Collective Intelligence in the Online Social Network of Yahoo!Answers and Its Implications

Ze Li, Haiying Shen, Joseph Edward Grant
Department of Electrical and Computer Engineering, Clemson University
Clemson, SC, 29631, USA
{zel, shenh, jegrant}@clemson.edu

ABSTRACT

Question and Answer (Q&A) websites such as Yahoo!Answers provide a platform where users can post questions and receive answers. These systems take advantage of the collective intelligence of users to find information. In this paper, we analyze the online social network (OSN) in Yahoo!Answers. Based on a large amount of our collected data, we studied the OSN's structural properties, which reveals strikingly distinct properties such as low link symmetry and weak correlation between indegree and outdegree. After studying the knowledge base and behaviors of the users, we find that a small number of top contributors answer most of the questions in the system. Also, each top contributor focuses on only a few knowledge categories. In addition, the knowledge categories of the users are highly clustered. We also study the knowledge base in a user's social network, which reveals that the members in a user's social network share only a few knowledge categories. Based on the findings, we provide guidance in the design of spammer detection algorithms and distributed Q&A systems. We also propose a friendship-knowledge oriented Q&A framework that synergically combines current OSN-based Q&A and web Q&A. We believe that the results presented in this paper are crucial in understanding the collective intelligence in the web Q&A OSNs and lay a cornerstone for the evolution of next-generation Q&A systems.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Question-answering (fact retrieval) systems

General Terms

Measurement, Performance

Keywords

Collective intelligence, On-line social networks, Knowledge networks

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

1. INTRODUCTION

Web search engine (e.g., Google and Bing) enables us to search information by keywords on the Internet. Recently, web search engines are improved by combining with social networks [1–5], enabling social friends to collaborate with each other to determine the relevance of the returned results to their queries. Users use web annotations or bookmarks to indicate the search results they are interested in, which helps their common-interest friends to quickly identify results useful to themselves.

However, picking up useful information from the overwhelming returned results still remains a challenge. Users sometimes prefer to directly receive the answers rather than going through a long tedious searching process. In addition, although the search engine based information retrieval performs very well in answering factual queries for information already existing in databases, it is not suitable for non-factual or context-aware queries, which are more subjective, relative and multi-dimensional in context, especially for information not existing in databases (e.g., suggestions, recommendations and advices). This remains as a formidable challenge facing current search engines without big breakthroughs in machine learning and natural language processing techniques.

Actually, people are the most "intelligent machines" that are capable of parsing, interpreting and answering questions, provided they are familiar with the questions. Each person has knowledge from his careers, education, life, experience, interests and so on, which forms his *knowledge base*. By collecting the intelligence of people to find information, Question and Answer (Q&A) websites such as Yahoo!Answers [6] (YA) and Ask.com [7] have naturally emerged as an alternative way for Q&A. These websites provide a platform where users can post questions and receive answers. If user A wants to frequently visit/track all questions and answers of user B, A adds B to its contact list by building a link to B. Then, A becomes B's fan. Thus, a knowledge-oriented online social network (OSN) with unidirectional links between nodes is formed in the Q&A system. YA classifies knowledge into 26 *general knowledge categories (KCs)* (e.g., Sports, Health). Each general knowledge categories has a number of *detailed KCs* (e.g., Golf, Tennis). Users with many points are recognized as *top contributors*, whose profiles indicate the general and details KCs they are knowledgeable in.

Although Q&A websites are becoming increasingly popular and can provide high quality answers [8], they have some shortcomings in satisfying users' needs. First, the latency for receiving a satisfying answer is high with the average equals

2:52:30 (hh:mm:ss) even when the number of the registered users is very large (290,000) [9]. This is because most users log in the Q&A website only when they have questions to ask. Even if some users may intend to answer others' questions, since all questions in one topic appear together in one forum, it is difficult for a user to identify the questions he can answer. Second, as Q&A websites are normally open to all anonymous users in Internet, spam is a difficult problem.

In recent years, OSN-based Q&A systems [8, 10–14] have been developed. Facebook launched a Q&A application in July, 2010. In an OSN-based Q&A system, users post and answer questions through the OSN in order to take advantage of the collective intelligence of their friends. Specifically, a centralized server identifies possible answers from the questioner's friends in his social network, and forwards the question directly to them. Expertise location systems [15–18] that search experts in specified fields share similarity with OSN-based Q&A in answerer location. Research [12] shows that the answerers in the OSN are willing to and able to provide more tailored and personalized answers to the questioners since they know a great deal about the backgrounds and preference of the questioners. However, the characteristics of the knowledge of the friends in a user's social network may affect the qualify of the answers for the user's questions. Factual questions such as "what is the time complexity of the X algorithm" need the answers from experts in the computing theory field, which may not be offered by the OSN-based Q&A systems.

By synergistically integrating the web Q&A system and OSN-based Q&A system, both systems' shortcomings can be overcome. To achieve this, it is important to understand the nature and impact of collective intelligence in the OSNs of both systems. However, no previous work has been devoted to studying the OSN in the Q&A websites, though previous research investigated the OSN-based Q&A systems. In this paper, we analyze the OSN in YA, a popular online Q&A website. For this effort, we have collected Q&A trace data during three months, and a large amount of personal data and their associated relationship in YA. The main contribution of this paper is an extensive trace-driven analysis of OSN structure, user behavior, user knowledge base and their relationships. Our analysis yields very interesting results and the highlights of our work are summarized as follows:

- Examination on the structural properties of the YA OSN shows that though it shares a common property with other previously studied OSNs in that the node indegree and outdegree exhibit power-law distribution, it has strikingly distinct properties: (1) It has low link symmetry; (2) It exhibits weak correlation between indegree and outdegree; (3) Users tend to connect to other users with different degrees from their own; (4) Users exhibit an extremely low clustering coefficient.
- Investigation on the knowledge base and behaviors of users in YA reveals that (1) A small portion of the users (i.e., 10%) contribute to the most of the high quality answers; (2) The 12 most popular general KCs account for 80% of all general KCs in the system; (3) The top contributors steadily contribute high-quality answers. Many top contributors focus on one general KC, and 56.5% of them have multiple general KCs, but all of the them have multiple detailed KCs; (4) There exists positive linear correlation between the number of fans and points of a user but no correlation between the number of contacts and points of a user; (5) The KCs of the users are highly clustered.
- Analysis on the relationship of knowledge base and OSN structure shows that (1) The size of the knowledge base within a user's one-hop OSN neighbors is small, and increases, though not significantly, within two-hop OSN neighbors; (2) Reciprocity (i.e., bidirectional connected) users share more common KCs than one-way connected users, who share more KCs than disconnected users.
- We finally discuss the implications of our findings on the design of spammer detection algorithms in Q&A systems and a distributed Q&A system that integrates both web Q&A system and OSN-based Q&A system.

# of nodes in the social network	119175
# of links in the contact network	1,786,036
# of links in the fan network	1,265,305
Ave. # of contacts per user in the contact network	14.98
Ave. # of fans per user in the fan network	10.61
Ave. # of general KCs in a user's contact network	2.1
Ave. # of detailed KCs in a user's contact network	4.2
Ave. # of general KCs in a user's fan network	2.2
Ave. # of detailed KCs in a user's fan network	4.2

Table 1: High level statistic of the crawled YA social network.

Our analysis provides critical insights regarding the different properties of the YA OSN and other friendship and/or knowledge oriented OSNs. The analytical results provide cornerstone for the performance improvement on current Q&A systems and the evolution of next-generation Q&A systems.

2. BACKGROUND AND MEASUREMENT METHODOLOGY

YA, as a knowledge market, was launched by Yahoo! on July 5, 2005. It has an OSN with unidirectional links between nodes. The nodes in a user's contact list are called *outdegree nodes*, which form *the node's contact network*, and the nodes in a node's fan list are called *indegree nodes*, which form *the node's fan network*. Thus, YA OSN incorporates two directional networks: *contact network* and *fan network*. The former includes all nodes and their outdegree nodes and the latter includes all nodes and their indegree nodes.

We wrote a crawler using Python. The crawler started from the first 4000 top contributors and inserted these users into an initially empty queue. It fetched the first user from the queue, recorded his profile information (i.e., total number of earned points, answers, best answers and questions), retrieved and inserted his public visible contacts and fans to the queue, and finally removed this user from the queue. This process repeated until the queue became empty. Crawling was started on Aug. 17 and ended on Oct. 19, 2011. As the crawled OSN data is seeded at 4000 different users with various KCs, it can well represent an actual knowledge-oriented OSN. In addition, for each user, we recorded its profile information for the activities during every week from Aug. 17 to Oct. 19 2011. In our trace data, about 8% of the users are top contributors. Table 1 shows the high level statistics of the crawled YA OSN.

3. ANALYSIS OF OSN STRUCTURE

In this section, we study the structural characteristics of the YA OSN. We also are interested in answering a ques-

Social network website	Reciprocity rate
Facebook [19]	100%
Flickr [20]	68%
Yahoo!360 [21]	84%
Digg [22]	39.4%
Yahoo!Answers	30.7%
Twitter [23]	22.1%

Table 2: Reciprocity rate of different OSN.

tion: does it show similar structural characteristics as other OSNs?

3.1 Reciprocity

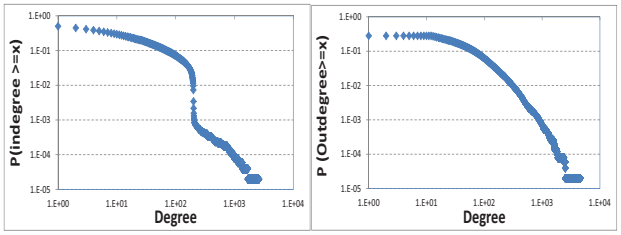
In an OSN, the pairwise bidirectional relationship between two nodes is called *reciprocity*. We define the *reciprocity rate* of an OSN as the number of reciprocity links over all links of all users. Table 2 shows the reciprocity rate of YA and a number of other OSNs from other studies [19–23]. We see that the reciprocity rate of YA is 30.7%. It is similar to the reciprocate rate (39.4%) of the content rating website Digg. Twitter also has a low reciprocity rate of 22.1%. In contrast, Facebook, Flickr and Yahoo!360 have high reciprocity rates, and they are 100%, 68% and 84%, respectively. In these OSNs, a large part of the users connect with each other by their real social ties (i.e., friendship) in their daily lives. Therefore, most links in these OSNs are bidirectional and their reciprocity rate is high. On the contrary, Digg, YA, Twitter are mainly information/knowledge sharing websites, in which people are mainly connected according to their interests. Therefore, most links in these websites are unidirectional and their reciprocity rate is low. Twitter generates the lowest reciprocity rate. This is because Twitter currently is treated as a social media by large companies and celebrities to publish information [23]. YA has the second lowest reciprocity rate. Our crawled dataset also shows that 16.7% of the users only have fans but no contacts. This implies that users prefer to connect to users who are knowledgable in certain categories, and knowledgable users can attract more fans.

3.2 Power-law Node Degree

One striking property of the general OSNs is that their node degree follows a power-law distribution. That is, the majority of nodes have small degree while a few nodes have significantly higher degree. The power-law distribution is caused by the *preferential attach process*, in which the probability of a user A connecting to a user B is proportional to the number of B's existing connections.

Figure 1 shows the indegree and outdegree complementary cumulative distribution functions (CCDF). The figures show that the indegree and outdegree conform to a distribution that is close to a power-law distribution. In other words, the *preferential attach process* also occurs in the YA knowledge sharing system. We also see that there is a sharp drop at Figure 1(a) at around x=200. In 2007, YA launched a new policy that each user can have maximum 200 contacts. As a result, only few old registered users have more than 200 contacts, and the size of most users' contact lists is close to 200, which produces the sharp decrease.

We ranked the users based on their number of earned points. Specifically, we sorted the users based on their number of points in a descending order and assigned a rank to each node sequentially; rank 1 was assigned to the top node.



(a) Outdegree (contacts) (b) Indegree (Fans)

Figure 1: Log-log plot of indegree and outdegree CCDF.

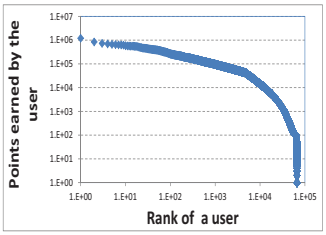


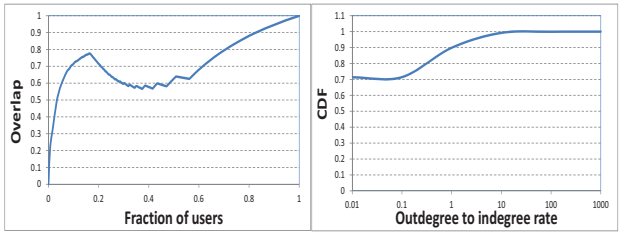
Figure 2: Distribution of points.

We then plotted the number of points a node has versus its rank in Figure 2. We see that the number of points of users also conforms to a power-law distribution. This implies that a small amount of users are very active in answering questions and the rest are not active. Also, some of these nodes may give high-quality answers so that they can earn more points quickly. This phenomenon explains the power-law distribution of node indegree, that is, users are likely to connect to the users that are active and knowledgable in their interested categories. We will show the detail of the reason in Section 4.4.

The power-law distribution of node degree is also caused by the popularity of the KCs, which affects the number of people involved in a KC. Users tend to connect to other users in popular KCs. Also, the users that are active in non-popular topics may not attract as much attention and has fewer fans. We will further investigate how the active answerers and category popularity affect the node degree in Section 4.

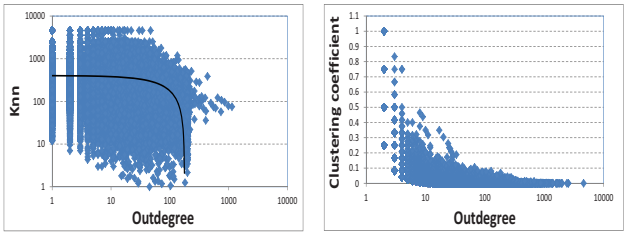
3.3 Correlation between Indegree and Outdegree

In general OSNs such as YouTube, Flickr, Digg and LiveJournal, the nodes with high outdegree tend to have high indegree. Specifically, the top 1% of nodes ordered by outdegree have a more than 58% overlap with the top 1% of nodes ranked by indegree [24]. To study the correlation between indegree and outdegree, we ranked nodes by indegree and outdegree, respectively, and generated two rank lists. We use L_{in} and L_{out} to denote the top x% of nodes in the ranked indegree list and ranked outdegree list, respectively. We define the *overlap* of L_{in} and L_{out} as $\frac{L_i \cap L_j}{L_i \cup L_j}$. Figure 3(a) shows the overlap between the top x% of nodes in the two ranked lists. We see the top 1% of nodes ordered by outdegree have a 29% overlap with the top 1% of nodes ranked by indegree. YA's overlap is much less than that of general OSNs. This means that some high-indegree nodes do not have high outdegree while some high-outdegree nodes do not have high indegree. In general OSNs, nodes connect to each other mainly for interaction and very socially active nodes should have both high indegree and outdegree. In YA, instead of aggressively making friends, the main purpose for user A to connect user B is to learn from user B in his interested KCs. Therefore, active and knowledgable nodes would have high indegree since many nodes connect to them and they may not connect to many nodes. Similarly, nodes who



(a) Overlap between outdegree nodes and indegree nodes (b) Outdegree-to-indegree ratio

Figure 3: Correlation between indegree and outdegree.



(a) Degree correlation function (b) Average clustering coefficient

Figure 4: Link degree correlation and clustering coefficient.

eager to learn would have high outdegree by connecting to many other nodes and they would not be connected by many nodes if they are not active in answering. The nodes in the 29% overlap are both eager to learn and are learned by many nodes. They may mutually establish relationships in order to exchange knowledge.

We further explore the indegree and outdegree of individual users in YA. Figure 3(b) shows the cumulative distribution function (CDF) of outdegree-to-indegree ratio of users in YA. The CDF in YA differs from those of YouTube, LiveJournal, Digg, and Flickr [22] in two ways. First, in YA, 71% of the nodes have an outdegree-to-indegree ratio lower than 1%, while that of the other four websites is less than 56%. The main reason is that in YA, the number of contacts a user has is limited to 200, and the number of fans of a user is not limited. In the general OSNs, the number of either indegree or outdegree of a node is not limited. Second, in YA, about 9.12% of nodes have an indegree within 20% of their outdegree, which is similar to the rate of 14.56% in Digg, while the percentage for the other three friendship-oriented OSNs is more than 50%. In addition to the 200 contact limit in YA, this is also because in friendship-oriented OSNs, users tend to aggressively make friends with others, while in the YA knowledge-oriented OSN, users selectively choose active and knowledgable users in their interested fields as contacts. We also see that YA has less than 10% of nodes whose outdegree-to-indegree ratio is around 1. Thus, it has much weaker correlation between indegree and outdegree than the other three OSNs. This can be explained by a much lower level of link symmetry in YA.

3.4 Link Degree Correlation

In the general OSNs (e.g., Flickr, LiveJournal and Orkut), high degree nodes tend to connect to other high degree nodes [24]. It implies that highly social nodes tend to connect with each other. We are interested to see whether this phenomenon also exists in YA. If so, it means knowledgable users tend to share knowledge between each other.

To answer this question, we examined how often nodes of different degrees connect to each other represented by *joint degree distribution*, which can be approximated by the degree correlation function K_{nn}. The K_{nn} of outdegree d is measured as the average indegree of all nodes connected to nodes with d outdegree [24]. An increasing K_{nn} implies a tendency of higher-degree nodes to connect to other high degree nodes while a decreasing K_{nn} indicates the opposite trend. Figure 4(a) depicts K_{nn} for YA associated with its trend line, from which we see that as a user's outdegree increases to around 200, the K_{nn} exhibits a sharp decrease. Also, when the outdegree is lower than 200, the K_{nn} remains significantly higher than outdegree constantly. The watershed of 200 is caused by the outdegree limit of 200 in YA. Then, we can conclude that YA exhibits different behaviors from the general OSNs where K_{nn} increases as outdegree increases. This is caused by the celebrity-driven nature in YA, i.e., there are a few extremely active and knowledgable users to whom many inactive users link to. We can also see that some nodes have indegree much lower than their outdegree, which means that some inactive users connect to many other nodes but rarely linked by other nodes. These results are consistent with those in Figures 1 and Figure 2.

3.5 Clustering Coefficient

We then explore the connection density of the neighborhood of a node, which is quantified by the *clustering coefficient*. The clustering coefficient of a node with N neighbors is defined as the ratio of the number of directed links existing between the node's N neighbors and the number of possible directed links that could exist between these neighbors ($N(N-1)$). The average of individual nodes' clustering coefficients is 0.029 in YA. This value is much lower than those of YouTube, Orkut, Flickr, LiveJournal and Digg that range from 0.136 to 0.330 [22]. In the friendship-oriented OSNs such as Facebook and Flickr, users tend to be introduced to other users via mutual contacts, increasing the probability that two contacts of a single user are also contacts to each other. Other OSNs such as YouTube, Orkut, LiveJournal and Digg are oriented by both friendship and knowledge, and they should have lower clustering coefficient than the pure friendship-oriented OSNs. YA is a pure knowledge-oriented OSN, and a user adds contacts only when he finds the contacts are knowledgable in the fields he is interested in.

Figure 4(b) shows the clustering coefficient of each node with respect to its outdegree. Nodes of low outdegree have higher clustering coefficients, indicating significant clustering among low-outdegree nodes. High-outdegree nodes, on the other hand, show much lower clustering coefficients due to their large number of diverse contacts. We conjecture that the contacts of low-outdegree nodes are most likely in a limited number of KCs. Since users in the same KC tend to connect to each other, low-outdegree nodes have high clustering coefficient. In contrast, the contacts of high-outdegree nodes are likely to belong to many KCs. As users in different knowledge community are less likely to connect with each other, their clustering coefficients are small.

3.6 Summary

The YA OSN shares similar power-law structural property with other studied OSNs (Section 3.2). That is, a few power-law indegree nodes are active and knowledgable answerers that own many fans, and a few power-law outdegree nodes create many contacts and are active in learning others'

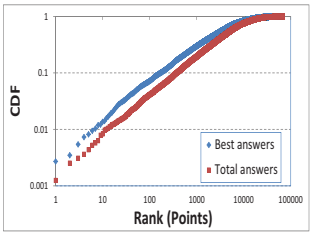


Figure 5: The CDF of the best answers and all answers.

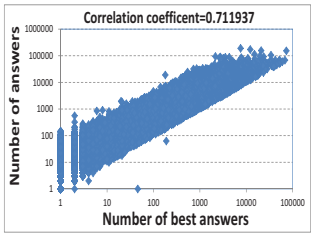


Figure 6: The correlation between best answers and all answers.

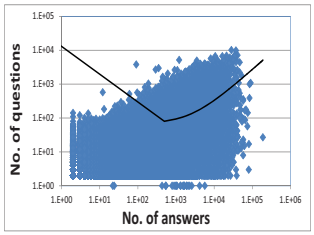


Figure 7: Correlation between # of questions and answers.

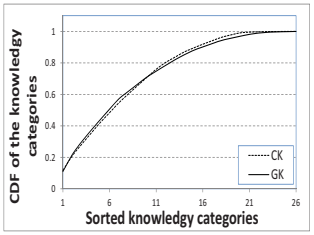


Figure 8: CDF of KCs.

knowledge. However, YA OSN has the following strikingly different properties from other general OSNs:

(1) YA shows a much lower level of link reciprocity, which means that the connection between two nodes tend to be unidirectional from an active learner to an active answerer. Since a fan-contact link means the fan's trust on the contact, the trust transitivity property along the links can be exploited to identify reputed sources and detect spammers in the Q&A system (Section 3.1).

(2) YA exhibits weaker correlation between indegree and outdegree. Nodes with high outdegree do not necessarily have high indegree, and nodes with high indegree do not necessarily have high outdegree. This means active knowledgable answerers are not necessarily active learner, and active learners are not necessarily active answerers (Section 3.3).

(3) YA does not have a tendency of higher-degree nodes to connect to other high degree nodes. Instead, nodes with a high indegree are connected by nodes with various outdegree due to celebrity-driven nature, in which many nodes tend to connect to a small number of active and knowledgable nodes (Section 3.4).

(4) The users in YA exhibit an extremely low clustering coefficient comparing to other friendship-oriented major OSNs due to its tendency of unidirectional connections to active and knowledgable answerers (Section 3.5).

4. ANALYSIS OF KNOWLEDGE DISTRIBUTION AND USER BEHAVIOR

As the Q&A OSN is knowledge-oriented, it is very important to examine the user knowledge distribution and associated user behaviors.

4.1 User Behavior

Figure 5 shows the CDFs of the best answers and all answers versus user rank based on the number of points. We see both CDFs follow a power-law distribution. 80% of the best answers are provided by 7628 users who are ranked in the top 10% of all users. Similarly, 80% of the answers are provided by 15739 users who are ranked in the 19% of all users. We also notice that all of the top contributors are within the top 10% users, which means that the best answers are from them. Therefore, in YA, a small portion of the users (i.e., 10%) contribute to most of the high-quality answers.

Figure 6 shows the number of all answers versus the number of best answers of each user. We calculated the Pearson correlation coefficient between these two numbers of all users, which is around 0.712. We can see that there is a positive linear relationship between the number of answers and the number of best answers and the correlation coefficient is very high. This is because as the number of answers provided by a user increases, the number of the best answers also increases.

Figure 7 shows the correlation between the number of questions and the number of answers from each user with the log-log scale. We also plot the trend line for the data based on linear regression. From the trend line, we see that users with a small number of answers have a large number of questions. However, as the number of answers increases, the number of questions decreases linearly and then increases linearly at the point of x=1000. We use the ratio of the number questions to the number of answers r_{qa} to reflect the querying and answering activities of the users. $r_{qa} > 1$ means a user's asked questions are more than his answered questions. Our data shows the average r_{qa} is 0.437, the variance is 5.61. 23.1% of the users have $r_{qa} < 0.01$, which are the selfless nodes that answer much more questions than the questions they ask. 13.6% of the users have $r_{qa} > 100$, which are likely to be free-riders that ask many questions while answer only a few questions. All top contributors are in the 23.1% of the selfless nodes. It is also very interesting to see that in the top 1409 users who answer more than 10,000 question, 110 (7%) of them did not ask any questions. We conjuncture that YA hires experts to answer others' questions in order to improve the quality of Q/A service.

4.2 Distribution of Knowledge Categories

We study the knowledge base of users by examining the KCs of the top contributors and normal users. Since the system does not specify the KCs in the profiles of normal users, we study their KCs through the questions in all the system's general KCs. This is reasonable because as Figure 5 shows, most of the normal users in the Q&A system are knowledge consumers, and they either provide low qualify answers or provide only a few answers. We call the KCs appearing in the top contributors' profiles *contributor's knowledge (CK)*. We notice that the KCs in CK include all *general knowledge (GK)* in the system.

We ranked the KCs in CK based on the appearance frequency of each KC in CK, and ranked those in GK based on the number of questions posted in each KC in GK. Figure 8 plots the CDF of the category rank in CK and GK, respectively. The figure shows that 80% of all questions are in the top 12 KCs in GK, and 80% of all contributors' KCs are also in the top 12 KCs in CK. This result means that users in the system are interested in the top 12 KCs in GK, and the active and knowledgable answerers also answer questions focused on the top 12 KCs in CK.

Each KC i has a pair of CK value and GK value, denoted by (v_{c_i}, v_{g_i}). A KC's CK value is defined as the percent

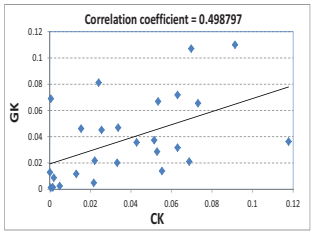


Figure 9: Distribution of KCs.

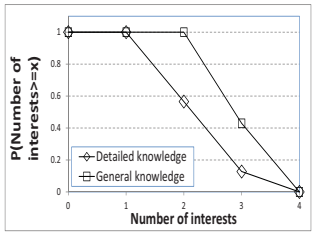


Figure 10: CCDF of KCs.

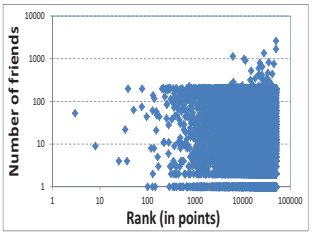


Figure 11: Number of contacts vs. rank.

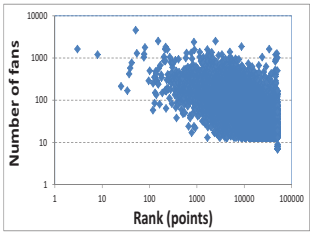


Figure 12: Number of fans vs. rank.

of its appearance frequency in the sum of the appearance frequencies of all categories in CK, and its GK value is the percent of its number of questions in the total number of questions. Our measurement shows that the Pearson correlation coefficient between the two values of all KCs equals 0.4988, which shows a strong correlation between CK and GK. Each point in Figure 9 shows (v_{c_i}, v_{g_i}) of each knowledgable category $i \in [1, 26]$. We see that the KC that has a large CK value tends to have a large GK value. The KCs that are popular in top contributors are also popular in general knowledge of all users, because the KCs in which top contributors frequently answer questions are also the KCs, in which users frequently ask questions.

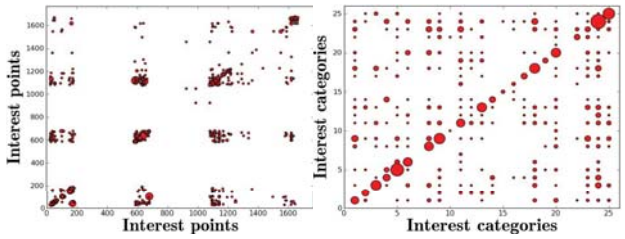
4.3 Behavior and Knowledge Base of Top Contributors

Section 3.2 shows that the node indegree exhibits a power-law distribution. The behavior of high-indegree users may greatly affect the attractiveness of the application as these users contribute significantly more than normal users. We like to study these users' behaviors including answering frequency and earning points, which also indicate the effort needed to attract application users. We quantified the number of answers submitted and points earned by the 4000 top contributors that have the highest indegree from Aug. 17th to Oct. 19th, 2011. Table 3 and Table 4 show the maximum, average and minimum numbers of the answers submitted and points earned by these users during each week during the time period. We see that the average number of submitted answers (around 40) and earned points (around 300) during each week remain nearly constant. Also, a few users are very active in answering questions, the largest number of questions answered per week is over 1100. In addition, because the users that provide more best answers earn more points, the quality of the answers from some users is also very high. The maximum number of earned points in the week of maximum 1524 submitted questions is 16742. The highest points earned by a user is 19975 in a week with 1405 maximum submitted questions.

Given the KCs of users, we are very interested in how knowledge and expertise are spread across different domains. Figure 10 shows the CCDFs of the general KCs and the detailed KCs of the top contributors. We see all of the top contributors have 2 or more detailed KCs. Only 56.5% of the top contributors have 2 or more general KCs. The result shows that many top contributors like to answer questions within one general KC, in which they may participate in answering questions in at least two detailed categories.

4.4 Relationship Between Degree and Rank

Figure 11 and Figure 12 show the number of each user's contacts and fans versus his rank based on the number of points, respectively. Figure 11 shows that in the contact network, there are no correlations between the number of a user's contacts and his points. This result implies that how active a user is in learning is not determined by how active he is in answering questions. Also, most users have less than 200 contacts, and some outliers have more than 200 contacts. This is because YA constrains the number of the contacts of each user within 200 since 2007. We found that the outliers' account creation times are all in 2007, while all other users' account creation times are after 2007. From Figure 12, we see a user with higher rank is likely to have larger number of fans. This is because active and knowledgable nodes having many points are more likely to attract fans. This is one of the most important reason for the power-law distribution of user indegree.



(a) Detailed KCs (b) General KCs

Figure 13: Correlation between detailed KCs and between general KCs.

4.5 Relationship Between Knowledge Categories

We assigned a numerical ID to each detailed KC so that the detailed KCs in the same general KC have close numerical IDs. We use matrix A[x][y] to represent the coexistence of two detailed KCs with ID x and ID y in one top contributor. Figure 13(a) shows the relationship between detailed KCs represented by the points of A[x][y]. We see that the detailed KCs are highly clustered. The KCs with IDs in [0,200], [600,700], [1000, 1200] are very likely to coexist with each other. However, KCs with IDs in [200, 600], [700, 1000], [1200, 1500] are seldom interested by top contributors because there KCs have extremely low popularity. Using the same way, we plot Figure 13(b) to show the relationship of the general KCs, which are assigned with ID from 1 to 26. We see that the top contributors are likely to have knowledge within the same category. It is also very likely for other kinds of category combinations to exist in a top contributor's specialized field.

4.6 Summary

(1) In YA, a small portion (10%) of the users (i.e., top contributors) contribute to most of the high quality answers. There is a strong correlation between best answers and all answers for a user with correlation coefficient equals

Period	Aug.17-Aug.22	Aug.23-Aug.30	Aug.31-Sep.08	Sep.09-Sep.14	Sep.15-Sep.21	Sep.22-Sep.29	Sep.30-Otc.06	Otc.07-Otc.13	Oct.13-Otc.19
Max	1367	1105	1306	1410	1500	1405	1235	1445	1524
Ave.	28.4	43.7	52.0	41.25	40	36.9	38.7	49.2	51.2
Min	0	0	0	0	0	0	0	0	0

Table 3: The number of answers submitted by top contributors during each week.

Period	Aug.17-Aug.22	Aug.23-Aug.30	Aug.31-Sep.08	Sep.09-Sep.14	Sep.15-Sep.21,	Sep.22-Sep.29	Sep.30-Otc.06	Otc.07-Otc.13	Oct.13-Otc.19
Max	8449	15808	13875	14435	15401	19975	14532	15643	16742
Ave.	302.8	302.1	301.3	274.1	300.0	322.4	321.2	314.4	309.7
Min	0	0	0	0	0	0	0	0	0

Table 4: The number of points earned by top contributors during each week.

0.712. At least 13.6% of the users are very likely to be free-riders (Section 4.1).

(2) In both contributor's knowledge and general knowledge, the top 12 KCs account for 80% of all knowledge. Meanwhile, there is a strong correlation between CK and GK with correlation coefficient equals 0.4988, which means the distribution of KCs in top contributors' profiles can represent the distribution of KCs of questions of all users (Section 4.2).

(3) The top contributors steadily and selflessly contribute knowledge to the system. 56.5% of the top contributors have multiple general KCs, and all of the top contributors have multiple detailed KCs (Section 4.3).

(4) There is no correlation between the number of contacts and the number of points of a user, but there is a positive linear relationship between the number of fans and the number of points of a user (Section 4.4).

(5) The KCs of the users are highly clustered, and users are likely to have knowledge within the same general KC. Different kinds of general category combinations are still likely to exist in a top contributor's specialized field (Section 4.5).

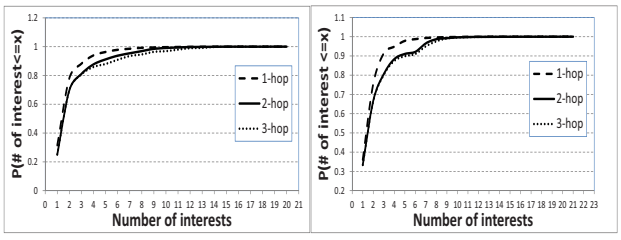
5. ANALYSIS OF KNOWLEDGE BASE IN A USER'S SOCIAL NETWORK

Users interested in the same KC tend to connect to each other as contacts and fans to facilitate knowledge sharing. In this section, we are interested in answering two questions: "how many different KCs exist within a certain hops of a user's contact network and fan network?" and "how shared KCs affect the link establishment between users?"

5.1 Relationship between Knowledge Base and Social Network Scope

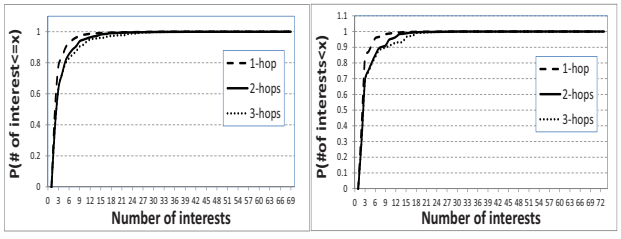
We are interested in answering a question: "how many KCs are there in a user's contacts or fans?" We define the size of the general (or detailed) knowledge base of a user within x hops in his contact (or fan) network as the size of the union of all general (or detailed) KCs of the contacts (or fans) within x hops in his contact (or fan) network. Figure 14(a) shows the CDF of the size of the general knowledge base of users within i ($1 \leq i \leq 3$) hops in their contact networks. We see that 80% of the users have a knowledge base with size <2 within 1 hop, and have a knowledge base with size <3 within 2 hops in their contact networks. The knowledge size distribution within 3 hops is approximately the same as that within 2 hops. Figure 14(b) shows the CDF of the size of the general knowledge base of users within i ($1 \leq i \leq 3$) hops in their fan networks. The distribution of the knowledge base in fan networks exhibits the same pattern as that in contact networks. Although a few users can have a knowledge base with size up to 21, 80% of the users have a knowledge size <3 within 1 hop, and have a knowledge size < 4 within 2 hops. Figures 15(a) and (b) show the CDF of the size of the detailed knowledge base of users within i ($1 \leq i \leq 3$) hops in their contact networks and fan networks, respectively. The results exhibit the same pattern as in Figure 14.

The small knowledge base size is caused by the reason that users are clustered at KCs. As YA OSN is knowledge-oriented, users with the same knowledge interest are likely to connect to each other. Also, as some of the KCs are highly correlated as shown in Figure 13, some users tend to share multiple KC interests. The knowledge base size for 3-hop scope is not significantly increased for both general KCs and detailed KCs. This is because the knowledge-oriented clusters are likely to be disconnected to each other and the 3-hop neighbors are still likely to be within the same knowledge cluster.



(a) Contact network (b) Fan network

Figure 14: Number of general KCs in the neighbors.

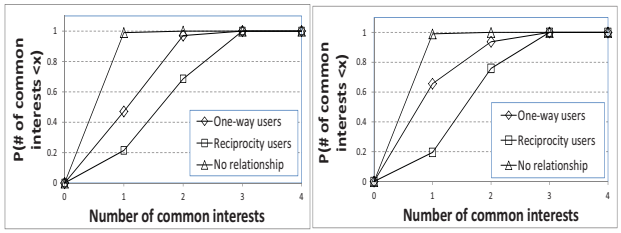


(a) Contact network (b) Fan network

Figure 15: Number of detailed KCs in the neighbors.

5.2 Homophily

Homophily is a tendency that "a contact between similar people occurs at a higher rate than among dissimilar people" [24]. In this section, we examine the pattern of homophily among users in the YA system by investigating the common KCs between each top contributor with his one-way connected contacts and fans, reciprocally connected users, and users without any relationship. Figure 16(a) shows the CDF of the number of common general KCs for users with one-way relationship, reciprocal relationship and no rela-



(a) General KCs. (b) Detailed KCs.

Figure 16: Common KC interests between users.

tionship, respectively. We see 80% of the reciprocity users share more than 2 common KCs; 80% of the one-way users share more than 1 and less than 2 KCs; 80% of the users without relationship share less than 1 KC. Figure 16(b) shows the CDF of the number of common detailed KCs for users with one-way relationship, reciprocal relationship and no relationship, respectively. We can also see that the distribution of detailed KCs for different users is similar to that of the general KCs. That is, no matter for detailed KC or general KC, reciprocity users share more interests than users with one-way relationship, which share more interests than users without relationship.

5.3 Summary

(1) Users in the social networks of the YA are clustered and centered by KCs. Some of the KC clusters are likely to be disjointed. (Section 5.1).

(2) Reciprocity users share more common KCs than one-way users, who share more common KCs than users without relationship (Section 5.2).

6. FURTHER DISCUSSIONS

In this section, we discuss some implications of our findings. While our findings are applicable to many different purposes and applications, we concentrate on spammer detection and distributed Q&A system design.

6.1 Implications to Spammer Detection in Q&A Systems

In YA, every registered users can post answers. Spammers might post commercial spam to earn attentions for their products.

Our study on YA presents two implications in the spammer detection algorithm design.

Best answer percent. Summary 4.6 shows that there is a linear relationship between the number of best answers and the number of all answers of a user with correlation coefficient equals 0.712. A spammer tends to post many answers but few of which would be selected as best answers. Therefore, by monitoring the ratio of the two numbers of a user, we can quickly identify the users with high ratios as suspicious spammers. Although the spammers can collude to rate their own answers as best answers, as the best answers are highlighted in the Q&A forum with high visibility to many other users, the false best answers can be easily identified using the abuse report policy.

Trust transitivity-based reputation. Summary 3.6 indicates that YA shows a very low level of link symmetry. Also, nodes with high indegree do not necessarily have high outdegree. i.e., user A connects to B only when user A trusts B's knowledge. Based on this property, we can evaluate node reputation based on the rationale that the users with many best answers should have a high reputation value and the users in the contact lists of high-reputed nodes should also be trustable and have high reputations. Similarly, in the HITS [25] and PageRank [26] algorithms, a webpage that is linked to by many webpages with high PageRank receives a high rank itself. Leveraging these algorithms, we can calculate the reputation value of users in order to detect the spammers:

$$R(u_i) = \frac{1-d}{N} + d \cdot \sum_{u_j \in S(u_i)} \frac{R(u_j)}{N(u_j)}, \quad (1)$$

where $R(u_i)$ denotes the reputation value of user u_i, d is a weight parameter, $S(u_i)$ denotes the set the users in u_i's fan network, and $N(u_j)$ denotes the outdegree of user u_j.

We use *Pagerank* to denote the above reputation calculation method, and use *Percentage* to denote the method that directly uses the percent of a user's best answers in his all answers as his reputation. We then ranked the user in the descending order of user reputation. Figure 17 shows the distribution of scaled reputation ([0,1]) of the users in *Pagerank* and *Percentage*. We see that *Pagerank* can more accurately reflect the reputations of users. Users with high best question percentage have high *Pagerank* reputation values and vice versa. However, *Percentage* results in approximately the same reputation values regardless of their best question percentages. The result indicates the effectiveness of *Pagerank* in reflecting node reputations.

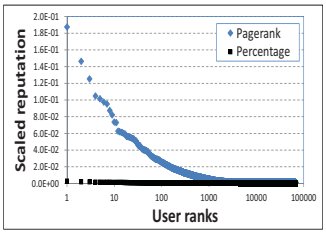


Figure 17: Reputation distribution of users.

6.2 Implications to Distributed Q&A Systems

Distributed Q&A systems [12–14, 27] identify the possible answerers in a questioner's social network in a centralized server and directly forward questions to the possible answerers. Google spent 50 million dollars to buy the Aardvark distributed Q&A system [14] on February 11, 2010. However, Google announced it would discontinue the Aardvark service in September 2011. Though we do not know the reasons, but our findings from YA can help enhance the performance of distributed Q&A systems, including Aardvark.

Embrace load imbalance. In order to balance the load between experts, the distributed Q&A systems [12–14, 27] use load balancing algorithm to evenly distribute the traffic among different experts. However, the assumption that every expert is willing to answer questions does not hold true. Summary 4.6 indicates that most users in the Q&A system are not actively in answering questions or are not be able to provide satisfying answers, while a small number of nodes (10%) are very willing to answer questions and able to offer satisfying answers. Therefore, rather than aiming to achieve load balance, forwarding more questions to those selfless answerers should be more effective in performance enhancement. Meanwhile, effective incentives such as reputation system or service pricing system are needed to encourage users to participate in question answering.

Bridge disjoint clusters. Summary 5.3 indicates that in the knowledge-oriented OSNs, some of the social network clusters centered on KCs are likely to be disjointed. Therefore, a user may not receive the answers for his questions in the distributed Q&A system because his connected users

have small knowledge base and they cannot reach other parts of the social network. Therefore, we need to create bridges between social network clusters to prevent the isolation of some users' social networks.

Hierarchical searching. Summary 4.6 indicates that users tend to have knowledge within the same general KC, and have several detailed KCs. To facilitate answerer search, users can be first indexed by their specialized general KC and then by detailed KC. To search an answerer, we can first identify the general knowledge cluster, and then use detailed KC to identify the experts.

Global index for unpopular topics. Summary 4.6 shows that the number of KCs interested or specialized by users conforms to the power-law distribution. If the Q&A activity is conducted in a distributed manner in YA, since a user prefers to connect to experts, it should be easy to find the experts to answer questions in popular KCs, but may take a long time to identify answerers in unpopular KCs. Therefore, we can use a global index (e.g., distributed hash table) for fast expert identification in unpopular topics.

In the framework, users invite their friends and knowledgable and active answerers to connect to. Such a hybrid friendship-knowledge oriented framework can leverage the advantage of the friendship-oriented OSNs that can provide trustable and personalized answers and knowledge-oriented OSNs that guarantee a small delay for answerer identification for both factual and non-factual questions. Bridges are added to isolated users' social networks to form a connected network. Thus, questions can be uninterruptedly forwarded along the connected friends to find answerers in a distributed manner. During the forwarding process, the probability that a user is identified as answerer should be determined by the user's both willingness and ability to answer the question based on his historical answering activity in the KC of the question. In addition, the experts in unpopular topics form a DHT structure for easy identification.

7. RELATED WORK

Online social networks. The rising popularity of OSN services has spurred a larger amount of research on OSNs. Most researches studied network structure and growth patterns. Backstrom *et al.* [28] investigated the evolution of network structure and group membership in MySpace and LiveJournal and showed that homophily can be used to improve predictive models of group membership. Zhu [22] measured and analyzed an online content voting network, Digg. He studied the structural properties of Digg OSN and the impact of OSN on user digging activities, and investigated the issues of content promotion and content filtering. Kwak *et al.* [23] studied the OSN structures in Twitter. Viswanath *et al.* [29] studied the network structure of Facebook, with an emphasis on the evolution of activity between users. Mislove *et al.* [24] analyzed the structures of multiple OSNs: Flickr, YouTube, LiveJournal and Orkut, and found they share some similar features. For example, the indegree of user nodes tends to match the outdegree, networks contain a densely connected core of high degree nodes.

Yahoo!Answers (YA). A number of the researches have been conducted on YA on other aspects. Adamic *et al.* [30] studied the content characteristics of the answers, based on which, they try to predict whether a particular answer will be chosen as the best answer. Su *et al.* [31] studied the quality of human reviewed data on the Internet using the answer ratings in YA. By using content analysis and human coding, Kim *et al.* [32] studied the selection criteria for best answers in YA. Cao *et al.* [33] proposed a category-based framework for search in YA. The framework uses language models to exploit categories of questions for improving answer search. Liu *et al.* [34] presented a general prediction model with a variety of content, structure, and community-focused features to predict whether a question author will be satisfied with the answers submitted by the community participants. As far as we know, our work is the first to study the structure, user behavior, user knowledge in the YA OSN from the perspective of knowledge sharing oriented OSN.

Knowledge sharing. Knowledge sharing has been studied for a long time. Initially, it was largely studied within organizational settings (e.g., Davenport [35]). The Internet gave rise to OSNs that aim at facilitating collaboration between people by providing an environment for mutual sharing and interaction (e.g., Wikipedia). Expert location systems [15–18] have been proposed to facilitate users to identify the experts of interests. Numerous online Q&A systems also have emerged in the Internet [6, 7], in which the anonymous users post and respond to others' questions. However, the latency in receiving a satisfying answer to a question is high. Some works studied Q&A behaviors in OSNs. Morris and Teevan [10, 11] studied how people use status messages in an OSN to ask questions. Similar to the status message, Hsieh *et al.* [8] proposed a market-based Q&A service called MiMir, in which all questions are broadcasted to all users in the system. However, by using status messages, only direct friends of a user can see the questions. Also, the broadcasting generates high overhead. White and Richardson [12, 13] developed a synchronous Q&A system called IM-an-Expert, which automatically identifies experts via information retrieval techniques and facilitates real-time dialog via instant messaging without broadcasting. However, IM-an-Expert focuses on the direct friends of a user. Meanwhile, the synchronous communication may face challenges with interruption costs and the availability of knowledge at the question time. Aardvark [14] tries to automatically route the question from a user to the most appropriate person in the Aardvark community. Yang *et al.* [27] proposed a social network-based system for supporting interactive collaboration in knowledge sharing over a peer-to-peer network. They found that applying social network-based collaboration support to knowledge sharing helps people find relevant content and knowledgable collaborators.

8. CONCLUSIONS

Regarding YA as a knowledge-oriented OSN, we have investigated the collective intelligence in the YA OSN in terms of OSN structure, user behavior and knowledge, and the knowledge base in a user's social network. Our study on the OSN structure shows that compared to other major OSNs, the YA OSN has some very distinct features. It has low level link symmetry, exhibits weak correlation between indegree and outdegree, and nodes tend to connect to nodes with different degree from their own. By studying the knowledge base and behaviors of users, we find that 10% of the users contribute to 80% of the best answers and 70% of the all answers. The first 12 most popular KCs include 80% of the questions among all questions. The top contributors steadily and selflessly contribute knowledge to the system. The KCs of the users are highly clustered since users are

likely to have knowledge within the same general KC. By studying the knowledge base in a user's social network, we find that the knowledge base of a user's social network is small because common-interest users are likely to be clustered. Also, a strong pattern of homophily is observed. We have outlined how these observed properties can be leveraged for spammer detection and distributed Q&A system design. In the future, we will further study the knowledge base of the non-top contributors and investigate the relationship between their knowledge base and behaviors.

Acknowledgements

This research was supported in part by U.S. NSF grants OCI-1064230, CNS-1049947, CNS-1156875, CNS-0917056 and CNS-1057530, CNS-1025652, CNS-0938189, CSR-2008826, CSR-2008827, Microsoft Research Faculty Fellowship 8300751, and U.S. Department of Energy's Oak Ridge National Laboratory including the Extreme Scale Systems Center located at ORNL and DoD 4000111689.

9. REFERENCES

[1] B. M. Evans and E. H. Chi. An elaborated model of social search. *IPM*, 2009.

[2] E. Amitay, D. Carmel, N. Har'El, Ofek-Koifman, and et al. Social search and discovery using a unified approach. In *Proc. of HT*, 2009.

[3] D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, and et al. Personalized social search based on the user's social network. In *Proc. of CIKM*, 2009.

[4] S. Kolay and A. Dasdan. The value of socially tagged urls for a search engine. In *Proc. of WWW*, 2009.

[5] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *Proc. of WWW*, 2007.

[6] Yahoo!Answer. http://answers.yahoo.com.

[7] Ask. http://www.ask.com.

[8] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan. Predictors of answer quality in online q&a sites. In *Proc. of SIGCHI*, 2008.

[9] G. Hsieh and S. Counts. mimir: A market-based real-time question and answer service. In *Proc. of SIGCHI*, 2009.

[10] 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. of CHI*, 2010.

[11] 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. of AAAI*, 2011.

[12] 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.

[13] M. Richardson and R. W. White. Supporting synchronous social q&a throughout the question lifecycle. In *Proc. of WWW*, 2011.

[14] D. Horowitz and S. D. Kamvar. The anatomy of a large-scale social search engine. In *Proc. of WWW*, 2010.

[15] H. H. Chen, L. Gou, X. Zhang, and C. L. Giles. Collabseer: A search engine for collaboration discovery. In *Proc. of JCDL*, 2011.

[16] C. Y. Lin, N. Cao, S. X. Liu, S. Papadimitriou, J. Sun, and X. Yan. Smallblue: Social network analysis for expertise search and collective intelligence. In *Proc. of ICDE*, 2009.

[17] H. Kautz, B. Selman, and M. Shah. Referral web: combining social networks and collaborative filtering. *Communications of the ACM*, 1997.

[18] D. W. McDonald and M. S. Ackerman. Expertise recommender: a flexible recommendation system and architecture. In *Proc. of CSCW*, 2000.

[19] Facebook. http://www.facebook.com.

[20] A. Mislove M. Cha and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proc. of WWW*, 2010.

[21] R. Kumar, J. Novak, and et al. Structure and evolution of online social networks. In *Proc. of KDD*, 2009.

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

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

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

[25] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of ACM*, 1999.

[26] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[27] 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. *IJHCS*, 2008.

[28] L. Backstrom, D. Huttenlocher, and et al. Group formation in large social networks: membership, growth, and evolution. In *Proc. of SIGKDD*, 2006.

[29] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In *Proc. of WOSN*, 2009.

[30] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo!answers: everyone knows something. In *Proc. of WWW*, 2008.

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

[32] 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.

[33] X. Cao, G. Cong, B. Cui, C. S. Jensen, and C. Zhang. The use of categorization information in language models for question retrieval. In *Proc. of CIKM*, 2009.

[34] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. In *Proc. of SIGIR*, 2008.

[35] T. Davenport and L. Prusak. Working knowledge: how organizations manage what they know. *Harvard Business School Press*, 1998.