CS224W Project Milestone Report Upon the Advent of Eternal September: a Case Study on Reddit Communities' Latent Networks

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 attract a substantial number of users every day and the trend has not shown a sign of decline. Did this surge of newcomers truly bring those subreddits increased popularity or it actually tore those communities down from inside? How did the older users react to newcomers? How did the newcomers reshape the community's latent network structure as well as other community characteristics, if any at all? How does old and new users interact with each other after defaulting happens? We plan to look into these problems from a network perspective.

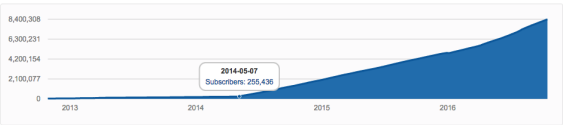

Figure 1. Subscriber number of subreddit */r/nottheonion*, which became a default subreddit on May 07, 2014. Data and visualization is from http://redditmetrics.com/r/nottheonion

II. Related Work

In the past people have conducted many researches that study behavior of different online communities. Danescu-Niculescu-Mizil et al.[4] proposed a framework to track linguistic change in online communities over time by analyzing data from two beer review communities. 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. In [5], Sanjay, et al. studied how a community's network features predispose its future growth, whereas Backstrom et al.'s work [6] focused on analyzing how a user's structural properties in the network can affect whether he will join a particular community. Even though both works analyzed communities using their network structures, the authors didn't consider what will happen after new members enters the community, i.e. after the community grows.

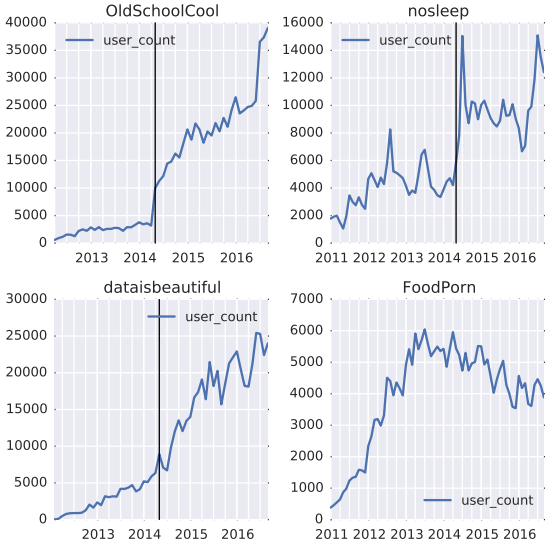

Figure 2. Number of monthly active users. An active user in a month is defined as an user who has commented within that month. The vertical line in each graph indicate the defaulted time for corresponding community.

III. Dataset and Representation

A. Dataset

In this project, we use Reddit comments data available on Google BigQuery[3]. We selected 10

[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

defaulted and 10 non-defaulted popular subreddits from the top 100 subreddits as our potential data candidates.

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 much 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.

In this milestone report, We select a handful of representative subreddits to conduct our experiments on so that we can receive quick feedback on results and again more insight within the limited time we have this quarter. Particularly, we look into subreddit *nosleep*, *OldSchoolCool*, and *dataisbeautiful*, which all got defaulted on May 07, 2014. Additionally, we conducted the very same experiments on *FoodPorn*, which is a non-defaulted subreddit, for comparison.

The data in Google BigQuery is stored in tabular form. We keep a subset of the table's fields (listed in table I) 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
author_flair_css_class	STRING

Table I
REDDIT COMMENT DATA FIELDS IN GOOGLE QUERY

B. Common Interest Graph

Given the dataset described above, we would like to construct a network that is able to represent interaction between users. 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 correlation. 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.

C. Common Interest Graph Characteristics

1) Degree Distribution: Figure 3 demonstrates degree distribution of subreddit OldSchoolCool on December 2014, with k=1 and 2. Recall from Section III-B that k is the parameter that measures how many posts the two users should both comment on in order to form an edge between them. While the distribution with k=2 follows power law, the distribution with k=1 looks quite different: there are a significant portion of users that have the same high degrees. We think the reason is that the network with k=1 is noisy since there are users who only comment on one very popular post, and in this case they will have edges formed with every other user who commented on the same post. As a next step we will work on adding weight on edges based on number of users who comment on a particular post.

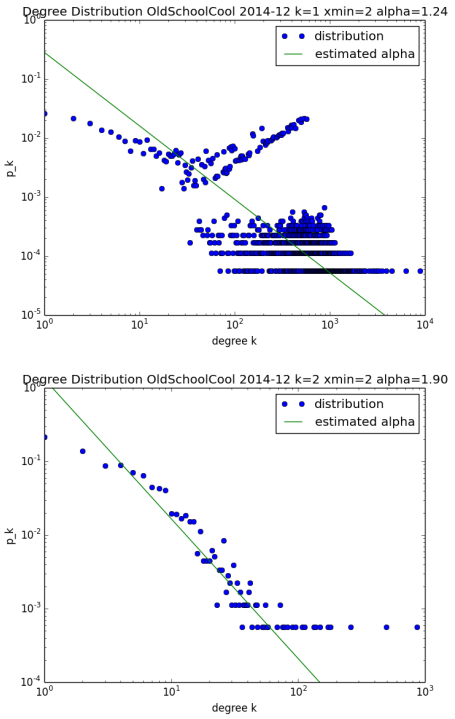

Figure 3. Degree distribution of subreddit OldSchoolCool on December 2014, with $k = 1$ on the top and $k = 2$ on the bottom

2) Relative Activeness: For each graph, we divide users into two disjoint groups: assimilated users and new users. Assimilated users are defined to be users that have made comments more than 3 months earlier than when the snapshot of the graph is taken and new users are just the complement of old users.

With such categorization of users, we can get three kinds of user interaction:

- Assimilated user-assimilated user interaction
- Assimilated user-new user interaction
- New user-new user interaction

For each of these three interactions, we can define the corresponding relative activeness. The relative activeness of a certain type of user interaction is defined to be the number of actual interactions(defined by having an edge between users) divided by the total number of possible interactions between users. For example, if we have 5 new users and there are 11 edges between them, then the relative activeness of new user-new user interaction is $\frac{11}{5\times4/2}$

By computing how relative activeness of all three kinds of user interaction changes over time for various subreddits, we discover that most of the subreddits follow one single pattern: Assimilated user-Assimilated user interaction is always at the top while new user-new user interaction is always at the bottom. We call this 'Assimilated-user-dominated' pattern. There is only one outlier, the subreddit nosleep, that has exactly opposite pattern: Assimilated user-assimilated user interaction is at the bottom while new user-new user interaction is at the top. Figure 4 demonstrates these two patterns.

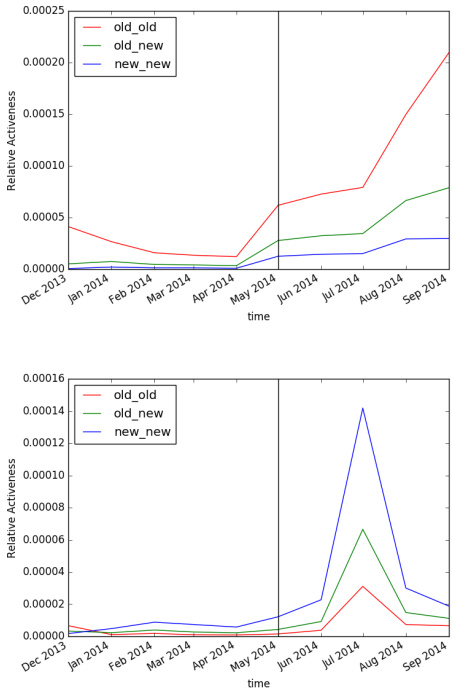

Figure 4. Two patterns of relative activeness of user interactions: Upper is assimilated-user-dominated and bottom is new-user-dominated. The black vertical line indicates the month when the subreddit gets defaulted.

IV. Approach and Results

In order to measure changes in online communities, we focus on several community and graph characteristics. They include:

1) Shared common interest level
2) Inter/intra user group activity
3) Content quality
4) Linguistic topic change

In this section, we introduce four propositions regarding subreddits with explosive growth based on our observation and experiments. In each of the following subsections, we state the proposition, explain our approach, and present some of our preliminary experiment results.

The last two propositions are not strictly related to the graph structure of subreddits. However, we believe they can provide useful contextual information about the overall changes taking place in those communities and hence keep them in this section.

A. *Users Share More Common Interests*

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 date of a subreddit, a large group of users are subscribed to it 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. the degree distribution should be more skewed. Surprisingly, our experiment shows neutral and even opposite results, where people share even more common interests after default.

1) Approach: In order to measure the skewness of a degree distribution, the conventional approach would be to compute the α value of the distribution given that the distribution abides power law, where larger α value represents less skewness and heavier tail. As mentioned in Section III-C1, the degree distribution with k=1 is quite noisy, so we use the distribution with k=2 in the calculation. In order to compute α, we use Maximum Likelihood Estimation. 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,

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

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

To conduct the experiment, we select 4 subreddits: *nosleep*, *dataisbeautiful*, *OldSchoolCool* and *FoodPorn*. Within these subreddits, *nosleep*, *dataisbeautiful* and *OldSchoolCool* were all defaulted in 05/2014, whereas *FoodPorn* is never defaulted.

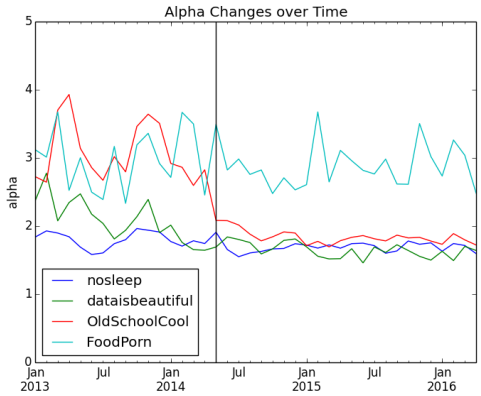

Figure 5. α changes over time in the span of 01/2013 and 04/2016. The black vertical line indicates the default date of *nosleep*, *dataisbeautiful* and *OldSchoolCool*.

2) Preliminary Results: Figure 5 demonstrates how α changes over time for the three subreddit, especially before and after the default date. It is not surprising and α of *FoodPorn* oscillates but is always around 3, since *FoodPorn* is never defaulted and there shouldn't be any change. We are able to witness that α of *OldSchoolCool* drops from 3 to 2 immediately after the default date, and is never able to rise up again. It contradicts our hypothesis that degree distribution becomes more skewed after default. Furthermore, default changes this community permanently. We can also see α of subreddit *dataisbeautiful* decreases over time but there is no drastic decrease at the time of default. On the other hand, α of subreddits *nosleep* remains the same even after default. This suggests that the effect of defaulting differs from subreddit to subreddit.

B. *Different Relative Activeness Pattern Reacts to Default Differently*

In Section III-C2, we showed two patterns of user interaction relative activeness, the 'assimilated-user-dominated' pattern and the 'new-user-dominated' pattern. Our further experiment shows that these two patterns react to defaulting very differently: After being defaulted, the 'assimilated-user-dominated' community managed to maintain users and all three types of user interaction grow significantly. On the other hand, the 'new-user-dominated' community has a 'roller-coaster' like experience, where it firstly enjoys a short-lived prosperity in user interaction and then drops back to its origin status.

1) Approach: We prove our argument by calculating the time series of user interaction relative activeness for the two different types of communities and compare the communities' user interaction before and after being defaulted.

2) Preliminary Results: Figure 4 shows the two types of communities' obvious different reactions before and after being defaulted. It is clear that after being defaulted, 'assimilated-user-dominated' subreddit enjoys a significant growth in user interaction relative activeness and is able to maintain such growth without dropping. On the contrary, the 'new-user-dominated' subreddit at first has significant growth in user interaction while then suffer from a significant drop.

We can reason about these experimental findings as follows: In the assimilated-user-dominated subreddits, the community is more likely be capable of retaining users; Otherwise the assimilated user-assimilated user interaction can not dominate in the community. Based on this analysis, it is reasonable that after being defaulted, these communities will manage to appeal and maintain a relatively large amount of the defaulted new users and transform them into 'loyal' assimilated users. On the other hand, in the 'new-user-dominated' subreddit, user interaction will eventually drop because of the community's lack of user-maintaining ability.

C. *Content Quality Increases*

One interesting non-graph-related observation we made is that the overall quality of content increases.

1) Approach: In order to capture the overall content quality of subreddits, we employed two statistics from Reddit comment dataset to examine both comment-level quality and post-level quality.

To evaluate comment-level content quality, we use comment monthly average scores resulted from user voting. The comment scores are expected to follow an exponential distribution with mean $\beta = \frac{1}{\lambda}$, and it is easy to show that the maximum likelihood estimator of β is the sample mean of comment scores. It suggests that the higher the average score is, the higher proportion the highly-scored comments constitute, which implies an overall more harmonious community. Note that this also implies that the community should not be a dependent variable of number of active users. In fact, *FoodPorn*'s average score increases while its monthly active user population's size declines.

To measure post-level content quality, we measure the percentage of comments containing complaint keywords generated by Empath[7] with keywords "shitty post" and "repost". As comments are generally in response to the top-level post, higher percentage of complaining comments indicates lower post-level content quality.

2) Preliminary Results: Originally by intuition, we were expecting to see a plummeting in content quality in a form of lower average score and increased percentage of complaints. To our surprise, as shown in figure 6, in all communities average score increases overtime without being negatively impacted presumably lower-quality content introduced by large influx of new users. On the other hands, complaints levels for defaulted subreddits either maintain the same or decrease overtime in figure 7.

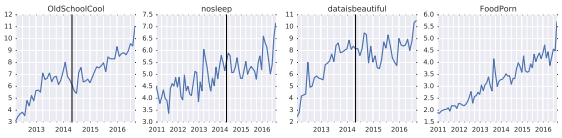

Figure 6. The monthly average score change overtime of subreddits *OldSchoolCool*, *nosleep*, *dataisbeautiful*, and *FoodPorn*. (from top left to bottom right)

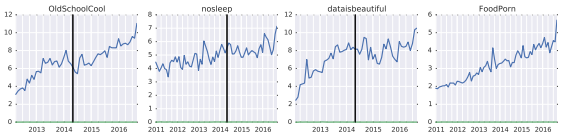

Figure 7. The monthly percentage of complaining posts of subreddits *OldSchoolCool*, *nosleep*, *dataisbeautiful*, and *FoodPorn*. (from top left to bottom right)

D. *Community Topic Drifts*

Another noteworthy, but not strictly graph-related finding is that what users talk about (topics) within the subreddits drift away further than before, but the topic's specificity is not necessarily weakened. In other words, even though new users bring changes to those defaulted subreddits, they simply shift the community's focus instead of diffusing the community's focus.

Figure 8. Monthly LDA topic distribution cosine distance between current month and December 2012. From left to right are: *OldSchoolCool*, *nosleep*, *dataisbeautiful*, and *FoodPorn*. *FoodPorn* is in fact an outlier here among all those non-defaulted subreddit, and shows some unusual pattern here to be investigated into.

We measure the subreddit's topics by leveraging the online version[8] of the well-know LDA model[9]. We evaluate the extent to which the topic drifts in a given month by comparing the cosine distance between the topic distribution in this month to a fixed month in the past before default, which we arbitrarily choose to be December 2012.

To assess a subreddit's specificity in a month, we trained another LDA model using a set of randomly sampled posts across all subreddits and compare the cosine distance between the general Reddit sample and the subreddit's comments in this month output by this general LDA model.

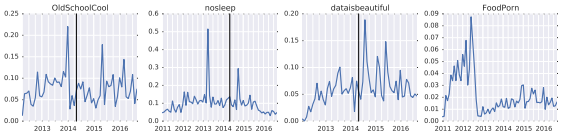

Figure 9. Monthly LDA topic distribution cosine distance between the community and general Reddit topic. From left to right are: *OldSchoolCool*, *nosleep*, *dataisbeautiful*, and *FoodPorn*.

V. Preliminary Conclusion and Todo

Our preliminary results refute the intuitive argument that a surge of new comers will exert a negative impact on online communities. On the contrary, our results indicate that those defaulted subreddits, despite facing large influx of new users, seem to not only have survived the Eternal September, but thrived even more.

Nonetheless, a few major concerns need to be addressed and some extra work needs to be done before we can draw a definitive conclusion from our observations.

1) Do those drastic changes in community characteristics caused by defaulting? Or they are simply caused by having a large influx of newcomers or by reaching a certain community size threshold (which may be triggered by being defaulted on Reddit)? We will review all these statistics' change as community size increases (measured by monthly active user count) and compare (both defaulted and non-defaulted) subreddits when they are at similar sizes.
2) How do our different experiment results correlate? We will run correlation tests between variables and draw conclusions based on statistically significant results.
3) Why do different subreddits behave differently in our experiments? What's the underlying nature of a subreddit that causes the difference? We have not come up with a concrete approach to answer this question right now but we believe the correlation test and review over community's absolute size will shed lights on this question.
4) How can we better represent the interaction between users in the graph? For example, we will try adding weight on edges based on number of users who comment on a particular post.
5) Do our preliminary results generalize well? We can easily confirm that by running experiments on more datasets.

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] 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.

[5] 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.

[6] 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.

[7] E. Fast, B. Chen, and M. Bernstein, "Empath: Understanding topic signals in large-scale text," *arXiv preprint arXiv:1602.06979*, 2016.

[8] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," in *advances in neural information processing systems*, 2010, pp. 856–864.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.