis, as in Figure 3b, 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 the tempting conclusion that users become more active the longer they exist in Reddit, and don't explain the very rapid increase in posting activity in the first few months. An alternative hypothesis, inspired by the "Wikipedians are Born, not Made" paper [], 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. 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.

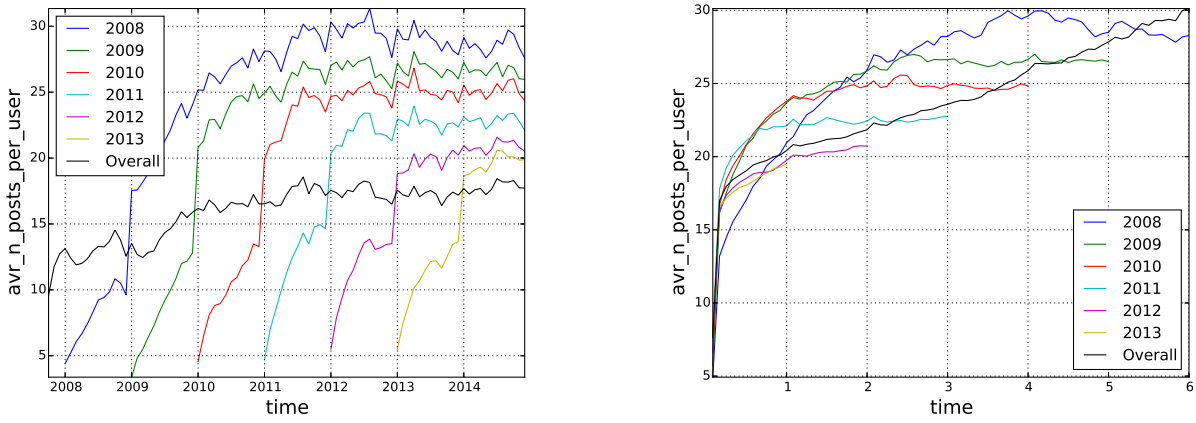
Figure 4 shows this analysis for the 2010, 2011 and 2012 cohorts[9]. Across all cohorts and yearly survival sub-cohorts, users who leave earlier come in with a lower initial posting rate. Thus, the rise in average posts per active user is driven by the fact that users who have high posting averages throughout 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 2b isn't that longer-lived users post more. It's that users who post more—right from the beginning—live longer.

Combining Figure 4'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 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 "causal" users that stick around.

5. EFFORT: COMMENT LENGTH

Activity as measured by the average number of posts per

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



(a) (b)

Figure 3: 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 as the year the user made his/her first post. For comparison, the black lines represent the overall averages from Figure 2.

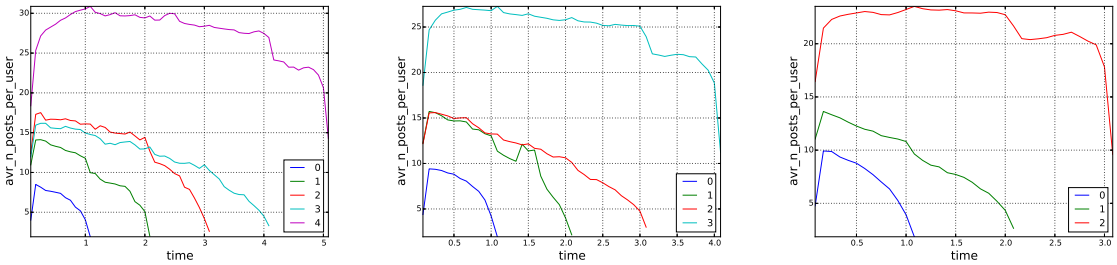


Figure 4: Each Figure corresponds to one cohort, from 2010 to 2012, left to right. The users for each cohort are further divided in groups based on how long they survived: users that survived up to 1 year are labeled 0, from 1 to 2 years are labeled 1, and so on. For all cohorts, longer-tenured users started at higher activity levels than shorter-tenured ones.

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 investigate how comment length has changed in the community over time, both overall and by cohort.

5.1 Comment length drops over time

Figure 5a shows the overall comment length in Reddit over time (the black line) and the overall length per cohort. Based on the downwards tendency of the overall comment length in Figure 5a, one might infer that users' commitment to the network is decreasing over time, or that there is some community-wide norm toward shorter commenting.

This, however, might not be the best way to interpret this information. Figure 5b shows the comment length per cohort based on the user referential time. An important thing to notice here is that younger users start from a lower baseline comment length than older users. Together with the fact that recent Reddit has experienced exponential growth, the weight when evaluating the overall average for Figures 5a and b as the years go by is shifted towards the size of the ever-growing younger generation; this younger generation brings the average down since they writing less on average.

5.2 Simpson's Paradox: the length also rises

Let us go back to Figure 5a, which shows the overall average comment size on Reddit over time. We see a clear trend towards declining sizes of comments 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 incentivize or promote longer comments on Reddit.

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

Table 1 provides some clues to what might be going on. When we move down the rows, we observe an increasing ten-

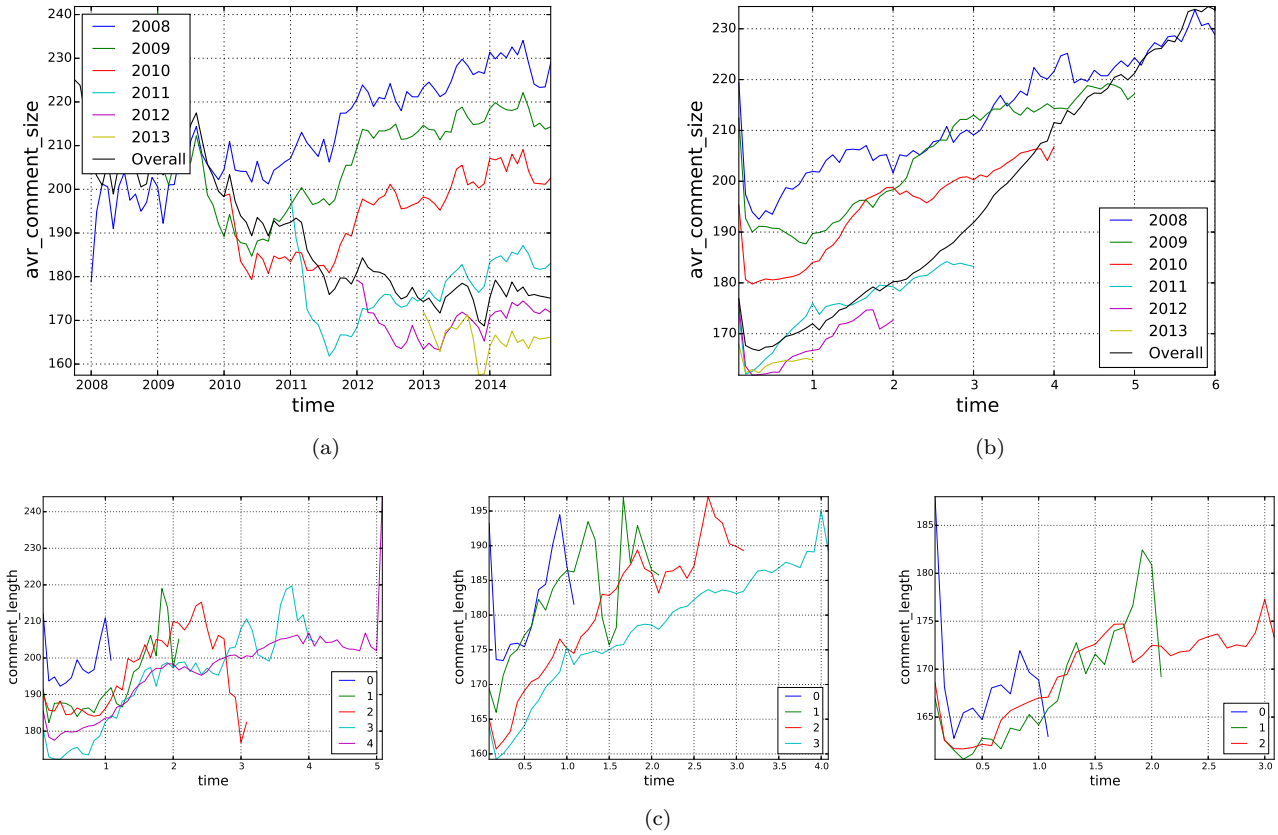


Figure 5: Figure (a) shows the average comment length over clock time and Figure (b) from the user-referential time. Both figures show the cohorted trends and the overall users trends. The overall average length per comment decreases over time, although for any individual cohort, it increases after a sharp initial drop. Figure (c), similar to Figure 4, 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.

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

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

dency in each column. It means that the average comment length increases for these users. However, when we move right through the columns, people in later cohorts tend to write less per comment. If we were to average each row, we would still get 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 increase year by year, we have a much larger contribution from them towards comments, compared to users of earlier cohorts. This uneven contribution leads to the paradox we observe in the Table 1.

Without the decision to condition on cohorts, one would have gathered an entirely wrong conclusion. People are not writing less as they survive, rather those who tend to write less are joining the community in much larger numbers. Knowing this, one may focus on better onboarding processes for newcomers, or try to learn why users in later cohorts tend to write smaller comments on average. Figure 5a suggests an interesting hypothesis: the initial month of each cohort year, which consists of data only from users who joined in that month, is generally very close to the overall line from the prior month—suggesting that new users are observing others' behavior and adopting what they see as the current norm of the community around post length.

5.3 New users burn brighter

As with the posting per user, we can not say if the increase in the curves seen in 5b are due to the lower effort users dying first or because users are writing more as they live on the

network. To answer this, 5c 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 PEOPLE MAKE

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

While submissions can be considered new content that an author generates, a comment can be considered as a contribution to an existing content from another author. Since the total number of comments always surpasses the number of submissions, Figure 6a shows the evolution of the overall and cohorted ratio of comments per submission over time from 2008 until 2013. Here we see that current recent Reddit main commenters are users from 2009, 10 and 11. It is important to highlight here that we are not talking about the average number of comments a submission gets, but how many comments a user authors for each of his/her submissions. We observe an increasing trend in the overall and cohorted ratios.

Again, we analyze our data from the user-time referential, as seen in Figure 6b. It shows a clear pattern for users in older cohorts to have a lower comment per submission ratio than the younger ones for the same surviving time. Also, cohorts from 2010 onwards show a much more similar evolution when compared with 08 and 09. Our observation that 2009, 10 and 11 are the main comenters therefore is mainly due to the fact that users in those cohorts are the ones that survived longer. To control for the surviving users, Figure 6c shows each cohort with users segmented by their surviving years. For this particular analysis, we include the years of 2008, 09, 10 and 11 because there is a change in user evolution in these years. 2010 and 11 are consistent with the following years, with users that survive longer rapidly leveling at higher values while users that survive less also rapidly leveling at lower values. For 2008 and 09, however, this is not the case. Users actually increased their commenting behavior in the early years. Still, users that survived longer show a higher commenting behavior than the ones that survived a shorter period.

This suggests that the increasing tendency in Figure 6b for the cohorts of 2010 and later is mainly due to the death of low-ratio users, but the earlier cohorts also increased due to an actual increase in the commenting behavior of the users, and not only low commenting users dropouts. Figure 6c also indicates that this ratio, conditioned on the cohort, can be a good predictor of user survival.

7. DISCUSSION

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

7.1 Why more low activity users are surviving?

We have seen that users from latter cohorts have a lower posting average than in earlier cohorts. One plausible explanation is that users who find a community suffer from a self-selection problem: users that find Reddit early in its life are also more likely than average to be those who will be attracted to it. Previous work has shown that online book reviews have a self-selection problem, since early reviews tend to be positively biased [16]. This would mean that the mixture of users joining in the early stage of the community are more likely to be the most active ones and the latter ones are more likely to be less active. A higher number of less active users joining the network would also account for their longer survival, although other mechanisms might also be in play.

Another is an argument based on cumulative advantage, status, and attention-seeking: surviving users from earlier cohorts might be more capable of producing content that gets attention from other users. This would lead to them getting more comments and votes for their content, and people who get positive attention are more likely to return [4,10,25].

7.2 Why are comments getting shorter?

One hypothesis that we might consider to explain the decreasing comment length of the users is associated to an "initial value problem". We can imagine that users, as they join the network, tend to produce content according to the norms of what they see [6,12]. 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.

7.3 Why comments per submissions are increasing?

The majority of users in social networks are known to be lurkers: users that only seek information and passively observe, not engaging and contributing to the network [20, 24]. It is reasonable to expect the same from Reddit. On the other hand, social networks often have a small number of "power contributors" [11, 21].

When we consider the evolution of the number of comments per submission, we observe that older cohorts have a much slower increase than earlier ones. Just as lurkers are the majority in the community and are attracted to it in search of information, we can imagine a set of users that are in the community in search of content and are whiling to engage in commenting, but hardly have the drive of a "power contributor" that make submissions and brings new content into the network. Since in early Reddit the amount of existing submissions was significant smaller than in the next few years, limited space existed for information seeking lurkers and "commenters". As the community grew, more content

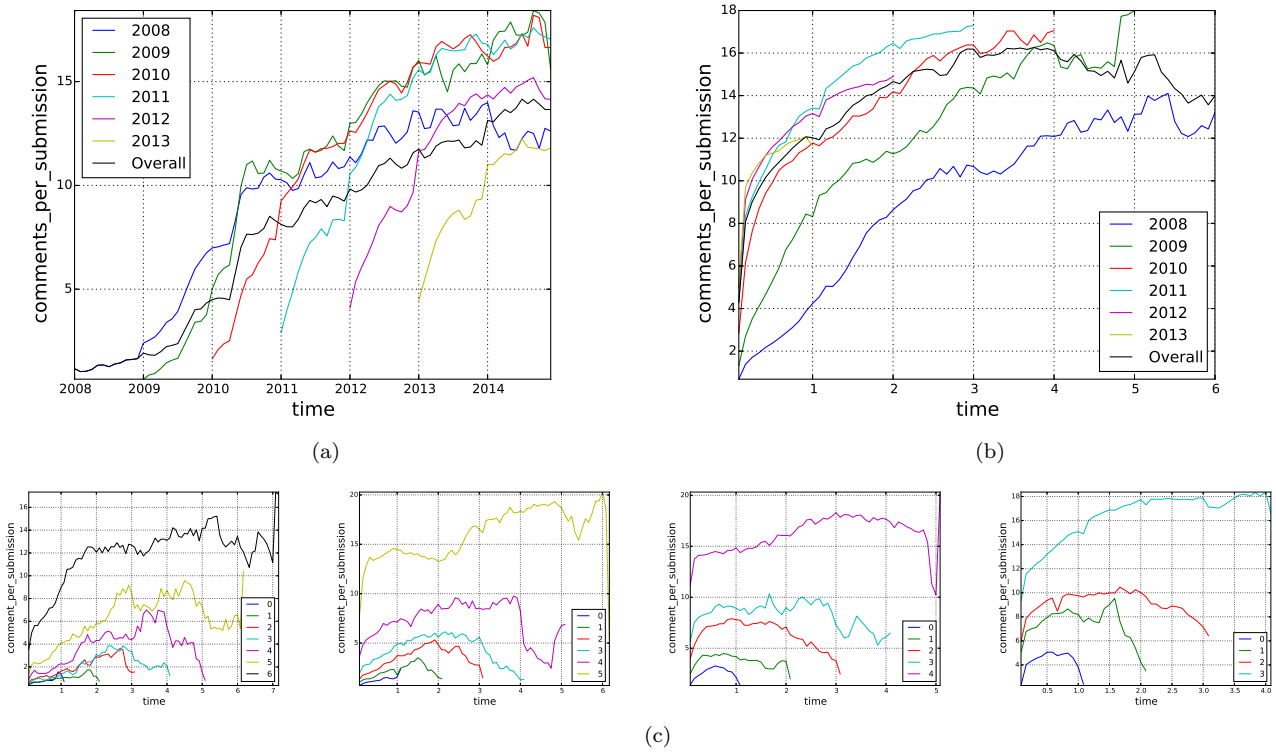


Figure 6: Figure (a) shows the average comment per submission ratio over clock time and Figure (b) from the user-referential time. Both figures show the cohorted trends and the overall users trends. Figure (c), just as in Figure 4, correspond to the cohorts of 2010, 11 and 12. They show the average comment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

from an absolute point of view was present in the social network, therefore activities of information seek became more prominent and could explain the increase in commenting activity.

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 a time-considering method with a naive overall averaging one. 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 behaviors that we would otherwise not notice.

From the user perspective, we found that user posting activity for surviving users is actually significant higher than a naive average. Also, controlling for survival, we found that the number of surviving low-posting users is increasing in the younger cohorts and low posting users die earlier than high posting users.

Similarly, we analyzed user effort based on average comment length. We found that, while the overall average in the network seem to decrease, users actually write longer comments as they survive, independently of the survival time. Also, later cohorts of users that joined the network are writing smaller comments. This is the main reason why the overall average decreases while a cohorted analysis show an increase. This is also an instantiation of the Simpson's paradox, that states that a existing trend in segmented groups disappears or reverts when these groups are joined together.

Finally, we analyze the type of activities users engage, differentiating comments and submissions. We observe that users from 2009, 10 and 11 are the main commenters in Reddit and that cohorts from 2010 onwards have a very similar evolution. 2008 and 09 cohorts, however, changed significantly their behavior as the network evolved. We discussed a possible explanation for this observation based on commenting being a information seeking task and contribution that is similar to lurking to some extent, and that early Reddit was not a fertile environment for this activity.

9. FUTURE WORK

Although many observations were made regarding Reddit, the main goal of this paper is not to study Reddit in depth. Therefore, many observations still remain without an explanation. Being aware that latter cohort users decrease the size of their comments and tend post less allow us to question ourselves about the nature and motivation of these users. Are users commenting less because newer users they are less interested into engaging with other user or are they simply "lazier" and writing is becoming less attractive for the newer generations? If it is a matter of writing being a less preferred method of communication, could Reddit improve its users' interactions providing alternative ways to

post media, such as pictures, audio and video? These are some questions that we can ask and needs further investigation of the data, and might evolve into new models of how user attention, effort and interaction in social networks are shifting through time.

One interesting question that recurrently arises when we were performing survival analysis was regarding the users' being active or not. Potentially, users "breaks" from the network can influence our results. In the same way, "death" in a social network is not well defined — users can delete their account, which is a clear sign of death, but they can also simple stop using the network. 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. Understanding how your network fare in terms of user burstiness is essential to understand how the users use the network and to set goals to improve or change the user experience. Also, a better definition of "death" would allow us to investigate the "rebirth" of users, that is, users that come back to the network. Whereas it can be considered a right censorship problem, it might pose a much more central issue, as network survival might not depend only in its ability to attract and retain users, but also in the ability to "resurrect" old users.

10. ACKNOWLEDGMENTS

Acknowledgments.

11. REFERENCES

[1] O. P. Attanasio. A cohort analysis of saving behavior by us households. Technical report, National Bureau of Economic Research, 1993.

[2] S. Beldona. Cohort analysis of online travel information search behavior: 1995-2000. *Journal of Travel Research*, 44(2):135–142, 2005.

[3] 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.

[4] B. Choi, K. Alexander, R. E. Kraut, and J. M. Levine. Socialization Tactics in Wikipedia and Their Effects. *Cscw*, pages 107–116, 2010.

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

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

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

[8] D. Gruhl, D. Liben-Nowell, R. Guha, and A. Tomkins. Information diffusion through blogspace. *ACM SIGKDD Explorations Newsletter*, 6(2):43–52, 2004.

[9] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. E. Zhao. Analyzing Patterns of User Content Generation in Online Social Networks. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 369–378, 2009.

[10] A. Halfaker, A. Kittur, R. Kraut, and J. Riedl. A Jury of Your Peers : Quality, Experience and Ownership in Wikipedia. *Proceedings of the 5th International Symposium on Wikis and Open Collaboration - WikiSym '09*, page 1, 2009.

[11] a. Kittur, E. 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. *Algorithmica*, 1(2):1–9, 2007.

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

[13] 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.

[14] E. M. Levy, C. M. Viscoli, and R. I. Horwitz. The effect of acute renal failure on mortality: a cohort analysis. *Jama*, 275(19):1489–1494, 1996.

[15] 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.

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

[17] W. M. Mason and S. Fienberg. *Cohort analysis in social research: Beyond the identification problem*. Springer Science & Business Media, 2012.

[18] J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 897–908, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

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

[20] B. Nonnecke and J. Preece. Lurker demographics: Counting the silent. *Proceedings of the SIGCHI conference on ...*, 2(1):1–8, 2000.

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

[22] 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*, pages 1917–1926. ACM, 2010.

[23] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating , Destroying , and

Restoring Value in Wikipedia. 2007.

[24] 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. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004.*, 00(C):1–10, 2004.

[25] C. Sarkar, D. Y. Wohn, and C. Lampe. Predicting length of membership in online community "everything2" using feedback. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion - CSCW '12*, page 207, 2012.

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

[27] J. L. B. K. S. Tomkins. Microscopic evolution of social networks. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 462–470, 2008.

[28] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith. Finding social roles in Wikipedia. *Proceedings of the 2011 iConference on - iConference '11*, pages 122–129, 2011.

[29] J. Yang, X. Wei, M. S. Ackerman, and L. A. Adamic. Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities. In *ICWSM*, 2010.

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