CITS4403 Computational Modelling SEM-1 2019

Project Report-----Loyalty in Online Communities

Group:

1. Qianwen Lu (22167601)

2. Xin Wang (22167361)

3. Yuan Shi (22167488)

1. Introduction

Loyalty is an important part of network community. Now with the development of the Internet, there are more and more different kinds of online communities. When users browse and post comments on those various communities, they generally tend to be loyal to one community and casual to others. However, it is unclear how loyalty is expressed in user behavior, or what the characteristics of communities with large numbers of loyal users are.

According to the research, we found that although there are many articles on the relationship between users, there are few articles focusing the relationship between users and communities. A basic example of this relationship is loyalty. For example, on a multi-community platform like Reddit, users have many alternative communities to browse, which makes it less likely that users will be loyal to a community. However, the fact is that many users are always loyal to a particular community. Therefore, understanding the cause of this phenomenon has become a crucial issue.

The first objective of this project is to provide a comprehensive description of user loyalty on Reddit, a multi-community platform. To this end, a measure is proposed to evaluate user loyalty based on user preference and commitment. Loyal users prefer a

community over other communities and will maintain this preference for a long time. Reddit's large-scale, multi-community features make it an ideal place to study loyalty. By applying this measure, the loyal users and loyal communities is defined based on their characteristics. The second objective is to study which characteristics of community affect user loyalty. If we know the common traits of a high loyalty community, we can build a community that is more user-friendly based on these traits.

2. Data Processing

We get dataset from the webpage provided by the paper, http://snap.stanford.edu/data/web-RedditNetworks.html. Based on our objectives, we use the dataset "reddit_chain_networks.tar.gz", it is the networks constructed from comment chains.

The original data is saved in JSON format. By analyzing the dataset, we find that each file represents to a subreddit, the file name is the name of the subreddit. The total number subreddit is 2046, that means there are 2046 JSON files. According to the research, the dataset contains 10^8 comments made by 10^7 users. To implement our project, we choose 10 loyal communities and 10 non-loyal communities as our dataset when running our codes in Jupyter notebook. We will implement the experiment based on Section 4 of the research paper. Since there is no code can be found, we can just implement the experiment by our understanding.

After getting the dataset, we should process these data. There are several steps in data processing. We generate several folders to save the processing files.

- Firstly, we get the keys and values of each json file. Keys represents to the users who post a topic in this subreddit. Values refers to the users who comment on the topics. Each key corresponds to one or more values. Users are indicated as username.
- Secondly, we save all the keys and values to a CSV file and named the CSV file the same as the subreddit. Each file contains all the users of one subreddit. The

files are saved in “dataprocessing” folder.

- Thirdly, there should be some repetitive users in a file. One purpose of our study is to calculate a user’s loyal rate to a specific reddit. Therefore, we should have usernames and the number of times each username appears. In the rest of contents, we will use times to represent the number of times each username appears. We also create a CSV file to save usernames and times for each subreddit and it is also 2046 files. The files are saved in “userinfo” folder. Then we generate a “usernamelist” file based on all the files in “dataprocessing” folder and drop duplicates. The “usernamelist” is saved in root dictionary.
- Finally, we generate a comprehensive CSV file contains all the usenames and times of all the subreddits. The file is named as “usertimes” and it is saved in root directory. The columns means subreddits and rows represent users. This is an important file and several parameters are computed on this file.

3. Experiment

Firstly, we define a function to check whether a user is loyal to a community. The function is called “isloyal”. There are two input variables username and subreddit. In the code, we firstly get a whole row based on the input usename. The row contains all the times of the user in all the subreddit. When you input a username and a subreddit name, you will get the use’s loyal rate to this subreddit. If the loyal rate is greater than 0.5, it means the user is a loyal user to the subreddit. Figure 3.1 is an example of how to use the function and the output of the function.

```
In [7]: isloyal('Could_Care_Corrector','Silverbugs')

        Could_Care_Corrector has 22.0 comments in Silverbugs
        Could_Care_Corrector has 139.0 comments in all the Reddit community
        Could_Care_Corrector loyalty to Silverbugs is 0.16
        Could_Care_Corrector is not a loyal user to  Silverbugs

Out[7]: '0.16'
```

Figure 3.1 An example to “isloyal”

In addition, we define a function to detect whether a community is a loyal community. The function is called “isloyalcommunity”. It contains only one input variable: name of the subreddit. With a input variable, the function find out the subreddit. Then traverse all the usernames and call “isloyal” function to get a loyal rate. We have a counter, if the loyal rate is greater than 0.5, the counter plus one. After traversing, counter will be the output. If the counter is larger than 25, the subreddit is a loyal community. Figure 3.2 is an instance of how to use the function and the output of the function.

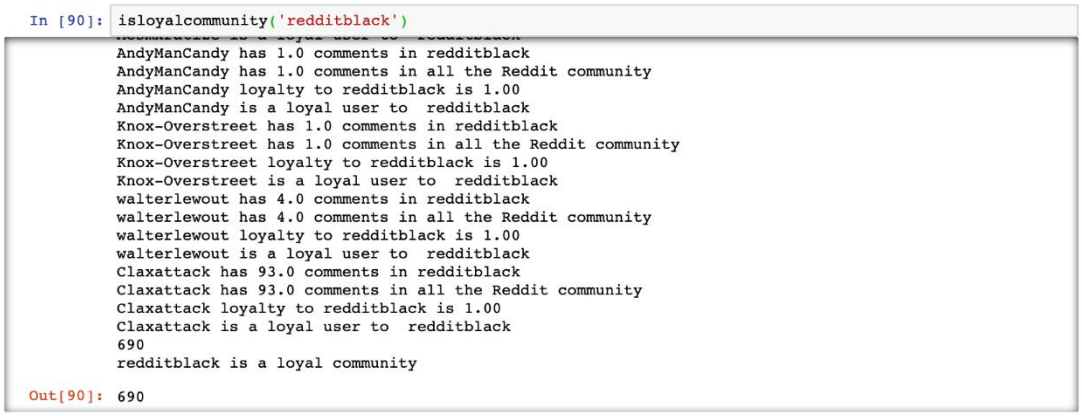

Figure 3.2 An example to “isloyalcommunity”

We implement two experiments and we find that loyal communities exhibit consistent structural features in their user-community interaction networks, and that these structural features are predictive of loyalty, across communities with vastly different topical interests.

For the first experiment, we choose two communities separately from loyal communities and non-loyal communities. We choose loyal: radditblack & non-loyal: Battlefield, both dedicated to particular video games, suggesting that differences in structural markers of loyalty exist between communities with similar topics. To generate graphs of the two community, we need to ensure nodes and edges of the graphs. We define two function as “nodes” and “edges”. In the “nodes” function, the input variable is the name of the subreddit. Since we just generate the graph based on the data in one month, we should do some parts of the data processing step to generate

a table contains usenames and times of this subraddit, which is saved in “eachmonuser” folder. Then nodes of the graph are the usernames of the subreddit. In addition, for the “edges” function, we read the original dataset directly and get the key_list and value_list for one month. For each element in key_list, it will correspond to one or more value. We set each value and each value in pairs and each pair is an edge. For example, “Sam” is an element in key_list, and the correspond value_list contains three elements “Jack, Tim and Tom”, there will be three pairs: (“Sam”, ”Jack”), (“Sam”, ”Tim”) and (“Sam”, ”Tom”). Therefore, we get two graphs. They are shown in Figure 3.3.

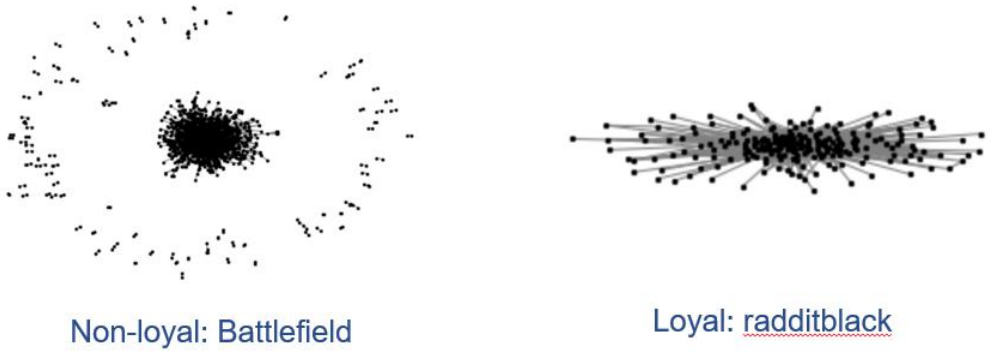

Figure 3.3 The example networks of loyal and non-loyal communities

We can easily find that peripheral nodes connect back to center in loyal network. However, in non-loyal network, peripheral nodes connect to each other. Furthermore, we calculate some structural features of the two communities: Density, Average clustering, activity assortativity. The results are shown as follow.

	Density	Average clustering	activity assortativity
radditblack	0.00262	0.597	-0.188669
Battlefield	0.07868	0.435	0.01605

Table 3.4 Some structural features of the two communities

Loyal networks have significantly higher edge density, which means that the average

user in a loyal community interacts with a greater number of other users. However, we do not find any significant difference in the average clustering coefficient for those two networks. The most obvious signal in the networks is a difference in activity assortativity. The differences in activity assortativity means highly active users tend to engage with others who have a wide variety of activity levels in high loyal communities, while in non-loyal communities users tend to comment near other users of similar activity levels.

Also, we generate the mean number of the degree for each network. It is shown in Figure 3.5. There is a significant difference between the networks. Loyal community has a very high mean degree, which means the user in loyal community in more activity. Then I also compute the PMF for the network. The graph in shown in Figure 3.6. Based on the graph, we realized that the networks also follow power law and in heavy-tailed distribution.

```
np.mean(degrees(graph_nonloyal)), np.mean(degrees(graph_loyal))
```

```
(4.412811387900356, 17.938864628820962)
```

Figure 3.5 The mean of degree for each network

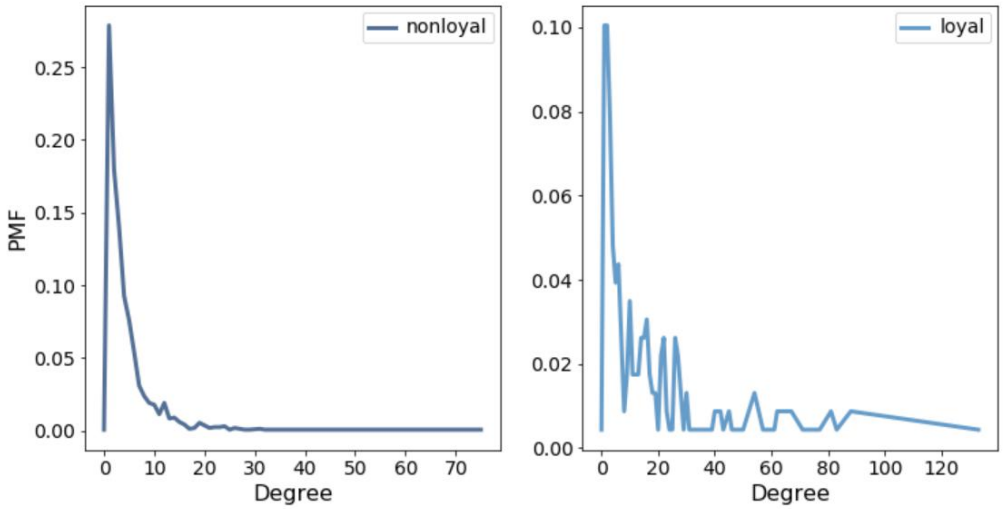

Figure 3.6 The distribution of PMF for each network

In contract with the original paper, we find nearly the same results. In that paper, the

function of wiring edges is more complex than our function. The authors explained in the paper that users are connected if they comment in the same linear comment chain within three comments of each other. The graph they generated is shown in Figure 3.7.

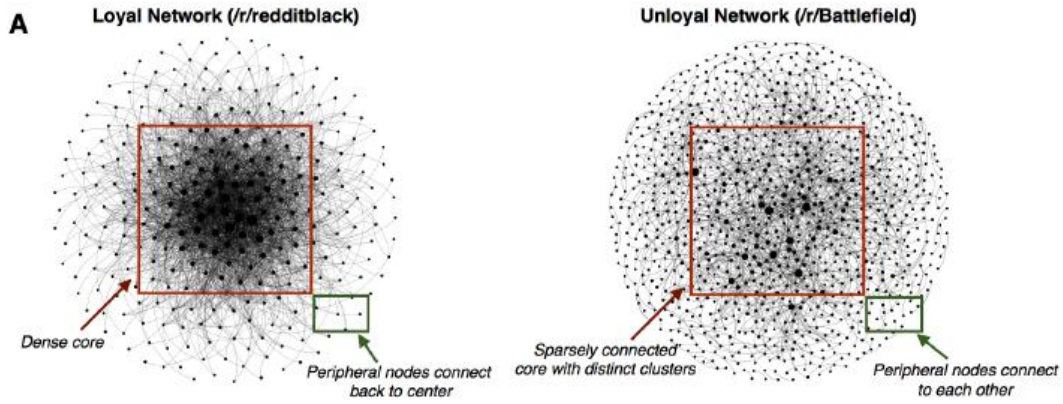

Figure 3.7 The authors' graph

Our graphs seem quite different from the authors' graphs. However, our graph of loyal community also appears peripheral nodes connect back to center, and peripheral nodes connect to each other in non-loyal community graph. For the statistic features, we get nearly the same tendency with the authors. In addition, we generate the degrees of the networks, which could support our result that the users in loyal community is more activity.

4. Discussion

Conclusion. In this report, firstly the loyalty is characterized at the community level. It is denoted that there are often more loyal users on topics of common hobbies, such as sports and video games. In addition to community topics, there are some social network diagrams that can be used to identify communities with high loyalty rate. We also find that loyal communities tend to have more intensive user-user interaction networks. And the loyal community's interaction network also includes more bridge relationships between active and inactive users. Overall, the findings suggest that more inclusive and cohesive communities tend to achieve higher user loyalty.

Discussion. In this project, loyalty is oprationalized as a relation between users and communities. In this way, we can understand how loyalty is reflected in the interaction structure between users and how it is reflected in consistent user behavior in different communities. Ultimately, we can use this consistency to predict user loyalty in various communities and build a more user-friendly community.

Also, this project can have an impact on community maintenance. The findings highlight the important role that loyalty plays in community development. We found that the difference in behavior between loyal users and vagrant users was consistent both in the content they generated and in the content they participated in, and it was discovered that these differences occurred early in the user interaction with the community. Therefore, if community maintainers want to turn vagrant users into loyal users, an entry point needs to be carefully designed. More generally speaking, the consistent difference between the two users suggests that for communities that are already mature, the maintainers should focus on maintaining content that appeals to core loyal users, instead of trying to attract outside users.

5. Future Work

So far, we have studied user-community relationship and have demonstrated that loyalty communities exhibit a pattern of features in the interactive network between users.

Next steps. What we haven't studied yet is the typical behavioral pattern exhibited by loyal users at the individual level. According to the articles we have studied, the behavioral characteristics of loyal users have been reflected in the language and the types of topics discussed in his initial contribution to the community. Therefore, we can use this feature to predict user loyalty, which is very important for community maintenance.

Bigger picture. In addition to the characteristic behavior at the individual level, there is a longer-term work waiting for us to do, which is to verify the applicability of the research results on platforms other than Reddit.Many of these platforms offer a variety of content, such as rewards and reputation mechanisms, to increase user loyalty. Further analysis can explore the impact of these different mechanisms on driving and shaping loyalty.

Reference:

[1] Hamilton, William et al. “Loyalty in Online Communities.” arXiv.org (2017): n. pag. Web.

[2] https://cs.stanford.edu/people/jure/pubs/loyalty-icwsm17.pdf

[3] Dataset: http://snap.stanford.edu/data/index.html

[4]Downey, Allen. et al. Think Complexity. 1st ed. Sebastopol, Calif: O′ Reilly, 2012. Print.