CS224W Project Final Report
Upon the Advent of Eternal September: A Case Study on Reddit Communities

Zhiyuan Lin, Yiqi Chen, Bowen Yao

I. Introduction

Although online communities often regard membership growth as an important goal of their development, research has shown that large influx of new users may interrupt a community's wellness by introducing information overload and lowering content quality [1], [2], which is also know as "Eternal September"[1] from the infamous case of Usenet's fall[3].

Founded in 2005, Reddit, as one of the most popular online communities[2], has accumulated a large user base over time. Among those popular subreddits, a handful of them have been made default for users throughout the past several years. Upon defaulted, those subreddits started to have a substantial number of new subscribers every day and the trend has not shown a sign of decline1.

In this project, we plan to look into the effect of large influx of new users on Reddit. In particular, how did the newcomers reshape the community's latent network structure as well as other community characteristics, if any at all? Is the community overall more connected upon defaulted? Does the community get better or worse in terms of both quantity and quality?

We hope our work can help people understanding such community growth pattern and characteristic change well and future online community designers and participants can create and maintain a healthier and more thriving online environment.

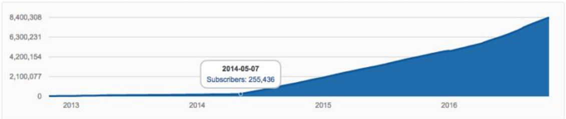


Figure 1: Subscriber number of subreddit */r/nottheonion*, which became a default subreddit on May 07, 2014. Data and visualization is from http://redditmetrics.com/r/nottheonion

II. Related Work

In the past people have conducted many researches that study behavior of different online communities.

In [4], Sanjay, et al. studied how a community's network features predispose its future growth. They categorized two types of community growth, and utilized previous growth rate and the structural features of a graph to predict its final size and longevity. On the other hand, Backstrom et al.'s work [5] focused on analyzing how structural properties in the network can affect whether a user will join a particular community, and whether the community will grow. They treated the two questions as binary classification problems, fitted the model on a decision tree, and found out that network related features are near the root of the tree. Both works analyzed communities growth using their network structures, and interpreted the research questions as prediction problems. However, the authors only answered why a community grows, but not what will happen after new members enters the community.

Danescu-Niculescu-Mizil et al.[6] proposed a framework to track linguistic change in online communities over time by analyzing data from two beer review communities. The authors discovered that at user level, the user started to adopt the community's specific language as they spent more time in the community, and at community level even the community's language changed overtime, its unpredictability decreased. To quantify linguistic change, Danescu-Niculescu-Mizil et al. used several post-level measures, including cross-entropy of posts, Jaccard self-similarity between adjacent posts, etc. Although it is an insightful investigation and extensive evaluation of linguistic change over time, the authors did not explain the cause of such linguistic change. The underlying cause can be studied through investigating interactions between newcomers and existing users. In particular, the latent network structure of the community may contribute a lot in the community.

III. Dataset and Representation

A. Dataset

In this project, we use Reddit comments data available on Google BigQuery[3]. We selected in total 10 popular subreddits from the top 200 subreddits with 7 defaulted and 3 non-defaulted subreddits

[1] https://en.wikipedia.org/wiki/Eternal_September

[2] http://www.alexa.com/siteinfo/reddit.com

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

as our dataset. Table I shows the name, default date and number of subscribers for these 10 subreddits.

Name	Default Date	Subscribers
OldSchoolCool	2014-05-07	8,902,702
nosleep	2014-05-07	8,718,721
dataisbeautiful	2014-05-07	9,102,965
nottheonion	2014-05-07	9,149,337
EarthPorn	2013-07-17	10,934,398
television	2013-07-17	10,367,213
books	2013-07-17	10,571,767
facepalm	N/A	464,229
FoodPorn	N/A	459,739
shutupandtakemymoney	N/A	289,277

[1] subsribers data is as of Dec. 11^{th}, 2016

Table I: Subreddits data specification

Those comments come from a wide range of time from 2005 to 2016. Since our selected defaulted subreddits are made defaulted in 2013 and 2014, we limit our data range to from January 2011 to September 2016. The size of the data subset will simplify our computation by letting us avoid spending time in sharding data or setting up distributed system, so that we can focus on the analysis.

The data in Google BigQuery is stored in tabular form. We keep a subset of the table's fields (listed in table II) as our dataset.

Name	Type
body	STRING
score_hidden	BOOLEAN
author	STRING
created_utc	INTEGER
link_id	STRING
parent_id	STRING
score	INTEGER
id	STRING
subreddit	STRING

Table II: Reddit comment data fields in Google Query

B. Graph Construction

Given the dataset described above, we would like to construct a network that is able to represent interaction between users. As opposed to social networks like Facebook, Reddit users generally focus more on the content rather than other users. Instead of using edges to represent friendship, we construct graphs representing common interest among users by connecting users who comment under the same post.

Additionally, we take all comments from each month as a *monthly snapshot* as we are studying dynamic communities that evolve over time.

Here for each monthly snapshot and each subreddit, we construct a Common Interest Graph as the following:

- Nodes: Users that commented at least once in the given month and subreddit
- Edges: Two users commented on at least k common posts

We believe that this construction could capture the signal of correlation between users. The parameter k represents the robustness of the connection. Here, we are not constructing graph based on that an edge between two users exists if one user replies to another user's comment because we believe that users are commenting/replying because of the content itself rather than the content creator.

Figure 2 demonstrates degree distribution of subreddit *OldSchoolCool* on December 2014, with k=1 and 2. We can see that when $k = 1$, there are a significant portion of users that have the same high degrees, which violates the power law. This is not surprising due to how we construct the graph. It is likely that two users both only comment on the same very popular post, and in this case they will have the same high degree, since they will form edge with every other user who commented on the same post. Therefore, the graph construction with $k = 1$ is too noisy since it takes weak connections between users into account. On the other hand, choosing $k = 2$ will only account for strong connections between users, and the degree distribution fits power law (see Figure 2). Therefore, we will use $k = 2$ in all experiments in the following sections.

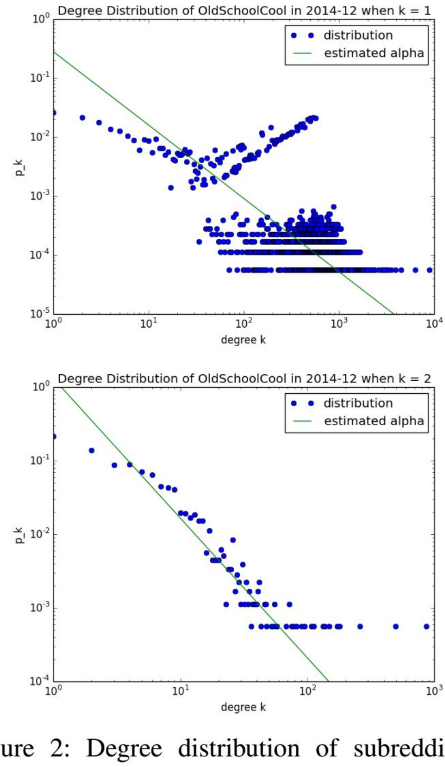


Figure 2: Degree distribution of subreddit *OldSchoolCool* on December 2014, with $k = 1$ on the top and $k = 2$ on the bottom

C. Monthly Active User (MAU)

1) Definition: Upon defaulted, the number of subscribers to the community will definitely grow by nature. However, we are more interested in those 'active' users who actually participated in the community rather than simply being subscribed by the system by default. Therefore we introduce the notion of Monthly Active User, which is defined as an user who has commented within that month. For succinctness, we will be using MAU to denote the number of Monthly Active User throughout this paper.

2) Two Patterns for MAU Growth: We observed that default boosts MAU across all of defaulted subreddits. However, not all subreddits' MAU follow the same pattern as their subscribers, which all skyrockets after default. We identify two different types of MAU growth patterns, the 'Surge' pattern and the 'Jump' pattern. Upon defaulted, both of the two patterns enjoy an enormous MAU growth but then they differ in the later period growth pattern, where 'Surge' manage to keep growing consistently while 'Jump' loses the power of further expanding. Despite of this difference in growth persistence, we can definitely be sure that one of the main results for default is a dramatic growth in MAU. Figure 3 shows how the 7 defaulted subreddits in our dataset fall into those two categories.

IV. Method and Experiments

To restrict the scope of our broad research question "what would *happen* to the online community after large influx of new users?", we focus on the changes of two properties of the communities before and after default:

1) Community Structure
2) Content Quality

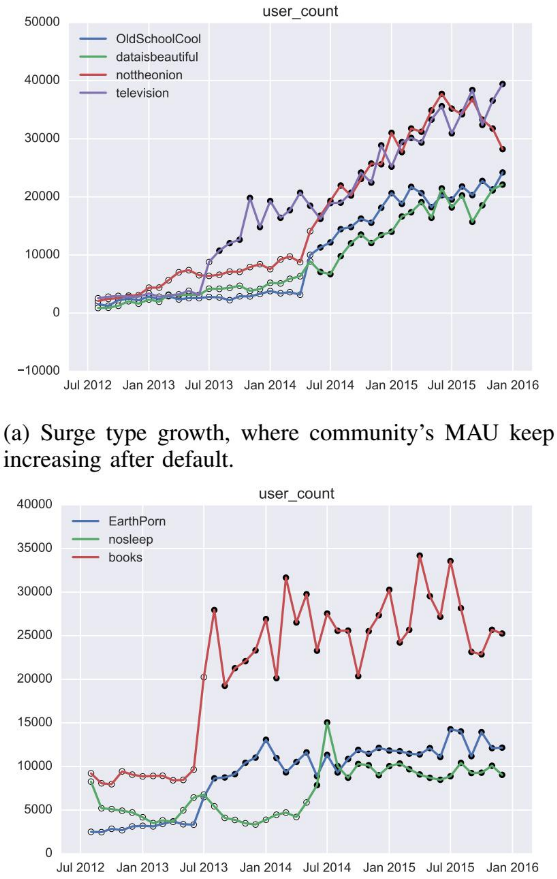
For the first question, we resort to the constructed common interest graph and study its degree distribution, which, in essence, reflects changes for users' common interests.

For the second question, we study the change of average comment score within the community, where the score for a comment is generated through votes of users. User can either up-vote or down-vote a certain post.

A. Community Structure

We resort to the common interest graph constructed in section III-B to study how common interest among users change before and after default.

1) Motivation: Intuitively, users subscribe to a subreddit because they are interested in the content, whereas users that subscribe to a subreddit by default are less interested. Therefore, after the default, a large group of users are subscribed to the subreddit but they share little common interest. In the Common Interest Graph, this group of users will have small degree, and therefore the degree distribution of the entire graph will be less heavy-tailed. i.e. user interest will be less concentrated.



(a) Surge type growth, where community's MAU keep increasing after default.

(b) Jump type growth, where community's MAU increases from a lower level to a relatively stable higher level.

Figure 3: Two monthly active user number growth type upon default: surge and jump. Black dots indicate data points when corresponding subreddits have already been defaulted.

2) Methodology: In order to measure the concentration of user interest, we compute the α value of the distribution given that the distribution abides power law [7], where larger α value represents a heavier tail, which implies stronger connection between users (more users have common-interest edge with others). In order to estimate α from observed discrete data points, we make use of Maximum Likelihood estimator since the same method for estimating α was proposed in [8] and it gave accurate estimation. An example of α estimation using MLE is shown in Figure 2. The process of computing α using MLE is outlined below:

Let x_i be degree of user i, $L(\alpha)$ be the log-likelihood of the degree distribution.

$$\begin{aligned}
L(\alpha) &= log(\prod_{i=1}^{n} p(x_i)) \\
&= \sum_{i=1}^{n} log(p(x_i)) \\
&= \sum_{i=1}^{n} log(\frac{\alpha-1}{x_{min}}(\frac{x_i}{x_{min}})^{-\alpha}) \\
&= \sum_{i=1}^{n} (log(\alpha-1) - log(x_{min}) - \alpha log(\frac{x_i}{x_{min}}))
\end{aligned}$$

In order to maximize α, we set $\frac{dL(\alpha)}{d\alpha} = 0$. Therefore,

$$\frac{dL(\alpha)}{d\alpha} = \frac{n}{\alpha-1} - \sum_{i=1}^{n} log(\frac{x_i}{x_{min}}) = 0$$

$$\alpha = (\sum_{i=1}^{n} log(\frac{x_i}{x_{min}}))^{-1} * n + 1$$

Figure 4 demonstrates how α changes over time in different subreddits. Contradicting to our hypothesis, we see that in most subreddits, there is no significant α change before and after the default date. There are also exceptions such as *OldSchoolCool*, where α drastically decreases at the time of default, which means that the community is even more connected.

We can learn from the above observation that there is no obvious correlation between defaulting and α value, or connectiveness of the common interest graph. However, whether defaulting affects α value in a more obscure way still remains unknown. We will dive deeper into the problem in the following subsections.

3) Relationship between Monthly Active User and α: Monthly Active User (MAU) is one of the most fundamental metrics of a community. In Figure 5a, we plot α value changes as MAU increases. It indicates a negative correlation between MAU and α value. Furthermore, when MAU is large, α value decreases much slower with MAU. There seems to be diminishing loss for α as a function of MAU. So we may make the assumption that the relationship between α and MAU abides the power law

$$\alpha = an^{-c}$$

where MAU is denoted as n for simplicity and a, c are positive constants.

In order to test the above assumption, we conduct log transformation on both sides of the equation and get

$$\log(\alpha) = -c\log(n) + \log(a)$$

This function shows that if we plot α versus n on log-log scale, the points should form a clear straight line. Figure 5c shows such log-log scale plot for all 10 subreddits and plots that pass Person's R correlation test with a significance level of 0.05 or below are colored blue. We can see that all but one subreddits pass the test and plotted points show clearly linear pattern, which proves the legitimacy of the model $\alpha = an^{-c}$. For the single outlier *shutupandtakemymoney*, we argue that the community is relatively small compared to other communities and communities can be unstable at the early stage with small size.

The above argument reveals a relationship between α and MAU that is consistant before and after default. Then what is the effect of default on α and hence, the common interests of the community? We can resort to results in section III-C to answer this question, where we showed that MAU changes significantly at default date. Also we can observe in Figure 5a that there is a big gap in x-axis which is consistent with effect of default (there are much more active users after default). Therefore, **though the relationship between α and MAU remain the same before and after default, default manages to insert substantial effect on the community's common interests through drastically boosting MAU**.

4) Intuition under the Hood: We can further interpret the relationship between α and MAU from a growth perspective. We regard both α and MAU as functions of time t. Then take derivative with respect to t for both sides of the equation

$$\log\alpha(t) = -c\log n(t) + \log(a)$$

we get

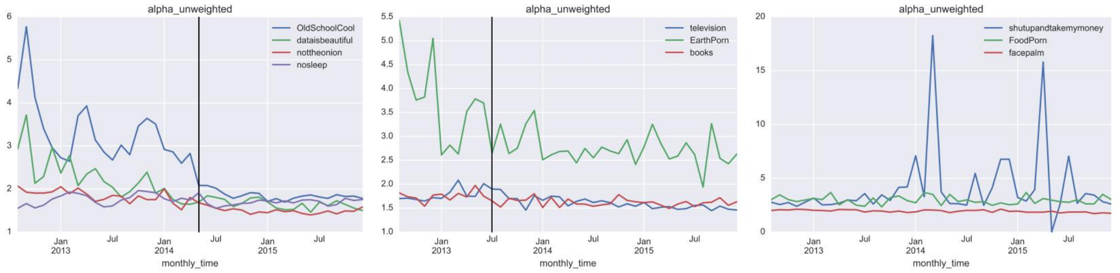
$$\frac{1}{\alpha}\frac{d\alpha}{dt} = -c\frac{1}{n}\frac{dn}{dt}$$

Then we discretize both sides with respect to time and get:

$$\frac{\alpha_t - \alpha_{t-1}}{\alpha_t} = -c\frac{n_t - n_{t-1}}{n_t}$$

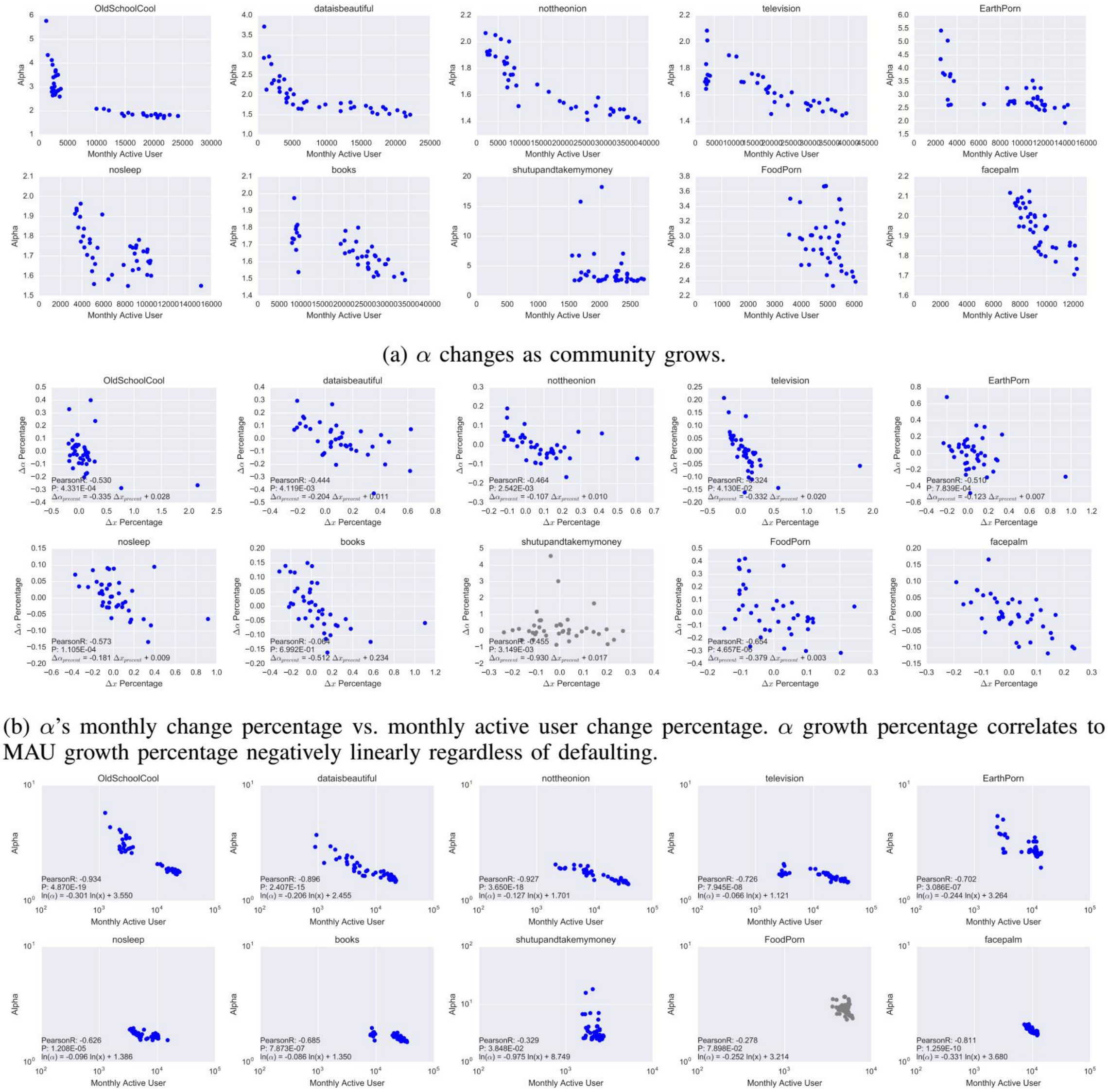
This means that the decay rate of α is proportional to the the growth rate of MAU. Figure 5b shows that such linear relationship indeed exists in our data sets, which further back-up our argument in section IV-A3. Intuitively, this linear relationship between two rates indicates that for Reddit communities, a faster growing community is more likely to become more concentrated in common interests.

5) Additional Reasoning based on Observed Post Size: We can further back-up our argument in section IV-A3 from another angle. We observe that the average number of comments per user (denoted by cu) remains constant for a certain community while the average number of comments per post (denoted by cp) increases as MAU increases. Figure 6 shows typical examples of constant cu and increasing cp. We argue that these observations indicate that a user is becoming more and more likely to join a current posts rather than open



(a) Subreddits got defaulted in 2014 (b) Subreddits got defaulted in 2013 (c) Non-defaulted subreddits

Figure 4: α changes over time. The vertical line represents the default date. It remains unclear whether defaulting has a significant effect over α.

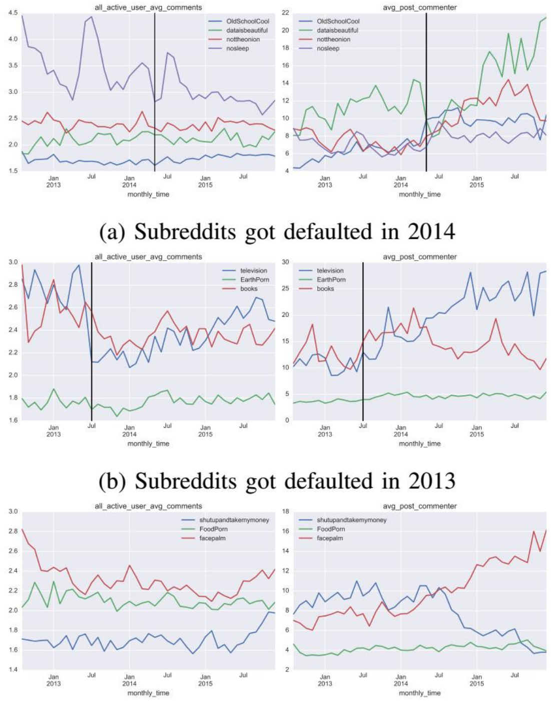


(a) α changes as community grows.

(b) α's monthly change percentage vs. monthly active user change percentage. α growth percentage correlates to MAU growth percentage negatively linearly regardless of defaulting.

(c) α changes as community grows plotted in log-log style. We can see a negative correlation between them.

Figure 5: Relationship between α and monthly active user (MAU). In figure 5b and 5c, least square fitted line is described is described at bottom left of each plot. In addition, results of Person's R test is also listed with insignificant results (p>0.05) grayed out (*shutupandtakemymoney* in 5b and *FoodPorn* in 5c).



(a) Subreddits got defaulted in 2014

(b) Subreddits got defaulted in 2013

(c) Non-defaulted subreddits

Figure 6: Constant average number of comments per user and increasing average number of comments per posts

its own posts, which implies a more concentrated interests throughout the community. To clarify this argument, we can consider the following simple model: Suppose p is the probability of a user uses one of its comments to open a new post to 'diverge' the community's common interests. Then the expected number of posts in the community is $p \cdot n \cdot cu$, where cu is constant, where MAU is denoted as n. Then we have

$$cp = \frac{cu \cdot n}{p \cdot n \cdot cu} = \frac{1}{p}$$

Therefore the increasing cp implies a decreasing p, namely users are less likely to open new posts to 'diverge' the interest, indicating a more concentrated common interests throughout the community.

B. Content Quality

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 score represents a piece of well-received content in the community and hence implies better content quality. In order to measure the content quality change in Reddit, we study posts' scores of each subreddits.

It is natural to occur to one that score might be a positively dependable variable of number of users in a subreddits and therefore increases only as the number of user grows. In figure 7, we can see that it is in fact not the case. As a counter example, for communities like *FoodPorn* (figure 7c), the average score even decreases as MAU increases. In addition, in communities with jump growth pattern, even for months with similar MAU, they may be temporally far away, which could introduce even more noise in the data. Hence, we focus on how scores changes overtime, especially the difference before and after default.

1) Observation and Hypothesis: By plotting scores overtime (gray dots in figure 8), even though communities' scores generally have increasing trends over time, we can see that the community score takes a quick plummet after default. With that said, we propose a hypothesis that community scores increase overtime, but a large influx of new users will negatively impact this trend.

2) Experiments: To test this hypothesis, we employ regression discontinuity design [9]. Assuming there is a linear correlation between time and score as observed and Pearson-R-tested in figure 8, the score s can be expressed as a function of t: $s = \beta_0 + \beta_1 t$. When defaulting event takes place, we expect to see a statistically significant change in either β_0 or β_1 (or both). To capture such change, we introduce the dummy variable $\mathbb{1}_{t \geq t_0}$, which is an indicator function with condition *the current month is after or equal to the default month*. Then, we are able to obtain a new linear equation as

$$s = \beta_0 + \beta_1 t + \beta_2 \mathbb{1}_{t \geq t_0} + \beta_3 t \mathbb{1}_{t \geq t_0}$$

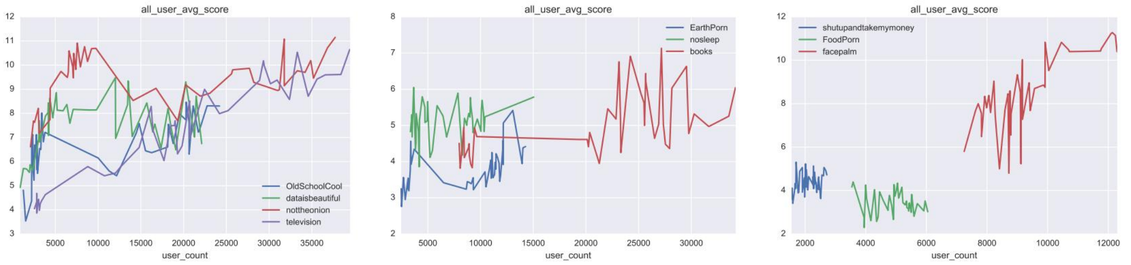
After linear regression, if β_2 or β_3 is statistically significantly non-zero and negative as indicated by t-test, we shall be able to claim that there is a statistically significant impact on the community score. A negative β_2 implies the defaulting event lowers the average community score while a negative β_3 suggests the score increasing gets slower after defaulting.

Dataset	β_2	p_{β_2}	β_3	p_{β_3}
OldSchoolCool	-1.8024	< 1e-3	-0.0124	0.686
dataisbeautiful	-0.9231	0.052	-0.2299	< 1e-3
nottheonion	-2.8387	< 1e-3	-0.0890	0.014
television	1.4901	0.002	0.1708	0.006
EarthPorn	-0.9071	0.008	-0.1065	0.018
nosleep	0.0001	1.000	-0.0586	0.017
books	-0.3700	0.355	0.0194	0.713

Table III: Regression discontinuity

Presented in table III, most of the defaulted subreddits do have a statistically significant negative value for β_2 or β_3.

Note that the only subreddit with statistically significant positive β_2 and β_3 value is *television*, which responded to defaulting with strong (about 3x more deleted comments) and long-lasting (around 2 years) moderation (figure 9b).



(a) Score change as user grows for surge-growing subreddits

(b) Score change as user grows for jump-growing subreddits

(c) Score change as user grows for nondefault subreddits

Figure 7: Score change vs. monthly active user number (MAU). Score is not strictly increasing as MAU grows.

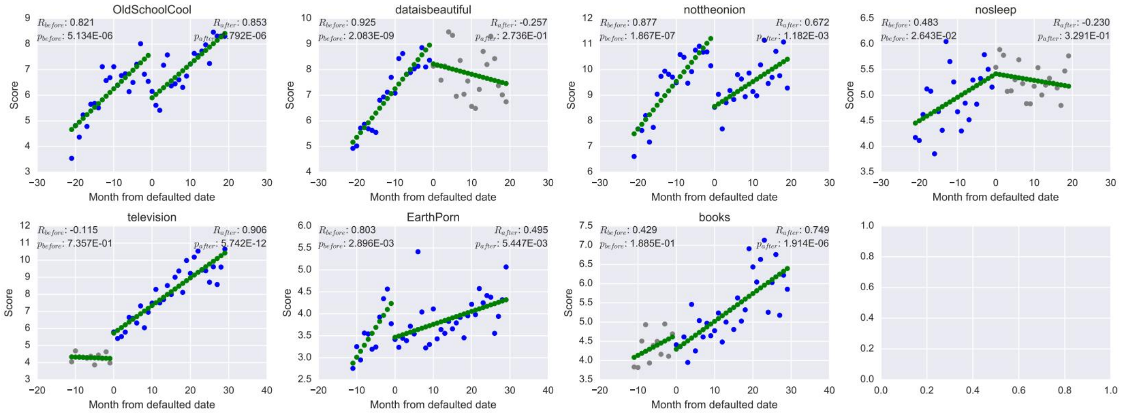
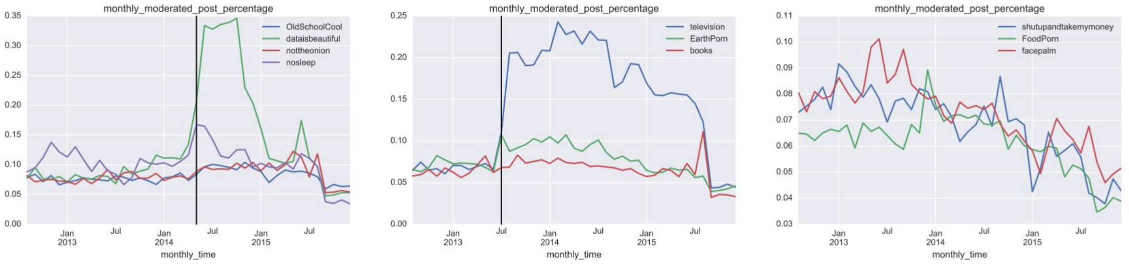


Figure 8: Regression discontinuity. X axis represent the number of months since default. Positive and negative values mean month after and before default respectively. Pearson's R is calculated for data points before and after default individually. Statistically significantly ($p < 0.05$) linear data is plotted in blue while insignificant ones are in gray. Green dots are predicted values given by $s = \beta_0 + \beta_1 t + \beta_2 \mathbb{1}_{t \geq t_0} + \beta_3 t \mathbb{1}_{t \geq t_0}$.



(a) Subreddits defaulted in 2014

(b) Subreddits defaulted in 2013

(c) Non-defaulted subreddits

Figure 9: Moderation over time, measured by the percentage of deleted comments in the datasets per month.

As a result, we present a proposition as following: **Community scores increase overtime, but a large influx of new users will negatively impact this trend by bringing down the score and/or slowing the increase speed.**

V. Conclusion

In this paper, we describe our research on the effect of large influx of new users on online communities by investigating 10 defaulted and non-defaulted subreddits.

With mathematical proof as well as statistical and visual analysis on empirical data, we summarize our main contributions as follows:

A. Defaulting Drastically Boosts MAU

As discussed in section III-C2, all subreddits experience a giant grow in MAU at the time of default. Furthermore, we are able to capture two different types of MAU growth pattern - surge and jump, as shown in figure 3.

B. Consistent Correlation between α and MAU

As shown in IV-A, at community structure level, we did not see obvious correlation between α and default. But we manage to discover the deeper consistent relationship between α and MAU and make the conclusion that default can still impose great impact on the community's common interest structure through dramatically boosting MAU.

C. Large Influx of New Users Negatively Impacts Content Quality

At content quality level, we found out a significant negative impact on average score of comments after default through regression discontinuity method. Strong and long-term moderation can mitigate such effect.

D. Online Community is Getting Better Overtime

Regardless of defaulting, we observed a general trend where communities are getting better overtime reflected by decreasing α and increasing score. Regular growth of the community size neither hurts the connectivity between users nor downgrades quality of content. From IV-A, we see that MAU negatively impacts α value, which indicates that more users have strong common interest with large number of users, and the community becomes more connected. Also, average score of comments is generally getting higher over time which reflects improvement of content quality. Even though average score of comments decreases at the time of default, communities with strong and long-term moderation survive it.

VI. Future Work

A. Interpret Two Types of MAU Pattern

In section III-C2, we show that default boosts MAU across all defaulted subreddits and there are two types of growth pattern. However, it remains unknown why default necessarily leads to growth in MAU and why growth patterns differ. The categorization is a very strong finding and is very consistent across all subreddits. However, since this work only focuses on the default event, the differentiation is out of scope of this work but it may lead to very interesting finding.

Nevertheless, we intuitively propose a model of MAU changes over time which may 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 captures how users are leaving the community and c describes the community's ability of appealing new users. Under equilibrium, $\text{MAU}_e = c/p$. Upon default, c becomes larger because the community naturally gets exposed to more users so more users are likely to join the community. Therefore MAU_{t+1} will be huge. Then for the 'jump' pattern community, they quickly reach the equilibrium stage and fail to grow further more, while for the 'Surge' pattern community, they still don't reach equilibrium and manage to keep growing. The model is merely primitive and requires data to back it up. But again, verifying this model and estimating the parameters are outside of the scope of this project.

B. Experiment using Weighted Common Interest Graph

As discussed in section III-B, if we consider two users have common interest when they commented on at least one same post ($k = 1$), the resulting Common Interest Graph will be noisy, and we have to let $k = 2$ in graph construction in order to filter out weak common interest. As another perspective, constructing a common interest graph weighted by size of post (e.g. two users have common interest edge of weight $1/n$ if they both commented on a post of size n) can also mitigate this problem. We have conducted preliminary experiments using weighted common interest graph, but did not find significant correlation between the network and other characteristics of the communities. Given that we have answered our research question using unweighted common interest graph discussed in section III-B, further investigation using weighted common interest graph is outside of the scope of this project. Nevertheless, it may lead to more findings around our research topic.

References

[1] Q. Jones, G. Ravid, and S. Rafaeli, "Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration," *Information systems research*, vol. 15, no. 2, pp. 194–210, 2004.

[2] B. S. Butler, "Membership size, communication activity, and sustainability: A resource-based model of online social structures," *Information systems research*, vol. 12, no. 4, pp. 346–362, 2001.

[3] C. Kiene, A. Monroy-Hernández, and B. M. Hill, "Surviving an" eternal september"-how an online community managed a surge of newcomers," *arXiv preprint arXiv:1605.08841*, 2016.

[4] S. R. Kairam, D. J. Wang, and J. Leskovec, "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*. ACM, 2012, pp. 673–682.

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "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*. ACM, 2006, pp. 44–54.

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

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *ACM SIGCOMM computer communication review*, vol. 29, no. 4. ACM, 1999, pp. 251–262.

[8] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[9] D. L. Thistlethwaite and D. T. Campbell, “Regression-discontinuity analysis: An alternative to the ex post facto experiment.” *Journal of Educational psychology*, vol. 51, no. 6, p. 309, 1960.