Better When It Was Smaller?
Community Content and Behavior After Massive Growth

Zhiyuan Lin
Stanford University
zylin@cs.stanford.edu

Niloufar Salehi
Stanford University
niloufar@cs.stanford.edu

Bowen Yao
Stanford University
boweny@stanford.edu

Yiqi Chen
Stanford University
yiqic@cs.stanford.edu

Michael S. Bernstein
Stanford University
msb@cs.stanford.edu

Abstract

Online communities have a love-hate relationship with membership growth: new members bring fresh perspectives, but old-timers worry that growth interrupts the community's social dynamic and lowers content quality. To arbitrate these two theories, we analyze over 45 million comments from 10 Reddit subcommunities following an exogenous shock when each subcommunity was added to the default set for all Reddit users. Capitalizing on these natural experiments, we test for changes to the content vote patterns, linguistic patterns, and community network patterns before and after being defaulted. Results support a narrative that the communities remain high-quality and similar to their previous selves even post-growth. There is a temporary dip in upvote scores right after the communities were defaulted, but the communities quickly recover to pre-default or even higher levels. Likewise, complaints about low-quality posts do not rise in frequency after getting defaulted. Strong moderation also helps keep upvotes common and complaint levels low. Communities' language use does not become more like the rest of Reddit after getting defaulted. However, growth does have some impact on attention: community members cluster their activity around a smaller proportion of posts after the community is defaulted.

Introduction

Usenet members would bemoan Septembers: the start of the academic year meant that a new set of university freshmen were getting access to Usenet for the first time. These new members were unaware of Usenet community standards and would regularly annoy the old-timers. However, over time, the September newcomers would acculturate or leave, and the communities would return to normal. Then, in September 1993, the popular online service America Online gave all of its members access to Usenet. This massive flood of new users permanently overwhelmed the Usenet community, earning the event the title of "Eternal September": the September that never ended.

The Eternal September raises the question: do online communities lose their core character when they are faced with sudden popularity? It is not uncommon to see comments on sites such as Reddit, Slashdot, or HackerNews asserting that the communities were better when they were smaller. And the "MTV Effect" threatens as well: the more that a community grows, the more it must broaden its appeal and become mainstream if it is going to continue growing, alienating the old-timers who made it what it is.

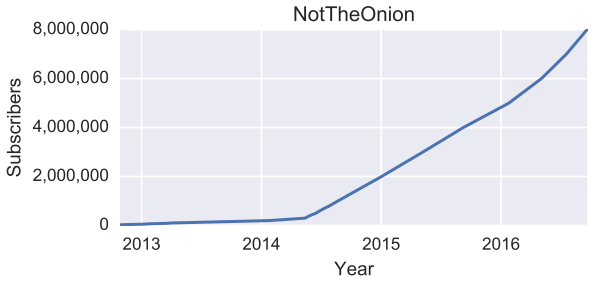


Figure 1: Subscriber numbers of the subreddit *NotTheOnion*, which was defaulted in May 2014. All subreddits in our dataset share similar subscriber growth patterns.

Research does reinforce a theory that a large influx of new users interrupts a community's success (Jones, Ravid, and Rafaeli 2004; Butler 2001; Kraut et al. 2012). However, some communities manage to survive sudden membership growth (Kiene, Monroy-Hernández, and Hill 2016). In this paper, we seek to arbitrate between these two theories by applying a quantitative lens to several communities that underwent massive exogenous growth.

We focus our study on Reddit, one of the most popular online communities. Reddit consists of a number of subcommunities, or subreddits, which are specifically themed. Example subreddits include *EarthPorn* for pictures of nature, *ShowerThoughts* for pithy reflections on life, and *Books* for reading enthusiasts. A small handful of subreddits have been *defaulted* by the central Reddit admins, which means that they were added to the default set of subreddits that all new users are automatically subscribed to and see when they are not logged in. The decision to be defaulted is exogenous to the subreddit, typically made by Reddit administrators. Upon being defaulted, subreddits begin gaining many new subscribers every day (Figure 1). For instance, created in October 2008, subreddit *NotTheOnion* had 239,805 subscribers in May 2014, right before being defaulted. Its subscribers

grew by more than an order of magnitude to over 3 million subscribers a year later.

In this paper, we capitalize on the natural experiment induced by the exogenous shock of defaulting creating a large influx of new users. Specifically, we test for changes in user reception, content, user interaction pattern and related underlying network structure of the community.

Firstly, studying user reception via upvote patterns asks whether it really was "better when it was smaller". However, subjective user reception to community content is only part of the picture. Secondly, we ask whether the subreddit's topic and language use shift as a result of new users' unfamiliarity with the community or broader set of interests. Thirdly, underprovision caused by popular voting when facing information overload (Gilbert 2013), may lead users to interact with a smaller portion of community content and therefore cause the community structure to become increasingly clustered. In particular, we would like to answer the following research questions:

RQ1 **User Reception:** How does users' subjective reception to the content change? Do people complain more about the content, or downvote it more?

RQ2 **Content:** Does the content become more generic? Does the community retain its topic focus, or does its language and content start drifting more toward the rest of Reddit?

RQ3 **Interaction Patterns:** How do users' commenting patterns and interactions change? Do users comment on a wide variety of posts or do they cluster under a smaller set of content?

We take a quantitative approach to study these questions. We analyzed over 45 million comments from 10 subreddits that were defaulted in July 2013 and May 2014. In the remainder of the paper, we provide answers to these questions based on our data analysis and discuss the implications.

Related Work

There is a vast research literature on how online communities grow and how they can attract new users. Additionally, researchers have studied how newcomers behave differently from established users and what effect newcomers can have on a community.

Why Communities Grow

Researchers have extensively studied success of online communities and how to attract new users (Kraut et al. 2012). For example, Zhu et al. have looked into online community success in relation to other communities (Zhu et al. 2014) and membership overlap (Zhu, Kraut, and Kittur 2014). Prior work has also studied how online communities grow organically and the factors that cause new people to join. For instance, a community's network features predispose its future growth: by utilizing previous growth rate and the structural features of a graph, researchers were able to predict an online community's final size and longevity (Kairam, Wang, and Leskovec 2012). Additionally, structural properties of the network were among the top features affecting whether a user will join a community and whether the community will subsequently grow (Backstrom et al. 2006). Other work has shown that information exchange, friendship, and social support are also key reasons why people join online communities (Ridings and Gefen 2004).

In studying organic growth, however, it is impossible to separate the effect that the influx of newcomers has on the community from the factors that caused the growth in the first place. For instance, if a community grows and its content becomes less specific at the same time, the causal direction is not immediately obvious. Instead, we rely on a natural experiment that caused sudden, large growth in communities to study the effect of this growth on the communities and the content that they share.

Newcomers

Newcomers to a community behave differently from established users. Newcomers are generally more energetic and are interested in a broader range of discussions (Jeffries et al. 2005). On the other hand, newcomers are also less motivated to help (Kraut et al. 2012) or display characteristics of good organizational citizenship (Organ and Ryan 1995).

Newcomers do effect the communities that they join. They may weaken a community's wellness by increasing information overload (Jones, Ravid, and Rafaeli 2004; Butler 2001), breaching community norms (Kraut et al. 2012), and lowering content quality (Gorbatai 2011). This prior work studied the effect of newcomers gradually arriving to an established community. But massive growth may have effects beyond those described here. For instance, massive growth may fundamentally change the nature of the community by overwhelming prior established norms.

Relatively little research has been done on massive growth in online communities. A qualitative study of the defaulting of the subreddit *NoSleep* identified three factors that helped the subreddit survive the influx of members: consistent enforcement by leaders, moderation by community members, and technological systems maintaining norms (e.g., a voting system and an automatic moderation tool) (Kiene, Monroy-Hernández, and Hill 2016). In this paper we take a longitudinal, quantitative approach to the same question of massive growth. We show that surviving the "Eternal September" is not a special case of *NoSleep*, but rather a common outcome of large new user influxes into an online community.

Community Evolution over Time

There is an abundant body of research on online communities' evolution and changes over time, which helps guide our research direction and methodology.

One of the key questions in community evolution is how the roles and contributions of a single user changes over time. Velasquez et al. introduced the concept of latent users: users who have learned how to participate in a given community, but no longer actively contribute content (Velasquez et al. 2014). Over time, some long-term users are converted into latent users whose community participation is much more selective. We extend this work by taking a more global view and studying the changes in the community as a whole.

Subreddit Name	Default Date	Subscribers
ExplainLikeImFive	2013-07-17	10,406,778
EarthPorn	2013-07-17	9,983,791
Books	2013-07-17	9,652,881
Television	2013-07-17	9,422,953
ShowerThoughts	2014-05-07	8,567,742
TIFU	2014-05-07	8,369,247
NotTheOnion	2014-05-07	8,203,342
DataIsBeautiful	2014-05-07	8,177,404
OldSchoolCool	2014-05-07	7,961,808
NoSleep	2014-05-07	7,833,173

Table 1: Subreddits included in our dataset. Subscriber counts are as of October 2016.

Another area of research studies the linguistic changes in online communities over time. By analyzing data from two beer reviewing communities, researchers proposed a framework to track linguistic change in discussions over time (Danescu-Niculescu-Mizil et al. 2013). They discovered that new users adopted the community's specific language as they spent more time in the community. Language changed at the community level too, becoming more predictable over time. To quantify linguistic change, the authors used several post-level measures including cross-entropy of posts and Jaccard self-similarity between adjacent posts. Prior work has also employed a language model, specifically Latent Dirichlet Allocation (LDA), to track linguistic change over time in order to quantify and evaluate mental illness severity (Chancellor et al. 2016). In this paper, we also use the LDA model to compare topic distribution between posts in sub-communities and general Reddit post samples to track linguistic change in sub-communities over time.

Another research area studies the evolution of network structures over time. For instance, the exponent α in the network's power law degree distribution may decrease over time as a result of graph densification (Leskovec, Kleinberg, and Faloutsos 2007). Similarly we use α to examine the interaction network's structural change after a massive growth.

Data

In this paper, we use the Reddit comment data available on Google BigQuery[1]. We selected ten popular subreddits defaulted in 2013 and 2014. Table 1 describes the name, default date, and number of subscribers for these 10 subreddits. We gathered comment data from roughly six years spanning January 2011 to September 2016, centered around the subreddits' default dates of 2013 and 2014. This resulted in 45,681,234 comments from ten subreddits.

In addition to studying community changes after being defaulted, we are interested in how community growth affects interactions between community members. To study these patterns, we use the concept of Monthly Active Users.

[1] https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments

Subreddit Name	Annual Growth Rate (%)
DataIsBeautiful	202.964
ShowerThoughts	136.733
NotTheOnion	124.942
OldSchoolCool	79.434
Television	76.894
TIFU	60.843
ExplainLikeImFive	41.674
NoSleep	7.881
EarthPorn	7.646
Books	-8.524

Table 2: Annual MAU growth rate after default, defined as the growth achieved in the year following the default. There is a clear stratification between *Surge* and *Jump* communities, with Surge communities growing faster in this time.

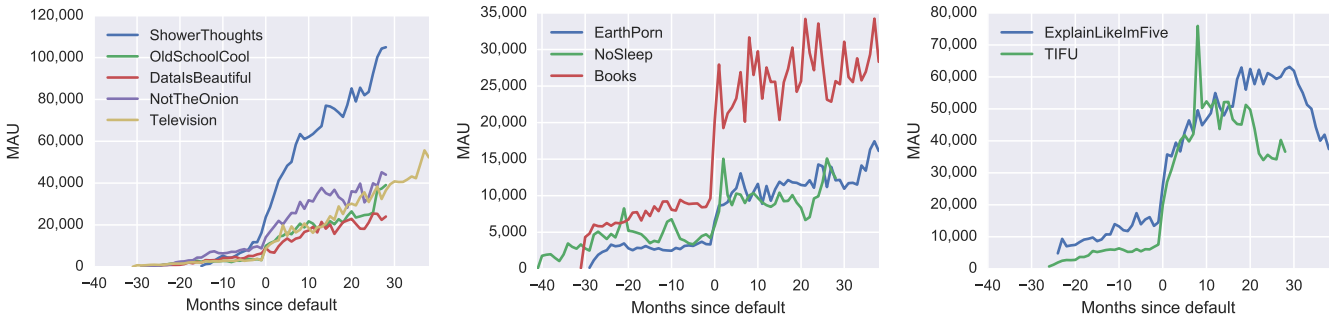
Monthly Active User (MAU)

Definition Upon being defaulted, the number of subscribers skyrockets as expected (Figure 1). However, we are more interested in active users: those who participate in the community rather than simply being automatically subscribed by Reddit. We instead use Monthly Active User (MAU) counts to measure the community's size. The MAU for a subreddit in a given month is the number of users who posted at least one comment in the subreddit that month.

***Surge* and *Jump*: two MAU growth patterns** While all ten subreddits' subscriber numbers permanently accelerate after being defaulted (Figure 1), not all subreddits' MAU growth follow the same pattern. We see two different types of MAU growth patterns: a *Surge* pattern (Figure 2a) and a *Jump* pattern (Figure 2b).

Upon being defaulted, both patterns enjoy enormous MAU growth. They differ, however, in their later growth pattern: Surge communities keep growing consistently over at least a year, while Jump communities stop growing almost immediately and then grow far more slowly or even begin shrinking. In other words, while Jump communities may be growing in terms of subscribers, a relatively static number of people actually contribute. Numerically, as shown in Table 2, Jump type communities have smaller annual growth rates (sub-10%) after being defaulted. A t-test on the two groups of annual growth rate also verifies this observation with $p < 0.05$.

Figure 2c shows two communities (*ExplainLikeImFive* and *TIFU*) whose MAUs initially display a Surge pattern but eventually peak and even started to decline. As they display a similar growth pattern as Surge and annual growth rate closer to Surge group, we categorized such growth pattern as Surge. However, there remains an open question of what separates them from the other Surge subreddits. Later, we describe a simple model that might explain those two growth patterns and why some Surge communities (e.g. *ExplainLikeImFive* and *TIFU*) might exhibit slowed growth.



(a) *Surge* type growth, where the MAU keeps increasing after the default.

(b) *Jump* type growth, where the MAU increases but then flattens out again.

(c) *Surge* type growth that later stopped growing.

Figure 2: Two monthly active user number growth type upon default: *Surge* and *Jump*. The x-axis is the number of months since the date of default, which is represented by $x = 0$.

Analysis

In this section, we report the communities' feedback patterns changes, linguistic changes, and interaction pattern changes as a result of being defaulted.

Communities Display Limited Negative Reaction to Massive Growth

Prior work offers conflicting accounts of whether newcomers have a negative effect on communities (Jones, Ravid, and Rafaeli 2004; Butler 2001; Kraut et al. 2012) or not (Kiene, Monroy-Hernández, and Hill 2016). The weight of evidence prior to this work suggests a negative effect, so we adopt *Hypothesis 1: Subreddits will have a lower percentage of upvotes, and a higher percentage of complaint comments, after being defaulted.*

However, in the analysis to come, we will report data that does not support Hypothesis 1: instead, our findings suggest that communities react to massive growth with only short-term dip in upvote scores, but no significant impact on long-term scores or complaint rates.

Method As suggested by previous work (Gorbatai 2011), new comers may introduce lower-quality content to online community. How does the community react to the content that these newcomers produce? If people begin becoming more cynical about the content of the subreddit, we might expect to see them downvoting more than before, or complaining about the subreddit more often. Hence, we use *average score* and *complaint comment percentage* to measure user reception to community content.

Average score: The voting mechanism in Reddit is a direct reflection of the community's reaction to a piece of content. The score of a comment or a post is the sum of upvotes (+1s) and downvotes (-1s) on it. Higher scores indicate that the post or comment was well received by the community. We calculate the average score of the month as the arithmetic mean of the final scores for all comments within that month.

Complaint comment percentage: In order to measure how much users complain about content of the subreddit, we used Empath (Fast, Chen, and Bernstein 2016) trained on a Reddit corpus to generate a set of complaining phrases[2] using the seed words *repost* and *shitty post*. Example phrases in the set generated by Empath include *shitpost*, *stupid post*, *downvoted* and *troll post*. Complaint comment percentage of a subreddit in a given month is calculated as the percentage of comments that contain at least one of these complaining phrases. A higher complaint comment percentage suggests that the community is reacting poorly to the content that is posted there.

We leverage a regression analysis to study the effect of defaulting on these two measures. We run a least square linear regressions on average score (real numbers) and complaint comment percentage (real numbers between 0 and 1). Since average score and complaint comment percentage are calculated monthly, we have one observation per month for each subreddit.

To focus on the immediate effect of defaulting on average score and complaint comment percentage, we limit the time range to 30 months consisting of 3 phases: before, during, and after, with each spanning 12, 6, and 12 months respectively[3]. We create a categorical *phase* variable in our regression analysis, with $phase_{before}$ used as the reference level.

Inspired by Kiene et al. (Kiene, Monroy-Hernández, and Hill 2016), we expect to see positive effect of moderation volume on community reception. As indicated in Kiene et al.'s work and other prior work, moderators primarily practice rule enforcement and content sanctions via comment deletion. So, like in prior work (Chancellor, Lin, and De Choudhury 2016; Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015), we measure moderation levels by looking at the proportion of comments that were deleted in that month. Different subreddits applied different moderation strategies in reaction to being defaulted — most applied a light hand, but some of the subreddits moderated much more heavily. We thus introduce a binary vari-

[2]Generic keywords such as *Reddit* and *subreddit* are manually removed.

[3]The results on average score, complaint comment percentage, and topic specificity we report are robust to this exact choice of time window

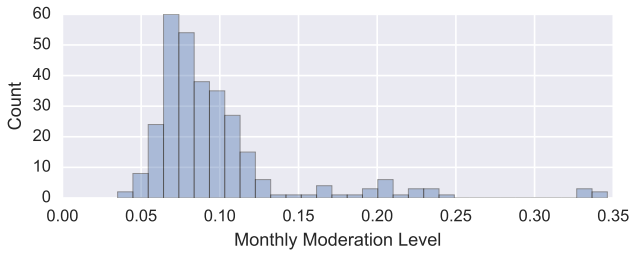


Figure 3: Distribution of monthly moderation levels of all subreddits in the dataset. The 3^{rd} quartile ($\approx 10.4\%$) is set as the $high\_moderation$ threshold.

Variable	β	SE	t
Linear Regression on Average Score			
$phase_{during}$	-0.3457^{*}	0.149	-2.313
$phase_{after}$	0.3045^{*}	0.120	2.542
$high\_mod$	0.6103^{***}	0.154	3.964
intercept	4.8795^{***}	0.177	27.614
Linear Regression on Complaint Comment %			
$phase_{during}$	-0.0011^{***}	< 0.000	-3.809
$phase_{after}$	-0.0010^{***}	< 0.000	-4.328
$high\_mod$	-0.0004	< 0.000	-1.342
intercept	0.0121^{***}	< 0.000	36.199

Table 3: Results of linear regression on average score and complaint comment percentage. ***: $p < 0.001$, *: $p < 0.05$. $N = 300$, Adj. R^2 are 0.898 and 0.541 respectively.

able $high\_moderation$ indicating whether the percentage of moderated posts that month is above the 3^{rd} quartile (moderating $\approx$ 10.4% of posts) (Figure 3).

Additionally, we control for individual differences in subreddits by including a categorical subreddit variable in our regression.[4]

Results Table 3 shows the results of both regressions: average score and complaint comment percentage.

In the regression on average score, $phase_{during}$ has a statistically significant negative impact on average score compared to the baseline ($phase_{before}$), while $phase_{after}$ tends to have a higher average score than the baseline. The result implies that the average score dropped immediately after defaulting, but recovered later to an even higher level during $phase_{after}$ (Figure 4). Moderation also has a positive effect on keeping the score high, consistent with prior work (Kiene, Monroy-Hernández, and Hill 2016).

In the regression on complaint comment percentage, both $phase_{during}$ and $phase_{after}$ have significant negative coefficients, which suggests during and after defaulting, communities are experiencing fewer complaints.

One possible explanation for the discrepancy between the temporary dip in average score and the lowered complaint comment percentage is that after being defaulted, there is a relatively higher percentage of new users. As it is expected to have more latent users among old-timers, their dissatisfaction with content and complaints might be overshadowed by new-comers, but reflected by the popular voting mechanism, which requires much less effort to participate in.

Together, these regression results indicate that defaulting does negatively impact the average score in the community, but only temporarily. In the long run, as reflected by the increased scores and fewer complaints, user reception to community content improves despite the large influx of new users.

Communities Keep Their Linguistic Identities

The second research question we would like to answer is whether what the community talks about becomes more generic, or more similar to the rest of Reddit, after the newcomers join, as a result of newcomers' unfamiliarity with the community and generally broader interests. Prior work suggests that, at an individual level, newcomers begin linguistically distinct from existing community members (Danescu-Niculescu-Mizil et al. 2013). So, we adopt *Hypothesis 2: subreddits' linguistic character will shift more toward the average subreddit following a default.*

However, our results will suggest that Reddit communities do not become more generic linguistically.

Method To understand linguistic topic drift, we employ the online version (Hoffman, Bach, and Blei 2010) of the Latent Dirichlet Allocation (LDA) model (Blei, Ng, and Jordan 2003) to create a topic distribution of the content on Reddit. LDA has previously been used to study social site data (Paul and Dredze 2011; Chancellor et al. 2016). To train the LDA model, we randomly sample 1,000 comments across all subreddits each month from January 2011 to September 2016, resulting in $1,000 \times 69 = 69,000$ general Reddit comment samples as our LDA training corpus.

The trained LDA model enables us to evaluate the topic distribution of each subreddit in a given month. Intuitively, if the subreddit is shifting away from its former identity, this would arise linguistically as the topic distribution of the subreddit changing relative to the topic distribution of the rest of Reddit that month. Following prior work (Hong and Davison 2010; Quan et al. 2015), we create a single mega-comment to analyze for each month by concatenating 1000 randomly sampled comments from that month. We also evaluate the monthly topic distribution of the general Reddit comment sample for comparison using the same mega-comment technique.

We then measure *topic specificity* of a subreddit in a given month by calculating the cosine distance between its topic distribution vector and the general Reddit topic distribution vector of that month. The higher topic specificity is (i.e., the further the cosine distance is between two topic vectors), the less alike those two vectors are, and therefore the more "specific" the subreddit's content is.

As before, we run a linear regression on topic specificity over 30 months around default dates with variables for *phase* and *high_moderation*, as well as controls for subreddits.

[4]This control for subreddit factor variable is omitted in the regression result tables.

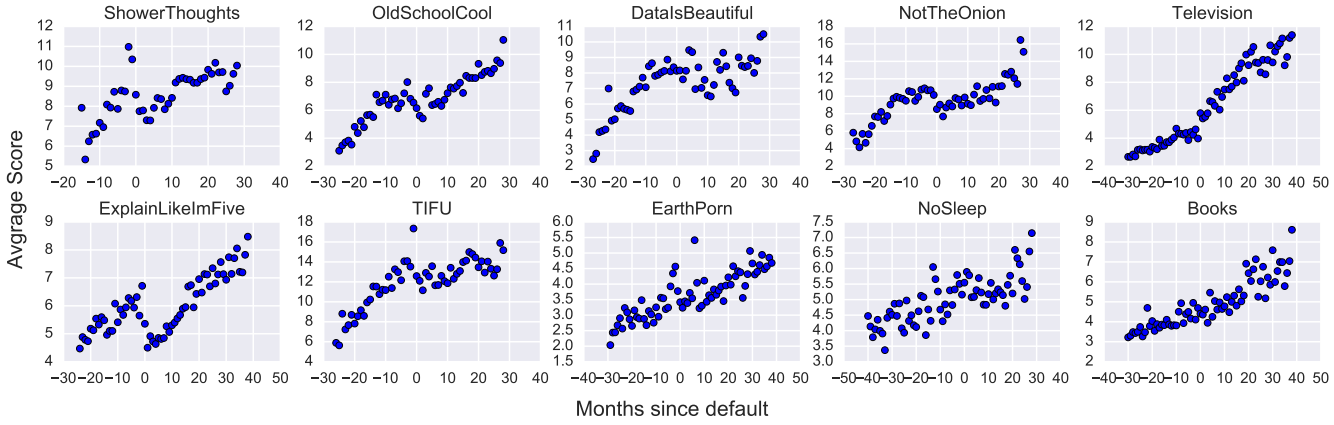
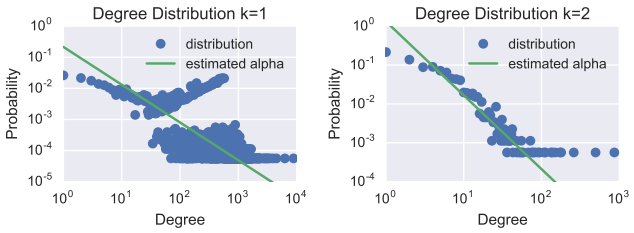


Figure 4: Average score change over time. The x-axis is the number of months since defaulting, with $x = 0$ as the date of defaulting. There is a overall increasing trend, with a temporary dip around defaulting.

Variable	β	SE	t
$phase_{during}$	-0.0028	0.013	-0.214
$phase_{after}$	0.0049	0.011	0.461
$high\_mod$	0.0078	0.014	0.0576
intercept	0.2591***	0.016	16.672

Table 4: Result of linear regression on topic specificity. ***: $p < 0.001$, otherwise $p \geq 0.05$. $N = 300$, Adj. R^2 is 0.768.



(a) (b)

Figure 5: Degree distribution of subreddit *OldSchoolCool* in December 2014. Where k is the minimum edge weight threshold for an edge to be counted, $k = 1$ is on the left and $k = 2$ is on the right.

Results Results are presented in Table 4. There was no significant topic specificity change caused by either defaulting nor moderation. The communities do not tend to start sounding more like the rest of Reddit after they are defaulted — they maintain their (linguistic) identity.

Users Cluster Around a Smaller Proportion of Community Content

There is a widespread of underprovision of attention in Reddit communities (Gilbert 2013). Presumably this underprovision of attention is worsened by having more posts competing for limited top spots. This logic produces *Hypothesis 3: community interaction will cluster into a smaller proportion of posts after defaulting.*

Our findings support Hypothesis 3, indicating that such attention underprovision becomes even severer after massive growth.

Method With limited bandwidth, a user is only able to view and comment on a relatively small proportion of posts in a subreddit. As the community grows, it is reasonable to expect the situation to get even worse: a higher percentage of users will comment under a smaller and smaller proportion of posts as a result of information overload and over-relying on others' voting on posts. To study such behavior change at community scale, we construct a co-comment graph that captures the shift in how densely users cluster around posts.

We would like to construct a network that represents interactions between users. Unlike social networks like Facebook, Reddit users generally focus more on the content rather than other users. Thus, instead of using edges to represent friendship, we construct a graph representing co-commenting behaviors among users: we connect users who comment on the same post.

For each monthly snapshot and each subreddit, we construct a co-comment graph by creating nodes for each user that has commented at least once in that month in the subreddit. Edges connect any two users who have commented on at least k common posts. For example, if $k = 1$, the edges connect pairs of users who commented on the same post at least once; if $k = 3$, the edges connect pairs of users who both commented on the same three or more posts. This construction helps us capture correlations between user behavior. The parameter k represents the strength of the connection.

Figure 5 demonstrates degree distribution of subreddit *OldSchoolCool* in December 2014 with $k = 1$ and $k = 2$. When $k = 1$, many users have the same high degrees (Figure 5a), which does not follow the expected degree power law distribution (Faloutsos, Faloutsos, and Faloutsos 1999). It is very likely that pairs of users will both only comment on the same popular post, and in this case they will have the same high degree, since they will form edges with every

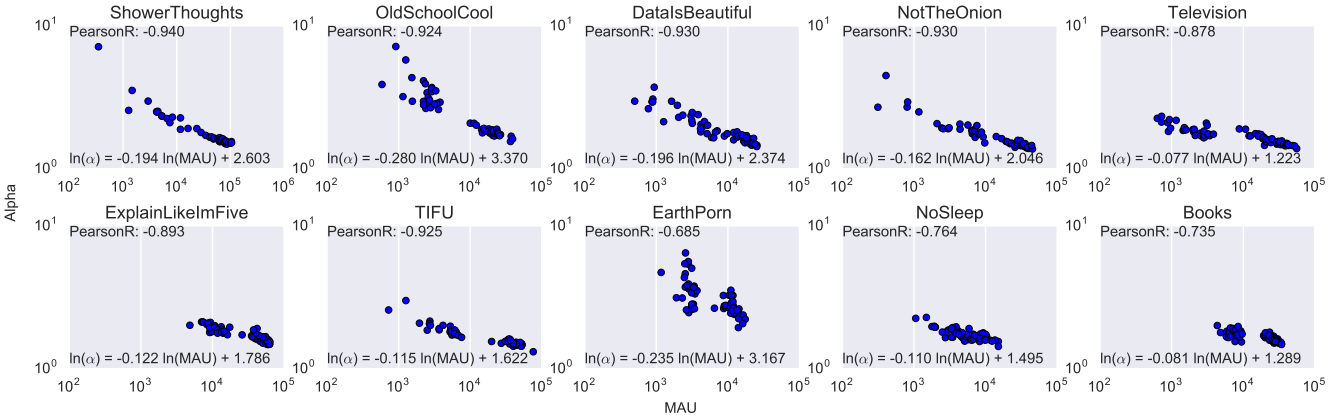


Figure 6: Relationship between α and Monthly Active Users (MAU) on a log-log plot. At bottom, the least square fitted line is described at bottom left of each plot with $0.077 \leq c \leq 0.280$ and $3.397 \leq a \leq 29.079$. $log(\alpha)$ and $log(MAU)$ have negative correlations: in all 10 datasets, Pearson's r (top left) ≤ -0.685 and over half of them ≤ -0.875.

other user who has commented on that post. Therefore, as Figure 5a shows, the graph construction with $k = 1$ is too noisy since it consists of many weak connections between users. On the other hand, choosing $k = 2$ will only account for stronger co-comment relationship between users while keeping many edges. Figure 5b shows the degree distribution of graph with $k = 2$ fits a power law. So, we use co-comment graphs with $k = 2$ in all analyses in this paper.

By analyzing the degree exponent α's change over time in the co-comment graph with $k = 2$, we can gain insight into how users distribute across posts. If users tend to cluster under a small proportion of posts, they will all have high degree in the co-comment graph. The result will be a more heavy-tailed or flatter degree distribution, which translates into a smaller exponent α in the power-law distribution. On the other hand, if users tend to spread out and comment on different posts instead of clustering around a few popular ones, there would be much more weight on the lower end of degree distribution in the graph and a larger α. In the extreme case where we have N pairs of users who have co-commented under k posts each, and each post in this case only has 2 comments, users are very sparsely distributed across many posts and the degree distribution would see $2N$ nodes all with degree k, leading to an infinitely large α.

Since our hypothesis about user commenting behavioral change is that it is *community growth* that worsens the extent of underprovision, it would be more appropriate to study how α change as community size changes instead of directly conducting regression analysis on α like what we do in previous subsections. Here, we use Monthly Active User, or MAU, as a measure of community size.

We track α as a function of Monthly Active Users (MAU). In order to estimate α from observed discrete data points, we make use of a Maximum Likelihood Estimator, or MLE (Clauset, Shalizi, and Newman 2009). A example fitted line with MLE-estimated α is drawn in Figure 5.

When MAU is large, α decreases much more slowly with MAU: in other words, there is a diminishing loss for α as a function of MAU. We model this relationship via the power law $\alpha = aMAU^{-c}$ where a and c are positive constants. If the model holds, and we take $\log$ on both sides, we will have $\log(\alpha) = -c\log(MAU) + \log(a)$. This equation suggests that if we plot α versus MAU on a log-log scale, the points should form a straight line with negative slope. Figure 6 shows visually how α and MAU are negatively linearly correlated in a log-log scale plot. Empirically, in the 10 subreddits, least square linear regressions result in $0.077 \leq c \leq 0.280$ and $3.397 \leq a \leq 29.079$. All $\log(\alpha)$ and $\log(MAU)$ have Pearson's r ≤ -0.685 in our datasets and over half of them have Pearson's r ≤ -0.875.

Variable	β	SE	t
$log(MAU)$	-0.1557^{***}	0.005	-34.024
intercept	1.8908^{***}	0.042	45.095

Table 5: Result of linear regression on $\log(\alpha)$. ***: $p < 0.001$. $N = 599$, Adj. R^2 0.821.

To further quantitatively verify this results, we run a linear regression on $\log(\alpha)$ with $\log(MAU)$ as independent variable. Similar to the regression analysis in previous subsections, we use controls for subreddits as independent variables. The result (Table 5) indicates a significant linear relationship between $\log(\alpha)$ and $\log(MAU)$ with a negative coefficient, which shows α is a power function of MAU.

Then what is the effect of defaulting on α, and subsequently on the community structure? Since MAU changes significantly upon defaulting (Figure 2), defaulting lowers α substantially by drastically boosting a community's MAU. According to the power law relationship suggested by the linear regression on $log(\alpha)$, α is more susceptible to MAU change when MAU is small.

Recall that a smaller α implies more users cluster around a smaller proportion of posts. This finding supports Hypothesis 3: a larger portion of users cluster around a smaller set of posts after defaulting, when the community size increases

drastically. Furthermore, because α decreases much slower when MAU is large, it suggests that the community structure is most susceptible to large influxes of new users when MAU is small, which is consistent with intuition.

Discussion

In this paper we have taken a quantitative approach to show that communities remain high-quality and similar to their previous selves even after a massive growth. While we did find a temporary dip in upvote scores right after the communities were defaulted, the communities quickly recovered to pre-default or even higher levels of upvotes. Additionally, complaints about low-quality posts did not rise in frequency after the subreddit was defaulted. These results show that overall user perception remained positive. We also found that high levels of moderation helped maintain the positive perception of community content after getting defaulted and that the communities' language did not become more generic or more similar to the rest of Reddit after the massive growth. Finally, we found that growth did have an impact on users' attention: community members reacted by clustering their activity around a smaller proportion of posts. This shift in the interaction pattern is most notable when the community undergoing the defaulting is small.

Interpreting the Two Types of MAU Growth

Defaulting boosts MAU (monthly active users) across all defaulted subreddits. We have identified two different patterns for growth after defaulting, Surge and Jump. However, more research is needed to understand why subreddits' growth patterns differ. A full investigation is beyond the scope of this paper, but we offer a simple model that can help guide future research.

We propose a stochastic process of MAU changes over time which can capture the difference in the two types of growth:

$$\text{MAU}_{t+1} = (1-p)\text{MAU}_t + c$$

where t refers to time, p is the probability that a given user will leave the community at time t and c is the number of newcomers at time t. Under equilibrium, the number of old users leaving is equal to the number of new users joining, i.e., $p \times \text{MAU}_e = c$ or $\text{MAU}_e = c/p$. Upon being defaulted, c becomes larger because the community gets exposed to more users and therefore more users are likely to join. Therefore MAU_{t+1} will be large. When p is too small to balance c, the community would display a Surge pattern. When p is large enough so that user inflow equals user outflow, it results in a Jump community. The subreddits *ExplainLikeImFive* and *TIFU* (Figure 2c) are two examples of communities that went beyond the new equilibrium and started to fall back.

This model also illustrates the factors that play an important role in determining a community's growth type. For instance, cultivating loyal users may influence user retention rate $1-p$ and an inclusive community with broader appeal could affect the number of new users attracted c. We have proposed this stochastic model to help explain the different growth patterns that we observe in the data, however more research is required to further explore the accuracy and implications of this model.

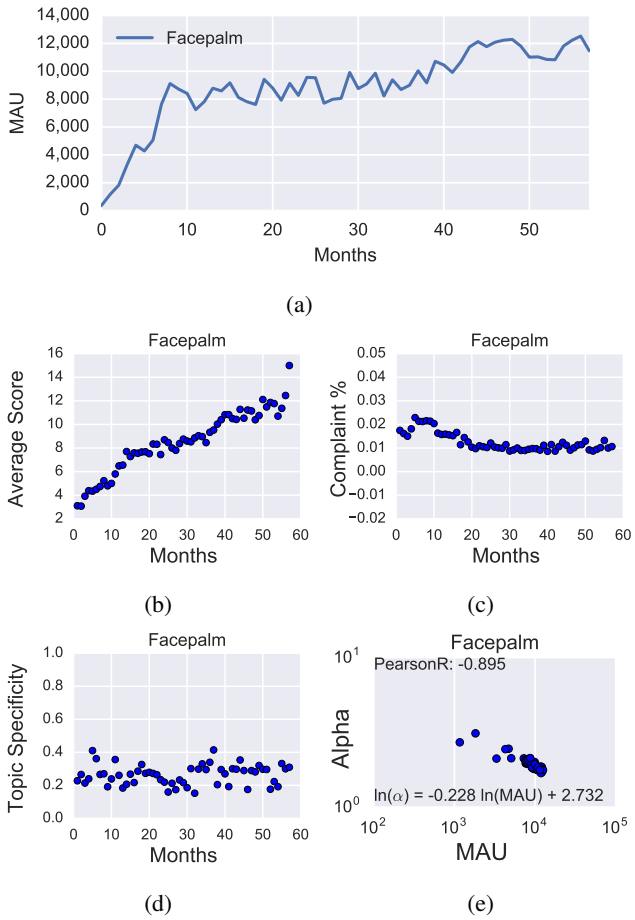


Figure 7: MAU, average score, complaint comment percentage, topic specificity, and MAU vs. α of Facepalm, a non-defaulted community. Month 0 corresponds to December 2011, which is the first month we have data for Facepalm.

Comparison with Non-Defaulted Communities

In a basic comparison between the defaulted subreddits in our study and a number of similar but non-defaulted communities we found that our observations still hold. For example, for one of these subreddits with similar MAU as the defaulted ones, Facepalm, Figure 7 shows the MAU, average score, percentage of comments that were complaints, topic specificity, and MAU vs. alpha. In line with the trends that we found in our study of defaulted communities, Facepalm has steadily increasing voting scores, slightly decreasing complaint comment percentages, and a relatively stable level of topic specificity. Unlike the defaulted subreddits, Facepalm does not have any major dips in scores, which is in accordance with our findings. Furthermore, the power law relationship between MAU and α also holds for Facepalm. Other non-defaulted communities display similar patterns.

Limitations and Future Work

In this section we detail several limitations in our work and discuss directions for future research.

Generalizability to Other Online Communities

We took advantage of the natural experiment that occured with defaulted subreddits to study the effect of a large new user influx to an online community. While this approach gives us more of an anchor to discuss the causality of our results, it also carries limitations.

We are only able to study subreddits that were eventually defaulted. This causes a selection bias in the communities that we can study because many good subreddits are never selected for defaulting, and many more never grow enough to be considered. This limitation also affects the types of communities that we can study. For instance, having a relatively generic topic that can attract a wider audience may have an effect on whether a subreddit is chosen to be defaulted. Thus, smaller or niche communities may exhibit different patterns in the face of an influx of newcomers. Future work will extend our study to other types and sizes of subreddits.

Additionally, we have focused on the effect of an influx of new users in a Reddit community and we expect that our findings are most well-equiped to explain growth in content-oriented communities with vote-based ranking systems such as Quora, Digg, and Hacker News. On the other hand, for communities such as Facebook and Twitter, inter-personal tie strength and ranking algorithms can both significantly impact how user interact with each other as well as how the community as a whole handles massive user inflow. For example, underprovision of attention as communities grow can be addressed by using appropriate content ranking algorithms. Future research will study how these different types of online communities handle growth.

Differentiating User Groups

In this work we have focused on community-level changes and have evaluated the community as a whole. However, prior work suggests that users can be differentiated based on the roles that they play in a community and how long they have been around (Velasquez et al. 2014). These user groups may exhibit different behaviors in the community. Therefore, our results may differ if we distinguish various user groups, particularly if we separate old-timers from newcomers.

For instance, it may be the case that satisfaction is high at the community level, but old-timers really are dissatisfied and are complaining more. In our work we observe that the average score recovers to high levels soon after defaulting. Nonetheless, we did not separate whether that recovery is mostly explained by new users outnumbering old-timers in votes, or by the content actually improving. Future work can study the roles that old-timers and newcomers play in each of the patterns that we have observed after massive growth in an online community.

The Perception of Community Deterioration

Members of online communities often anecdotally feel that the community became worse after it grew. However, contrary to this common opinion, we see that for the most part things remain stable. This raises an interesting question: what leads to this perception of community deterioration? Is it just a natural ennui that develops over time, even if the community were to stay small? Or is it a result of some other phenomena, such as a perception of reduced influence on the community? Future work can examine this question in more depth.

Conclusion

New members often bring fresh perspectives to an online community. But they can also interrupt the community with too broad a range of interests, unfamiliarity with community norms, and relatively lower content quality. The result is a threat: what happens to online communities after a large influx of new users? In this paper, we look into the effect of massive growth on Reddit by examining voting, linguistic, and network properties of subreddits after a substantial number of new users join. Our analyses indicate that despite a temporary dip in upvote score and increased activity clustering around a smaller proportion of posts, these online communities are not massively impacted by large influxes of new users. In the long run, post-growth online communities display a continued improvement of user reception and their languages do not become closer to generic Reddit discussions. Our results also show that strong moderation helps to keep up user reception to community content. In summary, online communities in general are not being devastated by large influxes of new users. These results provide further evidence that online communities can in fact survive the "Eternal September".

References

Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; and Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 44–54. ACM.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Butler, B. S. 2001. Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Information systems research* 12(4):346–362.

Chancellor, S.; Lin, Z. J.; Goodman, E. L.; Zerwas, S.; and De Choudhury, M. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM.

Chancellor, S.; Lin, Z. J.; and De Choudhury, M. 2016. This post will just get taken down: Characterizing removed pro-eating disorder social media content. In *Proceedings of the*

2016 CHI Conference on Human Factors in Computing Systems, 1157–1162. ACM.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial behavior in online discussion communities. In *Proceedings of ICWSM 2015*.

Clauset, A.; Shalizi, C. R.; and Newman, M. E. 2009. Power-law distributions in empirical data. *SIAM review* 51(4):661–703.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, 307–318. ACM.

Faloutsos, M.; Faloutsos, P.; and Faloutsos, C. 1999. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, volume 29, 251–262. ACM.

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4647–4657. ACM.

Gilbert, E. 2013. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 803–808. ACM.

Gorbatai, A. 2011. Aligning collective production with demand: Evidence from wikipedia. *Available at SSRN*.

Hoffman, M.; Bach, F. R.; and Blei, D. M. 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, 856–864.

Hong, L., and Davison, B. D. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, 80–88. ACM.

Jeffries, R.; Kiesler, S.; Goetz, J.; and Sproull, L. 2005. Systers: Contradictions in community.

Jones, Q.; Ravid, G.; and Rafaeli, S. 2004. Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Information systems research* 15(2):194–210.

Kairam, S. R.; Wang, D. J.; and Leskovec, J. 2012. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 673–682. ACM.

Kiene, C.; Monroy-Hernández, A.; and Hill, B. M. 2016. Surviving an "eternal september": How an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 1152–1156. New York, NY, USA: ACM.

Kraut, R. E.; Resnick, P.; Kiesler, S.; Burke, M.; Chen, Y.; Kittur, N.; Konstan, J.; Ren, Y.; and Riedl, J. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.

Leskovec, J.; Kleinberg, J.; and Faloutsos, C. 2007. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1(1):2.

Organ, D. W., and Ryan, K. 1995. A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. *Personnel psychology* 48(4):775–802.

Paul, M. J., and Dredze, M. 2011. You are what you tweet: Analyzing twitter for public health. *ICWSM* 20:265–272.

Quan, X.; Kit, C.; Ge, Y.; and Pan, S. J. 2015. Short and sparse text topic modeling via self-aggregation. *Proc. ICWSM* 2270–2276.

Ridings, C. M., and Gefen, D. 2004. Virtual community attraction: Why people hang out online. *Journal of Computer-Mediated Communication* 10(1):00–00.

Velasquez, A.; Wash, R.; Lampe, C.; and Bjornrud, T. 2014. Latent users in an online user-generated content community. *Computer Supported Cooperative Work (CSCW)* 23(1):21–50.

Zhu, H.; Chen, J.; Matthews, T.; Pal, A.; Badenes, H.; and Kraut, R. E. 2014. Selecting an effective niche: an ecological view of the success of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 301–310. ACM.

Zhu, H.; Kraut, R. E.; and Kittur, A. 2014. The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 281–290. ACM.