Loyalty in online communities

Dveep Chaudhari (201501172)* and Smit Thakkar (201501440)†
Dhirubhai Ambani Institute of Information & Communication Technology, Gandhinagar, Gujarat 382007, India
IT-454, Complex Networks

I. ABSTRACT

In multi-community platforms, loyalty of user to a particular community is a crucial measurement in this age of internet. When given multiple resources of content, user tends to be loyal to only few of them. Reddit is an one such online platform where people share information in related subreddits. Users subscribe to different subreddits based on personal interests and stays loyal to few of them. Despite topical interests, there are other structural parameters in a user interaction graph which governs the loyalty of a user towards a particular community. In this paper, we are analyzing the effect of those such parameters on loyalty of a community. User is loyal to a community if he prefers one community over a period of time. Community's loyalty is measured by the retention of loyal users. User interaction graph is a network of users where, users are connected if they directly reply to each other's comment. We used data of 2046 subreddits for 11 months for the experimentation. From the results, we concluded that communities with higher interaction, higher clusters, lower assortativity and higher inequality in terms of content distribution tends to be more loyal.

II. INTRODUCTION

This era of internet and productivity of people has led to an outburst of communities to explore. Raw data is organised and served to mankind using various algorithms, concepts and sheer commitment. This results in big communities, data giants and content providers such as Facebook, Youtube, Instagram, Reddit, Twitter and many more. With these many choices and limited time, a user cannot allot every bit of time to every community and chooses to allocate this limited proportion of time to specific communities only.

A user chooses to explore this wide range of options and it is up to the user to either commit and be loyal to the community or look out furthermore. A user usually prefers to stay and be loyal to a specific community or two and splits their frames of time on the internet accordingly. There could be various reason behind this time division and priorities such as; a community might attract the users area of interest, a communitys way of presenting hot topics, latest news and ongoing trends and a structure of a community.

A structure of a community can affect the loyalty of a user to that particular community. Interaction within a community plays a major role in building a users loyalty and trust towards that community. The interaction between the new users and the old ones should be like a constant flow of breeze. Formation of close groups interacting only with each other and their low outcomes and outputs towards the community can sabotage a community and loyalty of a community drops. Inequality in terms of user's activity distribution over a community affects the community's growth and user retention.

Our aim is to show user loyalty and loyalty of a community through a multi-community platform, Reddit. Reddit is a collection of entries submitted by its users. The site's content is divided into sections called subreddits or sub-communities. Subreddits are user-created areas of interest. We have analyzed the structural properties of the monthly user to user interaction of different subreddits graphs and how the structural parameters are related to the loyalty of communities. The five structural parameters used as a basis to judge are Density, Clustering Coefficient, Assortativity, Number of triads formed and inequality.

Further sections describe the dataset, the definitions, the methodology, the final conclusion. Section 2 describes the dataset used. This dataset is collection of monthly user interactions. Section 3 describes the definitions of terms widely used in this paper and their theoretical and practical implications. The Methodology section is the focal point of this paper which acutely describes the study with empirical results, discussions, graphs, explanation of structural parameters and comparisons between loyal and unloyal users and communities through these parameters. Next Section ends the paper with concluding remarks.

III. DATASET

For the purpose of analysis on loyalty measurement of a user and communities, we have used content from a multi-community platform, Reddit. Reddit is a social class website where each user posts an image, video, url, text, etc. on topical communities called subreddits. User can comment on posts and reply to each others comment in a thread-based interface. Votes, comments on the posts caters rewards in form of upvotes and downvotes from the user community. There are ample of sub-

*Electronic address: 201501172@daiict.ac.in
†Electronic address: 201501440@daiict.ac.in

reddits with different topics covering games, sports, science, nature, news and many more. Every topic contains variety of subreddits with plenty of users. Reddit therefore; provides a large dataset to work our way through. As a user with limited time and so many communities, he stays loyal with only a few over a long span. And so Reddit serves as a perfect dataset with multi-community structure and serves the purpose of researching user and community loyalty.

The dataset of user-user interaction of reddit contains information about the comments made by the user over the year 2014, from January to November. These includes data of total 2046 subreddits for 11 months of year 2014. For each subreddit there are two types of user interaction given: first is chain network where user is connected to other user if they have commented in a linear manner within three comments of each other (i.e., separated by at most two comments) where we assumed that user who have commented this much closely are highly likely to interact with each other. The second dataset contains reply network where two users are connected if they one has directly replied to others comment. Each dataset contain total 3 million users and 10^8 comments.

In this paper, we have analyzed our experiment only on reply network considering the stringent computational power.

IV. DEFINITION

Loyal user: In a multi-community platform, user has many options for information gathering and sharing. Though the availability of contents and chance of sharing from different communities, user prefers to spend time with only few communities. A loyal user of a community is a user who prefers the same community over time. For practical usage, we define loyal user in different terms. User X prefers community A at month t if he writes maximum number of comments in community A while considering his comments in all the communities at month t. If user prefers community A in month t and also preserves his preference for community A at month t+1 then we say that this user X is loyal user to community A.

Loyal community: Loyal community is a community which retains higher percentage of users over months. As number of active users differs from community to community, it is not a good measure for loyalty of community and we rather choose to look our for proportion of users community can preserve over time.

Loyalty rate: Loyalty rate of community is defined as the expected proportion of users who stays loyal at month t and also shows loyalty at month t+1. We ignore the user who leave the reddit at month t+1 because we want inter-community loyalty.

User interaction graph: In a reply network, every user is a node and if a user A directly replies to the comments of user B in a given month then we consider a directed edge between user A and B, pointing from A to B. This creates a directed graph with number of nodes equal to active users and number of edges equal to number of replies.

In this paper we have analyzed few structural properties of user interaction graph which includes density, clustering coefficient, assortativity, number of triangles. And at last we have considered activity distribution in a given graph by calculating ginis coefficient - inequality coefficient. Definitions of these terms are given below.

Density: Density of a directed graph is a proportion of possible edges present in a given graph.

$$d = \frac{E}{V(V-1)}$$

Clustering coefficient: clustering coefficient is the proportion of triplets which are closed triplets -3 x triangles.

$$C = \frac{Number\ of\ closed\ triplets}{Number\ of\ triplets}$$

Assortativity: Assortativity, or assortative mixing is a preference for a network's nodes to attach to others that are similar in some way. The assortativity coefficient is the Pearson correlation coefficient of degree between pairs of linked nodes. The assortativity coefficient is measured by eq. 21 in [3].

Inequality: Ginis coefficient measures inequality in frequency distribution of comments/ activity in a given community. This coefficient will range between 0 to 1, where 0 represents totally equality and 1 represents totally inequality. Ginis coefficient can be measured as half of the relative mean absolute difference.

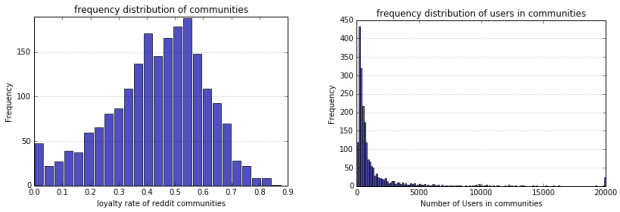
$$g = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n\sum_{i=1}^{n}x_i}$$

V. METHODOLOGY

In a reply network database, we have data of total 2046 communities spanning over 11 months of year 2014. We found the loyalty rate of each community by first gathering the list of loyal users of each community at each month followed by counting of proportion of user who stays loyal over succeeding months. This gives us the loyalty rate of all communities. Fig 1 shows the distribution of loyalty rate of all communities.

Structural parameters of the user interaction graph provides an insight about the loyalty of the communities. For that purpose we found the structural parameters of the user interaction graph. The parameters measured were density, clustering coefficient, assortativity, number of triangles and inequality. Fig 2. shows the distribution of nodes and edges in all the user interaction graph user interaction graph for 11 months. This results in total of 2046x11, which is equal to 22506 user interaction graph. For all the above generated graph we then measured structural parameters mentioned earlier. For the ease of analysis we then took average of parameter values over all 11 months for each community.

We further analyzed structure of the loyal communities and unloyal communities. The correlation between the structural parameters and the loyalty rate of the community has been found by measuring correlation coefficient. The correlation coefficient reveals the information on how structural parameters affect the loyalty rate of community. Finally we have made a comparison between the values of parameters of loyal community and unloyal communities and measured the significance of each parameter in determining the loyalty rate. We have considered loyal communities which have loyalty rate higher than the median loyalty rate whereas unloyal are the remaining.



(a) (b)

FIG. 1: Frequency distribution of loyalty rate and users in reddit communities

VI. RESULTS

Fig 3 shows the loyalty rate of all 2046 communities in decreasing order considered over 11 months. There are many subreddits which share similar topic but lies in a different region of loyalty rate. One such example of this scenario is online video games based subreddits r/redditblack and r/Battlefield. Both the communities have content of war related games but their loyalty rate are .82 and .31 respectively. This provides a perfect example of how structural properties of user interaction graph affect loyalty despite sharing the same topic of information.

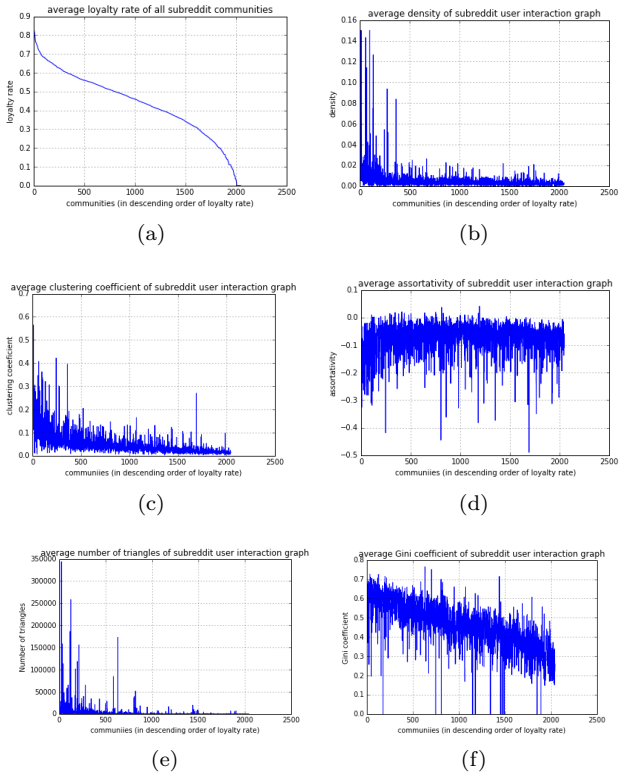
Fig 4A-E shows the density, clustering coefficient, assortativity, number of triangles and inequality for all communities. The communities are taken here on descending order of their loyalty rate. We can see from the graph that density is higher for communities with greater loyalty rate and decreases with reduction in loyalty rate. The same scenario stays for the clustering coefficient. Community with higher loyalty rate shows high clustering coefficient compared to communities with lower loyalty rate. The values of assortativity are relaitvely consistent with the change in loyalty rate. There is only slight increase in the assortativity values with decrease in loyalty rate. The number of triplet shows the same behaviour as clustering coefficient as both of them represents almost similar information about graphs. Finally the ginis coefficient coefficient of inequaltiy shows a difference in value for loyal and unloyal communities.

structural parameters	CC with Loyalty rate
Density	0.2301257
Clustering coefficient	0.558415
Assortativity	-0.092739
Inequality	0.7402283
Number of triangles	0.208842

TABLE I: Correlation coefficient of different structural parameters of user interaction graph and loyalty rate

Ginis coefficient steadily decreases with the loyalty rate unlike other predictors density, clustering coefficient and number of triangles which shows exponentially decay or sudden decrease with the declining of loyalty rate.

Fig 5A-D shows the comparison of the parameters values for loyal and unloyal communities. Fig 5A represents the difference between density values, 5B represents clustering coefficient, 5C represents assortativity and 5D



(a) (b) (c) (d) (e) (f)

FIG. 2: A. Loyalty rate of communities in descending order. B-F. Density, clustering coefficient, assortativity, number of triangles and inequality of communities with their decreasing order of loyalty rate

represents inequality. Further, in table 1 we have shown the correlation coefficient of all the parameters with loyalty rate. With highest correlation coefficient stays for inequality and only assortativity has negative correlation with loyalty rate but with negligible magnitude. All other parameters have positive correlation with loyalty rate with somewhat significant manner.

VII. ANALYSIS

Positive correlation of density with loyalty rate shows that user interaction network of loyal communities tends to have higher density compared to unloyal communities. Density represents the number of edges presented in a graph and edges in a user interaction graph represents a comment made by a user thus, higher density means

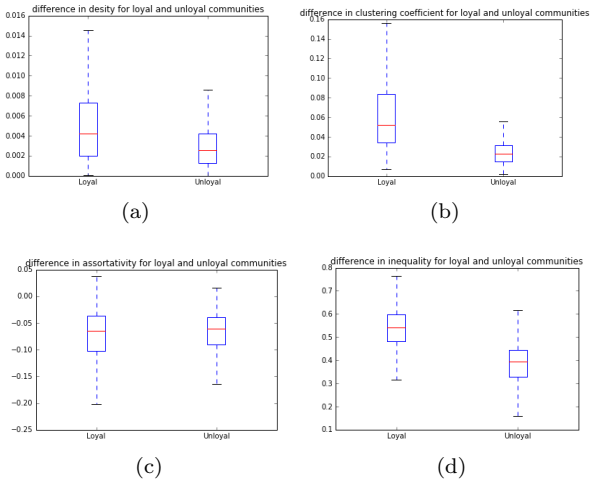


FIG. 3: Difference of density, clustering coefficient, assortativity and inequality between loyal and unloyal communities

higher interaction between users in a community. When interaction between user increases, users become more active in a community and chain of comments increases the chances of user to stay more active and thus more loyal community.

The clustering coefficient shows higher correlation with loyalty rate. This shows the importance of clustering in online communities. Higher clustering coefficient represent higher number of close groups in network. In loyal communities, user tends to bind with each other, making small groups inside community makes them to know some users more personally, enhancing their chance of being loyal to that community. Number of triangles provides the same insight for user interaction graph.

Assortativity of a graph is measured on structural bases considering degree of the nodes. Assortativity shows the preference of node to connect with others which are similar to itself. In user interaction graph, if a active user who made most comments in graph tends to reply on other active users comments and same for the case of non-active user then we see a positive assortativity in graph. For all the communities, assortativity seems to be unaffected by loyalty rate. Most assortativity values lies in range of -0.2 to 0.1, indicating that most users in online community dont have any preference for their comments and connects to any user without considering his activeness. But there is a slight difference in assortativity values for loyal networks. Loyal networks has little negative assortative value, indicates that active user replies to non-active users, making non-active users little proactive towards their contribution and thus increases the interaction between users, which leads in higher loyalty rate.

Ginis coefficient is one another parameter used for our analysis which measures inequality in a certain distribution. In this paper, we have measured the inequality in the frequency distribution of comments of certain graph. If all the users in a given community comments nearly equal number of times then ginis coefficient will get 0 value. If only few users have their shares on most of the comments then in this case inequality coefficient will be closer to 1. This inequality coefficiemt has the highest correlation coefficient with loyalty rate, which dictates that in loyal communities there are few users who comments most of the time, while many users comments less frequently. The users who comments more, shares more contents in community while those who comments less are the users who receive those contents. In the case of Reddit, most of the people join a community to gather the information, not to upload it or share it. While only few users provide the most proportion of content, leading to the perfect balance of users and thus increases the loyalty of community.

VIII. CONCLUSION

In this paper, we have analyzed the loyalty of a user in multi-community platform. Despite the bias for a topic, there are certain structural parameters of a graph of user interaction in a given community which affects the loyalty of community. Loyal communities are the one which can retain their users over long period of time. Communities which have higher interaction tend be more loyal. Communities in which users tend to make cluster are highly likely to have high loyalty rate. In loyal communities, active users have more preference to have interaction with in-active user compared to unloyal community. At last, in loyal communities, there is a set of few active users who shares most of the content while other bigger set of inactive users just receive those contents and this mixture makes the community to retain their users. In sum, in this paper we provide a greater insight on how structural properties of user interaction graph is correlated to loyalty of community and up to which rate these parameters affect the loyalty rate of community.

[1] SNAP: Web data: Reddit interaction networks *http://snap.stanford.edu/data/web-RedditNetworks.html*

[2] Loyalty in online community : *http://adsabs.harvard.edu/cgi-bin/bib_query?arXiv* 1703.03386

M. E. J. Newman, Mixing patterns in networks, Physical Review E, 67 026126, 2003