Can Dynamic Knowledge-Sharing Activities Be Mirrored From the Static Online Social Network in Yahoo! Answers and How to Improve Its Quality of Service?

Haiying Shen, *Senior Member, IEEE*, and Guangyan Wang

***Abstract*—Yahoo! Answers is an online platform where users can post questions and answer other users' questions. Our previous work studied the online social network (OSN) of Yahoo! Answers by analyzing information from the profiles (including fans, contacts, and interests) of top contributors and their related users. Rather than using the static profile information from the top-contributor-centered dataset, in this paper, we particularly analyze the actual questioning and answering (Q/A) behaviors of normal users. We build a Q/A network that unidirectionally connects each asker to his/her answerers. We analyze the structural characteristics of the Q/A network, user Q/A activities, and knowledge base of all users. In addition to the observations similar to our previous study, which indicates that the OSN of Yahoo! Answers can reflect user Q/A activities to a certain extent, we additionally observe that: 1) a large portion of users only ask questions without answering others' questions; 2) users are active in more knowledge categories than those indicated in their profiles; and 3) the knowledge categories of the top-contributor-related users cannot represent those of normal users. Finally, we analyze the characteristics of questions and answers in different knowledge categories. This paper not only provides an understanding of actual Q/A activities of users but also showcases the aspects of Q/A activities that the OSN of Yahoo! Answers can and cannot accurately reflect. Based on the insights gained from this paper, we propose a few methods to help improve the quality of service of Yahoo! Answers.**

***Index Terms*—Knowledge sharing, Question and Answer (Q&A) systems, Yahoo! Answers.**

I. Introduction

WEB search engines enable keyword-based search for information retrieval. They extract related information from large datasets and rank them by relevancy. Web search engines are suitable for information retrieval in enormous existing datasets on the Internet, but are not effective for nonfactual questions that do not have definite answers [1]. Also, they only return information for certain keywords, which would involve tedious work for a user to find what is truly needed. For example, if a basketball fan wants to know the Los Angeles Lakers roster when the Boston Celtics got their "big three," he may enter "lakers roster celtics big three" into the search engine, but can hardly find any useful information in the returned results.

Question and Answer (Q&A) systems such as Yahoo! Answers play a vital role in filling the gap of answering nonfactual questions and questions that are not easily searched by keywords in search engines [2]. These Q&A systems provide a platform where users can post questions and answer other users' questions. Users ask full questions instead of entering keywords, and the questions are answered by other users instead of by searching in the database. In this way, questions are better explained and better understood, since people are most capable in parsing and interpreting questions. Different people have different knowledge bases and their collective intelligence is comprehensive enough to provide answers to reasonable questions. Yahoo! Answers categorizes all questions into 26 general knowledge categories, with each general category consisting of a number of detailed knowledge categories. Leveraging the collective intelligence of their users, Q&A systems have become a favorable alternative to Web search engines. However, Q&A systems suffer from some major shortcomings such as long latency to receive answers, no answers for a question, and low trustworthiness of answers (e.g., spam). Understanding the questioning and answering (Q/A) activities of users is essential toward improving the performance of Q&A systems.

The motivation of this paper is to see if the dynamic Q/A activities can be reflected by the static online social network (OSN) in Yahoo! Answers (formed only by top contributors and their related users). If yes, instead of collecting and analyzing a huge amount of Q/A activity data during a long time, people only need to analyze the partial existing OSN in Yahoo! Answers to learn the actual or predict the future Q/A activities, which makes the formidable task much easier and faster. We present the details of our motivation below.

Yahoo! Answers incorporates an OSN, in which user A can connect to user B if A wants to subscribe to every

Manuscript received January 24, 2016; revised April 27, 2016; accepted June 3, 2016. This work was supported in part by the National Science Foundation under Grants NSF-1404981, IIS-1354123, CNS-1254006, in part by IBM Faculty Award 5501145, and in part by Microsoft Research Faculty Fellowship 8300751. This paper was recommended by Associate Editor F. Wang.

The authors are with the Department of Electrical and Computer Engineering, Clemson University, Clemson, SC 29634 USA (e-mail: shenh@clemson.edu; guangyw@clemson.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2016.2580606

answer and question from B. This knowledge-oriented OSN is a unidirectional network in that users can follow whoever they want without the confirmation from the one to be followed. Our previous work [3], [4] studied the OSN of Yahoo! Answers through user profile dataset that is collected by starting from the 4000 top answer contributors and following their OSN links to all the reachable users. With this top-contributor-centered OSN dataset, we have obtained the following findings: 1) the OSN of Yahoo! Answers has very low-level link symmetry with weak correlation between indegree and outdegree; 2) 10% of users contribute to 80% of the best answers and 70% of all the answers; 3) there exists a positive linear relationship between the number of answers and the number of best answers of a user; and 4) the knowledge categories interested by users are highly clustered. This previous work is the first to extensively study the OSN of Yahoo! Answers, which can help developers understand the nature and impact of collective intelligence in the OSN of Yahoo! Answers.

However, all users involved in our previous study have direct or indirect connections with top contributors in the OSN of Yahoo! Answers (related nodes of top contributors in short). This portion of users excludes those who use Yahoo! Answers only as a platform for Q/A activities rather than a social platform. Thus, our previous top-contributor-centered dataset may not represent the overall user Q/A behaviors in Yahoo! Answers. Also, our previous study extracted information from user profiles, which may not comprehensively or accurately reflect users' actual activities (e.g., user may not indicate all the knowledge categories they are interested in or keep them updated). Further, our previous study assumes that the static OSN relationship reflects their actual Q/A interactions, which may not be true. In this paper, we intend to investigate the following: 1) the actual Q/A activities of users in Yahoo! Answers and 2) whether the OSN of Yahoo! Answers reflects user actual Q/A activities; that is, whether the actual user Q/A activities in Yahoo! Answers follow our previous observations from the OSN of Yahoo! Answers.

Based on our crawled dataset of actual Q/A activities of users from Yahoo! Answers (i.e., Q/A dataset), we constructed a Q/A network that unidirectionally connects each asker to his/her answerers. We define indegree and outdegree of a node as the node's number of answers and questions, respectively. We analyze the structural characteristics of the Q/A network, user Q/A activities, and the knowledge base and behaviors of all users in our dataset. We also explore the knowledge distribution and coexistence of different knowledge categories in each user's interests and analyze the characteristics of questions and answers in different general knowledge categories.

After studying the structural properties of the Q/A network, we found that indegree and outdegree: 1) approximately follow the power-law distribution; 2) have low link symmetry; and 3) exhibit weak correlation. We also found that Yahoo! Answers has even lower reciprocity (i.e., bidirectional connection) rate in our Q/A dataset than in our previous OSN dataset.

By investigating the knowledge base and behaviors of all users in our dataset, we obtained the following findings: 1) the majority of best answers and answers are contributed by the top 10% of users; 2) a large portion of users ask only a few questions and do not give any answers; 3) there exists a high correlation between the number of best answers and the number of all answers of a user; 4) users are involved (ask or answer questions) in more categories than they indicated on their profiles; 5) the interests of top contributors and their related users cannot represent those of normal users; and 6) around 37% of the users provide no answers, in which 64% are one-time users (i.e., users with only one question).

This paper on the characteristics of questions and answers in different knowledge categories led to the following observations.

1) General knowledge categories with more factual questions receive fewer answers, while controversial and opinion-seeking knowledge categories (e.g., Pregnancy & Parenting, Society & culture, and Sports) receive more answers.
2) Social Science, Arts & Humanities, Health, and Science & mathematics are the knowledge categories with most verbose answers.
3) Politics & Governments is the obvious winner when it comes to the number of words to describe a question.

Comparing our observations from actual Q/A activities and our previous observations from the dataset of the OSN of Yahoo! Answers [3], [4], we can conclude that the static OSN relationship can reflect the characteristics of users' actual Q/A activities in Yahoo! Answers to a certain extent. Additional observations can be summarized below: 1) there are a large portion of users that are one-time knowledge consumers of the Yahoo! Answers platform; 2) real knowledge categories of normal users are more scattered than those indicated in the profiles of top contributors and their related users; and 3) factual questions tend to have fewer answers while controversial and opinion-seeking knowledge categories have more answers and longer answer lengths. Finally, from our analysis, we identify the challenges currently faced by Yahoo! Answers, and suggest several possible methods to improve the Yahoo! Answers system by leveraging our analytical results.

This is the first work that reveals whether the static OSN relationship (formed only by top contributors and their related users) can mirror the characteristics of users' actual dynamic Q/A activities in Yahoo! Answers. The rest of this paper is organized as follows. Section II gives an overview of related work. Section III introduces background and measurement methodology. Based on the users' actual Q/A activities, Section IV presents analytical results of the Q/A network and Section V presents the analytical results of knowledge distribution and user behaviors, and the features of different knowledge categories. Section VI presents our suggested methods to improve Yahoo! Answers performance. Finally, Section VII concludes this paper with remarks on our future work.

II. Related Work

This paper is aimed to see if the dynamic Q/A activities can be reflected by the static OSN in Yahoo! Answers (formed only by top contributors and their related users). If yes, instead of collecting and analyzing a huge amount of Q/A activity data during a long time, people can directly use the partial existing OSN in Yahoo! Answers to learn the actual or predict the future Q/A activities for improving the quality of service and the quality-of-user experience of Q&A systems. The topic of knowledge-sharing has been widely studied for many years. In the following, we classify the related work into three categories for discussion and will indicate the difference between this paper and the previous works in the end.

A. Q&A Systems

One research study on Q&A systems is about finding the best answerers for a question. Szpektor *et al.* [5] proposed a probabilistic representation of users and their matching questions. Ji and Wang [6] proposed to rank potential answerers on their expertise degrees for each question by using a learning model. Pal *et al.* [7] proposed a *k* nearest neighbor-based aggregation method to compute community scores in online community Q&A systems, which are used to route questions to the right set of communities. Zhao and Mei [8] first distinguished real questions from ordinary tweets with an automatic classifier, and then found that the questions on Twitter can predict the trends of Google queries through a comprehensive analysis. Qi *et al.* [9] proposed a probabilistic model to jointly assess the reliability of potential answerers in order to select good potential answerers for a question. Wang *et al.* [10] proposed an analogical reasoning-based approach that takes into account the relationship between the question and the quality of the answer to find the best answerer. Dror *et al.* [11] addressed recommending questions to appropriate users by exploiting the content and social signals that users provide regularly. The works in [12] and [13] have studied utilizing user expertise in answer ranking. The works in [14]–[16] have analyzed user activity in community question answering services. Furlan *et al.* [17] presented a survey of intelligent question routing systems.

Many other aspects of Q&A systems also have been investigated. Chan *et al.* [18] proposed to automatically classify the general questions into corresponding topic categories by using a hierarchical kernelized classification method. Liu and Nyberg [19] presented an answer ranking approach for Q&A systems that incorporates both cascade model and result voting model. Adamic *et al*. [20] analyzed the features of answer contents, and presented a prediction model to predict whether a particular answer will be chosen as the best answer. Gardelli and Weber [21] categorized questions in Yahoo! Answers into "informational" and "conversational." They used toolbar data to analyze the relationship between prequestion behavior and the types of questions a user would ask. Su *et al.* [22] used the answer ratings in Yahoo! Answers to study the quality of human reviewed data on the Internet. Kim *et al*. [23] studied the criteria for best answers by analyzing the best answer features in Yahoo! Answers. It improves answer search by using language models to exploit categories of questions. Liu *et al.* [24] analyzed the content, structure and community-focused features and gave an inclusive predictive model to predict whether an asker will be satisfied with the answers. Dearman and Truong [25] explored the reason why most users choose not to answer a question that they have browsed by taking a survey on 135 active members of Yahoo! Answers and showed several reasons such as subject nature and composition of the question, perception of how the questioner will receive, interpretation and reaction to their answers, and suspicion that their answers will be lost in the crowd of answers. Shtok *et al.* [26] proposed a method based on natural language processing to answer unanswered questions using the repository of solved questions.

B. Knowledge Sharing

Many Q&A systems have been proposed for knowledge sharing on the Internet. Harper *et al.* [27] proposed MiMir, where a question is broadcasted to all users in the system. White *et al.* [28] proposed IM-an-Expert that automatically identifies experts based on information retrieval techniques and uses instant messaging for real-time dialog. Horowitz and Kamvar [29] attempt to route the question from a user to all appropriate users in his/her social community. Yang and Chen [30] presented a system for supporting interactive collaboration in knowledge sharing over a peer-to-peer network by leveraging OSN. They found that by leveraging social network-based collaboration, it will help people find relevant content and knowledgeable collaborators who are willing to share their knowledge with. Wang *et al*. [31] introduced a framework that supports the entire pipeline of interactive knowledge harvesting. Their demo exhibits fact extraction from *ad-hoc* corpus creation, via relation specification, labeling, and assessment all the way to ready-to-use RDF exports.

C. General OSN-Based Q/A Systems

Previous research also studied the Q/A systems in general OSNs. Morris *et al.* [32] investigated the types of questions people ask and answer in a general OSN and the (dis)advantages of using OSN for information seeking in comparison with search engines. Teevan *et al.* [33] studied the factors that affect the quantity, quality, and speed of responses for questions through status messages in an OSN. This did their survey with 282 participants posting variants of the same question as status message on Facebook to analyze the affecting factors. Yang *et al.* [34] studied the cultural differences in people's question asking behaviors by conducting a survey among 933 people across four countries, and revealed that culture is a significant factor in predicting people's social Q/A behavior. Richardson and White [35] proposed prediction models to predict if a question will be answered, the number of candidate answerers for the question, and if the asker will be satisfied with the answer. They made prediction during the life cycle of a question to improve the Q/A process.

Unlike the previous works, this paper focuses on verifying if the OSN of Yahoo! Answers can reflect the actual user Q/A

TABLE I
HIGH-LEVEL STATISTICS OF OUR CRAWLED Q/A DATASET

# of questions	1,667,751
# of users involved in the questions collected	555,118
# of answers	5,555,920
# of the best answers	832,202
Ave. # of questions per user	3.004
Ave. # of best answers per user	1.499434
Ave. # of answers per user	10.008

activity. This paper can be leveraged to more effectively utilize the OSN of Yahoo! Answers, and more synergistically utilize both the OSN of Yahoo! Answers and Q/A activity information in Yahoo! Answers performance enhancement.

III. BACKGROUND AND MEASUREMENT METHODOLOGY

Yahoo! Answers, as a knowledge market, was launched by Yahoo! on July 5, 2005. It allows users to ask questions and answer the questions posted by other users. An asker's posted question is initially open to be answered for four days. The asker can choose to close the question after a minimum of 1 h or extend the active time for a period of up to eight days. A question cannot be answered after the open time period. After an asker receives answers, it can select the best answer. If a question has received answers and the open time period is elapsed but the asker has not selected the best answer, it is in the in-voting status, and there will be a two days period for users to vote for the best answer. When the best answer is selected for a question, this question is resolved.

In a user's profile, there are two lists of people: 1) fans and 2) contacts. Fans are those who follow this user and contacts are other users that this user follows. If user A wants to frequently visit or track all questions and answers of user B, A adds B to his/her contact list by building a link to B. Then, A becomes B's fan. These unidirectional links connect nodes to an OSN in Yahoo! Answers, with each node having OSN indegree and outdegree. The nodes in a user's contact list are its outdegree nodes, and the nodes in a node's fan list are its indegree nodes.

An asker needs to pay five points for asking one question. An answerer receives two points for answering a question and receives ten points if his/her answer is selected as the best answer. Points cannot be traded and only serve to indicate how active a user has been on the Yahoo! Answers website. Users with many points are recognized as top contributors by the system. A top contributor is a member of the answerer community who is considered knowledgeable in particular knowledge categories. Based on the point distribution among knowledge categories of the questions answered by a top contributor, the system determines up to three knowledge categories that the top contributor is knowledgeable in.

In this paper, we attempt to investigate the characteristics of the actual Q/A activities of users in Yahoo! Answers. We collected the questions from all knowledge categories in a two-month period from January, 2012 to March, 2012. A question without any answer was also collected. For each question, we recorded its general knowledge category, detailed knowledge category, asker and all answerers of the question. There are a total of 1 667 751 questions, 5 555 920 answers for these questions, among which 832 202 answers are the best answers. We call this dataset Q/A dataset. All of our collected questions are resolved. Table I shows the overall statistics of the Q/A dataset we crawled.

TABLE II
DIFFERENCES BETWEEN THE TWO DATASETS

	OSN dataset [3], [4]	Q/A dataset
Interests	retrieved from each user's profile	inferred from all questions each user asked or answered
User relationship	contact-fan relationship in the OSN of *Yahoo! Answers*	asker-answerer relationship of a given question
Questions	involving top contributors and their related users	involving normal users

Our previous work [3], [4] studied the dataset of the OSN of Yahoo! Answers. There are three major differences between our newly crawled Q/A dataset and the OSN dataset as listed in Table II. Our previous study assumes that the static OSN contact-fan relationship reflects the actual Q/A behaviors and the interests in a user's profile reflect his/her real interests. Also, OSN dataset only covers the top contributors and their related nodes. Due to these differences, it is important to analyze the actual Q/A interaction relationship rather than the static contact-fan relationship in the OSN, to infer users' more accurate interests from their Q/A activities, and to study the group of normal users instead of top-contributor-related users. Through this paper that more comprehensively and accurately showcases normal user Q/A activities, we can verify our previous assumptions and conclusions and also make additional observations. Further, the study on the general users rather than the top-contributor-related users can avoid the bias on the study user group.

IV. ANALYSIS OF Q/A ACTIVITIES

In this section, we construct the Q/A network in Yahoo! Answers and study its structural characteristics and user Q/A activities, and compare the results with previous studies on the OSN of Yahoo! Answers. In the Q/A network (V, E), V denotes all users in our Q/A dataset and link $e \in E$ connects asker A to user B if user B has answered at least one question from A. We define a user's indegree as the number of questions answered by the user and define a user's outdegree as the number of questions asked by the user. We call them Q/A indegree and Q/A outdegree in order to distinguish them from the OSN indegree and outdegree. Note that Q/A indegree and Q/A outdegree are not the indegree and outdegree of a node in the Q/A network. Q/A indegree and outdegree reflect not only the number of answers and questions of a user but also the frequency of the user in asking and answering questions as the Q/A dataset is for a certain time period, so they more accurately reflect the active degree of a user's Q/A activities compared to the OSN indegree and outdegree. Fig. 1 shows a snapshot of the Q/A network. We see that links are highly clustered with a few nodes having many links and many nodes having few links. The results indicate that a few

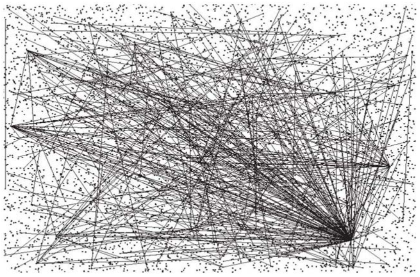


Fig. 1. Snapshot of the Q/A network.

users have many answerers for their questions and a few users are very active in answering many users' questions while most users are inactive in Q/A activities. We see that the Q/A network shares similar structural characteristics with the OSN of Yahoo! Answers network [4].

A. Reciprocity

As in the general OSN (e.g., Facebook and Twitter), we use reciprocity to reflect the pairwise bidirectional relationship between two nodes in our Q/A network. Although reciprocity in a unidirectional network is known to be much smaller than general OSNs, it is still a good measure of the social pressure or personal impulse of users to return the favor to those who have answered their questions, i.e., to answer back. Empirical evidence shows that users have the impulse to click on the answerer's profile when they saw a satisfactory answer and they feel obliged to give back an answer to the question from the answerer. To study this user behavior, we measured the reciprocity rate defined as the number of reciprocity links over all links of all users in the Q/A network and compare it with those in other unidirectional networks. The results are listed in Table III. In OSNs such as Flickr and Yahoo!360, real life friends are likely to connect with each other, which leads to a high percent of bidirectional links and hence high reciprocity rate. Digg, Yahoo! Answers, and Twitter have comparatively low reciprocity rates since they are mainly used for information sharing, in which users connect to others to share information. Among the three systems, Digg has a comparatively higher reciprocity rate. Although Digg acts like a news media, the news are "digged" and "buried" by users themselves and the users very actively communicate with each other through making comments on news and comments. Such communication may lead to an increased impulse for users to follow and follow back others. On the contrary, Twitter plays an important role in spreading first-hand news from sources such as celebrities and social media. Users are more likely to follow those that they are not acquainted with in real life such as celebrities, while the celebrities do not follow back in most of the time. This explains why the reciprocity rate in Twitter is lower than that in Digg. The reciprocity rate in our Q/A network is only 13%, which is much lower than that of the OSN of Yahoo! Answers (30.7%). This shows that actual user Q/A behavior has a much lower link symmetry than the established contact-fan social relationship, which indicates a significant lower social pressure on users for answering back to their previous answerers. This result conforms to a phenomenon in the general OSNs, in which user A may build or accept the friendship establishment with user B but does not have actual online interactions with B. Therefore, it is important to study the actual Q/A activities of an unbiased set of normal users. We see that the reciprocity rate of Twitter lies in the middle ground between those of the OSN of Yahoo! Answers and Q/A network. Twitter is featured by unidirectional follower–followee relationships and bidirectional friendships. This result implies that though some users wish to mutually benefit from each other in Q/A activities through building OSN connections, many users actually answer few questions from the answerers of their questions.

TABLE III
RECIPROCITY RATE OF DIFFERENT OSNs

Social network/*Yahoo! Answers*	Reciprocity rate
Yahoo!360 [36]	84%
Flickr [37]	68%
Digg [38]	39.4%
Twitter [39]	22.1%
OSN in *Yahoo! Answers* [3], [4]	30.7%
Q/A network in *Yahoo! Answers*	13%

B. Node Q/A Degrees Distribution

The node degree of general OSNs often follows a power-law distribution, which is indicated by the fact that most nodes have small degrees while a small portion of nodes have much larger degrees. This is due to the preferential attachment process [40], in which users with more existing connections with other nodes are more likely to be connected with other nodes. In our previous study on the OSN of Yahoo! Answers, we confirmed that the OSN indegree and outdegree approximately conform to a power-law distribution [41]–[43]. In other words, the preferential attachment process also exists in the OSN of Yahoo! Answers. This means that some nodes with high OSN indegree attract more nodes to connect to them, and some nodes with high OSN outdegree are easily attracted by more nodes. In this paper, we expect to determine if this preferential attachment process also exists in the Q/A activities. That is, if users that have asked or answered more questions have higher probability of asking or answering questions.

We draw the complementary cumulative distribution functions (CCDF) of the Q/A indegree and outdegree of each user in Fig. 2. Fig. 2(a) and (b) show that both indegree and outdegree in the Q/A network approximately conform to the power-law distribution [41]–[43], which means the existence of the preferential attachment process in both Q/A activities of users in Yahoo! Answers. This result means that users who already answered many questions tend to answer more questions while users who already asked many questions have a high probability to ask more questions.

C. Q/A Indegree–Outdegree Correlation

In general OSNs such as YouTube, Flickr, Digg, and Twitter, high outdegree nodes also tend to have high indegrees. In fact, top 1% highest outdegree nodes overlap about 58% with top 1% nodes ranked by the indegree [41]. With the goal

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

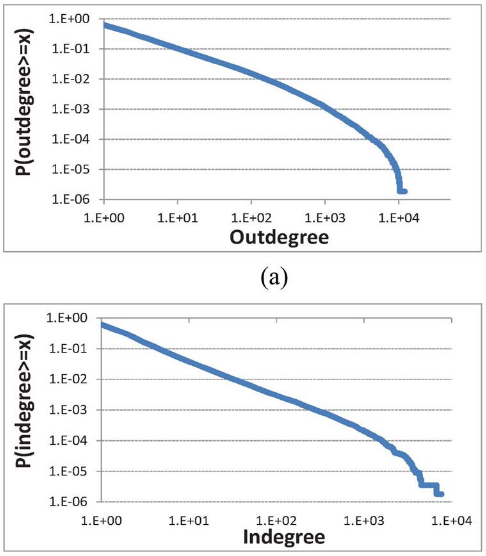


Fig. 2. CCDF of the number of questions and answers of each user. (a) Outdegree (number of questions). (b) Indegree (number of answerers).

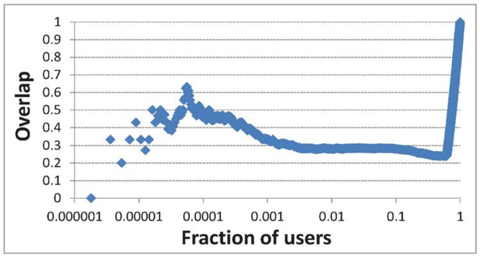


Fig. 3. Overlap of two node lists ranked by indegree and outdegree.

of studying the correlation between the number of answers and the number of questions (Q/A indegree and outdegree) of each user, we generated two ranked lists of users L_{in} and L_{out} by each user's Q/A indegree and outdegree, respectively. We use L_i and L_j to denote the groups of the top $x\%$ of nodes in the two ranked lists, and define the overlap percent as $(L_i \cap L_j/L_i \cup L_j)$. Fig. 3 shows the overlap between the top $x\%$ of nodes in L_{in} and L_{out}. We see that the top 1% of the two lists have about 28% overlap, which is very close to the result of 29% overlap in our previous study on the OSN of Yahoo! Answers but is much lower compared with that in general OSNs [3], [4].

This result confirms our previous conclusion from the OSN of Yahoo! Answers that users asking many questions would not answer as many questions and users answering many questions would not ask as many questions. In the OSN of Yahoo! Answers, active and knowledgeable users would have high OSN indegrees since many nodes follow them but may not connect to many nodes, and active learners would have high outdegrees by connecting to many nodes and may not be connected by many nodes. This is confirmed by our observed user Q/A activities. Many users ask many questions without contributing much to answering others' questions, while many users answer many questions but ask few questions. Also, the users in the 28% overlap consider Yahoo! Answers as a forum to exchange opinions with others, so that they have as many questions as answers. We see that in Fig. 3, about top 0.01% of the two lists reaches a high overlap of 60%, which indicates that a small portion of the most active users are active in both answering and asking activity.

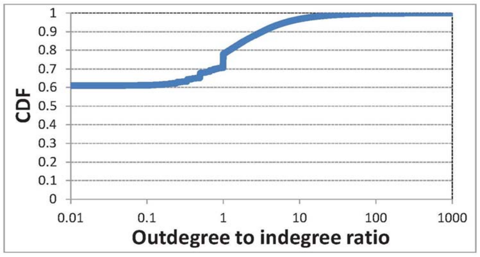


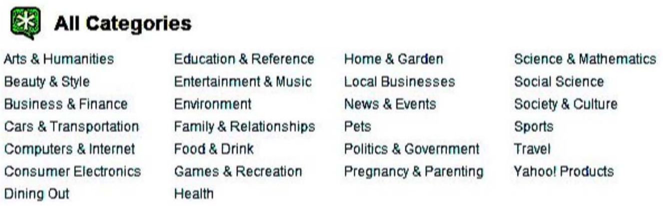
Fig. 4. Correlation between indegree and outdegree.

These results, to a certain extent, indicate the relationship between actual Q/A behaviors and contact-fan OSN relationship establishing behaviors in Yahoo! Answers. A user has high OSN indegree because the user has answered many questions (i.e., high Q/A indegree). Similarly, a user has high OSN outdegree because the user likes to learn by asking many questions (i.e., high Q/A outdegree). User A follows user B mainly because B has answered A's questions satisfactorily and A likes to learn. Therefore, we can say that the OSN relationship establishment is based on the Q/A behaviors in Yahoo! Answers.

In Fig. 4, we draw the CDF of the outdegree-to-indegree ratio to explore the relationship between the Q/A indegree and outdegree of individuals in Yahoo! Answers. In our previous study, on the OSN of Yahoo! Answers [3], [4], around 71% of nodes (compared to 56% in general OSNs) have an outdegree-to-indegree ratio below 0.01 and less than 10% of nodes have an outdegree-to-indegree ratio around 1. In our Q/A network, about 60% of users have an outdegree-to-indegree ratio below 0.01. This means that the number of answers received by most users is relatively low compared to the number of answers given by the users. We also see that about 20% of nodes have an outdegree-to-indegree ratio around 1. This means that the number of answers received by these nodes is similar as the number of answers given by them and they are not selfish or selfless nodes. It also explains the above outdegree-to-indegree ratio results in the OSN of Yahoo! Answers. The OSN relationship establishing behavior is driven by Q/A behavior. User A adds B as contact mainly because B has given a satisfying answer to A. As a user receives fewer answers than its posted answers, the number of its contacts is smaller than the number of its fans. Also, the trend in Fig. 4 is very similar to our previous result of OSN outdegree-to-indegree ratio, which also indicates that the outdegree-to-indegree ratio in the OSN of Yahoo! Answers can reflect the actual Q/A activities of user to a certain extent. The small discrepancy between the results from the OSN and Q/A network also implies that such reflection is not very accurate.

D. Summary

The Q/A network of Yahoo! Answers shares similar structural characteristics with the OSN of Yahoo! Answers. That is,



All Categories

Arts & Humanities
Beauty & Style
Business & Finance
Cars & Transportation
Computers & Internet
Consumer Electronics
Dining Out

Education & Reference
Entertainment & Music
Environment
Family & Relationships
Food & Drink
Games & Recreation
Health

Home & Garden
Local Businesses
News & Events
Pets
Politics & Government
Pregnancy & Parenting

Science & Mathematics
Social Science
Society & Culture
Sports
Travel
Yahoo! Products

Fig. 5. All general knowledge categories in Yahoo! Answers.

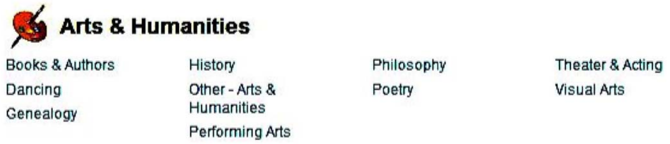
some nodes are active and knowledgeable answerers that give many answers and some nodes are eager learners that ask many questions, whereas most nodes are inactive. Also, as the OSN of Yahoo! Answers, our Q/A network shows a low reciprocity rate. This confirms that user A that has received an answer from user B does not necessarily answer B's questions. However, the Q/A network has the following different properties from the OSN of Yahoo! Answers and other general OSNs as follows.

1) Our Q/A network shows a significantly lower level of link reciprocity than that of the OSN of Yahoo! Answers, which means that the Q/A activity relationship between two nodes is more likely to be unidirectional rather than reciprocal. It implies that user A may build or accept the fan-contact relationship establishment with user B but does not have actual Q/A interactions with user B (Section IV-A).
2) Both of the Q/A indegree and Q/A outdegree distribution of the users have a power-law tail. This can be attributed to the preferential attach process that users who have asked or answered more questions have higher probability of asking or answering new questions. (Section IV-B).
3) Our Q/A network has an outdegree-to-indegree ratio that is not very close to that of the OSN of Yahoo! Answers, which means that though the OSN of Yahoo! Answers can reflect the actual Q/A activities to a certain extent, it cannot very accurately reflect it. In fact, the OSN relationship establishing behavior is driven by Q/A behavior (Section IV-C).

V. Analysis of Knowledge Distribution and User Behavior

Unlike many other friendship-driven OSNs that are centered on building social relationships, Yahoo! Answers is a Q&A site that is centered on sharing knowledge. In Yahoo! Answers, user A connects to other users that are knowledgeable in the topics that user A is interested in. Actually, social relationships are build in Yahoo! Answers in order to achieve better knowledge sharing. As the Q&A OSN is knowledge-oriented, it is very important to examine the user knowledge distribution and associated user behaviors. In this section, we analyze the distribution of user knowledge and associated user behaviors.

As explained previously, the knowledge in Yahoo! Answers is organized by general knowledge categories, each of which is organized by detailed knowledge categories. Fig. 5 shows a snapshot of the 26 general knowledge categories used in Yahoo! Answers. Fig. 6 shows a snapshot of the detailed knowledge categories of the Arts & Humanities general knowledge category. As mentioned, Yahoo! Answers regards the users that actively provide answers as top contributors. Our previously studied dataset involves 4000 top contributors (which constitute 8% of the users in the dataset) and their directly or indirectly connected users in the OSN. Thus, the knowledge categories derived from these users' profiles may be biased. In this section, we analyze the knowledge distribution and user behaviors based on our dataset consisting of normal users in the Q/A activities in Yahoo! Answers.



Arts & Humanities

Books & Authors
Dancing
Genealogy

History
Other - Arts & Humanities
Performing Arts

Philosophy
Poetry

Theater & Acting
Visual Arts

Fig. 6. Detailed knowledge category list of Arts & Humanities.

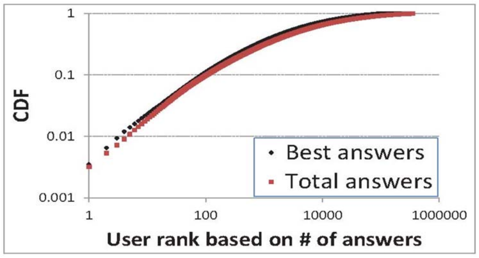


Fig. 7. CDF of the best answers and all answers.

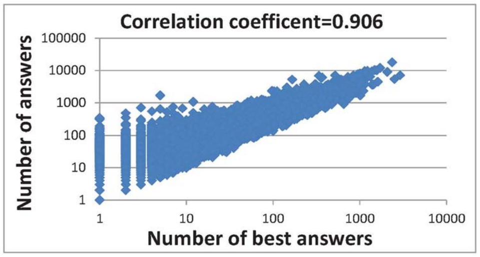


Fig. 8. Correlation between best answers and answers.

A. User Behavior

First, we study the knowledge distribution by analyzing the distribution of answers and best answers of all users. In Fig. 7, we rank all users by the number of their answers and draw the CDFs of the best answers and all answers versus the user rank based on the number of answers. Both CDFs approximately follow a power-law distribution [41]–[43]. Over 80% of all answers and best answers are given by the top 10% ranked users. Also, we see that over 80% of users have no more than one answer and best answer. These results are consistent with our previous observation from the OSN study that most high-quality answers are given by a small portion (i.e., 10%) of users.

Fig. 8 shows the number of all answers versus the number of best answerers of each user. The Pearson correlation coefficient between the two numbers is 0.906 here versus the

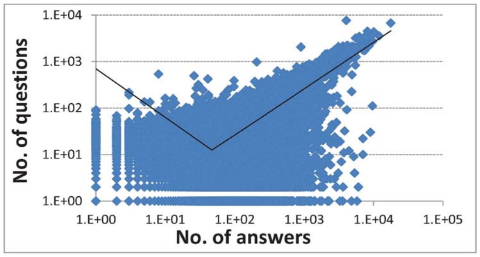


Fig. 9. Correlation between the number of questions and the number of answers of each user.

0.712 coefficient in our previous OSN study. As the number of answers that a user provides increases, the number of best answers from the user also increases. This result confirms the conclusion in our previous OSN study that there is a positive linear relationship between the number of best answers and the number of all answers given by each user but shows a stronger linear relationship. The higher coefficient in the Q/A dataset is caused by its large portion of users who have not provided any answers or best answers. This result implies that in the Yahoo! Answers system where many users do not provide any answer, if a user provides more answers, more of his/her answers will be selected as the best answers. Since, the correlation coefficient is very high, a user's number of provided answers can be a factor to be considered in best answer prediction. For example, the Yahoo! Answers system can use a linear regression model [44] to predict the number of best answers considering the number of answers provided by a user.

Fig. 9 shows the number of questions versus the number of answers of each user with the trend line. We can see that some users ask many questions but answer few answers (up left part) while some users answer many questions and post few questions (bottom right part). The number of questions starts from a high value when the number of answers is small, and it decreases as the number of answers increases. Then, at the point round $x = 80$, the number of questions starts to increase. This trend is similar to the trend in our previous OSN study. It means that the ratio of the number of questions to the number of answers (outdegree-to-indegree ratio) decreases and then increases. When $x \geq 80$, it means that the users are actively involved in the answering activity and we consider these users as active answerers. For these active answerers, the increasing trend line means that a user being more active in the answering activity also tends to be more active in the asking activity. When $x < 80$, it means that the users are less actively involved in answering activity and we consider them as inactive answerers. For these inactive answerers, some very inactive answerers ask many questions while some relatively more active answerers ask few questions, which leads to a decreasing trend line. The outdegree-to-indegree ratio can reflect if a user is more willing to answer questions than to ask questions. Nodes with lower ratios are more selfless nodes who answered more questions than the number of questions they asked, while nodes with higher ratios are more selfish nodes who answered fewer questions than the number of questions they asked. In our previous OSN study, the average outdegree-to-indegree ratio equals 0.437. In our Q/A dataset, except 37% of users that provide no answers, the average ratio of other users is 0.42, which is very close to the previous result of 0.437. This result indicates that a striking difference of Q/A dataset from OSN dataset is that Q/A dataset has a very large portion of users that provide no answers.

Our Q/A study shows that only about 6% of users have outdegree-to-indegree ratios less than 0.01 versus 23.1% in our previous OSN study. These users are selfless contributors who give much more answers than questions. We call them helpers. Also, our Q/A study shows that 37.6% of users have the ratios larger than 100 versus 13.6% in our previous study. These are selfish users who barely provide answers but ask many questions. The result discrepancy is caused by the fact that Q/A dataset covers normal users while the OSN dataset covers the top contributors and their related users. Comparing the results, we see that the OSN dataset includes more selfless users and fewer selfish users than the Q/A dataset. This is reasonable since the top-contributors-related users in the OSN of Yahoo! Answers are relatively active nodes that are more likely to be selfless than selfish compared to the normal user group. This observation also implies that the participants of the OSN of Yahoo! Answers are much more helpful to the Yahoo! Answers community than other nodes. Hence, it is important for the Yahoo! Answers system to incentivize users to join in the OSN of Yahoo! Answers and make contribution by answering questions constantly.

We further analyze the characteristics of user behavior among those who have not given any answers. The average number of questions they asked is 1.69 and 64% of them have exactly one question. This result shows that most of the users are "one-time users" that visit Yahoo! Answers only to look for answers of a specific question and never actually come back. One-time users account for about 23% of the total population of Q/A dataset. This number is strikingly large, although Yahoo! Answers has taken steps such as offering points to encourage users to answer questions. Users with large outdegree-to-indegree ratios (including one-time users) are considered to be the knowledge consumers in the Yahoo! Answers system, which consist of a large portion of overall population in Yahoo! Answers. Thus, how to encourage these knowledge consumers to become knowledge contributors and how to incentivize one-time users to join in the OSN of Yahoo! Answers as regular Yahoo! Answers users are critical challenges faced by Yahoo! Answers currently.

B. Behavior and Knowledge Base of Users

We study the knowledge base of users by examining their knowledge categories. Our previous OSN study only derived the knowledge category information from the profiles of top contributors and their related users. From the Q/A dataset, we use the general knowledge categories and detail knowledge categories from all questions a user asked or answered to denote this user's interested general knowledge categories and detail knowledge categories. Thus, this paper analyzes the real interests from users actual Q/A behaviors instead of those indicated in their profiles. In addition to

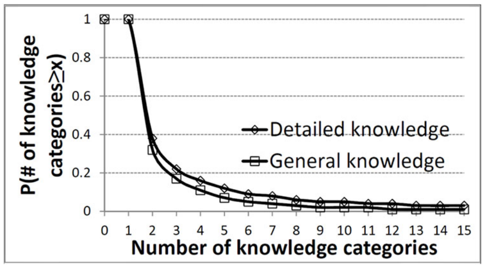


Fig. 10. CCDF of knowledge categories.

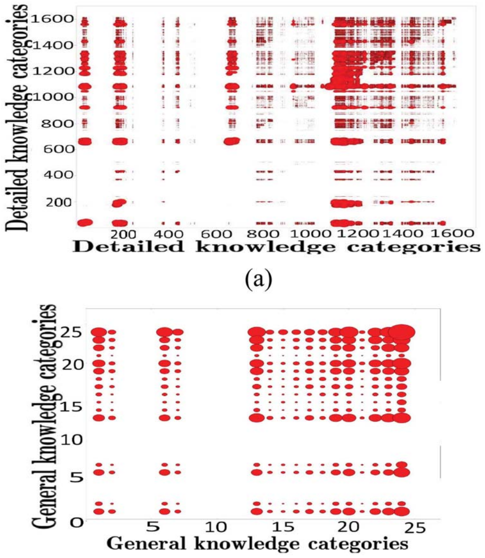
giving more accurate analysis, this paper can also verify if the interests in user profiles can accurately reflect their real interests.

Fig. 10 shows the CCDF of the number of general knowledge categories and detailed knowledge categories. We can see that about 60% of users have only one detailed knowledge category and about 70% of users have only one general knowledge category. We further explored the reasons for this phenomenon and found that most of these users have only one question and no answers; that is, they are one-time knowledge consumers. The number of either general knowledge categories or detailed knowledge categories of a user can be as large as 15. In our previous OSN study, the maximum number of knowledge categories of a user is 4, and only about 42% of users have only one general knowledge category and one detailed knowledge category. The differences are mainly due to three reasons. First, our Q/A dataset includes normal users that have more scattered interests while our previous OSN dataset only contains top contributors and their related users, who are supposed to be more focused in only a few knowledge categories. Second, a user may ask or answer questions in knowledge categories beyond those listed in their profiles, leading to more scattered interests. Third, a large portion of one-time users in Q/A dataset is the reason for the larger percent of users having only one detailed knowledge category and one general knowledge category. The differences between the results of two datasets indicate that the knowledge categories of top contributors and their related users cannot represent the knowledge categories of normal users, and the knowledge categories from user profiles cannot accurately represent their real interests.

C. Relationship Between Knowledge Categories

We then examine the relationship between knowledge categories of users. We expect to study the clustering features of categories that are more likely to coexist in a user's interested knowledge categories. To achieve this, we assigned close numerical IDs to the detailed knowledge categories in the same general knowledge category. For example, the detailed knowledge categories in the "Arts & Humanities" general knowledge category are given consecutive IDs from ID 27 to 36. Then, we generate a matrix $A[i][j]$, where i and j indices are the IDs in sequence. If two detailed knowledge categories with IDs i and j coexist in a user's detailed knowledge categories, the value of $A[i][j]$ is increased by 1, which finally is used as the radius of the point to be plotted in a figure. Fig. 11(a) shows the relationship between detailed knowledge categories represented by the points of $A[i][j]$ and the point radius equals $A[i][j]$. A larger radius means that the detailed knowledge categories i and j have a higher probability to coexist in a user's detailed knowledge categories. We see that the detailed knowledge categories are highly clustered and many clustered detailed knowledge categories are in the same general knowledge category as they have consecutive IDs. This means that if a user is interested in one detailed knowledge category, then the user has a high probability to be interested in its correlated detailed knowledge categories. Also, a node is very likely to be interested in multiple detailed knowledge categories in the same general knowledge category. The blank parts in the figure including the ID ranges in [10, 150], [200, 340], [420, 600], and [700, 750] mean that the detailed knowledge categories with these IDs are very unlikely to coexist with each other. These ranges are the subset of the ranges of the blank parts in our previous OSN study that only considered the knowledge categories in the profiles of top contributors and related users. This means a detailed knowledge category has a higher probability to coexist with another detailed knowledge category in the Q/A dataset than in the OSN dataset or a user tends to have more detailed knowledge categories in the Q/A dataset than in the OSN dataset. The result indicates that the detailed knowledge categories that users are actually involved in Q/A activities are more than those indicated in the profiles of top contributors and their related users. Also, there are more clusters in Fig. 11(a) than in the OSN study. Since this paper covers a large scope of normal users and uses the detailed knowledge categories from actual Q/A activities, which are more scattered and comprehensive, it shows more clustered detailed knowledge categories and fewer uncorrelated detailed knowledge categories, thus presenting more accurate characteristics of the relationship between detailed knowledge categories.

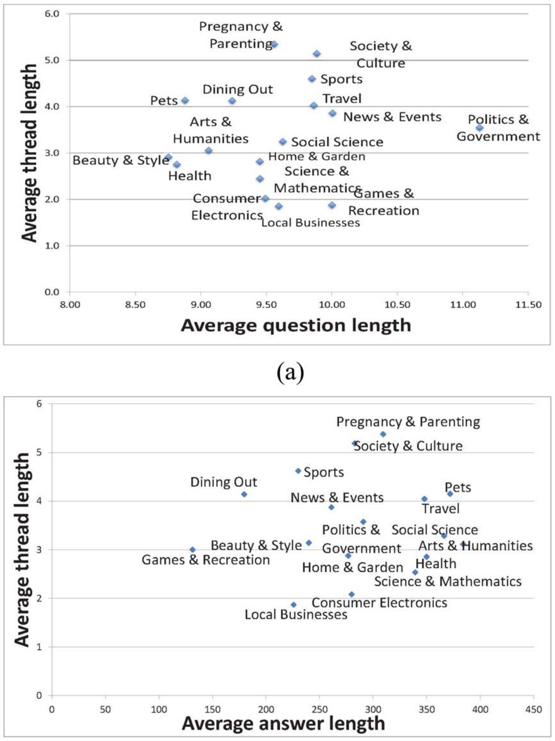
We then map the matrix of detailed knowledge categories to the matrix of general knowledge categories; that is, $B[x][y|]+ = A[i][j]$, where x and y are the general knowledge categories of detailed knowledge categories i and j, respectively. Fig. 11(b) plots the general knowledge category matrix with $B[x][y]$ as the point radius to show the relationship between general knowledge categories. The clustered general knowledge categories are correlated knowledge categories. For example, we found that many users who are interested in Health are also interested in Sports due to their correlation but are rarely interested in Games & Recreation. Unlike the figure in the OSN study that shows many big points in the diagonal line, Fig. 11(b) shows fewer points in the diagonal line. Fig. 11(b) also has more points in the upper-right part. The differences indicate that the profiles of top contributors and their related users mainly focus on detailed knowledge categories in one general knowledge category, while normal users are likely to be interested in multiple general knowledge categories in their Q/A activities. For example, the point of (1, 19) in the figure means many users are interested in both Arts & Humanities and Politics & Government.



(a)

(b)

Fig. 11. Correlation between detailed knowledge categories and between general knowledge categories. (a) Detailed knowledge categories. (b) General knowledge categories.



(a)

(b)

Fig. 12. Relationship between question length, answer length, and thread length of each general knowledge category. (a) Question length and thread length of each general knowledge category. (b) Answer length and thread length of each general knowledge category.

D. Question and Answer Characteristics

The actual user Q/A behavior is a good indicator for the overall characteristics of each general knowledge category in Yahoo! Answers. A better understanding of the various characteristics of each general knowledge category enables Yahoo! Answers to make more suitable rules to govern different categories, e.g., stricter rules should be made toward those categories that are controversial. We define question length, thread length and answer length as the number of words in a question, the number of answerers for a question, and the number of words in an answer, respectively. We then analyze the Q/A behaviors from the perspective of question and answer characteristics to characterize different general knowledge categories.

Fig. 12(a) is a scatter plot of the average thread length versus the average question length of each general knowledge category. We observe that the bottom four general knowledge categories including Science & Mathematics, Consumer Electronics, Games & Recreation, and Local Business have the least number of answers (with the average below 3). In Mathematics & Science, most questions are factual questions with definite answers. For example, the question "What is the algebraic expression for 12 fewer than *x*?" has only one correct answer and normally people will not have different opinions on the answer. In consumer electronics, a large portion of the questions are about the factual issues in computers, which can be answered easily and will not lead to further discussion. For example, for the question "How do I plug my blackberry into the computer?" the answer is simply "using a USB cable." Similarly, the Games & Recreation category is filled with questions that puzzle the askers when they are playing games and a single answer should be enough to help them.

The three general knowledge categories with the longest thread length are Pregnancy & Parenting, Society & Culture, and Sports. They mainly involve nonfactual questions, and users tend to have discussions due to different opinions. Pregnancy & Parenting involves many opinion and experience questions, which attract many answers. Society & Culture is rich of content and multivalued in nature, so its questions involve a lot discussion and disagreement. As to Sports, different people have different opinions and are easily to have arguments on questions such as "Who is the best player?" and "Which team is the best?" For example, "What was the greatest sports feat ever accomplished?" has attracted over five thousand users in discussion. The figure also shows that the average question length is the highest in Politics & Government, because its questions usually need more words to describe such as "If the USA used their military budget on internal improvements, would it help the citizens more or less?" In contract, Health has most concise questions such as "What causes goosebumps?" and "how to get rid of pimples?"

Fig. 12(b) is a scatter plot of the average thread length versus the average answer length. As we can see, the answers in Games & Recreation have the least answer lengths and have comparatively short thread lengths. We further divided the questions in this category into two kinds: 1) factual and 2) nonfactual. An example of factual questions is "When is the PS4 coming out?" and one answer is simply "Holidays 2013." An example for nonfactual questions is "What game should I play next?" with a few answers such as "Saints Row 3" and "Gears of War." We observe that the factual questions need only a few words to answer, and they occupy the majority of this category. Social Science, Arts & Humanities, Health, Science & Mathematics are among the categories with most verbose answers since these categories need long answers to explain clearly. For example, two of the four answers for the question "Age is just a number?" have hundreds of words.

An intriguing observation is that though pets has relatively low question length, it has long answer length and thread length. As pets involves more nonfactual questions seeking for opinions and experiences such as "How can I dry my dog after I have washed her?" there are quite a few answers and the answers tend to be long. As the pets category functions like a discussion board, its questions receive many answers.

E. Summary

We summarize our observations in our analysis of knowledge distribution and user behavior in this section.

1) In Yahoo! Answers, majority of answers and best answers (80%) are contributed by a small portion (10%) of the users. There is a strong correlation between the best answers and all answers of a user, with 0.906 Pearson correlation coefficient. About 37% of the users (including 23% one-time users) do not provide any answers, and the average number of their questions they have asked is under 2. About 6% of users (i.e., helpers) answer much more questions than they asked (Section V-A).
2) Although most users have only one knowledge category (around 60% for detailed knowledge categories and 70% for general knowledge categories), the number of general knowledge categories or detailed knowledge categories of a user can be as large as 15. Most users are just one-time knowledge consumers (Section V-B).
3) The knowledge categories that users are actually involved in Q/A activity are more than those indicated in the profiles of top contributors and their related users. Both general knowledge categories and detailed knowledge categories belonging to the same general knowledge category of the users are highly clustered (Section V-C).
4) Factual questions tend to have fewer answers and nonfactual questions tend to have more answers. Controversial and opinion-seeking knowledge categories have more answers and longer answer lengths (Section V-D).

VI. Discussion on Yahoo! Answers Improvement

As we have mentioned, Q&A system suffer from a very high latency before a question is answered. A major reason for this is that even with a large base of registered users, a large portion of them do not provide any answer. We can categorize these inactive users into two kinds: 1) one-time users who register, ask a question and never come back and 2) knowledge consumers who are only willing to ask questions but reluctant to give any answers. Therefore, Yahoo! Answers currently faces the challenges of encouraging knowledge consumers to become knowledge contributors, incentivizing users to answer more questions and incentivizing one-time users to be regular Yahoo! Answers users. Providing satisfactory quality of service to users, especially at the first time when they use Yahoo! Answers, is extremely important for keeping users. Latency of answer provision, quality/trustworthiness of the answers, and answer provision guarantee are main factors that determine the quality of service. We present several methods to leverage our analytical results to improve the Yahoo! Answers system.

A. Incentives for Answering Behaviors

It is important to make users feel more involved and attached to the Yahoo! Answers community, which make them play a better role in the community. Currently, the Yahoo! Answers system provides the point credit incentives, in which users gain points by answering questions and lose points when asking questions. To enhance the effectiveness of incentives, we can explore other methods. First, allowing users to exchange points for goods such as magazine subscriptions, games, music, and movies may further incentivize them to share their knowledge for economic awards. Second, we can use a reputation system that evaluates each user's reputation based on the outdegree-to-indegree ratio (which is used in our analysis). Then, users will try to provide more answers to increase their reputations. Third, motivated by our reciprocity analysis, we can create a list for each user indicating all other users that have been involved in Q/A activities with the user and their contributions to the user and the user's contributions to them. This method may stimulate the answering activities from users due to reciprocity.

B. Question Forwarding

Our previous results show that users who have answered more questions tend to contribute more answers, and these users are considered as unselfish, knowledgeable, and active users in Yahoo! Answers. Yahoo! Answers can proactively forward a question to the top contributors in the knowledge categories of the question. Since the OSN of Yahoo! Answers can reflect Q/A activity to a certain extent, it can be further leveraged in potential answerer selection to forward questions so that the asker can be more satisfied with and trust the offered answers. Specifically, in the potential answerer selection, we consider the candidate's OSN contact-fan relationship with the asker, and the similarities of the candidate's knowledge categories to the question and to the asker's knowledge categories. As the OSN of Yahoo! Answers cannot very accurately reflect Q/A activity, the actual Q/A interactions between candidates and the asker should be additionally considered. Finally, the selected potential answerers should have high probabilities to provide satisfactory answers. As knowledge categories have a clustering characteristic, to more precisely calculate the similarity of the knowledge categories, the knowledge category correlation degrees can be used as weights in the similarity calculation.

C. Use of Experts

Yahoo! Answers can also hire experts in the knowledge category fields to guarantee the quality of the answers and reduce the latency of the answer provision.

D. OSN Interest/Link Recommendations

We found that real knowledge categories of users are more scattered than those indicated in their profiles [45], [46]. This may be due to the reasons that a user did not comprehensively select interests in their profile or the user does not keep them updated. Recall that some knowledge categories are closely correlated. Thus, when a user selects an interest, Yahoo! Answers can recommend its correlated interests to the user. When Yahoo! Answers notices that a user more frequently asks or answers questions in a knowledge category not in his/her profile, Yahoo! Answers can recommend the user to add this knowledge category, along with its correlated knowledge categories. Yahoo! Answers can proactively notify a user his/her frequent answerer and suggests him/her to build an OSN link to the answerer. In this way, OSN can more accurately reflect Q/A activity, which in turn helps forward questions to better potential answerers.

E. Leveraging Category Clustering Feature

This paper shows both general knowledge categories and detailed knowledge categories belonging to the same general knowledge category of the users are highly clustered. That is, if a user is interested in a detailed knowledge category (e.g., Theater & Acting in Arts & Humanities), then the user is very likely to be interested in the correlated detailed knowledge categories in the same general knowledge category (e.g., performing arts in Arts & Humanities). If a user is interested in a general knowledge category (e.g., Sports), that user is very likely to be interested in the correlated general knowledge categories (e.g., Health). Thus, when a user is browsing questions and answers in a category, the Yahoo! Answers system can recommend relevant questions in correlated detailed knowledge categories and general knowledge categories to the user which is very likely to be interesting to the user. This function is like the function in YouTube that shows relevant videos of the video a user is watching. It makes it easy and convenient for users to find their interested topics without the need to search in a large number of categories.

F. Encouraging High-Quality Questions

In current Yahoo! Answers, users will lose points by asking questions. However, since a key part of Yahoo! Answers is questions that users may browse, Yahoo! Answers questions are indexed by search engines, and the question quality affects answer quality [47], we should not only encourage asking behaviors, but also incentivize users to ask high-quality questions. There were generally no significant correlations between answer quality and answer speed across all question types [48]. We propose to decide the points to be gained or lost by a user after his/her question is answered and closed. By that time, the quality of the question can be inferred from the number of answers, number of views, quality of answers and so on [49]. The quality of the question then determines whether to increase or decrease points and how many points should be increased or decreased.

G. Spam Detection

Since everyone can post answers, the Yahoo! Answers website can be easily exploited by users for their own benefits. Spammers could post fake questions and answers that are actually advertisements. For example:

Q: what is the capital of Mexico?

A: for a free apple iphone, go to www.thisisaspamsite.com.

Since questions in Yahoo! Answers rank high in search engines, this spamming behavior is quite profitable. Current spam reporting system in Yahoo! Answers can deal with this problem to a certain extent, but the latency after spam posting and before its removal is high. Since this paper shows that the number of best answers of a user has a very high correlation with the number of total number of answers of the user, and spam answers barely have chances to be selected as the best answer, we can build a regression model on the number of answers and the number of best answers. In this way, we can find the outliers who are most likely to be spammers.

H. Ranking Answers

Spam reduces the trustworthiness of answers in Yahoo! Answers [50]. For example, for the question “What’s wrong with me that I seem not able to digest anything,” “Our product can solve your problem in a minute” is a spam that will attract askers to click on its associated link. It is desirable to present answers with trust scores in the descending order of the scores. To evaluate the trust scores, we can consider the following: 1) the expertise, experience (reflected in the number of best answers and answers) and interested knowledge categories of the answerer; 2) the similarity of interested knowledge categories and OSN connection between the answerer and the asker; and 3) the Q/A interactions between the answerer and the asker.

I. Use of Videos and Pictures

As user-generated-contents are increasingly popular and a picture is worth a thousand words, videos and pictures can be used for knowledge categories that have long question lengths or answer lengths to improve user experience [51]. For example, an answerer can simply use a video to answer “How can I dry my dog after I have washed her?” to avoid tedious typing and make the answer easy to follow. To this end, the Yahoo! Answers system can provide relevant picture options for users to choose from when posting contents. However, this approach also brings some problems such as picture/video spams and much higher network traffic. We will explore methods to handle these problems in our future work.

VII. Conclusion

Understanding the Q/A activities of users in Yahoo! Answers is crucial to improving the quality of service of the system. Our previous study on the OSN of Yahoo! Answers discloses the characteristics of Q/A activities based on the profiles of top contributors and their related users. In order to investigate the actual Q/A activities of users over an unbiased scope of normal users, we studied over 1.6 million questions crawled from Yahoo! Answers. We built a Q/A network that connects askers to their answerers. We found that the Q/A network resembles the OSN of Yahoo! Answers network in terms of the distribution of node degree and low link symmetry. We found that the majority of answers are answered by the top 10% of users, which is consistent with our previous finding. We also found that around 37% of users have provided no answers, and the average number of questions of a user who has given no answers is barely larger than 1. Also, the knowledge categories from user OSN profiles cannot completely represent their real knowledge categories and the knowledge categories from top-contributor-related users cannot represent those of normal users. Furthermore, normal users have more scattered interested knowledge categories than top-contributor-related users. Also, there is a high correlation between the number of best answers and the number of all answers, which can be leveraged to detect spam users. We further analyzed the characteristics of answers and questions in different knowledge categories. By exploring the overall characteristics of each general knowledge category in Yahoo! Answers, we found that the factual questions tend to have less number of answers and Pregnancy & Parenting, Society & Culture, and Sports are the three categories that have the most verbose answers. Our study in this paper offers an in-depth understanding of actual Q/A activities of users and provides an insight of the relationship between the OSN of Yahoo! Answers and user actual Q/A activities. It serves as a basis for the performance enhancement on Yahoo! Answers such as reducing one-time users, providing answering incentives and offering quality scores of answers for askers to selectively read answers. The implementation of these approaches leaves as our future work.

References

[1] H. Sun, C. Jiang, Z. Ding, P. Wang, and M. Zhou, "Topic-oriented exploratory search based on an indexing network," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 46, no. 2, pp. 234–247, Feb. 2016.

[2] D. D. Wu, L. Zheng, and D. L. Olso, "A decision support approach for online stock forum sentiment analysis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 8, pp. 1077–1087, Aug. 2014.

[3] Z. Li, H. Shen, and J. E. Grant, "Collective intelligence in the online social network of Yahoo! Answers and its implications," in *Proc. CIKM*, Maui, HI, USA, 2012, pp. 455–464.

[4] H. Shen, Z. Li, J. Liu, and J. E. Grant, "Knowledge sharing in the online social network of Yahoo! Answers and its implications," *IEEE Trans. Comput.*, vol. 64, no. 6, pp. 1715–1728, Jun. 2015.

[5] I. Szpektor, Y. Maarek, and D. Pelleg, "When relevance is not enough: Promoting diversity and freshness in personalized question recommendation," in *Proc. WWW*, Rio de Janeiro, Brazil, 2013, pp. 1249–1260.

[6] Z. Ji and B. Wang, "Learning to rank for question routing in community question answering," in *Proc. CIKM*, San Francisco, CA, USA, 2013, pp. 2363–2368.

[7] A. Pal, F. Wang, M. X. Zhou, J. Nichols, and B. A. Smith, "Question routing to user communities," in *Proc. CIKM*, San Francisco, CA, USA, 2013, pp. 2357–2362.

[8] Z. Zhao and Q. Mei, "Questions about questions: An empirical analysis of information needs on Twitter," in *Proc. WWW*, Rio de Janeiro, Brazil, 2013, pp. 1545–1556.

[9] G. Qi, C. Aggarwal, J. Han, and T. Huang, "Mining collective intelligence in diverse groups," in *Proc. WWW*, Rio de Janeiro, Brazil, 2013, pp. 1041–1052.

[10] X.-J. Wang, X. Tu, D. Feng, and L. Zhang, "Ranking community answers by modeling question-answer relationships via analogical reasoning," in *Proc. SIGIR*, Boston, MA, USA, 2009, pp. 179–186.

[11] G. Dror, Y. Koren, Y. Maarek, and I. Szpektor, "I want to answer; who has a question? Yahoo! Answers recommender system," in *Proc. SIGKDD*, San Diego, CA, USA, 2011, pp. 1109–1117.

[12] M. A. Suryanto, E. P. Lim, A. Sun, and R. H. L. Chiang, "Quality-aware collaborative question answering: Methods and evaluation," in *Proc. WSDM*, Barcelona, Spain, 2009, pp. 142–151.

[13] C. Shah and J. Pomerantz, "Evaluating and predicting answer quality in community QA," in *Proc. SIGIR*, Geneva, Switzerland, 2010, pp. 411–418.

[14] Y. Liu and E. Agichtein, "On the evolution of the Yahoo! Answers QA community," in *Proc. SIGIR*, Singapore, 2008, pp. 737–738.

[15] K. K. Nam, M. S. Ackerman, and L. A. Adamic, "Questions in, knowledge iN? A study of Naver's question answering community," in *Proc. SIGCHI*, Boston, MA, USA, 2009, pp. 779–788.

[16] B. Li, M. R. Lyu, and I. King, "Communities of Yahoo! Answers and Baidu Zhidao: Complementing or competing?" in *Proc. IJCNN*, Brisbane, QLD, Australia, 2012, pp. 1–8.

[17] B. Furlan, B. Nikolic, and V. Milutinovic, "A survey of intelligent question routing systems," in *Proc. Intell. Syst.*, Sofia, Bulgaria, 2012, pp. 14–20.

[18] W. Chan *et al.*, "Community question topic categorization via hierarchical kernelized classification," in *Proc. CIKM*, San Francisco, CA, USA, 2013, pp. 959–968.

[19] R. Liu and E. Nyberg, "A phased ranking model for question answering," in *Proc. CIKM*, 2013, pp. 79–88.

[20] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, "Knowledge sharing and Yahoo! Answers: Everyone knows something," in *Proc. WWW*, Beijing, China, 2008, pp. 665–674.

[21] G. Gardelli and I. Weber, "Why do you ask this?" in *Proc. WWW Workshop*, Beijing, China, 2012, pp. 815–822.

[22] Q. Su, D. Pavlov, J.-H. Chow, and W. C. Baker, "Internet-scale collection of human-reviewed data," in *Proc. WWW*, Banff, AB, Canada, 2007, pp. 231–240.

[23] 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. ASIST*, Silver Spring, MD, USA, 2007, pp. 1–10.

[24] Y. Liu, J. Bian, and E. Agichtein, "Predicting information seeker satisfaction in community question answering," in *Proc. SIGIR*, Singapore, 2008, pp. 483–490.

[25] D. Dearman and K. N. Truong, "Why users of Yahoo! Answers do not answer questions," in *Proc. CHI*, Atlanta, GA, USA, 2010, pp. 329–332.

[26] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor, "Learning from the past: Answering new questions with past answers," in *Proc. WWW*, Lyon, France, 2012, pp. 759–768.

[27] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan, "Predictors of answer quality in online Q&A sites," in *Proc. SIGCHI*, Florence, Italy, 2008, pp. 865–874.

[28] R. W. White, M. Richardson, and Y. Liu, "Effects of community size and contact rate in synchronous social Q&A," in *Proc. SIGCHI*, Vancouver, BC, Canada, 2011, pp. 2837–2846.

[29] D. Horowitz and S. D. Kamvar, "The anatomy of a large-scale social search engine," in *Proc. WWW*, Raleigh, NC, USA, 2010, pp. 431–440.

[30] S. J. H. Yang and I. Y. L. Chen, "A social network-based system for supporting interactive collaboration in knowledge sharing over peer-to-peer network," *Int. J. Human Comput. Stud.*, vol. 66, no. 1, pp. 36–50, 2008.

[31] Y. Wang, M. Dylla, Z. Ren, M. Spaniol, and G. Weikum, "Pravda-live: Interactive knowledge harvesting," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manag.*, New York, NY, USA, 2012, pp. 2674–2676.

[32] M. R. Morris, J. Teevan, and K. Panovich, "What do people ask their social networks, and why? A survey study of status message Q&A behavior," in *Proc. CHI*, Atlanta, GA, USA, 2010, pp. 1739–1748.

[33] J. Teevan, M. R. Morris, and K. Panovich, "Factors affecting response quantity, quality, and speed for questions asked via social network status messages," in *Proc. AAAI*, Barcelona, Spain, 2011, pp. 1–4.

[34] J. Yang, M. R. Morris, J. Teevan, L. Adamic, and M. Ackerman, "Culture matters: A survey study of social Q&A behavior," in *Proc. ICWSM*, Barcelona, Spain, 2011, pp. 409–416.

[35] M. Richardson and R. W. White, "Supporting synchronous social Q&A throughout the question lifecycle," in *Proc. WWW*, Hyderabad, India, 2011, pp. 755–764.

[36] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Proc. KDD*, Philadelphia, PA, USA, 2006, pp. 611–617.

[37] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the Flickr social network," in *Proc. WWW*, Madrid, Spain, 2009, pp. 721–730.

[38] Y. Zhu, "Measurement and analysis of an online content voting network: A case study of Digg," in *Proc. WWW*, Raleigh, NC, USA, 2010, pp. 1039–1048.

[39] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. WWW*, Raleigh, NC, USA, 2010, pp. 591–600.

[40] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[41] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. IMC*, San Diego, CA, USA, 2007, pp. 29–42.

[42] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemp. Phys.*, vol. 46, no. 5, pp. 323–351, 2005.

[43] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.

[44] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[45] N. Zheng, S. Song, and H. Bao, "A temporal-topic model for friend recommendations in Chinese microblogging systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 9, pp. 1245–1253, Sep. 2015.

[46] S. Agreste, P. De Meo, E. Ferrara, S. Piccolo, and A. Provetti, "Analysis of a heterogeneous social network of humans and cultural objects," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 4, pp. 559–570, Apr. 2015.

[47] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proc. WSDM*, Palo Alto, CA, USA, 2008, pp. 183–194.

[48] A. Chua and S. Banerjee, "So fast so good: An analysis of answer quality and answer speed in community question-answering sites," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 64, no. 10, pp. 2058–2068, 2013.

[49] B. Li, T. Jin, M. R. Lyu, I. King, and B. Mak, "Analyzing and predicting question quality in community question answering services," in *Proc. WWW*, Perth, WA, Australia, 2012, pp. 775–782.

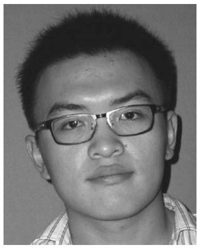
[50] E. Tan, L. Guo, S. Chen, X. Zhang, and Y. Zhao, "Spammer behavior analysis and detection in user generated content on social networks," in *Proc. ICDCS*, 2012, pp. 305–314.

[51] D. Davis, G. Figueroa, and Y.-S. Chen, "SociRank: Identifying and ranking prevalent news topics using social media factors," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

Haiying Shen (SM'07) received the B.S. degree in computer science and engineering from Tongji University, Shanghai, China, in 2000, and the M.S. and Ph.D. degrees in computer engineering from Wayne State University, Detroit, MI, USA, in 2004 and 2006, respectively.

She is currently an Associate Professor with the Department of Electrical and Computer Engineering, Clemson University, Clemson, SC, USA. 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.

Dr. Shen is a Microsoft Faculty Fellow of 2010, and a member of ACM.



Guangyan Wang received the B.S. degree in electronic and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2011, and the M.S. degree in computer engineering from Clemson University, Clemson, SC, USA, in 2014.

His current research interests include data analysis on question and answer systems.