IEICE TRANS. COMMUN., VOL.E104–B, NO.7 JULY 2021

PAPER Special Section on Future Directions of Research and Development on Communication Quality

Analyzing the Effects of Social Network Structure on the Growth and Survival of Online Communities in Reddit

Sho TSUGAWA[†a)] ***and*** **Sumaru NIIDA**[††], ***Members***

SUMMARY While online communities are important platforms for various social activities, many online communities fail to survive, which motivates researchers to investigate factors affecting the growth and survival of online communities. We comprehensively examine the effects of a wide variety of social network features on the growth and survival of communities in Reddit. We show that several social network features, including clique ratio, density, clustering coefficient, reciprocity and centralization, have significant effects on the survival of communities. In contrast, we also show that social network features examined in this paper only have weak effects on the growth of communities. Moreover, we conducted experiments predicting future growth and survival of online communities utilizing social network features as well as contents and activity features in the communities. The results show that prediction models utilizing social network features as well as contents and activity features achieve approximately 30% higher F_1 measure, which evaluates the prediction accuracy, than the models only using contents and activity features. In contrast, it is also shown that social network features are not effective for predicting the growth of communities.

key words: *online community, social network, Reddit, growth, survival*

1. Introduction

People join online communities for several purposes, such as sharing knowledge, socializing with each other, obtaining support for health-related issues, and developing software [1], [2]. These online communities serve as important platforms for both work-related and personal concerns [1]. The proper way of operating an online community is an important theme in the design of network services.

What makes an online community successful? Why do some communities successfully grow over time while others do not? The answers to these fundamental research questions have important implications for administrators managing their communities, for members wanting to join active communities, and for designers of online community platforms wanting to provide better services to users [3]. Unfortunately, many online communities die in the early stages of their lives [4]. For instance, on Facebook, users create over 100,000 new communities per day, but 13% of them produce no content after the first day and 57% of them have stopped all activity within three months [4]. Therefore, considerable effort by researchers in several disciplines has been devoted to clarifying factors affecting the success of online communities [4]–[13].

Thanks to the efforts of many researchers, several key factors affecting the success of online communities have been revealed. Example key factors include community size [14], diversity [5], member turnover [15], leadership [6], membership, and topic overlaps with other communities [7], [16]. However, since the success of online communities is a very complex phenomenon, we have only a partial understanding of its mechanisms, and analyzing the factors determining the success of online communities is still a hot research topic [1], [13], [15]–[17].

This paper continues the line of above-mentioned studies aimed at understanding the factors affecting the success of online communities, and comprehensively examines the effects of social network structure on the success of an online community. Among several aspects of the success of online communities, we particularly focus on two aspects of success: growth and survival, which have been actively studied in existing studies [8], [10], [13], [16], [18]. Analysis of a social network of individuals is useful for quantifying the characteristics of their communication patterns [19], and therefore, the effects of some features of the social networks of community members on the growth and survival of online communities have been already investigated [10], [18]. However, there still remain several uninvestigated features. A seminal work by Backstrom et al. [18] investigated social network features affecting community growth. The investigated features included the clustering coefficient [20] and the number of links in the social network of the community. Subsequently, Kairam et al. [10] examined the effects of density, clustering coefficient, clique ratio, and the fraction of the giant component on the growth and survival of communities. This paper extends the work by Kairam et al. [10], and makes an exploratory examination of the effects of a wide variety of social network features, which include centralization [21], features related to cluster structures [22], and features of the dyadic relation between nodes [23], [24].

We particularly focus on the popular online discussion forum Reddit*, and address the following three research questions related to the growth and survival of Reddit communities.

(RQ1) What are the structural characteristics of social networks affecting growth?

Manuscript received August 28, 2020.
Manuscript revised November 27, 2020.
Manuscript publicized January 8, 2021.

[†]The author is with Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba-shi, 305-8573 Japan.
[††]The author is with User Innovation Group, KDDI Research Inc., Fujimino-shi, 356-8502 Japan.
a) E-mail: s-tugawa@cs.tsukuba.ac.jp

DOI: 10.1587/transcom.2020CQP0006

*https://www.reddit.com/

(RQ2) What are the structural characteristics of social networks affecting survival?
(RQ3) How effective are social network features for predicting future growth and survival?

To answer these questions, we investigate the relation between the social network features of Reddit communities and their future growth and survival. Reddit consists of several *subreddits*, each of which is a forum for discussing a certain topic. We regard each subreddit as a community, and users who post messages and comments as community members. For each community, we construct a social network that represents commenting relationships in the community. We then investigate the relation between the features of the social network constructed for each community and its future growth and survival.

Our main contributions can be summarized as follows.

- We comprehensively investigate the effects of 15 social network features on the growth and survival of online communities. To the best of our knowledge, this is the first study to systematically investigate the effects of such wide variety of social network features.
- We reveal social network features that significantly affect the survival of online communities. Because we consider a wide variety of features, we find features not investigated in the previous studies that have significant effects.
- We construct classifiers to predict the growth and survival of online communities using network features as well as features related to contents and activities in the communities. Consequently, we demonstrate the effectiveness of social network features for survival prediction when combining them with activity and content features.

This paper is an extended and polished version of our previous conference paper [25]. We have added new results for growth and survival prediction using content features to clearly show the effectiveness of network features for predication. We have also added discussions related to the new results.

The rest of this paper is organized as follows. Section 2 gives a literature review. Section 3 introduces the dataset and methodologies for the analyses. Section 4 presents the results and gives answers to the three research questions. Section 5 discusses the implications of the results and the limitations of this study. Finally, Sect. 6 concludes this paper.

2. Related Work

Research of online communities can be categorized into two main categories: research on the factors affecting the behaviors of *individual* members in a community [9], [26]–[32] and research on the factors affecting the success of *communities* [1], [3]–[8], [10], [12], [14]–[18], [33]–[35]. The second category can be further divided into two categories: research focusing on *intra-community* features [1], [3]–[6], [8], [10], [12], [14], [15], [18], [33]–[35] and research focusing on *inter-community* features [7], [16], [17]. This study focuses on intra-community features affecting the success of online communities.

Analyzing the behavior of individual members in online communities is currently a hot research topic. Since continuous participation and contributions by individual members are important for the growth and survival of online communities, factors that motivate individual members to participate in communities have been extensively investigated [9], [26]. Because gaining new members and encouraging their continuous participation are also important, several studies particularly focus on new members and analyze the factors affecting the longevity of their participation [27]–[29]. Members who only consume the content of a community and do not contribute to the community are called *lurkers*, and the existence of such members is considered to be a problem for building sustainable online communities [30], [31]. Therefore, why members become lurkers has been investigated [30], and methods for detecting lurkers among community members have also been proposed [32].

Another line of studies investigates intra-community features that affect online community success. Butler [14] investigated the effects of activity and membership size on the sustainability of a community. Zhu et al. [6] examined the effects of leadership, and Kraut and Fiore [4] examined the role of the founder's activities on online community success. In WikiProjects, diversity [5] and member turnover [15] have been shown to affect the productivity of communities. Sharma and Choudhury [33] investigated linguistic features affecting social support in online mental health communities. Many other intra-community features affecting online community success have been discussed in survey papers [1], [3].

Among the intra-community features, the importance of social network features on the success of communities has been suggested [36], and the effects of some social network features have already been empirically shown [8], [10], [12], [18], [34], [35]. Seminal work investigating factors affecting community growth was conducted by Backstrom et al. [18]. They investigated the effects of the clustering coefficient and the number of links in a social network on the growth of online blogging communities and communities of researchers. Ducheneaut et al. [8] investigated the effects of several network features on the survival of online gaming communities. The features they considered included community size, density, centrality, the size of the giant component, and the number of connected components. Following these studies, Kairam et al. [10] investigated the effects of several network features (density, clustering coefficient, clique ratio, and the fraction of the giant component) on the growth and survival of communities in online social networking services. Patil et al. [12] investigated the effects of the clustering coefficient and degree distribution on the stability of online gaming communities and communities of researchers. Our work follows the line of these studies, and

investigates the effects of 15 social network features, to be introduced in Sect. 3, on the growth and survival of communities. While existing studies examine the effects of a few social network features on community success based on some hypotheses, we explore the effects of a wider variety of features.

While most studies, including this study, focus on intra-community features, recently, the effects of inter-community features on community success have been investigated. Zhu et al. [7], [16] investigated how the success of a community is influenced by relations with other communities. They showed that membership overlap [16] and topic overlap [7] with other communities have significant effects on the success of a community. Vincent et al. [17] investigated the relationships between content in Wikipedia and two other online communities (Reddit and Stack Overflow). Although these studies focus on a different type of factor to that in our work (i.e., external rather than internal factors), both approaches are necessary for a better understanding of online community success.

Recently, Reddit has been attracting the attention of many researchers because Reddit is a unique platform that is used by many users worldwide for a wide variety of purposes [37]–[39]. Choi et al. [40] analyzed conversations on Reddit, and found factors affecting virality. Lin et al. [11] investigated changes in Reddit communities after a sudden growth of the number of community members. Liang [41] constructed a model for predicting thread ratings in Q&A communities on Reddit. Buntain and Golbeck [42] proposed a method for identifying the roles of users on Reddit from their position in the social network structure. Although, many studies have analyzed Reddit data, to the best of our knowledge, this is the first study investigating factors affecting the growth and survival of Reddit communities.

3. Methodology

3.1 Dataset

For this study, we used publicly available comment data from a popular online discussion forum, Reddit [43]†. Reddit users can create subreddits, each of which can be considered as a community. In each subreddit, each user can make a post, comment on a post, or comment on another comment using a bulletin board system. From the comment data, we can extract commenting relationships between members, which can be used to construct social networks for each Reddit community. Thus, Reddit data can be used to investigate the relation between the social network structure of a community and its growth and survival.

From the available data, we determined target communities for analysis. Since the dynamics of community evolution is expected to be different between communities created in the early days when Reddit was just launched and those created in the mature stage of Reddit, we decided to analyze communities created during a restricted time frame to eliminate the effect of the community creation period. To analyze the growth and survival of communities, we require an observation period of a certain length. Considering this requirement, the time frame was determined to be the 6-month period from January 2013 to June 2013. Note that the creation date of a community is when the first comment is made by someone. From the comment dataset, we obtained all comments posted in all communities between January 2009 and November 2017. We then extracted communities created during the 6-month period from January 2013 to June 2013. We excluded communities where comments are observed in only one month or the number of comments is less than 100. This process gave 4,823 communities, all of which were the target communities used in the following analyses.

3.2 Overview of Analyses

We take monthly snapshots of each community and use them to analyze the relationship between social network features and measures for quantifying the growth and survivability of communities. Following Kairam et al. [10], we investigated the 2nd-month and 6th-month snapshots. Here, the m-th month snapshot of a community means the m-th month snapshot after the community is created.

We adopt ordinary least squares (OLS) regression analysis for investigating the effects of social network features on the future growth and survival of communities (RQ1–2). The dependent and independent variables will be explained in the following subsection. We then construct models to predict the future growth and survival of communities from the independent variables using machine learning, and investigate the prediction accuracy of the constructed models (RQ3).

3.3 Network Construction

For each snapshot of each community, we construct a social network representing relationships among community members. A social network is represented as a directed unweighted graph $G = (V, E)$. A set of nodes V represents the set of members who posted comments in the community during the month. A directed link $(u, v) \in E$ represents that member u comments on a comment or a post by member v. Following [10], the frequency of comments between members is simply ignored, so the social network is represented as an unweighted graph. Note that members whose posts do not receive any comments are not included in the social network. Also note that for calculating some social network measures defined for undirected networks, we ignore link direction.

†We obtained the dataset from the URL https://files.pushshift.io/reddit/comments/.

3.4 Measures

3.4.1 Dependent Variables

As the dependent variable of the regression analysis for community growth, we use the N-month growth rate for the m-th month. The N-month growth rate of a community for the snapshot of the m-th month is defined as

$$r(N,m) = \frac{\frac{c(m+N)-c(m)}{N}}{\frac{c(m)}{m}}, \quad (1)$$

where $c(m)$ is the total number of comments posted in the m months since community creation. This measure is the ratio of the number of comments during N-months period after the m-th month divided by the number of comments during the period before the m-th month. Since the lengths of the two periods may be different, the number of comments are normalized by the lengths of the periods. The growth rate measure in this paper follows that used by Kairam et al. [10]. While Kairam et al. focus on the growth rate of the number of subscribed community members, this study focuses on the actual activity of communities and, therefore, uses the number of comments rather than the number of subscribed members for quantifying community growth. We consider that a community with a lot of inactive members who do not post comments is not an active community whereas a community with a small number of active members who post many comments is an active community.

As the dependent variable of the regression analysis for community survival, we use *remaining lifetime*. The remaining lifetime of a community for the snapshot of the m-th month is defined as the number of months where at least one comment is observed after the snapshot is taken.

3.4.2 Independent Variables

Independent variables related to social network structures used in the regression analysis are shown in Table 1. These measures have been widely used for quantifying the structural characteristics of social networks (e.g., [44], [45]) and are selected from survey papers on the measurement of complex network structures [23], [46]. The relation of the variables *density*, *clustering*, *clique*, and *GC* to the growth and survival of online communities has been investigated by Kairam et al. [10]; the other variables have not previously been considered in this context. When calculating *modularity*, *num. cluster*, *diameter*, and *path*, we considered only the largest component of a graph and link direction was ignored (i.e., the graph was treated as an undirected graph). For obtaining *modularity*, and *num. cluster*, a popular cluster detection algorithm called the Louvain algorithm [47] was used. When calculating *ave. deg.*, and the three centralization measures, we ignored the link direction.

As further independent variables of the regression analysis, we also use basic features quantifying the activity-level of communities: *num. comments*, *num. members*, and *growth rate*. For the m-th month snapshot, the variable *num. comments* is defined as the number of comments posted in the month, the variable *num. members* is defined as the number of members who post at least one comment in the m-th month, and the variable *growth rate* is defined as $(c(m) - c(m-1))/(c(m-1) - c(m-2))$.

In addition to each feature for the m-th month snapshot, the average of the feature in the past m months is also used as an independent variable. Therefore, we have 36 independent variables: 15 social network features, 3 basic features for the m-th month snapshot, and the past averages of these 18 features.

In the following analyses, we excluded communities with less than 10 nodes in the network snapshots since it is meaningless to calculate network features of such small social networks. Several statistics of the communities used in the following analyses are shown in Table 2.

Table 1 Social network features used as independent variables of the regression analysis.

Variable label	Description
ave. deg.	average degree
density	edge density
clustering	clustering coefficient [20]
clique	the fraction of nodes belonging to the largest clique [10]
modularity	modularity [22] obtained with the Louvain algorithm [47]
num. cluster	number of clusters obtained with the Louvain algorithm [47]
GC	fraction of the giant component
num. comp.	number of connected components
assortativity	degree assortativity [24]
reciprocity	reciprocity [23]
diameter	maximum of shortest path lengths
path	average of shortest path lengths
deg. cent.	centralization based on degree [21]
bet. cent.	centralization based on betweenness centrality [21]
clo. cent.	centralization based on closeness centrality [21]

Table 2 Statistics of the target communities.

	2nd month	6th month
Number of communities	1226	926
Ave. number of comments	702.6	1201.4
Ave. number of nodes	76.8	126.7
Ave. 3-month growth rate	1.29	1.58
Ave. 6-month growth rate	1.63	1.21
Ave. remaining lifetime [month]	30.4	33.8

4. Results

4.1 Features Affecting the Growth of Online Communities

We first tackle **RQ1**: *What are the structural characteristics of social networks affecting growth?* To answer this question, regression analyses were conducted. The dependent variables were 3-month growth rate and 6-month growth rate. For constructing the regression models, backward stepwise linear regression based on Akaike's Information Criterion (AIC) was used since the independent variables are correlated with each other.

Tables 3 and 4 show the results of the regression analyses. The regression coefficients shown in the tables are standardized. Dashes indicate that the variable was not selected

Table 3 Results of regression analyses for the 3-month growth rate (*:$p < 0.05$, **:$p < 0.01$).

Dependent variable: 3-month growth rate		
	2nd month	6th month
	Std. Coeff.	Std. Coeff.
num. members	-	0.0603
density	-0.254**	-
density (ave.)	0.354**	0.0931*
clustering (ave.)	-0.112**	-0.167**
num. comp. (ave.)	-	-0.0972*
assortativity (ave.)	-	0.172**
reciprocity (ave.)	-	-0.200**
path	-	0.0701
deg. cent. (ave.)	-0.0975	0.268**
bet. cent. (ave.)	0.107*	-
Num. of observations	1226	926
R^2	0.0175	0.0430

Table 4 Results of regression analyses for the 6-month growth rate (*:$p < 0.05$, **:$p < 0.01$)

Dependent variable: 6-month growth rate		
	2nd month	6th month
	Std. Coeff.	Std. Coeff.
growth rate	-	0.0951**
num. members	-	0.126*
num. members (ave.)	-	-0.143*
clustering (ave.)	-	-0.131*
GC	-	0.0806
assortativity (ave.)	-	0.130*
reciprocity (ave.)	-	-0.143**
deg. cent. (ave.)	-	0.185*
Num. of observations	1226	926
R^2	n.s.	0.0439

via the stepwise method. For the snapshot of the 2nd month, we couldn't obtain a statistically significant model for the 6-month growth rate. Note that in the analyses for the snapshot of the 6th month, we manually excluded the variables *clique*, *clique (ave.)*, *diameter*, and *diameter (ave.)* from the independent variables to avoid multicollinearity and obtain interpretable results.

These results show that the constructed models achieve only low R^2 values, which suggests that future activity growth cannot be well explained by the independent variables used in our analyses. In particular, for the snapshot of the 2nd month, most of the variables are either not significant or not selected for both the 3-month growth rate and the 6-month growth rate.

Summary of answers to RQ1: There are no clear social network characteristics that have a large influence on the future growth of communities.

4.2 Features Affecting the Survival of Online Communities

We next tackle **RQ2**: *What are the structural characteristics of social networks affecting survival?* To answer this question, we conducted additional regression analyses. The procedures were the same as those in the previous subsection, but the dependent variable was *remaining lifetime*.

Table 5 Results of regression analysis for remaining lifetime (*:$p < 0.05$, **:$p < 0.01$).

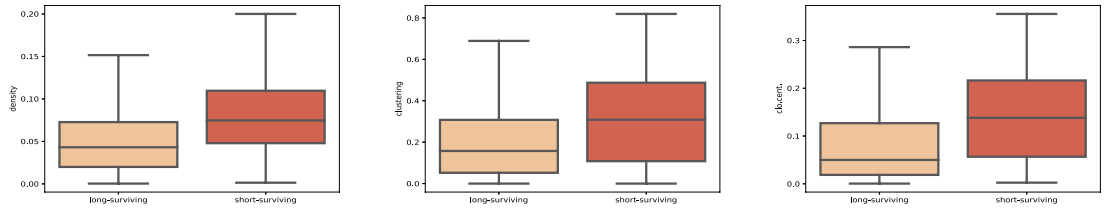
Dependent variable: remaining lifetime		
	2nd month	6th month
Independent variables	Std. Coeff.	Std. Coeff.
growth rate	-	-0.226**
num. comments	-	0.117*
num. members	-	-0.152*
ave. deg.	0.0964**	-
density	-0.313**	-0.179**
density (ave.)	-	-0.444**
clustering	-	0.0761
clustering (ave.)	-0.164**	-
clique (ave.)	0.108	0.273**
num. cluster	0.0982**	0.170**
GC (ave.)	-	-0.108*
reciprocity (ave.)	-	0.186**
deg. cent. (ave.)	-	-0.145**
clo. cent. (ave.)	-0.109**	-
Num. of observations	1226	926
R^2	0.175	0.245

Table 5 shows the results of regression analyses when the dependent variable is *remaining lifetime*. Note that in the analyses, we excluded the variable *clique* from the independent variables to avoid multicollinearity and obtain interpretable results.

In contrast to the results for growth rate, we can see that the obtained models achieve reasonable R^2 values, which suggests that social network features have a considerable influence on the remaining lifetime of communities. We should note that since there are many factors affecting the lifetime of communities, the value of R^2 is not so high, but it is comparable level to those in existing studies using regression analysis for exploring factors affecting community success (e.g., [7]).

Looking at the effect of each variable, the following findings are obtained (a more detailed discussion will be given in Sect. 5).

- Density has the largest effect among the variables. The effect is negative, suggesting that communities in which members are densely connected with each other tend not to survive a long time.
- Centralization has relatively large negative effects, suggesting that highly centralized communities tend not to survive a long time.
- The number of clusters has a relatively large positive effect, suggesting that communities composed of multiple groups of members tend to survive a long time.
- Clustering has a significant effect only for the 2nd-month snapshot. This effect is negative, which suggests that communities with high clustering in their early stages tend not to survive long.
- Reciprocity and clique ratio have significant effects only for the 6th-month snapshot. These effects are positive, which suggests that long surviving communities tend to form large cores and many reciprocal relationships as they evolve over time.



(a) density (b) clustering (c) clo. cent.

Fig. 1 Box plots for comparing structural characteristics between long-surviving and short-surviving communities

To clearly show the differences between long-surviving and short-surviving communities, we compare *density*, *clustering*, and *clo. cent.*, which have large effects on the remaining lifetime between them. Figure 1 shows the box plots for comparing *density*, *clustering*, and *clo. cent.* for the snapshots of the 2nd month between long-surviving and short-surviving communities. Here, communities whose remaining lifetime is more than or equal to 24 are classified as long-surviving, and others are classified as short-surviving. The boxes in the figure indicate the range of values from the first quartile to the third quartile. The lines within the boxes indicate the median. The ends of whiskers of the boxes are the lowest datum within 1.5 IQR (Inter Quartile Range) of the first quartile and the highest datum within 1.5 IQR of the third quartile. Outliers are not plotted. These results show that long-surviving communities tend to have lower density, clustering coefficient, and centralization based on closeness than short-surviving communities. Mann–Whitney U test shows that these differences are statistically significant ($p < 0.01$).

Summary of answers to RQ2: Several social network features are significant and have a considerable influence on the survival of communities. Long surviving communities tend to have the following characteristics: Their social networks are not densely connected, are not highly centralized, and are composed of multiple clusters; they have low clustering in their early days; and they have a large core (known as core-periphery structure [48]) with many reciprocal links in their later stages.

4.3 Prediction

We finally tackle **RQ3**: *How effective are social network features for predicting future growth and survival?* We conducted experiments to predict the future growth and survival of communities from the snapshots of the 2nd month and the 6th month. As features for the prediction, we used social network features and basic activity features used in the regression analyses. In addition to these features, we also used content features extracted from comment texts in the communities. By comparing content and basic activity features with social network features, we examine the effectiveness of social network features for the prediction tasks. We used Doc2Vec [49], which is a popular technique for obtaining vector representations of documents, to obtain content features of each community. For each community, all comments posted during the m-month-period from the community creation were regarded as a document for the community. Then, Doc2Vec was applied to the documents of all target communities for obtaining content features. The vector size (i.e., the number of dimensions) were 128, words with total frequency lower than 10 were ignored, distributed memory algorithm was used for training. We used Python gensim package[†] for Doc2Vec. We constructed classifiers using Random Forests [50] to predict future growth and survival of online communities from the features. The number of decision trees was 500, and each decision tree was trained with randomly selected $\lfloor \sqrt{f} \rfloor$ features, where f is the number of features used in the model.

For each experimental setting, we constructed five types of classifier: *Full*, *w/o Net*, *Net*, *Content*, and *Activity*. The classifier *Full* is constructed from all features, *w/o Net* is constructed without using network-related features, *Content* is constructed only from Doc2Vec features, *Net* is constructed only from network-related features, and *Activity* is constructed only from features related to basic activity (i.e., *num. comments*, *num. members*, and *growth rate*). Prediction accuracies of the constructed classifiers were evaluated by 10-fold cross-validation.

We first examine the growth prediction. The task here is to predict whether the N-month growth rate of a community will be over a threshold value. This task is intended to find growing communities. Here, we show the results for only the 3-month growth rate, but we obtained similar results for the 6-month growth rate. As the threshold values, we used the 3rd quartile of the 3-month growth rate. The threshold was 1.15 for the snapshot of the 2nd month and 1.41 for the snapshot of the 6th month. Table 6 shows the prediction accuracy of the constructed models using precision, recall, and F_1-measure [51]. Table 6 shows that the prediction accuracy of the Full model is lower than that of the Activity model. As suggested from the results of regression analyses, this result shows that social network features are not effective for predicting growing communities.

We next tackle the prediction of survival of communities. The task here is to predict whether the remaining lifetime of a community is under a threshold value. The aim is to identify dying communities in advance. As the threshold values, we used the 1st quartile of the remaining lifetime.

[†]https://radimrehurek.com/gensim/index.html

Table 6 Accuracy for growth prediction.

	2nd month			6th month		
	Precision	Recall	F_1	Precision	Recall	F_1
Activity	0.304	0.492	0.376	0.356	0.536	0.428
Net	0.136	0.447	0.208	0.126	0.453	0.197
Content	0.029	0.29	0.053	0.2	0.495	0.285
w/o Net	0.055	0.447	0.098	0.196	0.51	0.283
Full	0.058	0.474	0.104	0.226	0.565	0.323

Table 7 Accuracy for survival prediction.

	2nd month			6th month		
	Precision	Recall	F_1	Precision	Recall	F_1
Activity	0.419	0.212	0.281	0.504	0.245	0.330
Net	0.475	0.188	0.270	0.495	0.203	0.287
Content	0.489	0.086	0.147	0.593	0.203	0.302
w/o Net	0.610	0.141	0.229	0.644	0.245	0.355
Full	0.616	0.208	0.311	0.631	0.376	0.471

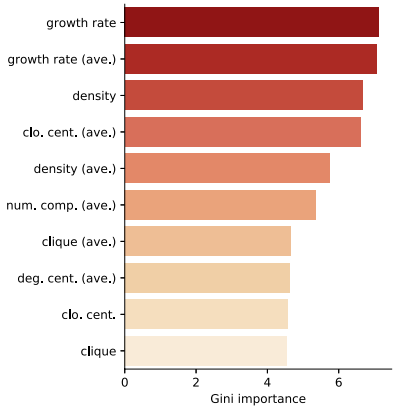


Fig. 2 Top 10 most predictive features in the full model.

The threshold was 16 for the snapshot of the 2nd month and 30 for the snapshot of the 6th month. Table 7 shows the prediction accuracy of the constructed models. These results show that the Full model achieves higher F_1 measure than other models, which suggests the effectiveness of social network features for predicting the survival of online communities. The *Full* model achieves approximately 30% higher F_1 measure than the *w/o Net* model. It can also be seen that the *w/o Net* model achieves lower F_1 than the *Activity* model for the 2nd month, which indicates that the content features for the 2nd month snapshot are not effective for community survival prediction. In contrast, network-related features are shown to be effective for survival prediction both for the 2nd month and the 6th month snapshots.

To confirm the contributions of network features for survival prediction, we investigate the importance of each feature in the Full model for the 6th month snapshot. Figure 2 shows the top 10 most predictive features in the Full model for the 6th month snapshot measured by the *Gini importance* (also called as *mean decrease Gini*) [50]. Using the Gini importance has been a common heuristic for evaluating the importance of each feature on the prediction accuracy in Random Forests [52]. Higher Gini importance of a feature implies that the feature is more useful for the prediction than other features. This result confirms that network features contribute to survival prediction.

Summary of answers to RQ3: Social network features are effective for predicting community lifetime when combining them with activity and content features but not effective for predicting community growth.

5. Discussion

5.1 Findings and Implications

Validation of findings in existing studies: Our results show that several features significantly affect the remaining lifetime of community, supporting the findings of previous studies. Communities with high clustering and high density tend to have a shorter lifetime, while communities with a high clique ratio tend to survive a long time. These results are consistent with the findings of Kairam et al. [10]. Therefore, these features can be expected to affect the survival of several types of online community.

Effects of centralization: Our results show that communities with high centralization tend to have a shorter lifetime. A highly centralized structure may cause specific members to be overloaded, a factor which is suggested to have a negative impact on the success of online communities [3], [53], [54]. We expect this is the reason why centralization has a negative effect on the lifetime of communities in this case. However, we should note that, for open source software development projects, different results are reported. Tsugawa et al. [34] found that centralization had positive effects on community success, while Toral et al. [35] found that centralization had no significant effects on community success. Although the measures of success in these studies [34], [35] are software-related metrics, which are different from those in our studies, this does suggest that centralization may have different effects for different types of community.

Features not investigated in existing studies: To the best of our knowledge, the positive effects of reciprocity and the number of clusters on the lifetime of communities have not been shown in previous studies. A higher number of clusters implies that these communities cover diverse topics. Therefore, our finding that communities with a higher number of clusters tend to survive a long time is consistent with the existing finding that diversity affects community success [5]. Theoretically, reciprocity is a determinant of sociability that affects the success of online communities [2]. Our finding is consistent with this theory.

Poor predictive power of social network features for activity growth: Although existing studies [10], [18] show that social network features are effective predictors for the growth in the number of subscribed community members, our results suggest they are poor predictors of activity growth. In online communities there are lurkers [31], who only consume community content and make no contributions. The existence of such members may be the source of the difference between the success of membership

growth prediction in existing studies and the failure of activity growth prediction in this study.

Effectiveness of social network features for survival prediction: Our results show that social network features are effective for survival prediction. Particularly in the early stage, content features are not effective for survival prediction whereas social network features contribute to improving prediction accuracy. This suggests that social network features are key factors that identify dying communities in their early stage. But we should note that the prediction accuracy of our constructed model should be not enough for practical use. More efforts are still needed for constructing accurate prediction models.

Practical implications: Our results reveal the social network structure of communities that tend to survive a long time. In summary, communities with a core-periphery structure, a high number of clusters, low clustering, low density, and low centralization tend to survive. These features can be used as criteria for checking the *health* of communities, and it may be useful for community administrators to monitor these features. Moreover, our findings may be useful for designing social bots [55] for activating communities. If interventions by social bots can control the communication patterns of community members, it might be possible to increase the lifetime of communities.

5.2 Limitations and Future Work

We recognize that there are some limitations of this study, and these suggest directions for future work. First, the generalizability of the results obtained in this study should be validated in future research. Our study design such as the observation period, the criteria for determining target communities, and method for constructing the social networks might affect the results. Validating our findings in different settings is an important work. The applicability of our findings in different types of communities should be also validated. Since different factors affect the success of different types of community [3], features affecting the success of Reddit communities may not have an effect in other communities. Moreover, the topics addressed in Reddit communities are highly diverse [37], [40], and features affecting growth and survival might differ by topic (e.g., the features affecting Q&A communities and those affecting sports news communities might be different). Exploring the relation between a community's topic and features affecting its growth and survival is an interesting future avenue of research.

Second, the prediction accuracy of constructed models should be improved for practical use. Although we show that social network features are effective for predicting the survival of online communities, the accuracy is not high enough for practical scenarios. More features, such as inter-community [7], [16] features, should be incorporated to further improve the model.

Third, why activity growth cannot be predicted from social network structure is still unclear. External factors or the topic of the community may have an impact on activity growth. Moreover, our definition of community growth may affect our results. There are other possible definitions for the measures of community growth. For instance, the growth rate can be defined as the ratio of the number of comments in the next one month divided by the previous one month. The growth rate also can be defined both considering the number of comments and the number of community members. Social network features may be related to other aspects of the community growth. More effort is still needed to clarify whether the definition of community growth affect the findings or not.

Fourth, obtaining the community lifetime from the finite lengths of the observation periods is a limitation of the methodology in this study. Communities that are active (i.e., have comments) on the last month of the observation period might have longer lifetime, but we couldn't know the *actual* lifetime of these communities. If actual lifetime was available, we could compare the characteristics of very long-surviving communities and other communities. This is a fundamental limitation of this study for investigating the lifetime of communities. Statistical techniques such as survival analyses used in [16] can be useful to address this limitation in future research.

Fifth, more focus should be given to the temporal aspects of communities. Following Kairam et al. [10], we considered snapshots of communities in only their 2nd and 6th months. However, since communities evolve over time, temporal and dynamical analyses [56] would give more information than snapshot-based analyses. Example research questions are: How do temporal changes of social network structure affect the growth and survival of communities? What are the network's structural characteristics immediately before the death of a community? Are there any structural characteristics that appear or must be present before the sudden growth of a community? Answering these questions would take us one step further toward understanding the dynamics of community evolution.

6. Conclusion

We have investigated how the social network structure of an online community affects its future growth and survival. In particular, we have investigated the effects of 15 social network features on the growth-rate and remaining lifetime of communities in Reddit. Our results have shown that social network features used in this paper do not have large influence on growth rate. In contrast, several social network features have significant and considerable effects on the lifetime of communities. We found that long surviving communities tend to have the following characteristics: Their social networks are not densely connected, not highly centralized, composed of multiple clusters, have low clustering in their early days, and have a large core and many reciprocal links in their late stage. We also conducted experiments to predict future community growth and survival from social network features as well as features obtained from the contents and activities in the communities. We have shown that social

network features are effective predictors of the future survivability of communities when combining them with activity and content features. In contrast, social network features make almost no contribution toward predicting the future growth of communities.

References

[1] R.E. Kraut and P. Resnick, Building Successful Online Communities: Evidence-Based Social Design, MIT Press, 2012.
[2] J. Preece, "Sociability and usability in online communities: Determining and measuring success," Behav. Inform. Technol., vol.20, no.5, pp.347–356, 2001.
[3] A. Iriberri and G. Leroy, "A life-cycle perspective on online community success," ACM Computing Surveys (CSUR), vol.41, no.2, p.11, 2009.
[4] R.E. Kraut and A.T. Fiore, "The role of founders in building online groups," Proc. CSCW'14, pp.722–732, 2014.
[5] J. Chen, Y. Ren, and J. Riedl, "The effects of diversity on group productivity and member withdrawal in online volunteer groups," Proc. CHI'10, pp.821–830, 2010.
[6] H. Zhu, R. Kraut, and A. Kittur, "Effectiveness of shared leadership in online communities," Proc. CSCW'12, pp.407–416, 2012.
[7] H. Zhu, J. Chen, T. Matthews, A. Pal, H. Badenes, and R.E. Kraut, "Selecting an effective niche: An ecological view of the success of online communities," Proc. CHI'14, pp.301–310, 2014.
[8] N. Ducheneaut, N. Yee, E. Nickell, and R.J. Moore, "The life and death of online gaming communities: A look at guilds in World of Warcraft," Proc. CHI'07, pp.839–848, 2007.
[9] J. Koh, Y.G. Kim, B. Butler, and G.W. Bock, "Encouraging participation in virtual communities," Commun. ACM, vol.50, no.2, pp.68–73, 2007.
[10] S.R. Kairam, D.J. Wang, and J. Leskovec, "The life and death of online groups: Predicting group growth and longevity," Proc. WSDM'12, pp.673–682, 2012.
[11] Z. Lin, N. Salehi, B. Yao, Y. Chen, and M.S. Bernstein, "Better when it was smaller? community content and behavior after massive growth," Proc. ICWSM'17, pp.132–141, 2017.
[12] A. Patil, J. Liu, and J. Gao, "Predicting group stability in online social networks," Proc. WWW'13, pp.1021–1030, 2013.
[13] T. Cunha, D. Jurgens, C. Tan, and D. Romero, "Are all successful communities alike? Characterizing and predicting the success of online communities," Proc. WWW'19, pp.318–328, 2019.
[14] B.S. Butler, "Membership size, communication activity, and sustainability: A resource-based model of online social structures," Inform. Syst. Res., vol.12, no.4, pp.346–362, 2001.
[15] B. Yu, X. Wang, A.Y. Lin, Y. Ren, L.G. Terveen, and H. Zhu, "Out with the old, in with the new?: Unpacking member turnover in online production groups," Proc. ACM on Human-Computer Interaction, vol.1, no.CSCW, pp.117:1–117:19, 2017.
[16] H. Zhu, R.E. Kraut, and A. Kittur, "The impact of membership overlap on the survival of online communities," Proc. CHI'14, pp.281–290, 2014.
[17] N. Vincent, I. Johnson, and B. Hecht, "Examining Wikipedia with a broader lens: Quantifying the value of Wikipedia's relationships with other large-scale online communities," Proc. CHI'18, pp.566:1–566:14, 2018.
[18] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: Membership, growth, and evolution," Proc. KDD'06, pp.44–54, 2006.
[19] S. Tsugawa, "A survey of social network analysis techniques and their applications to socially aware networking," IEICE Trans. Commun., vol.E102-B, no.1, pp.17–39, Jan. 2019.
[20] D.J. Watts and S.H. Strogatz, "Collective dynamics of 'small-world' networks," Nature, vol.393, no.6684, pp.440–442, 1998.
[21] L.C. Freeman, "Centrality in social networks conceptual clarification," Social Networks, vol.1, no.3, pp.215–239, 1979.
[22] M.E.J. Newman and M. Girvan, "Finding and evaluating community structure in networks," Phys. Rev. E, vol.69, no.2, p.026113, 2004.
[23] L.d.F. Costa, F.A. Rodrigues, G. Travieso, and P.R. Villas Boas, "Characterization of complex networks: A survey of measurements," Adv. Phys., vol.56, no.1, pp.167–242, 2007.
[24] M.E.J. Newman, "Assortative mixing in networks," Phys. Rev. Lett., vol.89, no.20, p.208701, 2002.
[25] S. Tsugawa and S. Niida, "The impact of social network structure on the growth and survival of online communities," Proc. ASONAM'19, pp.1112–1119, 2019.
[26] S. Malinen, "Understanding user participation in online communities: A systematic literature review of empirical studies," Computers in Human Behavior, vol.46, pp.228–238, 2015.
[27] R.P. Karumur, T.T. Nguyen, and J.A. Konstan, "Early activity diversity: Assessing newcomer retention from first-session activity," Proc. CSCW'16, pp.595–608, 2016.
[28] A. Halfaker, R.S. Geiger, J.T. Morgan, and J. Riedl, "The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline," American Behavioral Scientist, vol.57, no.5, pp.664–688, 2013.
[29] D. Yang, R. Kraut, and J.M. Levine, "Commitment of newcomers and old-timers to online health support communities," Proc. CHI'17, pp.6363–6375, 2017.
[30] N. Sun, P.P.L. Rau, and L. Ma, "Understanding lurkers in online communities: A literature review," Computers in Human Behavior, vol.38, pp.110–117, 2014.
[31] B. Nonnecke and J. Preece, "Lurker demographics: Counting the silent," Proc. CHI'00, pp.73–80, 2000.
[32] A. Tagarelli and R. Interdonato, "Time-aware analysis and ranking of lurkers in social networks," Soc. Netw. Anal. Min., vol.1, no.5, pp.1–23, 2015.
[33] E. Sharma and M. De Choudhury, "Mental health support and its relationship to linguistic accommodation in online communities," Proc. CHI'18, pp.641:1–641:13, 2018.
[34] S. Tsugawa, H. Ohsaki, and M. Imase, "Inferring leadership of online development community using topological structure of its social network," J. Infosocionomics Society, vol.7, no.1, pp.17–27, 2012.
[35] S.L. Toral, M. Rocío Martínez-Torres, F. Barrero, and F. Cortés, "An empirical study of the driving forces behind online communities," Internet Research, vol.19, no.4, pp.378–392, 2009.
[36] J.S. Coleman, "Social capital in the creation of human capital," Am. J. Sociol., vol.94, pp.S95–S120, 1988.
[37] P. Singer, F. Flöck, C. Meinhart, E. Zeitfogel, and M. Strohmaier, "Evolution of Reddit: from the front page of the Internet to a self-referential community?," Proc. WWW'14, pp.517–522, 2014.
[38] E. Gilbert, "Widespread underprovision on reddit," Proc. CSCW'13, pp.803–808, 2013.
[39] T. Weninger, X.A. Zhu, and J. Han, "An exploration of discussion threads in social news sites: A case study of the Reddit community," Proc. ASONAM'13, pp.579–583, 2013.
[40] D. Choi, J. Han, T. Chung, Y.Y. Ahn, B.G. Chun, and T.T. Kwon, "Characterizing conversation patterns in Reddit: From the perspectives of content properties and user participation behaviors," Proc. COSN'15, pp.233–243, 2015.
[41] Y. Liang, "Knowledge sharing in online discussion threads: What predicts the ratings?," Proc. CSCW'17, pp.146–154, 2017.
[42] C. Buntain and J. Golbeck, "Identifying social roles in reddit using network structure," Proc. WWW'14 Companion, pp.615–620, 2014.
[43] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift Reddit dataset," Proc. ICWSM'20, vol.14, pp.830–839, 2020.
[44] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," Proc. WWW'10, pp.591–600, 2010.
[45] B. Viswanath, A. Mislove, M. Cha, and K.P. Gummadi, "On the evolution of user interaction in Facebook," Proc. WOSN'09, pp.37–42, 2009.

[46] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.U. Hwang, "Complex networks: Structure and dynamics," Physics Reports, vol.424, no.4, pp.175–308, 2006.
[47] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," J. Stat. Mech., vol.2008, no.10, p.P10008, 2008.
[48] S.P. Borgatti and M.G. Everett, "Models of core/periphery structures," Social Networks, vol.21, no.4, pp.375–395, 2000.
[49] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," Proc. ICML'14, pp.1188–1196, 2014.
[50] L. Breiman, "Random forests," Mach. Learn., vol.45, no.1, pp.5–32, 2001.
[51] C.J.V. Rijsbergen, Information Retrieval, 2nd ed., Butterworth-Heinemann, Newton, MA, USA, 1979.
[52] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," Proc. Advances in Neural Information Processing Systems (NIPS'13), pp.431–439, 2013.
[53] D. Maloney-Krichmar and J. Preece, "A multilevel analysis of sociability, usability, and community dynamics in an online health community," ACM Trans. Comput.-Hum. Interact. (TOCHI), vol.12, no.2, pp.201–232, 2005.
[54] 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.
[55] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," Communications of the ACM, vol.59, no.7, pp.96–104, 2016.
[56] P. Holme, "Modern temporal network theory: A colloquium," Eur. Phys. J. B, vol.88, no.9, pp.234:1–234:30, 2015.

Sho Tsugawa received the M.E. and Ph.D. degrees from Osaka University, Japan, in 2009 and 2012, respectively. He is currently an assistant professor at the Faculty of Engineering, Information and Systems, University of Tsukuba, Japan. His research interests include network science, social network analysis, and computational social science. He is a member of IEEE, ACM, IEICE, and IPSJ.

Sumaru Niida received B.E. and M.E. degrees in electrical and computer engineering from Yokohama National University in 1994 and 1996, respectively. He also received Ph.D. in psychology from University of Tsukuba in 2017. He is currently a senior manager in the User Innovation Group of KDDI Research, Inc. His research interests include service quality assessment and service design method.