Analysis of Knowledge Sharing Activities on a Social Network Incorporated Discussion Forum: a Case Study of DISboards

Zhuozhao Li, Harrison Chandler and Haiying Shen*, *Senior Member, IEEE*

Abstract—DISboards is a discussion forum that provides a platform for knowledge sharing on planning and resources for Disney-related travel (Disney World, Disney Cruise Line, etc.). Since no previous work has been devoted to studying the online social networks (SNs) in the forums, we examine the SN and knowledge sharing activities in DISboards as a case study of discussion forums. Based on a large amount of data collected, we provide an in-depth study of DISboards. In particular, we analyzed SN structure, effect of SN in the forum, category characteristics and so on. We found that users with more friends are generally more active in the forum; teens are more active and constitute a significant part of the SN. We clustered the selected categories (e.g., resorts, dining, and hotels) into three groups: report, fact, discussion, and characterized their properties. Most users focus narrowly on only a few categories, while very few users participate in many categories. The development of SN should be able to attract more users to involve in the forum. We believe that the results presented in this paper are crucial in understanding SN and knowledge sharing in the forums. The paper also gives an instruction for the enhancement of SNs to incentivize users' activeness in the forums.

Index Terms—DISboards forum, discussion forum, social networks, knowledge sharing

1 Introduction

An Internet forum, or message board, is an online discussion site where people can hold conversations in the form of posted messages. Internet forums provide a platform for knowledge sharing and play an irreplaceable role in allowing users from across the world to discuss on a wide variety of topics and be heard by others. With over 1.8 billion Internet users worldwide, there are literally thousands upon thousands of forums [1]. Some of the most active forums today include Ultimate Guitar [3], Something Awful [2] and DISboards [12]. Forums tend to be for a special purpose, e.g., DISboards focuses on the Disney related issues.

Recently, forums become a popular platform on Internet for knowledge sharing. Research [34] shows that the replies in SNs are willing to and able to provide more tailored and personalized answers, since they know a great deal about the backgrounds and preference of the questioners. Therefore, by synergistically integrating the forum and SN, the users may be more likely to receive useful knowledge they are seeking from the forum. To provide high quality answers and knowledge sharing service for the users, it is important to understand the nature and impact of SNs on the forums. The research topic of knowledge sharing, especially the Internet scale knowledge sharing, has been lasting for at least 15 years. Everyday there is an enormous amount of knowledge sharing through network. Nowadays some QA websites (e.g., Yahoo! Answer [35]) and forums (e.g., DISboards [12]) try to provide platforms to share the knowledge through users. It is a culture of generosity. Indeed, if there is something that someone knows, there is opportunity to share it on a forum. We seek to understand the knowledge sharing activity on a forum. For example, with such a diversity of categories in which a user can participate in, we want to study how users focus their topics on. Since there is no previous work that focuses on investigating both social network and knowledge sharing on a forum, it is meaningful for us to study how SN affects the forum and how knowledge is shared across different categories on a forum that is incorporated with SN.

In this paper, we use DISboards as a case study to investigate the SN and knowledge sharing activities of users in a forum. DISboards is a Disney World discussion forums and information board. It provides Disney planning resource for Disney World, such as Disney Cruise Line and Disney World Vacations includes park hours, theme park descriptions and strategies. DISboards has more than forty-six thousands registered users and more than fifty-two millions posts. As a forum with SN, unlike other forums that may only have monotonous user group (mostly adults), DISboards generally attracts users of different ages, such as teens and adults because of the general fit of Disneyland to people of different ages. This feature allows us to do some extra analysis (e.g., teen and adult activities) that is not available in other SNs. Furthermore, we are able to crawl sufficient information we require from DISboards, such as user profile, users' friend list and recent posts, which are difficult to crawl in other forums. Since we expect to have a thorough

- * *Corresponding Author. Email: hs6ms@virginia.edu; Phone: (434) 924-8271, Fax: (434) 982-2214.*
- *Zhuozhao Li and Haiying Shen are with Department of Computer Science, University of Virginia, Charlottesville, VA 22903. E-mail: {zl5uq, hs6ms}@virginia.edu*
- *Harrison Chandler is with Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI 48108. E-mail: {hchandl}@umich.edu*

analysis of different aspects, we choose to use DISboards as our case study. Our general conclusions can be extended to other social network incorporated discussion forums though different forums for different purposes should have their own features. Users participate in forums to attain useful information, make friends and share fun with others. Specifically, in DISboards, many users who are interested in Disney may ask questions about travel plans, adventures, and transportation in corresponding categories. They will receive replies from other users or administrators for free. Moreover, SN helps users to achieve the goals, since SN is a good platform in which users can make friends and share fun. We have collected trace data from DISboards during one month, and a large amount of personal data and their associated relationship. The main contribution of this paper is an extensive trace-driven analysis of DISboards, with a focus on SN and knowledge sharing. We investigate the SN structure and user behavior. For knowledge sharing, we analyze ego network of user interactions and categories characteristics. We then use the concept of entropy and relationship between categories to measure knowledge sharing spread across categories. Our analysis yields very interesting results and the highlights of our work are summarized as follows:

(1) We study the SN structure in DISboards, and find that: (a) It exhibits a power-law distribution like other SNs. (b) The global clustering coefficient of the DISboards SN is low and the cluster coefficient decreases when the degree increases. (c) There is no clear correlation between the degree of a user and the degree of his/her friends, which is counter-intuitive to many other SNs.

(2) We examine the effect of SN in DISboards, which reveals: (a) The high-degree users are very active in the DISboards forums. (b) Though high-degree users are more active than low-degree users, they do not necessarily receive many replies in their threads, as DISboards lacks a mechanism for finding threads created by friends. (c) Though the majority of users are adults, teens constitute a significant part of the SN. Teens are very likely to be more active in DISboards. (d) To make the DISboards more attractive, the administrators may consider to develop more SN components of DISboards to stimulate user interaction activities.

(3) We characterize the properties of various DISboards categories and cluster the selected categories into three groups according to the thread length (i.e., the number of replies) and post length (i.e., the number of characters): (a) The report categories have both the longest thread and post length, while fact categories have the shortest. (b) The discussion categories have both long thread and post length, but shorter than the report categories.

(4) Analysis of the thread/post overlap, indegree and outdegree, ego networks, user entropy of each category and category similarities shows that: (a) The discussion categories have higher thread/post overlap, broader indegree and outdegree range. (b) The most active users in the discussion categories are "discussion persons", while the most active users in the report and fact categories are "answer persons". (c) The users in DISboards are quite concentrated on a few categories and few users post across several categories.

The remainder of the paper is organized as follows. Section 2 gives an overview of the related work. Section 3 presents the statistics of our data set. Section 4 displays the SN structure and effect of SN in DISboards. Section 5 describes the characteristics of categories. Section 6 presents how knowledge is shared across different categories. Section 7 presents the discussion and implication of this paper. Section 8 concludes the paper with remarks on future work.

2 Related Work

Many researches focused on the network structure and growth pattern of SN and QA websites [23]. For instance, after studying the evolution of network and group membership of MySpace and LiveJournal, Backstrom *et al.* [8] found that it is effective to use homophyly to improve the predictive model of group membership. Cha *et al.* [10] measured the users' influence by using different metrics in Twitter, and found that most influent users can have significant influence on a variety of topics. Kwak *et al.* [19] studied the topological characteristics of Twitter and its power as a new medium of information sharing. Burghardt *et al.* [9] conducted an empirical study on Stack Exchange and found out the factors that affect which answers are chosen as the best answers. Yao *et al.* [36] proposed a set of algorithms to detect high-quality posts in community question answering sites such as Stack Overflow and Mathematics Stack Exchange. Dong *et al.* [14] proposed an approach to predict the best answerer for questions in community question answering sites. Their approach is based on distributed representation of different words and considers both user activity and user authority. Viswanath *et al.* [31] studied the evolution of activity between users in Facebook to capture the trends of the links in the activity network, i.e. growing stronger or weaker. Digg is an online voting network. Its goal is to feature the most interesting stories on its front page, and it aggregates opinions of its many users to identify them. Zhu [38] analyzed the structural properties and the impact of SN on Digg and revealed that Digg has a totally different SN with a much lower degree of link symmetry and weaker correlation of indegree and outdegree. All the above works provide guidance on how to study the network structure in this paper. Moreover, the DISboards we investigate is more a forum-based SN, rather than voting or other SNs, such as Twitter and Digg.

Wang *et al.* [32] proposed a new analytical framework for understanding the knowledge sharing process in online QA communities. This framework can help to identify the important communication and helpful knowledge sharing in online QA communities. Yahoo! Answers (YA) is one of the biggest online QA websites for knowledge sharing and it has gained attention from many researchers. Adamic *et al.* [4] sought to understand YA's knowledge sharing activity and investigated the forum categories according to content characteristics and patterns of interaction among the users. Then, they proposed a method that combines user attributes and answer characteristics to predict whether an answer will be selected as the best answer. Kim *et al.* [18] studied the criteria to be the best answers. By evaluating the types of comments users left upon the selection of best answers for their own questions, the authors inductively derived the best answer selection criteria and grouped them into

seven value categories. By using the answer ratings in YA, Su *et al.* [29] studied the quality of human reviewed data on the Internet and the feasibility of using YA for human-reviewed data collection. Li and Shen [24] investigated the collective intelligence in the YA SN in terms of SN structure, user behavior and knowledge, and the knowledge base in a user's SN. Unlike YA which is a general question-answer forum, DISboards is a discussion forum specifically for issues related to Disney. Besides, we not only investigate the knowledge sharing in DISboards, we also analyze the SN in DISboards.

Some works focus on the study of SN properties. Gonzalez *et al.* [16] identified the main components of Google+ structure, characterized the key features of their users and their evolution over time, and compared them to those of Facebook and Twitter. Gong *et al.* [15] performed a study of the evolution of social-attribute (e.g., location, communities of interests) networks using Google+ and how attributes impact the social structure. Zhao *et al.* [37] studied the early evolution of the Renren SN, and analyzed its network dynamics at different granularities to determine their influence on individual users. There has been work concentrating on the thread and message level. Arguello *et al.* [7] found that posters are less likely to receive replies if they are newcomers. Posting on-topic, introducing oneself via autobiographical testimonials, asking questions using less complex language increase replies. Joyce and Kraut [17] studied if the type of newcomers' posts and related responses affect their continuances in this forum. Lakhani and Von Hippel [20] used Usenet posting patterns and questionnaires to determine users' motivations for providing voluntarily help to information seeker on this site. To the best of our knowledge, our work is the first thorough study on a forum (i.e., DISboards) specifically serving a realm, with a focus on its SN and knowledge sharing.

3 DISBOARDS DATA SET

DISboards is a forum in which users interact through posts and replies. This forum has 57 categories such as *Adventures by Disney* and *Camping*. In a forum's category, each discussion is called *a thread*. Users can reply in a thread, which we can call it a *reply.* A *post* of a user can be either a thread created by himself or a reply to another user's thread.

TABLE 1: The categories of DISboards.

Adventures by Disney	Budget Board
Camping	disABILITIES
Disney for Adults	Disney for Families
Disney Resorts	Disney Restaurants
Disney Trip Reports	Disney Weddings
Disney World Tips	DVC-Mousecellaneous
DVC-Operations	DVC-Planning
Gay at Disney	Orlando Hotels and Attractions
Teen Disney	The College Board
Transportation	Welcome Board

TABLE 2: High-level statistics of DISboards crawl.

Threads crawled	13,807
Users crawled	27,000
Number of threads	200,000
Number of registered users	476,442
Percent of users crawled	5.67

There are several subjects in DISboards (e.g., Disney Trip Planning Forums, Disney Vacation Club and Global Neighbours) and each subject has several categories. However, some subjects are not general and restricted to special topics. For example, some subjects contain topics only about the California and Canada Disneyland. In our analysis, we selected some general subjects, such as Disney Trip Planning Forums and Disney Vacation club. We crawled user IDs and thread data from 20 randomly selected categories in these general subjects (out of 27 categories in general subjects), as shown in Table 1, which we believe is sufficient enough for the analysis. As in many previous studies [4], [19], [26], [38] that studied online social networks or question-answer forums, we harvested one month of DISboards activity to study the SN structure and characteristics of DISboards. The crawl script went through every thread in the chosen categories that had received a post between 2011/05/13 and 2011/06/13; for each post on a thread, the post time and the user ID of the poster were collected. DISboards has a category called *Teen Disney* specifically for teenage users of this site; this category was crawled to obtain the user IDs of teenage users. In all, 13,807 threads were crawled. This yielded around 27,000 unique user IDs, representing 5.67% of DISboards registered users. Comparing to previous studies that have analyzed $\sim$0.08% of MySpace users [5], $\sim$0.3% of Orkut users [5], 0.77% of testimonial Cyworld network [5], and less than 1% of LiveJournal [8], our study was based on much larger samples of the user graphs, which should be sufficient for the analysis in this paper.

Each DISboards user has a user webpage that contains a profile along with activity statistics of the user including the total number of posts, user IDs of friends, and the forum name, date, and time of the user's 500 most recent posts. For each user ID found in the forum crawl, we fetched the user's activity statistics. Only the 500 most recent posts of each user were collected due to DISboards search limits; however, only 4.8% of the users have more than 500 posts. Then, for every user ID we crawl, we crawl their friends' user IDs. Table 2 shows a summary of our crawled dataset.

4 DATA ANALYSIS

4.1 Social Network Structure

Rather than serving as a posting and replying forum, DISboards provides users with the ability to establish a friend relationship with other users – the users can add other users as friends. This creates a SN in which two nodes are connected by an undirectional edge if they have friend relationship. The friend relationship on DISboards comes with advantages, such as the ability to access private notes and photo albums of friends. Besides, the user can see the post history of his friends. As we know, Disney is pretty attractive to children. For those families who have kids and like travelling to DISboards frequently, it is especially beneficial to join the SN since they can share more new insights and experiences about Disney with others. Note that there are not only adults but also many teens in DISboards. Therefore, the SN is a good platform for the families and teens to establish friendship and share fun with others.

We first analyze the user lifetime (i.e., the last post time date minus the first post date) of the crawled users. As

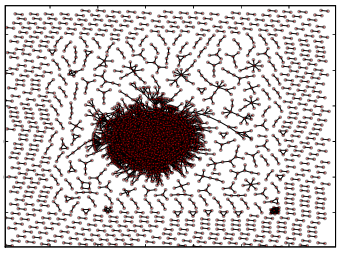


Fig. 1: Friendship connections in the SN in DISboards.

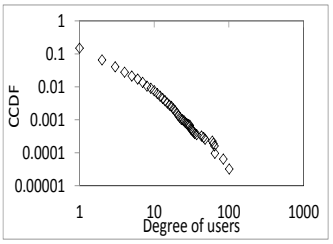


Fig. 2: CCDF log-log plot of the distribution of degree.

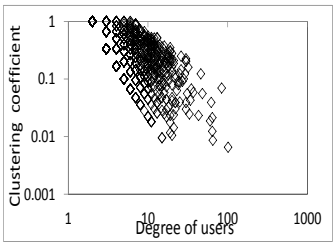


Fig. 3: Clustering coefficients versus degree.

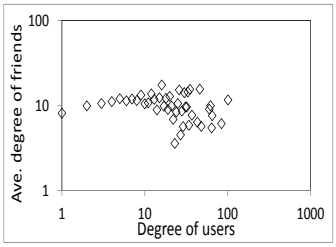


Fig. 4: Average degree of a user's friends versus the user's degree.

shown in Table 3, we see that around 50% of users have lifetime of 0 (no post) and 1 day (only post on one day). This result corresponds to the result of Buzznet [21], which also has a majority of users with lifetime 0 and 1. This is because in many forums, there are many users that only register to view threads without replying or only register to ask a question and then disappear. However, there exist around 30% of users that have lifetime more than 100 days. These long-lifetime users generally consist of users that are warmhearted users who answer others' questions frequently, families who have kids and would like to visit Disney quarterly, and business salesmen from nearby hotels and restaurants.

TABLE 3: User lifetime distribution.

Days (<=)	Percentage(%)
0	31.45
1	49.69
10	62.10
100	69.59
1000	98.97

In a power-law network, the probability that a node has degree x is proportional to $x^{-\alpha}$ for $x > x_{min}$ and $\alpha > 1$, where α is the exponent parameter of the power-law distribution and x_{min} is the minimum value. A power-law distribution has been observed in many SNs, including Digg, YouTube, and Facebook [26], where most users have a small degree and a few users have a very large degree. Some SNs such as Twitter have a non-power-law follower distribution [19]. We are interested in finding whether the SN in DISboards has a power-law distribution. We define *the degree of a user* as the number of the user's friends.

The SN in DISboards is shown in Figure 1. We see that the majority of nodes in the middle are interconnected with each other. However, many other nodes have only one or two friends with them. This indicates that the users in DISboards may follow a power-law distribution. We verify this observation by plotting the complementary cumulative distribution function (CCDF) of the degrees of users. Figure 2 shows the CCDF of the degree of DISboards users with logarithmic scale for both axes. We see that the degree of users follows a power-law distribution as the curve almost perfectly fit to a straight line. Because DISboards is primarily a forum, over 80% of forum users in our dataset have zero friends, opting not to participate in the SN of the forum at all. The results indicate that most users regard DISboards as a discussion board for seeking answers or sharing experience in Disney. Few users regard DISboards as a SN platform to connect with friends to routinely share experience and have discussions concerning Disney. These users may involve Disney activities more frequently, such as traveling to Disney. We use the maximum-likelihood method [11] to best fit the power-law distribution to estimate parameter α. We find that $\alpha = 2.14$, which is in the typical range $2 < \alpha < 3$ [11]. The Kolmogorov-Smirnov goodness-of-fit metric [25] of the power-law distribution is 0.0643, which indicates that the data deviates a little and the power-law coefficient α approximates the distribution very well.

Further, we look into who these high-degree users (i.e., degree > 10) are. Since the profile of a user does not reveal directly whether the user is regular user or business user (e.g., restaurant or hotel salesman), we infer the users' roles indirectly, that is, we analyze which categories the high-degree users mostly post on. A reasonable assumption is that most of the forums have some restrictions for the advertisement posting, such as what categories and how frequently they can post, because too many advertisement posts destroy the user experiences when they are surfing the forums for useful information. In the case of DISboards [13], posting the advertisements on a few specific categories results in a more organized manner for both the administrators and the business salesmen, which prevents the business salesmen from posting on the categories that are not allowed. On one hand, by this way, the administrators can well organize DISboards so that the user experiences are enhanced, as it allows the users to easily find what they want. On the other hand, the salesmen's advertisements are not overwhelmed by the discussions in the active categories.

Table 4 shows that the percentage of high-degree users who mostly post on each category. Interestingly, we find that many high-degree users (around 80%) are most likely to post on several categories, such as *Disney Trip Reports*, *DVC-Mousecellaneous*, *Adventures by Disney* and *Budget Board*. It is apparent from the names that the salesmen are less likely to post on these categories. Only less than 10% of high-degree users often post on categories such as *Disney Restaurants* and *Orlando Hotels and Attractions*, indicating that high-degree users are less likely to be the business salesmen.

Next, we look at the *shortest path length* and *diameter* of DISboards SN. The *shortest path length* between two nodes is defined as the smallest number of intermediate user nodes. The *diameter* is defined as the longest of all the shortest path length between two user nodes in the DISboards SN. Table 5 shows the average shortest path length and diameter of DISboards and some other SNs [26]. Note that in this analysis, we eliminate the users with zero friends since these users have infinity path length to other users. Unsurprisingly, we see that the average shortest path length and diameter of DISboards are both longer than other four SNs. This

TABLE 4: The percentage of high-degree users on each category.

Category	Percentage
Disney Trip Reports	21.93
DVC-Mousecellaneous	16.76
Adventures by Disney	15.79
Budget Board	14.17
disABILITIES	10.62
Welcome Board	9.00
Orlando Hotels Attractions	3.15
Disney Restaurants	2.75
Disney for Families	2.14
Transportation	1.00
Disney World Tips	0.67
Disney for Adults	0.64
Gay at Disney	0.57
DVC-Planning	0.56
Disney Resorts	0.11
The College Board	0.09
Teen Disney	0.01
Camping	0.01
DVC-Operations	0.00
Disney Weddings	0.00

is because unlike other SNs that are maturely developed, the SN on DISboards attracts fewer users. Therefore, the administrators of DISboards still need to incentivize the SN on DISboards.

TABLE 5: Average shortest path length and diameter of several SNs.

SN	Avg. shortest path length	Diameter
DISboards	14.56	46
Flickr	5.67	27
LiveJournal	5.88	20
Orkut	4.25	9
YouTube	5.10	21

Clustering coefficient is a measure of the tendency of nodes in a graph to cluster together, with a higher clustering coefficient meaning the users are more highly clustered. Assume a node k has n neighbors $(N_1, N_2, ..., N_n)$ connected with directed links. If two nodes i and j are connected with a link, we denote the directed link as L_{ij}. Then the clustering coefficient (local clustering coefficient) of a node k with n neighbors is:

$$C_k = \frac{|L_{ij} : i, j \in N|}{n(n-1)} \quad (1)$$

The clustering coefficient of a graph (network average clustering coefficient)is the average clustering coefficient of all its nodes. Assume there are in total K nodes in the graph. The network average clustering coefficient is,

$$\bar{C} = \frac{1}{K} \sum_{k=1}^{K} C_k \quad (2)$$

We clustered the nodes with the same degree into groups and calculated the clustering coefficient of each group. It is expected that the clustering coefficient decreases as the degree increases [26]. Figure 3 shows the clustering coefficients of users in each group versus the degree of the users in the group. The figure confirms our expectation, which is consistent with the observations in YouTube, Orkut, Flickr, and LiveJournal [26]. It indicates that the low-degree nodes are more highly clustered. High-degree users have lower clustering coefficient because they have friends with varying degrees. Since having too many low-degree friends would lead to a lower clustering coefficient, it is not surprising to see that the high-degree users tend to have lower clustering coefficient.

Excluding the nodes with no edges, the global clustering coefficient for the DISboards SN graph is 0.114, lower than the coefficients of Digg, YouTube, Orkut, Flickr, and LiveJournal (ranging from 0.136 to 0.330) [26], [38]. A high global clustering coefficient indicates that friends tend to find each other through mutual friends. In DISboards, users find each other in a different way, such as frequent posts in the same threads or forum categories. Also, with a limited number of users participating in the SN, most users are not clustered with other users to whom they communicate frequently via the forum, which results in lower clustering coefficient.

In Orkut and LiveJournal, high-degree users tend to have links with other high-degree users; the opposite is true in Digg, which has a few high-degree users followed by many low-degree users [26], [38]. We are interested in finding whether DISboards is like Orkut and LiveJournal or Digg. Thus, we clustered the users with the same degree together into one group. Then we calculated the average degree of friends of users in each group. Figure 4 shows the average degree of friends versus users' degree. Note that the number of data points in this figure is equal to the number of group. As the degree of users is significantly smaller in DISboards than in any of the other SNs mentioned, the difference between the highest-degree users and the lowest-degree users is small; similarly, the degree correlation effect is small to the point that no clear correlation can be observed. It indicates that this correlation effect in DISboards is counter-intuitive to other SNs, which have shown either positive or negative correlations between the degree of users and the degree.

In summary, we make three observations. (1) Like many other SNs, the SN in DISboards exhibits a power-law distribution. (2) The global clustering coefficient of the DISboards is low. This might simply reflect the low participation in the SN; alternately, it might indicate that users make friend relationships differently in DISboards than in other SNs. (3) There is no clear correlation between the degree of a user and the degree of his/her friends. This is not like other SNs, which have shown either positive or negative correlations.

4.2 Effect of Social Network on Forum Usage

The primary function of the DISboards website is a forum; it also includes SN to enhance the forum user experience. SN could be used to increase a sense of community in the forums, create loyalty in the users, and aid users in discovering relevant contents. In this section, we attempt to quantify the relationship between forum activity and the SN activity to discover the impact of the SN on the user activities in the forum.

First, we examine the correlation between the SN and posting activity. We expect that the SN participants tend to have much higher level of activity [19]. Figure 5 shows the CCDF for users' average number of posts per day, broken down by the participants (i.e., that have at least one friend), non-participants of the SN and all the participants. The figure proves our hypothesis that the SN participants tend to have much higher level of activity. Only 8.3% of non-participants post once or more per day on average, while 36.4% of participants post once or more per day on average. This result indicates a stark difference in activity and engagement for participants and non-participants. The result implies that the SN motivates users to be more active

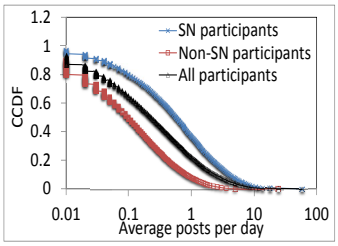


Fig. 5: Distribution of average posts per day.

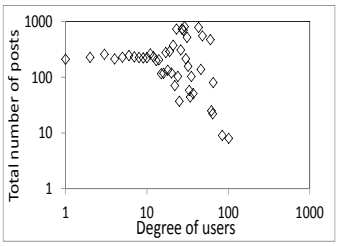


Fig. 6: Total number of posts versus degree.

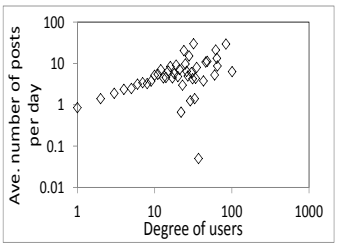


Fig. 7: Average number of posts per day versus degree.

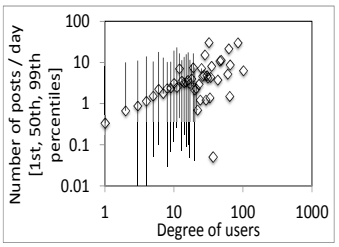


Fig. 8: The 1st, 50th, and 99th percentiles of the number of posts per day.

and establish stable friendship for information sharing or discussion, which drives users to increase their activities.

We grouped nodes with the same degree together and calculated the average of the total number of posts (including thread creation and replies) from each user. It is expected that the user with higher degree post more than the users with lower degree [19]. Figure 6 shows the total number of user posts in each group versus the degree of nodes in each group. The figure indicates that there is little correlation between the degree and the number of posts of users. The result even somehow deviate our expectation – the users with higher degree post fewer than the users with lower degree. This observation in DISboards is counter-intuitive to other SNs. There might be two reasons. First, the SN in DISboards is not well developed and hence the correlation between the degree and the number of posts of users is weak. Second, we suspect that this might be caused by the features of DISboards. Since the purpose of DISboards is to share information about Disney and to gather people/families to travel together, the users add friends to inquiry about Disney in detail or to discuss the plan when they gather to visit Disney. Therefore, the users who have many friends tend to contact their friends in private messages rather than posting on threads. Furthermore, the newly registered users who have low degree tend to have more questions about Disney and post more than others.

We then grouped nodes with the same degree together and calculated the average of the number of posts per day of the nodes in each group, with the result shown in Figure 7. We see that the two factors show a positive correlation; as the degree increases, the average number of posts increases linearly. This result indicates that users with more friends are generally more active in forums. The results imply that the combination of SN into the forum might entice users to be more engaged in the site. Figure 8 similarly shows the 1st, 50th, and 99th percentiles of the average number of posts per day of users in each group versus the degree of the nodes in each group. This figure also indicates a similar correlation as Figure 7. When the degree increases, the median number of post per day also becomes greater. This is because that SN can stimulate the users to be more active in the forum; if a user has more friends, (s)he is likely to be move active in posting. Therefore, it is important for the DISboards administrators to provide incentives to users to add more friends in the SN.

Next, we examine the correlation between the SN and a user's thread popularity, that is, we analyze whether the number of friends of a user in DISboards SN affects the user's thread popularity or not. As in many other forums (such as YA [35]), we measure a user's thread popularity by the number of replies to the threads that are created by the user. For example, once a topic in YA receives certain number of replies, it is regarded as "hot topic". This is because the threads with more replies generally provide more attractions to the users, and the users would like to post on the threads to discuss with others. We grouped the nodes with the same degree and calculated the average number of popularity of their created threads. We expect to see that if a user has more friends, the threads created by her/him tend to have higher thread popularity [19]. Figure 9 shows the average thread popularity and the average thread popularity per day of nodes in each group versus the degree of the users in the group, respectively. The figure shows little correlation between degree and thread popularity; nodes with low degree gain smaller thread popularity per day while nodes with high degree gain various thread popularity per day, which does not match as our expectation. This result indicates that if a user has few friends, (s)he receives few replies. Also, adding more friends does not necessarily attract more replies for a user's thread, even though users can see all recent posts of their friends by visiting their webpages. This could be because DISboards does not provide an alert service to users for their friends' new threads, such as the news feed on Facebook or the dashboard on Tumblr [30] that notify users when their friends post. Including such a feature can quickly notify the forum users of their friends' threads and posts in time for them to respond. Furthermore, recall the conclusions from Figure 6 that low degree users may post more to ask questions about Disney, while high degree users prefer to use private message rather than posing. Therefore, the low degree users may not have few replies while the high degree users may not be necessary to have more replies. Hence, the relationship between degree and thread popularity is not apparent. Figure 10 shows the CCDF of the fraction of threads' replies that come from SN friends of the thread creator. Only 2.5% of threads receive any replies from friends; this number is very low, and indicates that for most threads, the SN does not attract replies, as previously determined.

We grouped the users with the same degree together. In each group, we calculated the average fraction of replies by SN friends of all the users in the group. Figure 11 indicates the fraction of replies by friends versus the degree of users. The figure demonstrates that few replies of users' posts are from their friends and the SN of DISboards has little effect on attracting users to reply. The result is consistent with Figure 10. Therefore, to make the DISboards more attractive, the administrators may consider on developing more

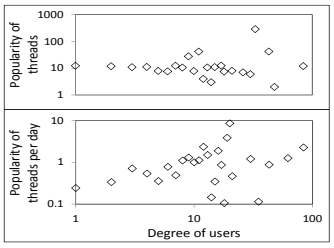


Fig. 9: The popularity of threads versus degree.

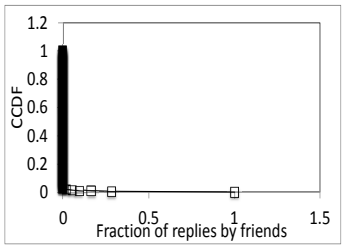


Fig. 10: CCDF of fraction of replies in a thread from its creator's friends.

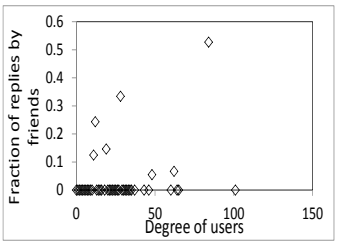


Fig. 11: The fraction of friends within users' replies

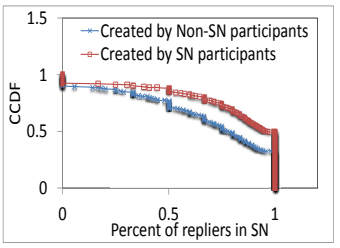


Fig. 12: CCDF of fraction of replies in a thread from SN participants.

SN components of DISboards to stimulate user interaction activities. Although there exist SNs for the users to see their friends' post history by clicking to each friend's profile page, it is better to have a new tool to notify the users when their friends make a new post.

Figure 12 shows the CCDF of the percent of replies from SN participants, broken down by threads created by SN participants and non-participants. From the figure, SN participants are more likely to post on threads created by SN participants. There are two potential causes for this result. First, SN participants recognize other social participants and post in their threads. Second, SN participants share interests with one another that are not shared by non-participants (e.g., participants are interested in discussing family or personal issues, while non-participants only want fact information on Disney).

Previous studies on other SNs have shown that users tend to explore the same contents as their friends [38]; this property gives some high-degree users heavy influence on the popularity of contents, as they can drive the interests of their friends. Next, we study whether this property exists in the SN of DISborads. The post similarity between two users is defined as the number of common threads that both users have posted on divided by the total number of threads both users have posted on. A user's post similarity is defined as the average of the post similarity values between this user and every his/her friend. It is a measure of how often the user posts on the same threads as his/her friends. Generally, a high post similarity indicates that a user tends to post on the same threads as his/her friends, while a low post similarity indicates the opposite. Figure 13 shows the CDF of each user's post similarity. The figure indicates that over 50% of users have no posts in common with their friends; this percentage is much higher in DISboards than in Digg [38] (40%). This indicates that the SNs in DISboards have lower influence on the number of replies between social friends. As before, the simplest explanation for such a low level of post similarity is the lack of a tool to actively alert a user of his/her friends' posting activities. DISboards does not have reminders to remind the users that their friends have posted. Unless the users click their friends' profile pages, they cannot see any post update of their friends.

In summary, we have drawn three conclusions. (1) High-degree users in the SN are also very active in the DISboards forums. (2) High-degree users in the SN do not necessarily receive many replies in their threads. This is because that DISboards lacks a mechanism to actively alert a user of his/her friends' posting activities. (3) Post similarity between friends in the SN is low for the same reason as above.

4.3 Activities of Different User Groups

In this section, we study the user activities on the DISboards forum in terms of activity time and different user age groups. Figure 14 shows the percentage of posts that occurred on each day of a week. It demonstrates that the forum is more active on weekdays than on weekends. This is simply because people generally like to travel to Disney during weekends, while they post questions and share their experiences for their trips during weekdays, according to Alexa [6]. Figure 15 (X: maximum=23) shows the average percent of posts that occur at each hour of the day. We see that most posts occur during the day from 8 a.m. to 5 p.m. with a peak around 10 a.m. Demographic information for DISboards users collected at Alexa [6] shows that the users tend to be women with children who browse from home; the peak times on the forum are when children are in school. Further, there is a drop in post activity around 4 p.m., a time when children go home from school; later, post activity rises as users relax at the end of the day.

The difference in a website usage patterns for different age groups may indicate future changes in how people use the website. For example, the use of Facebook was pioneered by college students; lately, however, the average age of Facebook users has been steadily increasing as more and more adults begin to utilize SNs [22]. We then study the usage patterns of different age groups on DISboards.

TABLE 6: Comparison of teens and adults.

	Adults	Teens
Number of users crawled	25951	1049
Percent of total users	96	4
Percent of participating users	14.3	36.3
Average clustering coefficient	0.109	0.164

Table 6 shows an overview of differences between teen and adult users of DISboards in our crawled dataset. The number of teen users (1049) is much smaller than the number of adults (25951). They constitute 4% and 96% of the total crawled users in our dataset, respectively. This is primarily because the use of DISboards tends to skew towards adults (especially those with children), who have the financial means and independence to plan a trip to Disney. The collected data reveals that the percentage of teens that participate in the SN is over twice the percentage of adults. Teens most likely have a greater interest in the SN, so their higher usage is not surprising. Teens also have a much higher average clustering coefficient, 0.164, compared to 0.109 for adults. This has two implications: 1) teens are more likely to find SN friends through their other SN friends, and 2) teens are more concentrated in certain forums, and thus establish friend relationships with

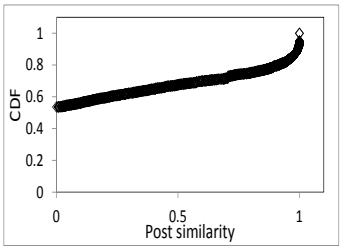


Fig. 13: CDF of post similarity.

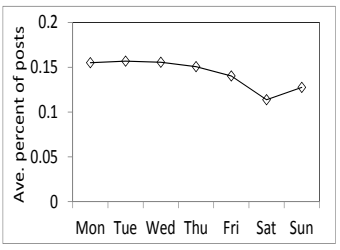


Fig. 14: Average percent of posts on each day of the week (Mon-Sun).

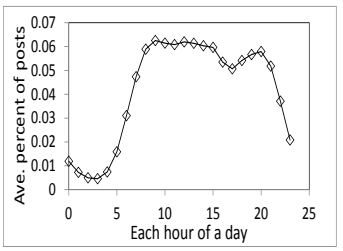


Fig. 15: Average percentage of posts in each hour of the day.

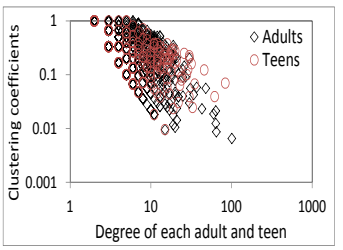


Fig. 16: Clustering coefficients versus degree of each adult and teens.

the users discovered through the forums. Recall that SN participants are more active in posting as observed in the previous section. Since adults constitute the major part of the users but mostly are not SN participants, if DISboards has a strategy to attract adults to join in the SN, much more users will be active in this forum.

Figure 16 shows the clustering coefficients of teens and adults against degree, respectively. Generally, teens and adults show the same trend; as the degree increases, the clustering coefficient decreases for both groups. This result is consistent with that in Figure 3 due to the same reasons. However, for teens, the drop is slightly less steep. This is because high-degree teens tend to be friends with high-degree teens, which leads to higher clustering coefficients.

We group the adults with the same degree together and calculate the average total number of posts per adult in each group. Similarity, we group the teens with the same degree together and calculate the average total number of posts per teen in each group. Figure 17 shows the total number of posts versus the degree for teens and adults, respectively. Teens and adults are roughly equal in this measure; both show little correlation between total posts and degree as in Figure 6. Figure 18 shows the average number of daily posts against the degree of adults and teens, respectively. As in Figure 7, both groups have a positive correlation between average daily posts and the degree, but the correlation is greater for teens. This result is a further evidence that the SN holds more importance for teens; those teens who more actively use the SN are also most active in the forum.

We create an interaction network, in which all users are nodes. If user A replies a thread created by user B, there is an edge from A to B. We define *indegree* to be the number of users that respond to a user's threads and *outdegree* to be the number of users to which a user has responded. Figure 19 and 20 show the CCDF of indegree and outdegree distributions, respectively. From Figure 19, we can see that the indegree of teens drops less sharply than the adults' indegree, which means that the teens are more likely to receive more replies. Similarly, Figure 20 shows that the outdegree of teens also drop less sharply than the adults. It turns out that the adults have less possibilities to reply to others than the teens. All the phenomena demonstrate that teens are more active in the interaction network in terms of replying and receiving replies than the adults.

In this section, we have made four observations. (1) Though the vast majority of users are adults, teens constitute a significant part of the SN. (2) Teens are more likely to find SN friends through their other SN friends. (3) Teens with greater participation in the SN tend to be more active in the DISboards forum, while adults have this correlation to a lesser degree. (4) Teens are more active in replying and receiving replies than the adults.

5 Characterizing Categories

In this section, we characterize user posting activities across different categories. The *thread length* is defined as the number of replies in a thread, and the *post length* is defined as the number of characters in a post, which indicates answer verbosity.

5.1 Basic Characteristics

Based on an initial scan of the threads on DISboards, the categories can be approximately classified into three types: reports, facts and discussion. While it is difficult to determine the strict type of each category without reading through the posts, we would like to study the category characteristics indirectly using the thread length and post length. We calculated the average thread and post length for each category by taking the average length of all the threads and posts in each category, respectively. Figure 21 shows a scatterplot of average post length and average thread length for each category and Table 7 presents each length in detail.

TABLE 7: Thread/post length, overlap and category types.

Category	Thread Length	Post Length	Thread /post overlap	Category Type
Adventures by Disney	3.83	54.39	0.34	Discus.
Budget Board	4.33	47.34	0.36	Discus.
Camping	4.49	44.06	0.34	Discus.
disABILITIES	4.27	93.75	0.40	Discus.
Disney for Adults	4.90	47.81	0.44	Discus.
Disney for Families	4.43	60.22	0.41	Discus.
Disney Resorts	3.77	44.97	0.42	Fact
Disney Restaurants	3.93	38.52	0.44	Discus.
Disney Trip Reports	6.25	97.57	0.19	Report
Disney Weddings	3.77	47.51	0.31	Discus.
Disney World Tips	4.03	46.36	0.49	Discus.
DVC-Mousecellaneous	4.60	40.82	0.35	Discus.
DVC-Operations	4.40	55.17	0.43	Discus.
DVC-Planning	3.56	35.89	0.43	Fact
Gay at Disney	5.01	29.48	0.25	Discus.
Orlando Hotels & Attractions	3.11	39.43	0.49	Fact
Teen Disney	5.30	33.86	0.26	Discus.
The College Board	4.09	49.13	0.40	Discus.
Transportation	3.42	38.27	0.44	Fact
Welcome Board	4.09	21.52	0.56	Fact

We observed from Figure 21 that the average thread length and the average post length are the longest in the report categories *Disney Trip Reports*. After reading many threads in this category, we found that most users created their own threads initially with a long report. It is intriguing to see that the report categories even have the longest thread length. In order to plan their own trips better, users

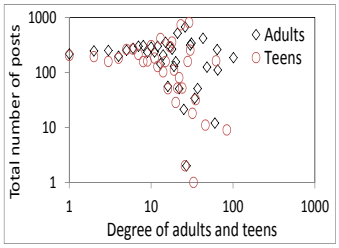


Fig. 17: Total number of posts versus degree of adults and teens.

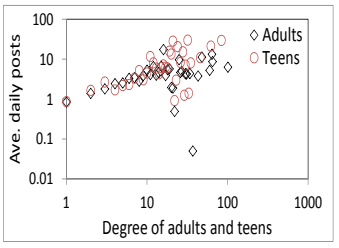


Fig. 18: Average daily posts versus degree of adults and teens.

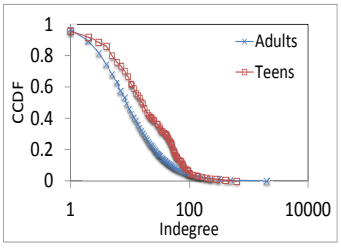


Fig. 19: Indegree of adults and teens.

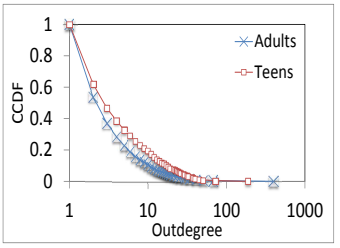


Fig. 20: Outdegree of adults and teens.

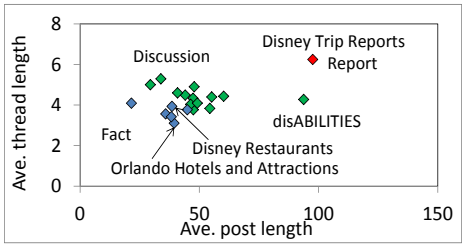


Fig. 21: Average thread length versus average post length of each category.

would like to attain information from others' trips. If they have questions about the reports, they would reply the threads to query. For instance, one report, named "The Solo Trip Which Led to a Vinylmation Addiction!", provides a detailed schedule of a user's trip events. Many other users have replied to the report with comments ("Thanks for this great trip report") and questions for the reporter, resulting in a long thread length. Even the reporters update their threads with new reports. Since the reports themselves have many characters, the report categories generally have the longest thread length and the longest average post length.

Fact threads are used to provide useful information about Disney. We can see from Table 7 and Figure 21 that fact categories have shorter thread lengths and shorter post lengths, such as *Disney Restaurants* and *Orlando Hotels and Attractions*. In these categories, the main topics center around restaurant ratings, hotels, services, and etc. For example, in *Disney Restaurants*, some brief introductions of the restaurants near Disney or deals of restaurants are posted as advertisements to attract people. Sometimes, it is sufficient for people to obtain information by only reading the threads without replying. A few users reply to appreciate the thread creators for sharing and some users replies with questions for more information, such as the availability of hotel rooms. Moreover, once those questions are replied, unless someone else does not agree with the answers, they are less likely to attract more replies. Therefore, fact categories have short thread lengths. Since the fact threads only describe fact information, they are not as long as the report threads. Many questions for facts are simple, which can be answered in a few words. For example, a question about the price of the hotels near Disney can be replied in a few characters. Therefore, fact categories tend to have relatively short post lengths, and their thread lengths vary in a range. From Figure 21, we can see that the discussion categories (e.g., *disABILITIES* and *Adventures by Disney*) attract many replies with shorter lengths compared to the report categories. In the discussion categories, users seek others' opinions and advices on some issues and receive replies with moderate lengths. But the replies are not very lengthy since the repliers offer brief ideas rather than plans in detail as in reports. An extreme example is *disABILITIIES* category, which also has a moderate thread length but very long post length (a little bit shorter than the report categories' post length). It is because *disABILITIES* is a place for sharing tips and information on touring Disney vacation destinations with mental or physical disabilities, including anything from allergies to broken legs to neuropathy. Complicated discussions are more likely to occur to meet the special needs, which generates longer replies. For instance, a user stated that he would bring his elderly parents to Disney for their upcoming trip and asked for advices. In the replies, a couple of users provided exhaustive suggestions to him. Therefore, the post length in *disABILITIIES* is very close to that in the report categories.

We can conclude our observations as follows.

(1) Based on the thread length and post length, we can indirectly infer the main thread type for each category. In Figure 21, the first cluster (red categories) consists of report categories, which have both the longest thread and post length. *Disney Trip Reports* is the only category in this cluster. The second cluster (blue categories) consists of fact categories. We observe that in this cluster, most categories are full of fact threads and replies, leading to the shortest thread length and post length. This cluster includes *Welcome Board*, *Orlando Hotels and Attractions*, and *Disney Restaurants* and so on. The third cluster (green categories) consists of discussion categories, in which users discuss with others and seek for suggestions. Most categories are in this cluster, such as *disABILITIES* and *Disney for Families* and so on.

5.2 User Roles

While thread length and post length are two related metrics used to gauge the levels of discussion in different categories, they both fail to take into account the different roles of users (i.e., as thread creators or repliers) in the forum. In the fact categories where users share facts, the majority of thread creators are fact askers (i.e., novices), and those who have expertise are primarily repliers. Therefore, the population of thread creators and repliers is rather distinct. In the discussion categories, a thread creator usually also replies, as a way of continuing discussions.

To measure how many users are both thread creators and repliers in a certain category, we characterize the categories with thread/reply overlap: whether the users who create threads are also the ones who reply in a category. Assume the number of users who have posted in category $\mathcal{A}$ is n. Let t_i and r_i be the number of threads created and the number of replies by user i in this category, respectively. In category $\mathcal{A}$, the number of threads and the number of

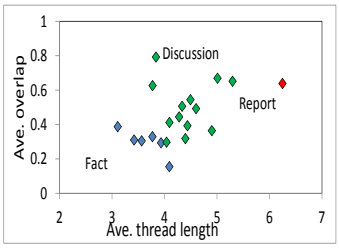


Fig. 22: Overlap versus thread length.

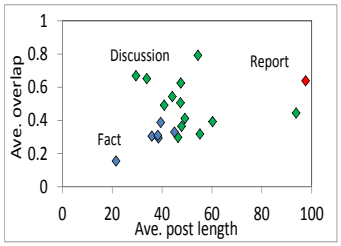


Fig. 23: Overlap versus post length.

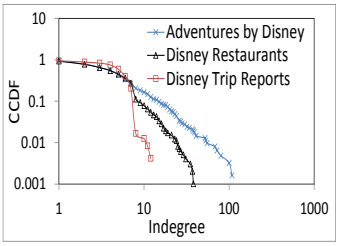


Fig. 24: Log-log distribution of indegree.

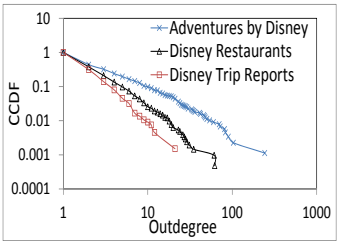


Fig. 25: Log-log distribution of outdegree.

replies for all the users can be grouped as two vectors $\vec{T} = (t_1, t_2, ..., t_n)$ and $\vec{R} = (r_1, r_2, ..., r_n)$, respectively. The thread/reply overlap of a certain category is defined as $\frac{\vec{T} \cdot \vec{R}}{|\vec{T}||\vec{R}|}$. A larger thread/reply overlap implies many users are both creators and replies.

Figure 22 and 23 show the thread/reply overlap versus thread length and post length for each category; this information is summarized in Table 7. We still use different colors to represent the different types of categories as in Figure 21. From these figures, we can see that the discussion categories have higher thread/reply overlap, because users tend to both create and reply and have more discussion on a topic. They post the issues they seek for advices and discuss with others, which results in higher overlap between thread creators and repliers.

It is not surprising that report categories have the third highest thread/reply overlap because in this category thread creators share their trips to Disney. Those reports attract many novice users to view and reply to the threads if they have questions about the reports, leading to a certain amount of discussion. Moreover, the report creators themselves reply to their own threads as updates of their reports, which also results in high thread/reply overlap. Therefore, the thread/reply overlap of the report categories is still among the highest.

As shown in the figures, we can see that the fact categories have the lowest thread/reply overlap, which is consistent with the result of the fact cluster in *Yahoo! Answers* [4]. The probable reason is that in the fact categories in DISboards, there are many novices who tend to ask more than reply. On the other hand, the repliers may be warm-hearted persons who have experiences to Disney. Therefore, the thread creators and the repliers in the fact categories are quite distinct, which corresponds to a low thread/reply overlap.

In summary, we have made two observations. (1) The discussion and report categories tend to have a high proportion of users who both create and reply threads. (2) The users in the fact categories tend to either create or reply threads, which means that the thread creators and the repliers are quite distinct.

5.3 Indegree and Outdegree Distribution

We draw an interaction network for a specific category. The *indegree* of a user is the number of users that respond to this user's threads, and the *outdegree* of a user is the number of users to which the user has responded. We examined three categories: *Disney Trip Reports*, *Disney Restaurants*, and *Adventures by Disney*. Each of these three categories are selected from the report, fact and discussion categories, respectively.

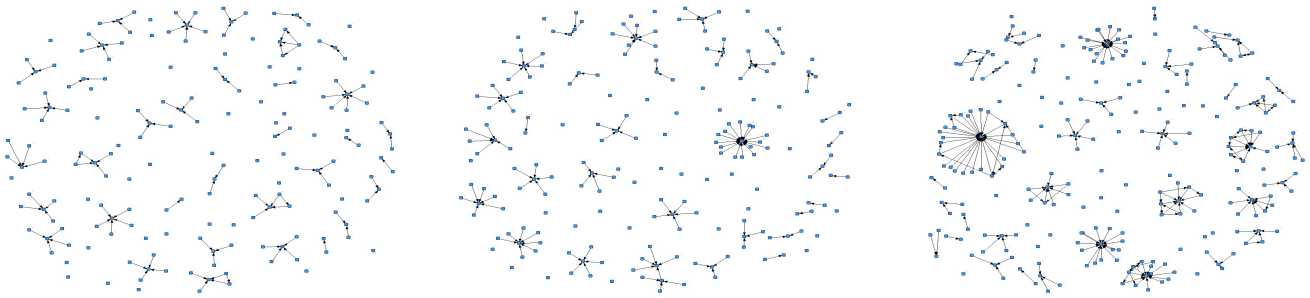
Figure 24 and Figure 25 show the CCDF of indegree and outdegree for these three categories. We can see that the users differ in their activity level in all three categories. We can see that *Adventures by Disney* has much larger indegree and outdegree ranges than the other two categories. It is because in the discussion categories, the users would tend to have more discussion, which means the users might reply more to others or receive more replies from others. This will result in a higher indegree and outdegree in these categories. Figure 24 and Figure 25 also show that the report categories (*Disney Trip Reports*) has a narrower indegree and outdegree distribution than the fact categories (*Disney Trip Reports*). It is because in the report categories, most of the thread creators (reporters) are more concerned about their own threads and update their own threads by replying or even have no replies. This will lead to a lower indegree and outdegree comparing to other types of categories.

Specifically, all the categories tend to have larger outdegree ranges than the indegree, which indicates that some users reply to more users than the number of users who reply to them or the number of threads created by themselves. In the report categories, some users make a reply to the reporters in order to attain useful information. In the fact categories, it might be the reason that a certain number of warm-hearted users like to help others voluntarily, but do not often ask for help themselves. Also, some users are hotels and restaurants that reply to many users but do not need to ask. In the discussion categories, some users regularly offer advices or join in discussion.

In summary, we have made three observations regarding the indegree and outdegree distributions. (1) The users in the discussion categories tend to receive and reply more than the report and fact categories, while the users in the report categories tend to have fewer replies from others and reply fewer to others than the fact categories. (2) All the categories have some users who reply more than create new threads or receive replies from others.

5.4 Analysis of Ego Network

The ego network of a user consists of the user, the ties to other users the person interacts with directly, and interactions between those users. That is, if a user A has replied to another user B, then A have a link point to B. Thus, the ego networks of all the users within a category provide a simple visualization of users' interactions; a densely connected ego network indicates discussion between users, while an ego network with few connections indicates more limited



(a) *Disney Trip Reports* (b) *Disney Restaurants* (c) *Adventures by Disney*

Fig. 26: Ego networks of three categories.

interactions. From users' ego network we can tell that a person is an "answer persons" or a "discussion persons" [33]. "Answer persons" are tied to many neighbors who have few ties. In contrast, "discussion persons" are highly tied to others who are highly connected.

In order to have a better understanding of individual users within different categories, we drew ego networks for three categories from the report, fact and discussion clusters, respectively: *Disney Trip Reports*, *Disney Restaurants*, and *Adventures by Disney*. Figures 26(a), 26(b) and 26(c) show the three selected categories' ego networks of 100 randomly sampled users of each category, respectively. We define the weight between two nodes as the number of times that a node has replied to the other node.

From these figures, we can see that the most active users of category *Disney Trip Reports* are "answer persons", because there are no interactions between their neighbors. It implies that in the report categories, the repliers would not reply to other repliers since the repliers are mostly askers. *Disney Restaurants* has an ego network very similar as *Disney Trip Reports*. The most active users in *Disney Restaurants* are also "answer persons", because the thread creators generally are novices in the fact categories. The ranges of the weights for these two categories both are $[1, 3]$, most of which are 1 for the links. This range of the weights is very small, which also further demonstrates that these two categories are defined into the correct clusters.

On the other hand, the neighbors of some of the highly active users in *Adventures by Disney* are themselves connected with others, which indicates that they are more likely to be "discussion persons". The reason is that discussion categories tend to have more discussion, which attracts users to both create and reply and make the ego network dense. The range of the weights is $[1, 10]$, which means some users in this category even reply to other users ten times. Moreover, more than 40% of the weights of the links are higher than 2. Therefore, all the observations above prove that the discussion categories tend to have a densely ego network, which means the most active users are more likely to be "discussion persons".

In summary, we have made some observations: (1) In the report and fact categories, most active users are "answer persons" because most of their neighbors have few interactions. (2) In the discussion categories, most active users are "discussion persons" because their neighbors have many replies to each other.

6 EXPERTISE AND KNOWLEDGE ACROSS CATEGORIES

In this section, we describe the knowledge sharing in DISboards using two metrics. The first measures users' entropy, namely the width of categories the user participates in. The second is the relationship between categories, which means that the portion of users who are actively answer questions in one category also are active in the other categories.

6.1 User Entropy

User entropy is a measurement that can capture the degree of concentration in a person's reply patterns to particular topics [4]; We utilize this measurement to analyze user activities on DISboards. In detail, the entropy [27] of a user k is calculated by:

$$E_k = -\sum_{i=20} p_k(x_i) * log(p_k(x_i)) \quad (3)$$

where $p(x_i)$ is defined as the probability that the user k posts in i^{th} category. This is calculated by the number of posts in the i^{th} category divided by the total number of posts on DISboards of the user, as the previous paper defined [4]. Thus, for a specific user k, we have $\sum_{i=20} p_k(x_i) = 1$.

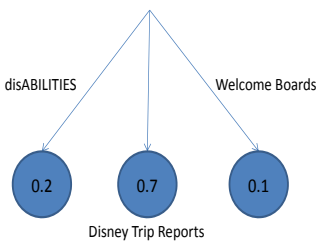


Fig. 27: Illustration of the hierarchical entropy calculation.

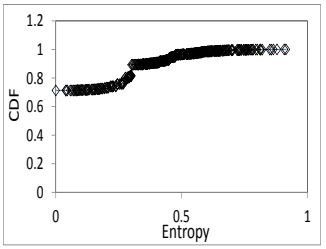


Fig. 28: The distribution of entropy.

Figure 27 gives an example on how to calculate entropy. For example, for a certain user, his posts are all in three categories *disABILITIES*, *Disney Trip Reports*, and *Welcome Boards* with 0.2, 0.7, and 0.1 proportion, respectively. The entropy values are $E_1 = -0.2 * log(0.2) = 0.140$, $E_3 = -0.7 * log(0.7) = 0.108$, and $E_3 = -0.1 * log(0.1) = 0.100$. Then, we sum up all the three entropies for this user, and obtain the total entropy for the user, which equals $0.140 + 0.108 + 0.100 = 0.348$. Note that if the user is very concentrated and replies in only a few categories, the entropy is small. On the other hand, when the entropy becomes larger, it means that the user replies in more

categories [4]. For example, if a user replies onece ever, the entropy of this user is 0, which means this user is only concentrated on one category.

Figure 28 shows the CDF of entropy distribution of all users. We can see that more than 70% of the users have an entropy that is zero or very close to zero,which means that they are highly concentrated on one category or only a few categories. About 10 users have an entropy that is larger than 0.8 and the largest entropy is 0.91. However, comparing to the SNs in *Yahoo! Answers* [4] that have over 80% of users with entropy larger than 1, this value is still relatively small, which also indicates that users are very concentrated on DISboards. It is because DISboards has a few categories of Disney, while *Yahoo! Answers* covers much categories. Since people may encounter diverse problems in life, users on *Yahoo! Answers* are more likely to post on several categories. This will lead to a higher entropy on *Yahoo! Answers*. Therefore, the results turn out that the users concentrate on fewer categories in DISboards compared to *Yahoo! Answers*.

Hence, in DISboards, most users are very concentrated since they might only care about a certain aspect of Disney. Similarly with other forums, users tend to concentrate only on the categories they are interested in.

6.2 Relationship between Categories

We then track the similarity between categories by tracking the post patterns. We use $\mathcal{A}$ and $\mathcal{B}$ to represent the set of users that have created threads in two categories, respectively. *Poster similarity* between the two categories is defined as $\frac{\mathcal{A}\cap\mathcal{B}}{\mathcal{A}\cup\mathcal{B}}$. We also use $\mathcal{C}$ and $\mathcal{D}$ to represent the set of users that have replied in two categories, respectively. *Replier similarity* between the two categories is defined as $\frac{\mathcal{C}\cap\mathcal{D}}{\mathcal{C}\cup\mathcal{D}}$. We assign the most similarity with the most grey and the least similarity with the least grey. Using this method, we generated a grey-scale map in Figure 29 to represent both poster similarity and replier similarity, where the higher similarity, the greyer map between two categories. The figure shows that the categories related to Disney Vacation club (DVC) have both the highest post similarity and replier similarity with other DVC categories. It is probably because in these categories, both the thread creators and repliers are almost the DVC members. As a club's members, they tend to be more active and have discussion in different DVC categories.

Besides, the *Transportation* category has high poster similarity and replier similarity with all the other categories, especially with the *budget board*. Most people would make plans before they travel to Disney. Transportation is the first issue they usually consider in plan, e.g., which season has the lowest flight tickets, vehicle renting issue, and so on. Moreover, budget is another important issue that they need to consider in their plan. Since the topics in these two categories are mostly correlated with users' concern about their plans, it is not surprising that these two categories have high similarities. Furthermore, the *budget board* has great poster and replier similarities with *Disney Restaurants* and *Orlando Hotels and Attractions*. It is because generally, besides transportation, the other two main concerns of trips are from food and hotels. Therefore, the users may visit all these categories and post in order to attain useful information for their plans, which results in high similarities among these categories. *Disney for Families* is also highly correlated with the categories of transportation, restaurant, budget and hotel. This is because families are most likely to plan their trips before they go to Disney. Also, families are more likely to travel to Disney for resorts, which makes *Disney for Family* similar with *Disney Resorts*. Many users might be interested in comprehensive vacation packages with transportation, hotel accommodations, etc. Therefore, DISboards can target users of these categories with special deals or discounts on packages to improve sales.

In summary, we have made three observations. (1) We find that most categories in DISboards do not have high poster and replier similarity, which indicates that most users are concentrated on one category or a few categories. (2) A user tends to post in multiple correlated categories such as primary trip planning categories (e.g., Transportation, Budget Board, Disney Resorts, Disney Restaurants). It is also true in real world that users tend to have interest on some related topics. (3)Many users might be interested in comprehensive vacation packages with transportation, hotel accommodations, etc. Therefore DISboards can target users of these categories with special deals on packages to improve sales.

7 Discussions and Implications

Understanding the motivations of people's participation in discussion forums has been widely studied recently. Unlike the traditional QA websites that adopt reputation-ranking policy to encourage users to share knowledge, discussion forums do not have any similar mechanisms. Therefore, it is critical to study the motivations of knowledge sharing to further motivate people to contribute their knowledge in discussion forums.

In this paper, we first analyze the user lifetime distribution. Our results indicate that many users (around 80%) only use discussion forums once to ask some questions and leave. It is essential to keep this part of users and make them switch from askers to answerers. One simple way to keep users staying longer may be to send them reminders or advertisements. DISboards can further reward such users with certain discount to their future travels to Disney if they can reply to others and share their previous travel experiences in the discussion and report categories.

We then analyze the SN structure and the effect of SN on forum usage. Our results indicate that SN does incentivize users to be active in the forums. However, high-degree users in the SN do not necessarily receive many replies in their threads. This could be because that DISboards lacks a mechanism to actively alert a user of his/her friends' posting activities. Hence, to stimulate user activities in the forum and attract more users, discussion forums such as DISboards can incorporate more SN components such as adopting the alert mechanism to actively notify users of their friends' new posts. Besides, adopting some reputation-ranking mechanisms may also help incentivize the users to be active in answering the new users' questions and sharing their previous travel experiences.

Our findings related to category characteristics may also be useful in incentivizing users. A user tends to post in

(a) (b) (c)

Fig. 29: Similarities between categories (the shades correspond to different scales). (a) overlap in users who replied in both categories. (b) each number represents one category. (c) overlap in users who created threads in both categories.

multiple correlated categories such as primary trip planning categories (e.g., Transportation, Budget Board, Disney Resorts, Disney Restaurants). It is also true in real world that users tend to have interest on some related topics. Many users might be interested in comprehensive vacation packages with transportation, hotel accommodations, etc. Therefore, DISboards can target users of these categories with special deals on packages to improve sales when they first join the SN of DISboards, and when they recommend a specific amount of friends to join DISboards.

Moreover, in the category characterizing analysis we observed that users in discussion and report categories tend to be more active. This is because knowledge sharing is another important motivation to keep users active. Users will learn from other users with different aspects of their trips. Introducing the best answer mechanisms may also help keep users active, as previous studies [28] demonstrated that being the best answerer is one motivation for users to keep sharing knowledge. For example, users can achieve awards from being the best answerer and the awards can be used to redeem some special discounts during their trips.

8 Conclusions

In this paper, we examine the SN and knowledge sharing activities in the DISboards forum. We analyze the SN structure, effect of SN on the forum, category characteristics, ego networks of user interactions, user entropy, and relationship between categories.

First, we find that the SN in DISboards exhibits a power-law distribution like other SNs. The global clustering coefficient of the DISboards SN is low and the cluster coefficient decreases as the user degree increases. Higher-degree users are more active in the forum but do not necessarily receive many replies to their threads. Though the vast majority of users are adults, teens constitute a significant part of the SN and they are more active in the forum.

Second, we cluster the selected categories into three groups (report, fact and discussion) and characterize their properties. We find that the discussion categories have more users who both create threads and reply to others. Also, the users in the discussion categories are more active than in the report and fact categories.

Third, most active users in the discussion categories are tied to others who themselves are highly connected, while most active users in the report and fact categories are tied to users who themselves have few ties. Further, the users are quite concentrated on a few categories and few users post across several categories.

Our observations suggest that to stimulate user activities in the forum and attract more users, DISboards can incorporate more SN components such as adopting the alert mechanism to actively notify users of their friends' new posts. In the future work, we will examine leveraging user expertise in the information sharing and enhancing SN to stimulate user activities. Based on the conclusions in this paper, we will further study the psychological effect on the discussion forums that incorporate SN.

Acknowledgements

This research was supported in part by U.S. NSF grants OAC-1724845, ACI-1719397 and CNS-1733596, and Microsoft Research Faculty Fellowship 8300751.

References

[1] Internet usage statistics. http://www.internetworldstats.com/stats.htm.
[2] Something awful. http://www.somethingawful.com/.
[3] Ultimate guitar. http://www.ultimate-guitar.com/.
[4] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and Yahoo Answers: Everyone knows something. In *Proc. of WWW*, 2008.
[5] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proc. of WWW*, pages 835–844, 2007.
[6] Alexa. http://www.alexa.com/.
[7] J. Arguello, B. Butler, E. Joyce, R. Kraut, K. Ling, C. Rosé, and X. Wang. Talk to me: Foundations for successful individual-group interactions in online communities. In *Proc. of SIGCHI*, 2006.
[8] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proc. of SIGKDD*, 2006.
[9] K. Burghardt, E. F. Alsina, M. Girvan, W. M. Rand, and K. Lerman. The myopia of crowds: A study of collective evaluation on stack exchange. *Robert H. Smith School Research Paper No. RHS*, 2736568, 2016.
[10] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10:10–17, 2010.
[11] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
[12] Disboards. http://www.DISBoards.com/.
[13] Disboards guidelines. http://www.wdwinfo.com/guidelines.htm/.

[14] H. Dong, J. Wang, H. Lin, B. Xu, and Z. Yang. Predicting best answerers for new questions: An approach leveraging distributed representations of words in community question answering. In *Proc. of FCST*, pages 13–18. IEEE, 2015.

[15] N. Z. Gong, W. Xu, L. Huang, P. Mittal, E. Stefanov, V. Sekar, and D. Song. Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *Proc. of IMC*, 2012.

[16] R. Gonzalez, R. Cuevas, R. Motamedi, R. Rejaie, and A. Cuevas. Google+ or google-?: dissecting the evolution of the new OSN in its first year. In *Proc. of WWW*, 2013.

[17] E. Joyce and R. Kraut. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11(3):723–747, 2006.

[18] S. Kim, J. S. Oh, and S. Oh. Best answer selection criteria in a social Q&A site from the user-oriented relevance perspective. In *Proc. of ASIST*, 2007.

[19] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc. of WWW*, 2010.

[20] K. Lakhani and E. Von Hippel. How open source software works: "free" user-to-user assistance. *Research policy*, 32(6):923–943, 2003.

[21] J. Lang and S. F. Wu. Social network user lifetime. *Social Network Analysis and Mining*, 3(3):285–297, 2013.

[22] A. Lenhart, K. Purcell, A. Smith, and K. Zickuhr. *Social media & mobile internet use among teens and young adults*. Pew Internet & American Life Project Washington, DC, 2010.

[23] Z. Li and H. Shen. Learning network graph of sir epidemic cascades using minimal hitting set based approach. In *Proc. of ICCCN*, 2016.

[24] Z. Li, H. Shen, and J. Grant. Collective intelligence in the online social network of Yahoo! Answers and its implications. In *Proc. of CIKM*, 2012.

[25] F. J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[26] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. of SIGCOMM*, 2007.

[27] C. E. Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

[28] Y. Shoji, S. Fujita, A. Tajima, and K. Tanaka. Who stays longer in community qa media?-user behavior analysis in cqa. In *International Conference on Social Informatics*, pages 33–48. Springer, 2015.

[29] Q. Su, D. Pavlov, J. Chow, and W. Baker. Internet-scale collection of human-reviewed data. In *Proc. of WWW*, 2007.

[30] Tumblr. https://www.tumblr.com/.

[31] B. Viswanath, A. Mislove, M. Cha, and K. Gummadi. On the evolution of user interaction in Facebook. In *Proc. of Sigcomm Workshop*, 2009.

[32] G. A. Wang, H. J. Wang, J. Li, A. S. Abrahams, and W. Fan. An analytical framework for understanding knowledge-sharing processes in online q&a communities. *ACM Transactions on Management Information Systems (TMIS)*, 5(4):18, 2015.

[33] H. T. Welser, E. Gleave, D. Fisher, and M. Smith. Visualizing the signatures of social roles in online discussion groups. *Journal of social structure*, 8(2):1–32, 2007.

[34] R. W. White, M. Richardson, and Y. Liu. Effects of community size and contact rate in synchronous social q&a. In *Proc. of SIGCHI*, 2011.

[35] Yahoo! answer. https://answers.yahoo.com/.

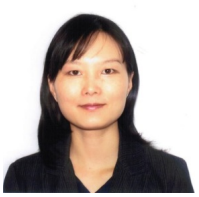
[36] Y. Yao, H. Tong, T. Xie, L. Akoglu, F. Xu, and J. Lu. Detecting high-quality posts in community question answering sites. *Information Sciences*, 302:70–82, 2015.

[37] X. Zhao, A. Sala, C. Wilson, X. Wang, S. Gaito, H. Zheng, and B. Y. Zhao. Multi-scale dynamics in a massive online social network. In *Proc. of IMC*, 2012.

[38] Y. Zhu. Measurement and analysis of an online content voting network: a case study of Digg. In *Proc. of WWW*, 2010.

Zhuozhao Li received the BS degree in Optical Engineering from Zhejiang University, China in 2010, and the MS degree in Electrical Engineering from University of Southern California in 2012. He is currently a Ph.D student in Department of Computer Science at University of Virginia. His research interests include data analysis and cloud computing.

Harrison Chandler received his BS degree in Computer Engineering from Clemson University in 2012. He is currently a Ph.D. student in the Department of Electrical Engineering and Computer Science at University of Michigan. His research interests include embedded and distributed systems.



Haiying Shen received the BS degree in Computer Science and Engineering from Tongji University, China in 2000, and the MS and Ph.D. degrees in Computer Engineering from Wayne State University in 2004 and 2006, respectively. She is currently an Associate Professor in the Department of Computer Science at University of Virginia. Her research interests include distributed computer systems and computer networks, with an emphasis on P2P and content delivery networks, mobile computing, wireless sensor networks, and cloud computing. She is a Microsoft Faculty Fellow of 2010, a senior member of the IEEE and a member of the ACM.