Eternal September:
Online Information-Sharing Community Analysis

Ehsan Gholami, John Karasinski, Jenette Sellin, Dmitry Shemetov
June 9, 2016

Activity in online communities varies greatly over time. Topics of conversation can shift dramatically, while users joining or leaving the community change the quality of discourse. We investigate the evolution of several online communities, Reddit, Hacker News, and StackExchange, to find how they change over time. We divide online users into cohorts based on the year of their first interaction with the community. These cohorts are tracked to compare the activities of newer users to older users, as well as the behavior of individual cohorts over time. We analyze time slices of the activity in these networks to create graphs from the users' interactions with one another. We find that users in older cohorts, though fewer in number, have a much larger centrality in the network than newer users. Additionally, older users are more active than newer users, and are far less likely to leave the community. In order for online communities to thrive, it is important for their moderators to incentivize the oldest users to remain active, as they have the most influence over the community.

Introduction and Overview

Motivation for this analysis comes from the problem of Eternal September[1]. The term originates from the 1990s Usenet community during the early Internet days. One of the earliest message boards, Usenet was a community of mainly academics and scientists privileged with university Internet access. As is common of all communities, Usenet had its own set of cultural expectations for interaction as well as distinct standards for quality of discussion. Adherence to these standards was interrupted every September when a new wave of freshman would gain access to university computers. The immediate effect was dilution of community culture by the incoming users, though, eventually, conformity-effects led the newcomers to adjust and assimilate. However, this cycle changed around 1993, when large-scale Internet Service Providers allowed Usenet access to a growing number of new PC owners. The result was an ongoing influx of new users, leading to a permanent culture shift. This phenomenon extends into modern internet communities, describing the dilution of community interests and standards caused by incoming users.

We investigate aspects of Eternal September—changes in user entry to the community, differences in user commenting behavior, community growth, and user centrality. Looking to

[1] See https://goo.gl/nA5dAQ for converstation of older users around the time Eternal September was being realized.

existing online communities including Reddit, HackerNews, and StackExchange, we inspect and compare these changes in the networks of users over time. Our main findings support the primacy of users joining early in a community. Across all the communities we studied, earlier users had, across time, higher commenting activity on average and higher values of network centrality. This remains despite the early users being outnumbered by the ones joining later. Naturally, these findings could be useful to designers of online social networks.

The Eternal September phenomenon was born out in some of the data. The /r/math subreddit experienced a loss of older member activity in 2010-2012, which corresponded to a complementary rise in the alternative community /r/puremathematics[2]. Community dissolution occurred.

Background

Recent research has studied online communities. The investigations have focused on the partitioning of users into subcommunities and inspecting their changing statistics over time. However, a missing feature of the literature is network-theoretic analyses, such as centrality measures of users in different 'cohorts'. 'Cohorts' are groups of users partitioned based on their arrival time to the network. The group of users who have arrived (posted their first activity) in the network during each year are considered to be (and remain) members of the same cohort. Our analysis provides a preliminary step in this direction.

Fire et al. [1] studies the importance of user arrival pattern in time in the network. This study is on the Reddit dataset which includes over 1.65 billion comments from 11,965 sub-communities. This study proves that users' arrival patterns affect the network topology. In fact, these features are so correlated that it is shown that one can uncover users' arrival patterns by investigating the network topology. Different sub-reddits (sub-communities) in the dataset represent various users' arrival curves (UACs), ranging from sublinear to superlinear, and from polynomial to event oriented style. With majority of curves following polynomial patterns, this study shows the importance of users' arrival pattern, namely cohorts in the network topology and the user behavior in the network.

Tang, John, et al. in [2] introduce new temporal distance metrics to investigate the information diffusion speed, considering the evolution of a network. They show the importance of accounting for the temporal changes in network connections. They investigate how previous static measures are ineffective for capturing network characteristics in their entirety. Similarly, Santoro, Nicola, et al. [3], and Casteigts, Arnaud, et al. [4], propose an approach under which one can investigate the evolution for temporal as well as atemporal indicators in time-varying and static graphs, respectively. Casteigts, Arnaud, et al. [4], introduces a new framework, called *time-varying graphs* (TVGs) to unify the existing

[2] See https://www.reddit.com/r/math/comments/zzopa/meta_this_subreddits_quality_has_sharply/ for an example of the 2012 dissatisfaction.

concepts and time varying formalisms into one structure. All of these recent studies emphasize the importance of temporal changes in networks.

The approach of dividing users into cohorts to study their large scale behavior was developed in Barbosa, et al. [6]. The authors use this simple approach of defining cohorts by temporal differences between users on Reddit data from 2007-2014. They found wide differences in user behavior, including variety in comment activity, effort, and survival rates. The authors found that older users who remain in the community are considerably more active than younger users, and that these newer users are unlikely to catch up. Similar results were found using other metrics, all suggesting that older users were significantly more influential than newer users.

Zhang et al. [5] studied an online community, the Java Forum, and compared different algorithms to measure user expertise. They characterized the network as having a bow tie structure, similar to how others researchers have characterized the web. They calculated the expertise of users based off simple statistical measures (ie, how many posts a user makes), z-score measures, a pagerank-esque algorithm, and HITS Authority. Two 'Java programming experts' manually ranked a random sampling of 135 users by their expertise on a 5 level scale, from 'Newbie' to 'Top Java expert'. They then compared the results of all of these algorithms, and found that the simple metrics perform as well or better than the more complicated methods, and that structural information in the network could be used to evaluate the expertise of individual users.

As for the user behavior and importance, centrality measures are always interesting. Many different measures have been proposed based on network and user characteristics. These measures vary from degree, betweenness to closeness, and from eigenvector, power, to reach. Many times there is a question of how to choose the best centrality measure for representing importance of nodes in a network. Thomas W. Valente et al investigate the correlation between eight various centrality measures, including for weighted graphs, in various situations to show how they evaluate similar. Andrea Landherr et al reviews three most popular centrality measures in social networks [8], which shows they can result in absolutely different conclusions based on the network situation. Tore Opsahl et al in [9] investigate how to break a tie between users with similar centrality measures in weighted graphs. They focus on three most common centrality measures. Erjia Yan show how four different centrality measures are correlated for the case of co-authorship network over time. There are in general various views one can choose or compare centrality measure for the purpose of the network.

As stated previously, we extend these by focusing on network-theoretic analyses, using measures such as closeness centrality, betweenness centrality, and degree distribution.

Technical approach and results

We investigate three datasets: Reddit, Hacker News, and StackExchange. Each data set contains a set of comments posted between 2006 and 2016. Comments are made in reply to other users' comments. Using this reply structure, we build a directed, weighted graph, placing an edge from User A to User B, when User A comments on User B's comment and weighing by the number of times User A replies to User B (Fig. 1). We divide our data into time slices on orders between months and years, depending on the data set. Users are placed into cohorts based off the date of the user's first interaction (comment or post) with the network. We investigate how the properties of these cohorts change over time, with respect to themselves and one another.

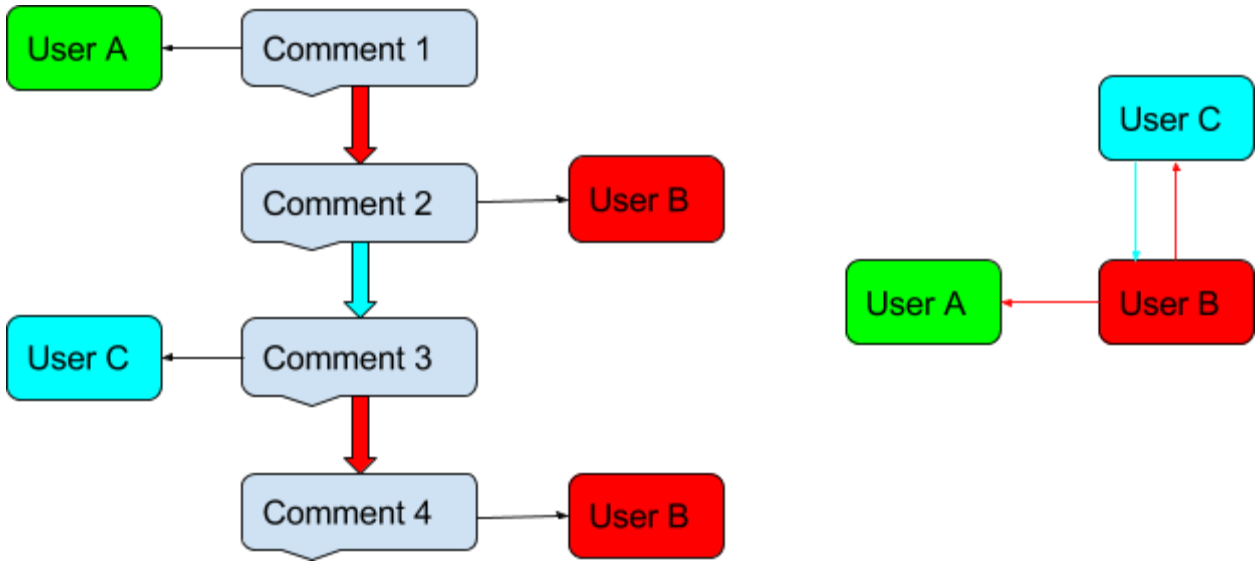

Fig 1. User-Comment network.

	Reddit (/r/math)	Hacker News	StackExchange (math)
Years	2008-2014	2006-2016	2011-2015
Users	38,832	251,715	66,702
Comments	326,355	9,082,312	1,358,756

Table 1. Year spans, user and comment counts for the three datasets we investigated.

The centrality measures used are betweenness centrality, degree centrality, closeness centrality and Eigenvector centrality.

- Degree centrality is the degree of the node: $D(x) = deg(x)$.
- Betweenness centrality: $b(x) = \sum_{j \neq x \neq k} \frac{\sigma_{jk}(x)}{\sigma_{jk}}$, where $\sigma_{jk}(x)$ is the number of shortest paths from node j to node k that pass through node n_i and σ_{jk} is the total number of shortest paths from node j to node k.
- Closeness centrality: $c(x) = \sum_j \frac{1}{d(x,j)}$ where $d(x,j)$ is the shortest distance from node x to node j.
- Eigenvector centrality: the centrality score of node i is the i^{th} component of the largest eigenvector v that solves $Ax = \lambda x$ where A is the adjacency matrix.

Reddit

Reddit is a social media website for sharing links to internet content and for discussion. The site is broken down into subcommunities called subreddits, where users can post thematically related content. User voting determines which posts get shown on the top of the page. Each post has its own discussion section, where users can comment and comment on the comments, in a tree structure, with the comments organized by votes.

The full Reddit data set we had access to was massive[3], containing a billion comments, across millions of users, and hundreds of thousands of subreddits. Due to computational limitations, we opted to study a small subcommunity of this site: the mathematics subreddit, or "/r/math". This smaller data set had 38,000 users and 326,000 comments. We split the users into cohorts, marked by the year of the first comment and then tracked each cohort's average statistics over time.

Analyzing the number of active users per cohort, we see that users join and slowly leave the community. Since joining a cohort is impossible, after the year has passed, so users can only leave. Each newly entering cohort is substantially larger than the previous years' reflecting reddit's growth in popularity.

[3]User /u/Stuck_In_The_Matrix scraped the data from Reddit and user /u/fhoffa made it available on Google Big Query.

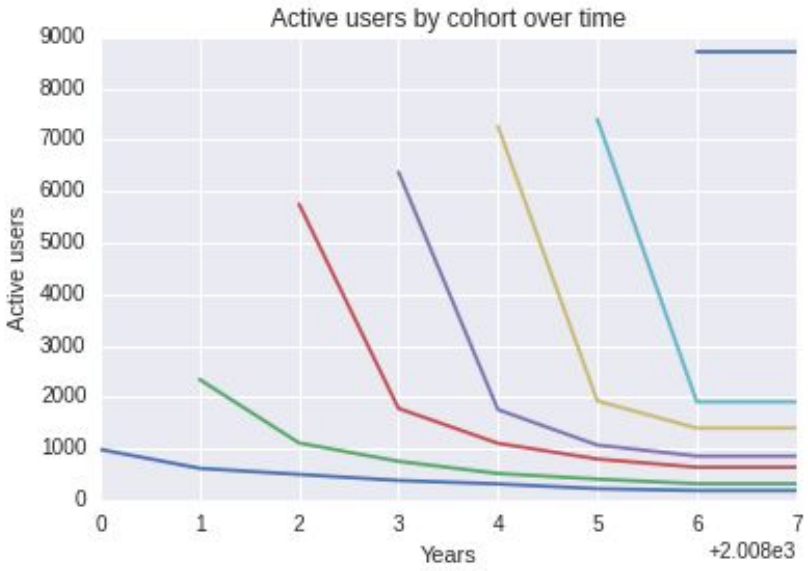

Figure 2. Number of active Reddit users in each cohort.

Next we looked at commenting behavior of each cohort, by averaging the frequency of comments. The earliest users (2008), though fewer in number, are the most active. Strangely, the next cohort (2009), didn't catch on and became the least active by 2014, along with the next 3 years of users. The (2010-2012) cohorts are similarly active by 2014.

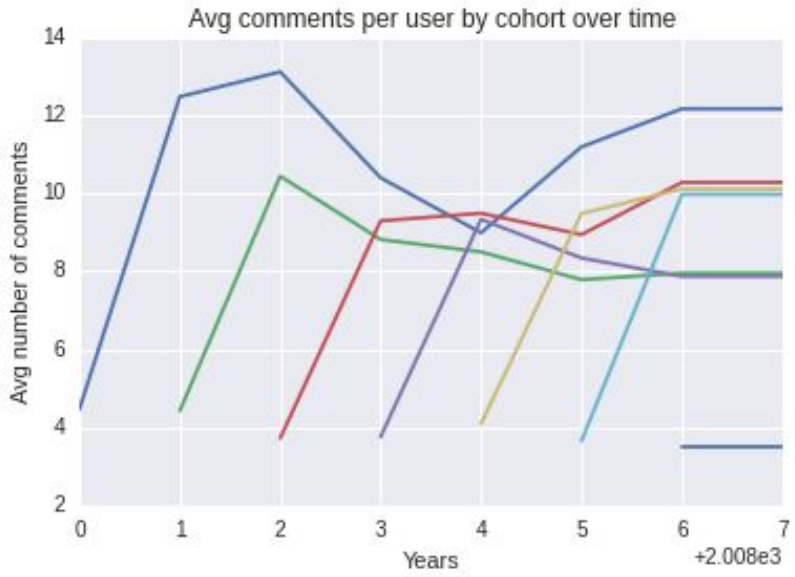

Figure 3. Average number of comments per active Reddit user.

The dip in the older user behavior corresponds to the rise in the number of comments on a sister subreddit called /r/puremathematics, which was created around that time for users who felt that discussions on /r/math had become too popularization oriented.

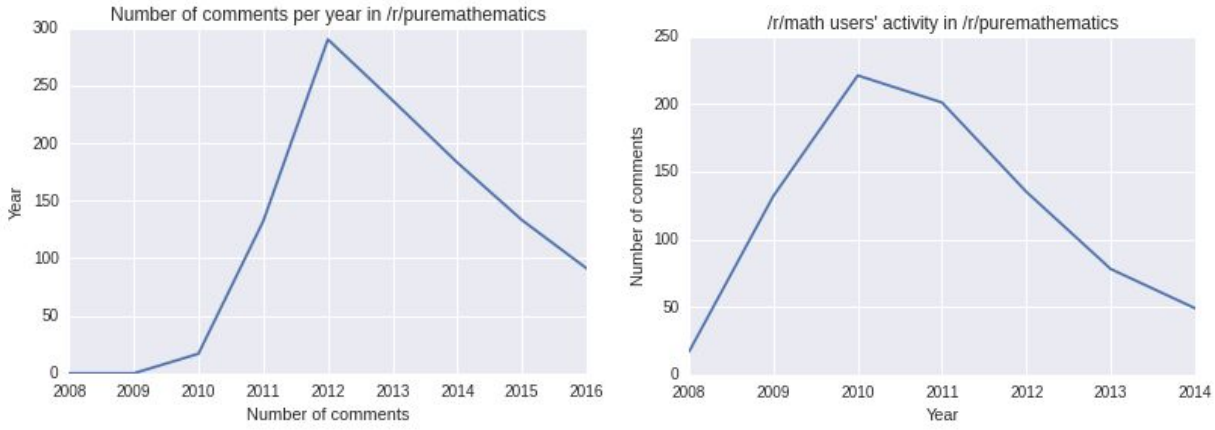

Figure 4. Activity on /r/puremathematics and specifically the activity of users that are also /r/math members.

Corresponding with the commenting behavior, the centrality of the 2008 cohort is the highest. The 2009 cohort, though less active, is more connected than the 2010-2012 cohorts. Most of the cohorts start out in their year as highly connected, then steadily dropping off. There could be many reasons for this: users could focus on speaking to other central in their first year or they could connect well to their cohort which proceeds to disperse over time.

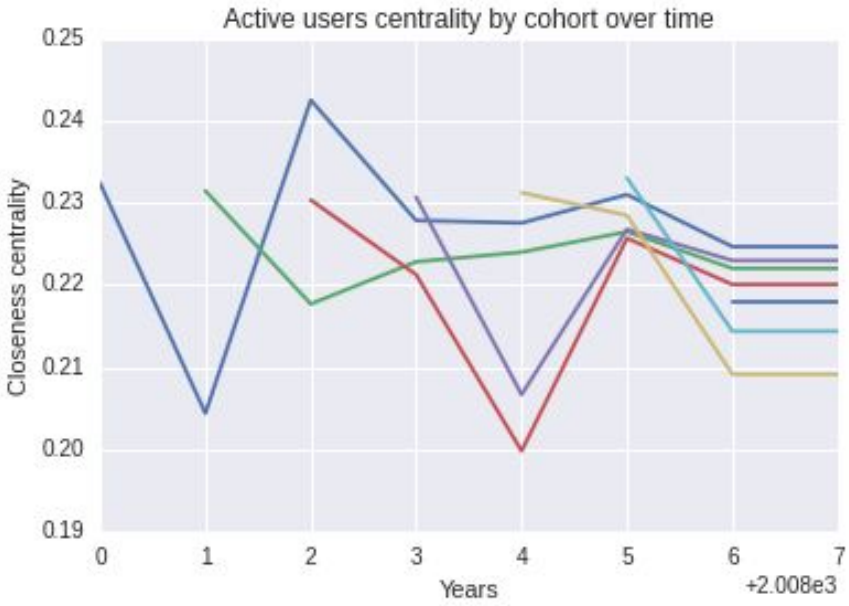

Figure 4. Average centrality of the active Reddit users in each cohort.

Thus, we find evidence of Eternal September in this community. We found that 833 of the total /r/puremathematics 1083 users had posted in the /r/math community. The top

commenters in /r/math however, did not correspond to top posters in the /r/puremathematics community.

Further analysis will involve combining the two subreddit data sets into a single network and see if we can recover the two subreddits out of clustering. We look to performing this analysis in the future.

Hacker News

Hacker News is a social news website focusing on computer science and entrepreneurship. In general, content that can be submitted is defined as "Anything that good hackers would find interesting. That includes more than hacking and startups. If you had to reduce it to a sentence, the answer might be: anything that gratifies one's intellectual curiosity." Hacker News will often focus on topics such as software and open source projects, law (especially with regard to Snowden, EFF, etc), and mainstream tech news.

Unlike the other data sets, the data for this analysis was scraped from Hacker News' API[4]. As with the other datasets, we divide the users into cohorts based off the year of their first interaction with the network. While all the data up to May 1st 2016 was captured, only data from January 1st 2007 to December 31st 2015 was used for this analysis in order to fully capture the behavior over the complete year. Additionally, as there were very few users in the 2006 cohort compared to the other cohorts, they are removed for the following analysis[5]. The data was broken down by month, forming a total of 108 time slices (twelve months for each of the nine years). By breaking the data into these slices, older users do not benefit simply from being around longer.

To get a rough understanding of the cohorts, we first considered the total number of active users in each cohort over each of the 108 networks, see Figure 5. From this, we see properties that are common across all the websites analyzed. Early users which joined the site are more likely to remain active than newer users. While newer cohorts are much larger than older cohorts, the behavior of the constituent users is very different. Newer cohorts see a sharp spike—it appears that many users will remain active for only a short time, never to return. The 2013 cohort, for instance, peaked at approximately 9000 users near the end of 2013, but had already declined to less than half of that number just a few months into 2014. While it appears that the site is growing in popularity, the acceleration of it's growth has peaked. 2013 was the largest cohort, and the 2014 and 2015 cohorts resemble the size and behavior of the 2010 cohorts. While the site is still growing in users, it is at a lower rate than before.

---

[4] Available at https://hn.algolia.com/api.

[5] Paul Graham (pg), one of the founders of Hacker News and is a very popular contributor to the site is part of this cohort. As a user he is an extreme outlier, and his individual behavior dominates the properties of the very small cohort.

With the number of active users in each time slice known, the average number of comments per active user was calculated, see Figure 6. The average number of comments per active user is dropping over time, both within each cohort and overall. Despite this, the earlier cohorts appear to comment on posts more often than the newer cohorts. Older cohorts appear to be more important to the network, as they are more likely to remain active, and more likely to comment frequently compared to newer cohorts. Despite the flux of new users to the site, by these metrics it is the oldest users that remain the most important and dedicated to the site.

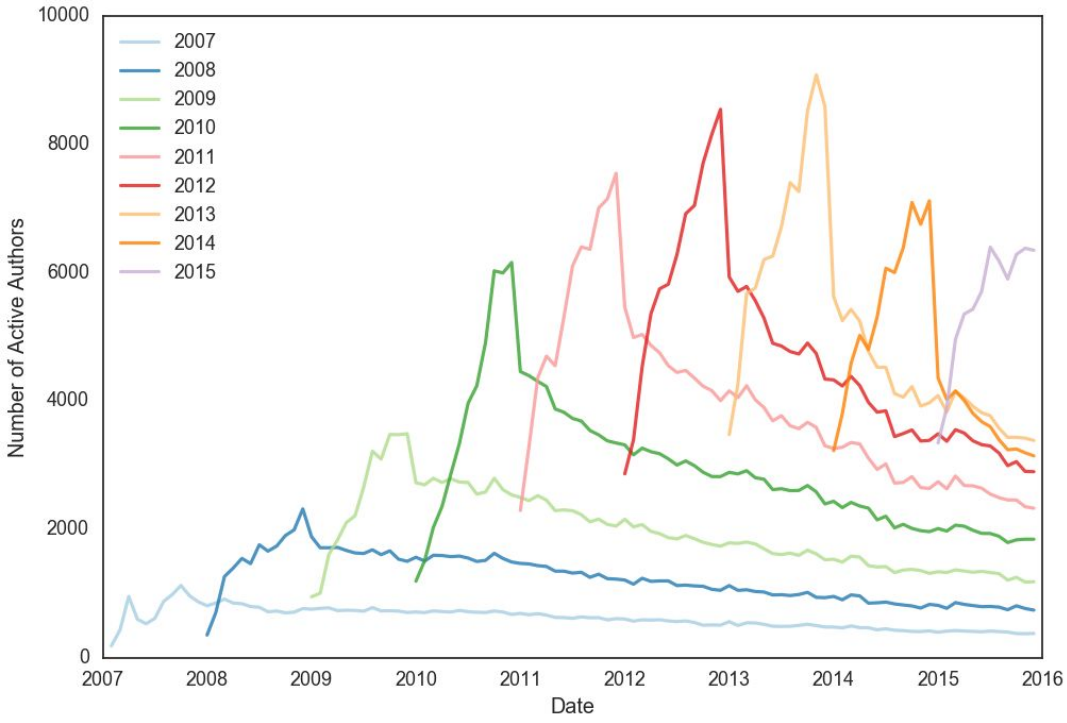

Figure 5. Number of active Hacker News users in each cohort.

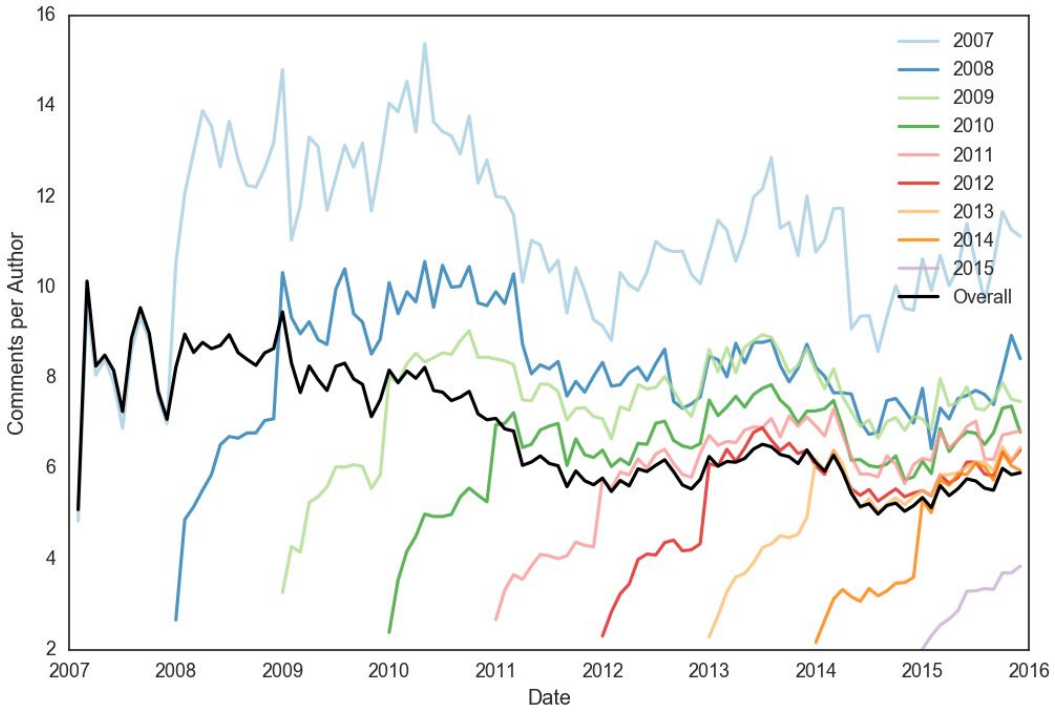

Figure 6. The average number of comments per active Hacker News user in each cohort.

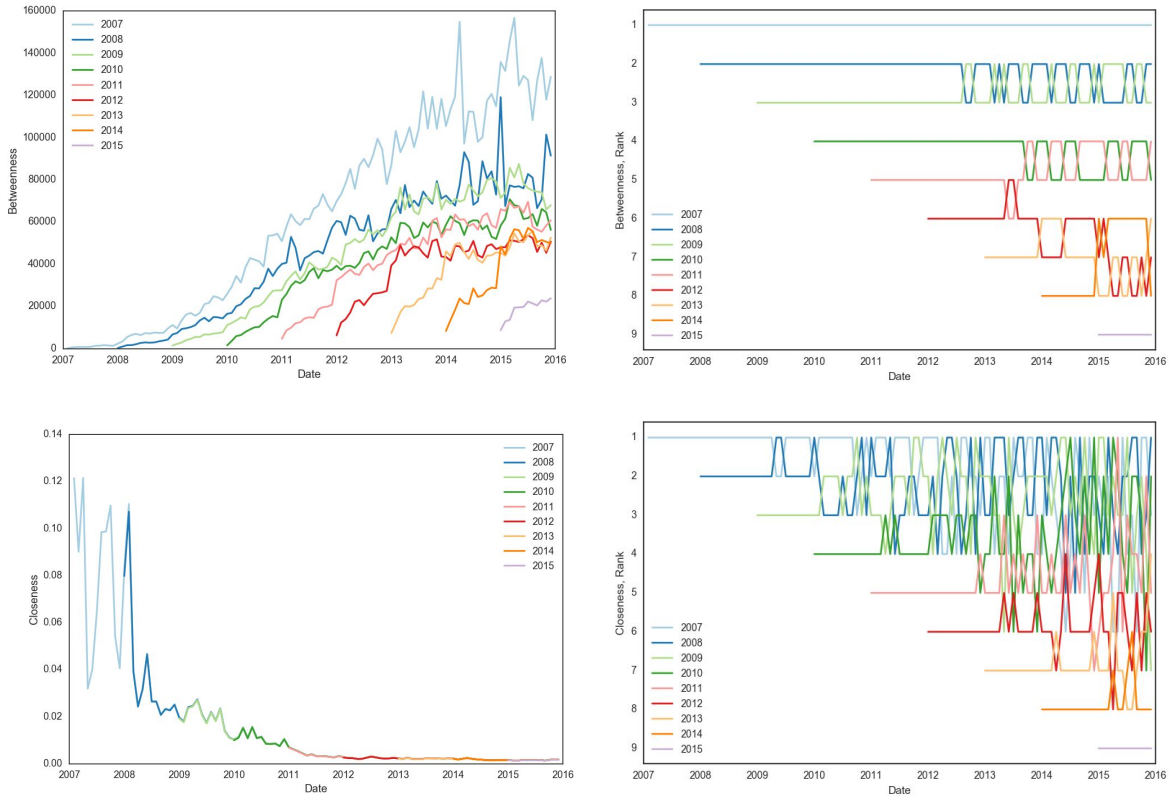

Figure 7. Upper left: The average betweenness of each cohort over time. Upper right: The rank of each cohort's average betweenness compared to the others. Here a lower rank

indicates a more central cohort. Lower left: The average closeness of each cohort over time. Lower right: The rank of each cohort's average closeness compared to the others.

Various centrality measures, including eigenvector centrality, betweenness, closeness, pagerank, degree, indegree, outdegree, and the shell index of each node in each graph was calculated, and the average value for the members in each cohort was calculated. These centrality measures were calculated in order to investigate the importance of different cohorts to the network. The betweenness metric shows that, similar to what we saw with the number of comments per user, older users are more important to the network, see Figure 7. The oldest cohorts have a larger number of shortest paths from all vertices to all others that pass through their nodes. Taking the rank of each cohort at each time slice, we see that the oldest cohorts consistently rank as the most central, though the exact order is not necessarily constant. A quick visual scan of the closeness appears to suggest that all cohorts have the same value at each time slice. Ranking these values, however, shows that older cohorts again have a higher centrality. Unlike the betweenness metric, however, closeness is much less stable with regard to which cohort is the most important.

While the degree of a node can be calculated relatively quickly, measures such as betweenness and centrality take much longer to run on large networks. After running various centrality measures, we find a very high degree of correlation between the metrics. The cohorts were ranked by each metric within each time slice, and the correlation between different metrics was calculated. With the exception of centrality, the metrics have a correlation factor greater than .97 with one another. Closeness also has a high amount of correlation with other metrics, with a value of around .85, though the authors are unclear why closeness is so different than other metrics. After running all of these centrality measures on the 108 networks and comparing their results, it appears that, for the Hacker News data set, simple metrics such as degree are satisfactory to determine the most central cohorts in the network. Additionally, the CPA (comments per active user) appears to be very effective at measuring network centrality, and has a correlation value >.98 with other centrality measures (with the exception, again, of the closeness metric).

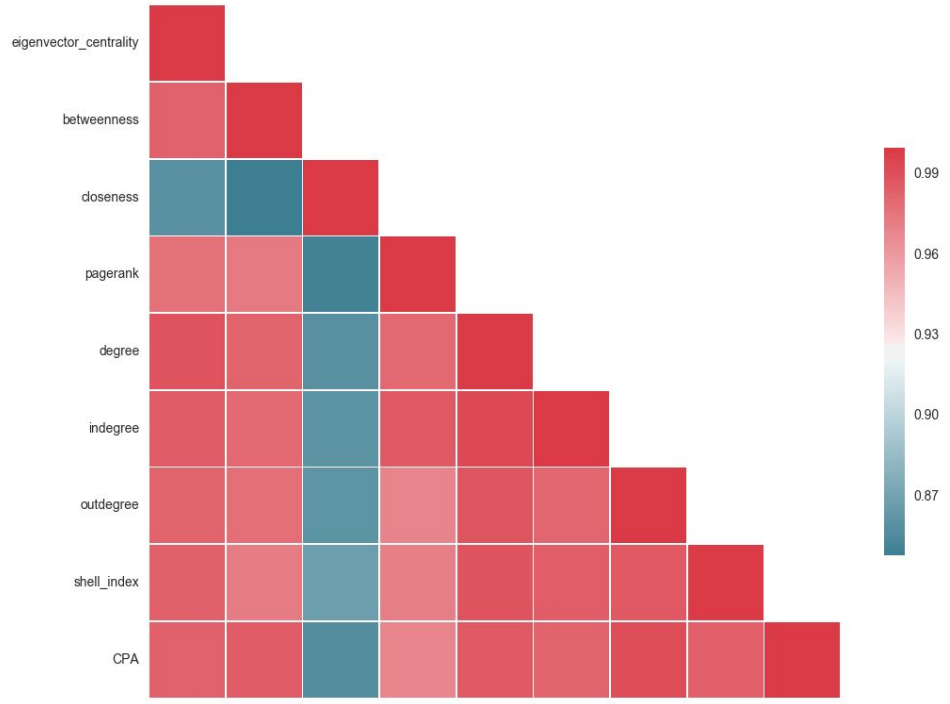

Figure 8. The correlations between each centrality measure calculated at each of the 108 Hacker News time slices. All metrics aside from closeness show a very high degree (>.97) of correlation. Closeness shows a correlation of ~.85 with other centrality metrics.

StackExchange

StackExchange is a series of question and answer sites, each relating to a specific topic. In particular, we look at Math StackExchange, which is a forum for anyone studying math to ask and answer questions in mathematics. Users come from all levels of math, and are primarily students. Content ranges from introductory homework problems to open ended theoretical discussion.

Nodes are defined to be users, while edges exist between two users if they have commented on the same Question or Answer post in Math StackExchange. The edges are weighted by the number of times two users have commented in the same post. We split the results into five cohorts based on the year of their first activity.

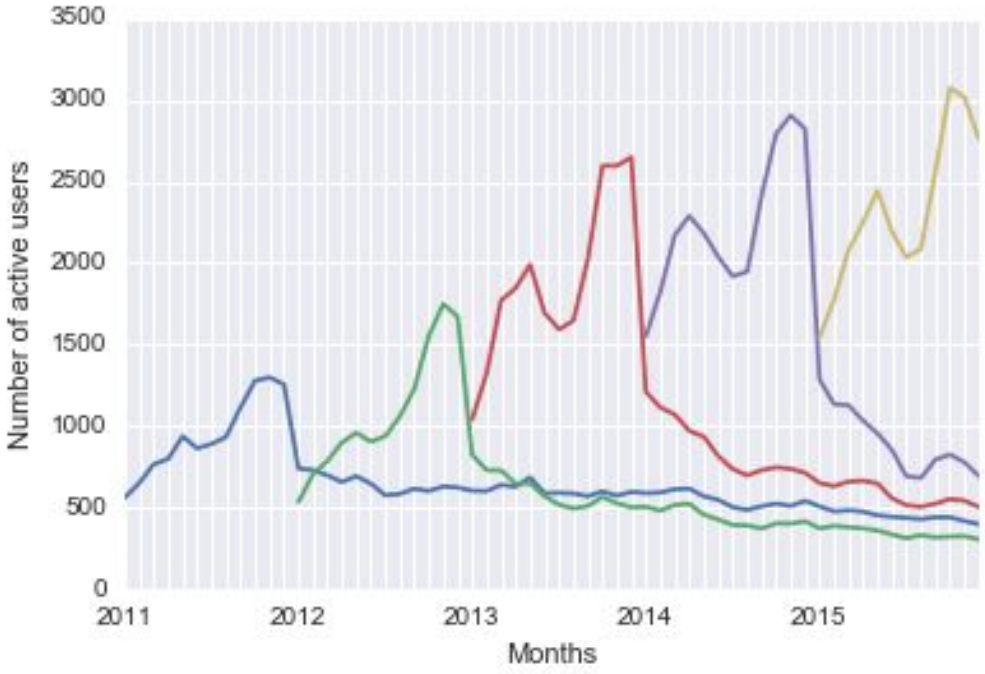

Figure 9. Number of active StackOverflow users in each cohort.

Analyzing the number of users joining Math StackExchange, we see that the community is growing. Each year, the influx of new users is larger than the previous year, so the size of each cohort surpasses the last. Users are considered to be active if they have posted in a given month. Since cohorts cannot grow after the year of their establishment, we see a decrease in users per cohort over time. Eventually, user the number of active users dies down to a similar level for all cohorts. The graph suggests that older users may dominate activity levels once the initial influx has died down.

Additionally, newer users do not stay active for a long time. In fact, most leave within one year of their first activity. The proportion of users leaving newer cohorts within a year is much higher than within older cohorts. This substantiates that older users are more loyal to the system.

Interestingly, the shape of the graph is significantly reflective of the website's audience. The majority of Math StackExchange users are students who join the site during the school year to help with classes. As such, we see a huge wave of new users joining in the fall, followed by another wave of students joining after winter.

Figure 10. Average number of comments per active StackOverflow user.

Next, we analyze the average number of comments per active user over time. In general, excluding spikes in new user activity, the oldest cohort dominates the average number of comments; despite being severely outnumbered by new users, older users post more frequently. In parallel with the number of incoming new users, we also see a spike in activity during a cohort's first year. This spike decreases over the year, just as the number of active users decreases during the first year.

Overall, the average number of comments per user is decreasing over time. However, the most recent cohort (2015) challenges this decrease with an overwhelming increase in user comments. This likely reflects a growing popularity of Math StackExchange.

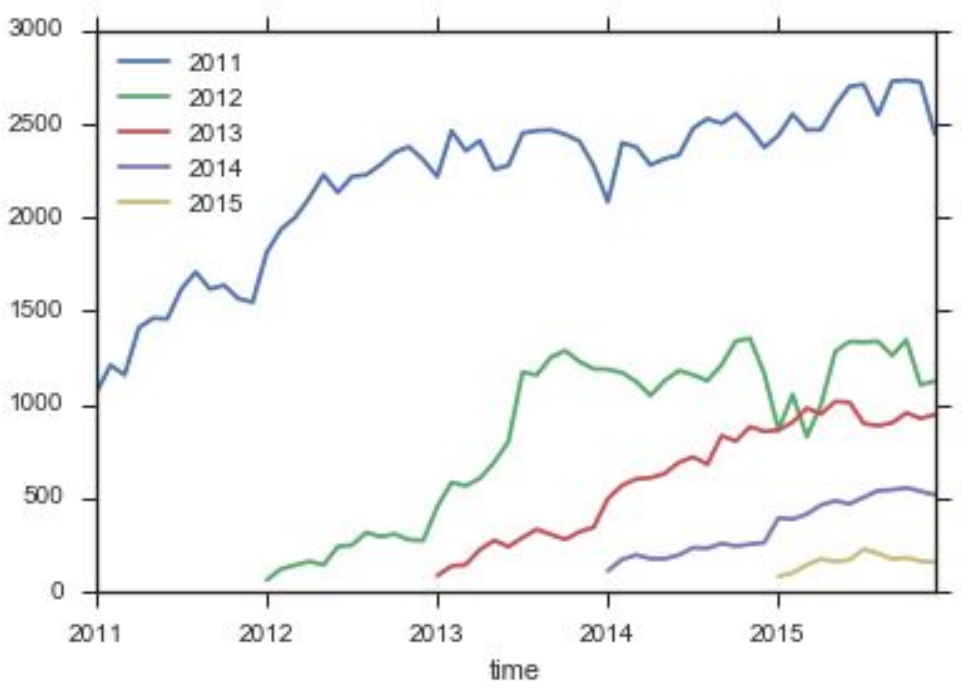

Figure 11. Average degree of StackOverflow users in each cohort.

Finally, we study user centrality among cohorts over time. Studying the average weighted degree of users within a given month, we find that older users have higher degree centrality than newers. This relationship is easily maintained, even when older users are vastly outnumbered. Even while cohort size decreases over time, average weighted degree within a cohort rises steadily. From this we conclude that the longer a user is active, the more likely it is that they will be central to the community. Therefore, to ensure community strength, moderators must incentivize older users to stay active.

Conclusions

Among all of the above communities, users from older cohorts tend to remain more active than users from newer cohorts. Newer users drop out of the community at a higher rate than older users, illustrating that older users are more loyal to the system. Additionally, despite being greatly outnumbered by newer users, the older users post more often and more consistently. Consequently, older users have higher centrality compared to newer users. Comparing the results of multiple centrality measures (betweenness, degree, closeness, eigenvalue), we find that the difference between the measures upon an online network is fairly insignificant-- if a cohort outranks all others by one centrality measure, that cohort will also tend to outrank others by a different centrality measure. Overarchingly, we find that older users are more valuable to the system in terms of centrality, posting activity, and system loyalty. To maintain a healthy online community, moderators must incentivise user retention.

An Eternal September phenomenon was witnessed in the Reddit dataset. Older users slowed their activity in 2010-2012, which corresponded with a rise in those users' activity in an alternative community. Future work will investigate the same split from a network theoretic perspective. Other subreddits should also be considered: /r/TrueReddit and /r/TrueTrueReddit (a sequence user communities disenchanted with the quality of content on the main pages).

Bibliography

[1] Fire, Michael, and Carlos Guestrin. "Analyzing Complex Network User Arrival Patterns and Their Effect on Network Topologies." *arXiv preprint arXiv:1603.07445* (2016).

[2] Tang, John, et al. "Temporal distance metrics for social network analysis."*Proceedings of the 2nd ACM workshop on Online social networks*. ACM, 2009.

[3] Santoro, Nicola, et al. "Time-varying graphs and social network analysis: Temporal indicators and metrics." *arXiv preprint arXiv:1102.0629* (2011).

[4] Casteigts, Arnaud, et al. "Time-varying graphs and dynamic networks."*International Journal of Parallel, Emergent and Distributed Systems* 27.5 (2012): 387-408.

[5] Zhang, Jun, Mark S. Ackerman, and Lada Adamic. "Expertise networks in online communities: structure and algorithms." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[6] Barbosa, Samuel, et al. "Averaging Gone Wrong: Using Time-Aware Analyses to Better Understand Behavior." Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2016.

[7] Valente, Thomas W., et al. "How correlated are network centrality measures?." *Connections (Toronto, Ont.)* 28.1 (2008): 16.

[8] Friedl, Dipl-Math Bettina, and Julia Heidemann. "A critical review of centrality measures in social networks." *Business & Information Systems Engineering*2.6 (2010): 371-385.

[9] Opsahl, Tore, Filip Agneessens, and John Skvoretz. "Node centrality in weighted networks: Generalizing degree and shortest paths." *Social Networks* 32.3 (2010): 245-251.

[10] Yan, Erjia, and Ying Ding. "Applying centrality measures to impact analysis: A coauthorship network analysis." *Journal of the American Society for Information Science and Technology* 60.10 (2009): 2107-2118.