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

Zhiyuan Lin, Yiqi Chen, Bowen Yao

October 20, 2016

1 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" from the infamous case of Usenet's fall[1] [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, 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.

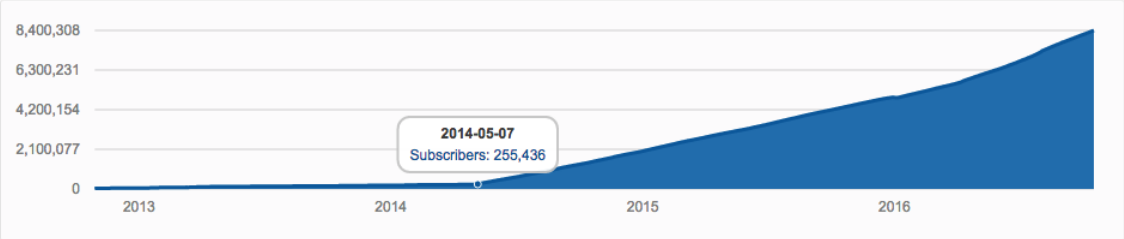

Figure 1: Subscriber number of subreddit */r/nottheonion*, which became a default subreddit on May 07, 2014. Data and visualization is from http://redditmetrics.com/r/nottheonion

2 Related Work

2.1 No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities [4]

2.1.1 Summary

Danescu-Niculescu-Mizil et al. proposed a framework to track linguistic change in online communities overtime for both users and the community as a whole by analyzing data from two beer review communities. Specifically, the authors focused their analysis on following parts:

1. User-level linguistic evolution

[1] https://en.wikipedia.org/wiki/Eternal_September
[2] http://www.alexa.com/siteinfo/reddit.com

2. Community-level linguistic evolution

3. The interplay between them

At user level, the authors discovered that the user started to adopt the community's specific language as they spent more time in the community. At community level, the authors found that even the community's language changed overtime, its unpredictability, which was measured by cross-entropy, decreased. The result indicates that the community's language evolves towards a more recognized while constantly changing norm. However, when looking at users in the context of the community, they stop to adopting the community's language change after roughly a third of their total time spent, or lifespan, in this community, regardless how long their lifespan actually is.

To test linguistic change's predictability for user lifespan, Danescu-Niculescu-Mizil et al. used several post-level measures, including cross-entropy of posts, Jaccard self-similarity between adjacent posts, adoptation of lexical innovations, first-person singular pronouns, and number of words as features for machine learning tasks. The experiments showed that those features improved the absolute F1 score by up to 12%.

2.1.2 Discussion

The paper of Danescu-Niculescu-Mizil et al. is an insightful investigation and extensive evaluation of linguistic change of both users and communities over time. Their proposed framework is a useful quantitative approach to study the interplay between user-level and community-level linguistic change, which we can employ in this project. Nonetheless, their work mostly focuses on user-level linguistic change under a context of constantly evolving community language norm. The authors did not explain the cause of such linguistic change. Potential causes of this phenomenon remain unanswered, which could be network-related factors such as new users fail to join older users' clique in the user interaction network. Future work is also needed to study on many other aspects of this problem at a more aggregate level such as network structure and topic shift as communities and users age concurrently, for which our project will pursue.

2.2 The Life and Death of Online Groups: Predicting Group Growth and Longevity [5]

2.2.1 Summary

In [5], Sanjay, et al. studied how a group's network features predispose its future growth. They considered two kinds of growth. The first one is diffusion growth, referring to the way by which a network attracts new members through the links the new members have with the old members of the network. The second one is non-diffusion growth, referring to that a new member is attracted by other factors of the network, such as the common theme or topic of the network, rather than through linkage with old members of the network. With this categorization, they proposed a hypothesis that resolves the seemingly contradicting fact observed in previous social science studies: High clustering in a network makes a fringe node of network(someone who is not a current member of the network but has link with members of the network) more likely to join the network while on the other hand negatively effect the network's overall growth rate. They resolve this by showing that though high clustering facilitate diffusion growth, groups whose growth relying more on diffusion tend to have smaller eventual sizes.

Then they formulated eight predictive models from three dimensions, using previous growth rate and the structural features of a graph to predict its final size and longevity. The output coefficients of these features coincide with and verified the discoveries they made in the first part.

At last, they hypothesized a core-periphery structure where groups have one or several highly connected cores and a periphery of members that are loosely connected or even entirely disconnected from the core. They proposed that such structure can best facilitates the network's growth in that it has tightly connected core that allows for swift information transmission while also have loose periphery to allow for structure holes in the graph and prevent the network from being too inwardly focused.

2.2.2 Discussion

The work of Sanjay, et al.[5] provides insight on factors that predispose the growth of the network. By categorizing growth of a graph into two parts, they provide a reasonable explanation of the contradictions encountered by previous social science study. The core-periphery structure they proposed gives a unified theoretical explanation of findings they made. However, their work didn't consider what will happen after the growth of a group, i.e. after a new member enters the community, which will be the focus of our project. In spite of this, their work sheds light on our project in the follow aspects:

1. The relationship between the structure of the old member's group and the response they have towards the new members;

2. The change of group's structure after a large influx of new member.

2.3 Group Formation in Large Social Networks: Membership, Growth, and Evolution [6]

2.3.1 Summary

Backstrom et al.'s work [6] focused on analyzing communities in a social network using network structural features, where each community is a subset of nodes in the network. In particular, the authors tried to investigate the following 3 questions.

1. Membership: what are the important features that determine whether a user will join a particular community?

2. Growth: what are the important features that contribute to the growth of a community?

3. Change: how does the topics of a community change over time, and how does it relate to users moving between communities?

To address those questions, the authors conducted the experiments based on two datasets: LiveJoural and DBLP. The LiveJoural dataset represents users within the free online blog post social network called LiveJournal, where users, friendship and communities are explicitly defined. The DBLP dataset consists of computer science publications, where each author is a node, edges are defined by co-authorship, and communities are defined as conferences.

The authors treated the Membership and Growth questions as binary classification problems, where they predicted the probability of each user joining each community, and the probability of each community growing significantly in a certain time frame. The authors trained the classification model using decision tree, and pointed out that the spliters near the root of the tree were the important features. The model indicated that whether a user will join a community is largely determined by the ratio of his or her friends who are in that community. Also, the growth of a community comes from number of users who has a friend in the community.

To address the Change question, the authors defined the topic of a community/conference in a particular year as the most frequently used words in papers accepted by the conference, and defined the movement of users between communities as authors publishing a paper to a conference in a year and publishing to another conference in a later year. The authors tracked the change of topics of a community and the movement of authors over years in order to see if they are aligned. The result demonstrated that the change of topics does not play a vital role in the movement of authors.

2.3.2 Discussion

Backstrom et al.'s work [6] addresses one of the most crucial components in social network – communities. Communities are more important today than the time when the authors wrote the paper 10 years ago, since social network has reached a completely different scale. This paper is the first work that investigates explicit communities in a social network using network structural features, and it indeed discovers some interesting properties regarding to communities. Though determining important features using decision tree is a simple but effective approach, other algorithms may give extra leverage and insight upon this problem. For example, logistic regression and neural network

can not only give a numerical weight to each feature, but also allow people to see whether each feature affect the result positively or negatively, by looking at the sign of the weight.

Investigating user transitions between communities from a macro perspective is far more complicated than predicting whether a user will join a community or whether a community will grow. This paper looked at this program through the angle of the relationship between topic transition within a community and user transition. Our project tries to investigate the same problem but with a different perspective: what's after users join a community, e.g., the interaction between new members and old members within a community. Backstrom et al.'s work can indeed help us understand the network patterns in user transitions between communities so that we are able to build our project on top of their work.

3 Proposal

3.1 Problem Formulation

Inspired by previous work and available data, we plan to take a close look at some popular (among top 100 by subscribers) defaulted and non-defaulted subreddits. We would like to answer the question "**what would happened to an online community after a large influx of new users?**", with a focus on latent networks underlying it. Specifically, we will look at the following aspects:

1. Structural properties change before and after default (overall graph properties)
2. Interaction between old and new users (components within the graph)

3.2 Data

In this project, we use Reddit comments data available on Google BigQuery[3]. We selected 10 defaulted and 10 non-defaulted popular subreddits from the top 100 subreddits as our potential data candidates. We will select a representative subset of them to conduct our research so that we can run more experiments given the limited time we have this quarter.

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 2012 to 2015. 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.

The data in Google BigQuery is stored in tabular form with fields listed in table 1.

3.3 Approach

3.3.1 Network Representation

Since there are no explicit friendship defined between Reddit users, we need to construct networks from the dataset. We will concentrate on the user interaction graphs. For a user interaction graph for a given subreddit S, we define it as

$$G =< V, E >$$

where

$$V = \{\text{all users who have posted in the subreddit from year 2012 to year 2015}\}$$

$$E = \{(v_1, v_2, w) \text{ if } v_1 \text{ and } v_2 \text{ have replied to each other's comments } \forall v_1, v_2 \in V\}$$

where w is the edge weight, defined as the count of such bidirectional replies

Alternatively, we can define an edge as (v_1, v_2) if v_1 and v_2 have replied to each other's comments more than ϵ times for some ϵ, in order to restrict our focus to strong interaction. Another interesting representation we can build is to define an edge as (v_1, v_2) if v_1 and v_2 have commented in the same post.

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

Name	Type
body	STRING
score_hidden	BOOLEAN
archived	BOOLEAN
name	STRING
author	STRING
author_flair_text	STRING
downs	INTEGER
created_utc	INTEGER
subreddit_id	STRING
link_id	STRING
parent_id	STRING
score	INTEGER
retrieved_on	INTEGER
controversiality	INTEGER
gilded	INTEGER
id	STRING
subreddit	STRING
ups	INTEGER
distinguished	STRING
author_flair_css_class	STRING
removal_reason	STRING

Table 1: Reddit comment data fields in Google Query

3.3.2 Network Monthly Snapshot

As we are studying the interaction graph change over time, we need to compare the graph's statistics and structures at different time. Consequently, we define the graph "snapshot" G_m for month m as the subgraph consisting of only interactions that took place in month m.

3.3.3 Statistics, Algorithms and Models

We intend to explore the change of interaction network with a series of statistics, algorithms and models, including:

1. Clustering coefficient
2. Size of old user set V_{old} and new user set V_{new}
3. Number of edges between old user set and new user set
4. Post topic (from LDA model) distribution divergence (measured by cosine distance, cross-entropy, Jensen-Shannon divergence, or Kullback–Leibler divergence) between V_{old} and V_{new}
5. Structural explanation of the reaction and structural change of the community
6. Neighborhood exploration algorithms (e.g. random walk) to see whether the result aligns with V_{old} and V_{new}

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.