Communities of Yahoo! Answers and Baidu Zhidao: Complementing or Competing?

Baichuan Li
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
bcli@cse.cuhk.edu.hk

Michael R. Lyu
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
lyu@cse.cuhk.edu.hk

Irwin King
AT&T Labs Research, USA &
The Chinese University of Hong Kong, Hong Kong
irwin@research.att.com
king@cse.cuhk.edu.hk

***Abstract*—Community Question Answering (CQA) attracts increasing volume of research on question retrieval, high quality content discovery and experts finding. However, few studies are focused on community per se of CQA services and also provide an in-depth analysis of them. This paper aims to enrich our knowledge on two of these CQA services, namely Yahoo! Answers and Baidu Zhidao through reviewing their communities, comparing similarities and differences of the two communities, together with analyzing their influence on solving questions. Six data sets are employed for comparative analysis. In this paper: (1) We analyze the social network structures of Yahoo! Answers and Baidu Zhidao; (2) We compare the the social community characteristics of top contributors; (3) We reveal the behaviors of users in different categories in these two portals; (4) We reveal temporal trends of these characteristics; (5) We find that the community of Yahoo! Answers and Baidu Zhidao complement each other in efficiency and effectiveness of answering questions.**

***Index Terms*—Community question answering, community, Yahoo! Answers, Baidu Zhidao, comparative analysis**

I. Introduction

Recently, Community Question Answering (CQA) services such as Yahoo! Answers[1] and Baidu Zhidao[2] have been developed to provide online users with more targeted and flexible online Question Answering services. Different from traditional Question Answering (QA) systems which automatically answer questions posed in natural language based on local database or data on the web, CQA services are featured with allowing users to answer questions asked by other users. As such, users are linked with each other and an online community of users is established. It is such kind of community which makes CQA distinct from QA.

The community in the CQA portals comprises of different users, who are further classified as askers, answerers and asker-answerers. Askers or answerers refer to those who asked/answered at least one question but have not answered/asked any others' questions. Asker-answerers are defined as those who have both asked questions and answered others' questions. Owing to so called "Community", users obtain what they need through interactions with other users. As shown in Fig. 1, users interact with others in direct or indirect ways. In this figure, we use blue, red, and black triangles to represent askers, answerers and asker-answerers. An answerer (red triangle) interacts with an asker (blue triangle) indirectly via answering the asker's question (dotted lines which link users, questions and answers in Fig. 1). Besides, users interact with answerers through rating up (if the user think the answer is good) or rating down (if the user think the answer is bad) their answers, and interact with askers by tagging their questions if they feel the questions are very interesting, or interact with other users by reporting abuse (if some users post prohibitive information such as advertisements in their questions or answers).

Increasing number of CQA portals are popular in the world, such as Baidu Zhidao, Quora[3], WikiAnswers[4], and Yahoo! Answers. Having reviewed aforementioned CQA portals, we find that questions and answers are all routinely put to different categories in these CQA services, and therefore sub-communities gradually appears as a result of such categorizations. However, few studies to date have been conducted to investigate the similarities and difference among communities of CQA, although community plays an essential role in promoting communications among users in CQA services. Investigating this topic will:

- show us a comprehensive knowledge of the characteristics and behaviors of users (especially the top contributors);
- give us a better understanding of efficient CQA community structure;
- facilitate us to construct or refine the current communities and therefore to provide a better platform for question resolving and knowledge sharing.

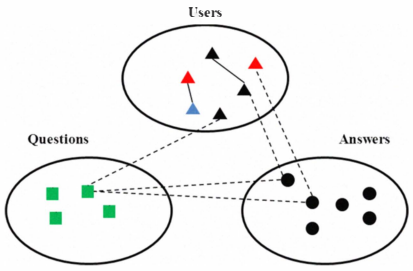

Fig. 1. User interactions in CQA

[1]http://answers.yahoo.com/
[2]http://zhidao.baidu.com/
[3]http://www.quora.com/
[4]http://wiki.answers.com/

To address the issue, Yahoo! Answers in USA (English CQA portal) and Baidu Zhidao (Chinese CQA portal) are chosen for a comparative study. We select these two CQA portals because of their great impact on netizens. Yahoo! Answers staff claim that 200 million users worldwide[5] and 15 million users visit daily[6]. Baidu Zhidao report that there are 0.25 billion net citizens using Baidu Zhidao, and every day more than 10,000 new questions and 100,000 new answers are posted[7].

Our contributions and notable findings in this paper include:

1) We find that users in CQA portals form sub-communities within one or a few categories.
2) We reveal the great differences in the communities and sub-communities' characteristics in Chinese CQA portal (Baidu Zhidao) and English CQA portal (Yahoo! Answers in USA) through social network and statistical analysis. To our knowledge, this is the first of its kind of studies made publicly available to demonstrate various characteristics of these CQA services.
3) We show the development of community (and sub-community) in CQA portals such as Yahoo! Answers and Baidu Zhidao through temporal trends analysis on the data in 2008 and 2010.

This paper proceeds as follows. In Section II, we provide a brief overview of recent related work in CQA research. Section III describes the data sets. Section IV provides the comparison between the community structures and top contributors' behaviors in Yahoo! Answers and Baidu Zhidao through social network analyses. In Section V, we conduct community detection to find sub-communities and analyze the behaviors of users in the corresponding sub-communities in Yahoo! Answers and Baidu Zhidao. In both Sections IV and V, we also present the change of community (sub-community) from a temporal perspective. In Section VI, we investigate the influence of community structure on efficiency and effectiveness in question solving. Conclusions are given in Section VII.

II. Related Work

Recent research in CQA mainly focuses on question retrieval, high quality answer finding and expert discovering.

Question Retrieval. Question retrieval is viewed as a special case of traditional information retrieval. In question retrieval, both queries and documents are questions which include question subject, content and/or additional information. In CQA domains, many state-of-the-art retrieval models are employed or developed for question retrieval, like the language model [1], [2], vector space model [3], Okapi BM25 [3], translation model [1], [4], and translation-based language model [5]. In addition, Wang et al. [6] propose a syntactic tree based algorithm to find similar questions from the perspective of NLP. Meanwhile, some characteristics of CQA are incorporated in current retrieval models, like category information [7], label ranking [8], question utility [9], and domain knowledge [10].

Answer Quality. Identifying the quality of one answer is of great importance to CQA portal. Since the quality of answers varies, distinguishing high quality answers from low quality ones help to select the best answers and identify spammers. Jeon et al. [11] propose a framework to predict the quality of answers with non-textual features. Agichtein et al. [12] leverage more features like community feedback to identify high quality answer. Recently, Wang et al. [13] propose an answer ranking algorithm by modeling question-answer relationships via analogical reasoning. In addition, Suryanto et al. [14] develop a series of models to find good answers considering both answer quality and answer relevance.

Expert Finding. Another important issue in CQA research is to discover experts (authorities). Estimating the authority of users could straight-forwardly provide a mechanism to judge the quality of answers under the hypothesis that experts always offer good answers. Jurczyk and Agichtein [15] [16] present link analysis of the link structure of CQA community to discover topic-free authority users. Zhang et al. [17] undertake another similar work using PageRank and HITS algorithms. Bouguessa et al. [18] argue that one major drawback of link analysis approach method is determining how many users should be chosen as authoritative. To address this problem, they propose a method to identify experts through a mixture model. Recently, a more personalized expert finding problem called "question routing", which finds experts for newly posted questions, is developed using language models [19], [20], [21] and topic models [22].

Communities of social network provides a large amount of useful information for us to explore, such as recommendation [23], [24], privacy issues [25], [26], etc. However, research on communities in CQA portals is still not well developed. Adamic et al. [27] analyzed the knowledge sharing activity in Yahoo! Answers and Rodrigues et al. [28] looked at individuals' objectives (socializing or knowledge sharing) when he/she posted a new question. But none of them have revealed the characteristics of communities and attempt a comparative analysis of community between different CQA portals.

III. Datasets

Six datasets are used in our experiments. The first dataset (YA08) is provided by [29], and contains 216,563 questions, 2,044,296 answers and 171,676 users crawled from Yahoo! Answers. In our experiments, we use part of the data whose questions were posted in January, 2008. Eventually, YA08 include 49,438 questions. The second dataset (BZ08) is crawled from Baidu Zhidao and all the questions were posted in January, 2008. As questions in Baidu Zhidao are ordered with the posting time, we crawled the questions every 10 from the question id "42900000" to "43900000" and finally 50,653 resolved questions are fetched.

[5]http://yanswersblog.com/index.php/archives/2009/12/14/yahoo-answers-hits-200-million-visitors-worldwide/

[6]http://yanswersblog.com/index.php/archives/2009/10/05/did-you-know/

[7]http://www.enet.com.cn/article/2010/0705/A20100705680331.shtml/

TABLE I
SUMMARY OF DATASET YA08, BZ08, YA10 AND BZ10

Dataset name	# of questions	# of answers	# of askers	# of answerers	# of asker-answerers	average length of questions	average length of answers	# of top categories
YA08	49,438	300,575	20,080	70,713	12,293	55.887	46.351	6
BZ08	50,566	158,388	38,289	68,196	4,984	66.139	113.005	14
YA10	162,175	838,807	47,593	107,375	36,181	49.865	44.235	26
BZ10	43,001	120,801	35,358	76,001	2,495	55.119	88.849	14
YA-Q	3,000	N/A	N/A	N/A	N/A	N/A	N/A	26
BZ-Q	3,000	N/A	N/A	N/A	N/A	N/A	N/A	14

TABLE II
THE DISTRIBUTION OF EACH COMPONENT IN BOW TIE STRUCTURE FOR YAHOO! ANSWERS, BAIDU ZHIDAO AND WEB PAGES.

	SCC	In	Out	Tendrils	DC
YA08	5.369%	39.053%	9.903%	37.652%	8.022%
BZ08	0.076%	8.646%	2.642%	29.296%	59.339%
YA10	12.135%	12.826%	46.394%	22.871%	5.775%
BZ10	0.036%	5.950%	5.123%	20.606%	68.285%
Web	27.740%	21.294%	21.207%	21.517%	8.243%

The third dataset (YA10) and forth dataset (BZ10) are crawled from Yahoo! Answers and Baidu Zhidao respectively whose questions are posted in July 2010. For YA10, we crawl the resolved questions across all categories posted from July 16 to July 22. For BZ10, we crawl the questions every 10 from the question id "167500000" to "169000000" and finally fetch 43,001 resolved questions.

The fifth (YA-Q) and sixth (BZ-Q) datasets each contains 3,000 questions posted in May 2010. We use these two datasets to analyze the efficiency of solving questions for the two portals' communities. Table I gives the statistics of the above six datasets.

IV. SOCIAL NETWORK ANALYSIS

To expose the community structure in Yahoo! Answers and Baidu Zhidao, we employ social network analysis on the YA08, BZ08, YA10 and BZ10 datasets. In Section IV-A, we analyze the bow tie structure of communities in Yahoo! Answers and Baidu Zhidao. Then we conduct centrality analysis to find top contributors (users who contribute a great number of answers or questions) and reveal the composition and characteristics of top contributors in Section IV-B.

A. Bow Tie Structure

Bow tie structure is used to represent "the ordered and recurrent structures that underlie complex technological or biological networks"[8]. A typical bow tie structure of one social network is composed of five parts (see Fig. 2): *Strongly Connected Component* (*SCC*), *In*, *Out*, *Tendrils* and *Disconnected Component (DC)*. For each node in *SCC*, there exists at least one directed path to any other node in *SCC*. In is the set of nodes which can reach any node in *SCC* in a directed path and Out is the set of nodes which can be visited by any node in *SCC* through a directed path. *Tendrils* is the set of nodes (besides *SCC*, *In* and *Out*) which can reach any node in *Out* or be visited by any node in *In*. Other nodes belongs to *DC*.

[8]http://en.wikipedia.org/wiki/Bow_tie_(biology)

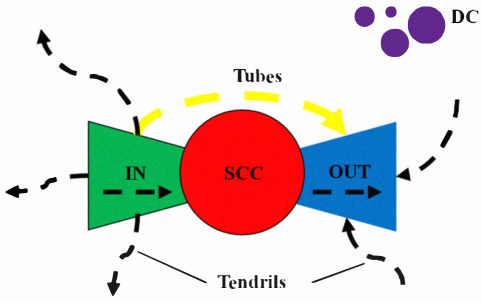

Fig. 2. The web is a bow tie (refer to [7]).

The community structure of CQA can be modeled by a directed graph $G(V, E)$ where each node $v \in V$ represents a user and each edge $e \in E$ (from answerer to asker) represents one ask-answer links (we can also add weight to each edge when considering the number of times asking-answering happens). Thus, in the constructed graph from CQA community, *SCC* is composed by users who frequently help each other since one can reach every other through the ask-answer links. *In* groups the users who mainly ask questions and *Out* groups the users who mainly answer questions.

We employ the algorithm described in [30] to analyze the bow tie structure of communities in Yahoo! Answers and Baidu Zhidao and report the result is in Table II. For comparison, we also include the result of web pages [30].

Observation 1 [Bow tie structure]. Users in Baidu Zhidao cannot construct a typical bow tie while users in Yahoo! Answers have the tendency to form a bow tie. In Baidu Zhidao, Less than 0.1% of users who frequently ask and answer questions. Furthermore, the ratio of *SCC* in Baidu Zhidao continues decreasing, which has impacted the community construction. The proportion of *In* decreases while the the proportion of *Out* increases, which shows more users prefer to answer questions and less users actively ask questions in Baidu Zhidao from 2008 to 2010. In YA08, The *In* and *Tendrils* components are dominant, which hold more than 75% users among all. In YA10, the ratio of users who mainly ask questions reduces much and the proportion of users who mainly answer questions increases greatly. In addition, the proportion of *SCC*, *In* and *Tendrils* are nearly the same. We believe that if the ratio of *SCC* keeps increasing in the future, the structure of community in Yahoo! Answers tends to be a bow tie like the structure of the web in the future. That would make the community more

balanced and active since the proportions of frequently askers, frequently answerers are equal with or a bit more than the proportions of the frequently asker-answerers.

B. Centrality

In a social network, the centrality measures the importance of a node within the graph. There are various measures of centrality, such as degree, closeness, betweenness and eigenvector. In this section, we apply the closeness centrality to find the top contributors and degree centrality to look for top askers and top answerers in Yahoo! Answers and Baidu Zhidao. By doing this, we aim to explore some characteristics of top contributors such as whether they are active in many categories or not, and their preferences in answering and asking.

1) Closeness: In network theory, closeness is defined as the mean geodesic distance (i.e., the shortest path) between a node and all other nodes reachable from it. In our experiments, we construct a weighted undirected graph to calculate the closeness centrality for each user in YA08, BZ08, YA10 and BZ10. The weight on each edge is the reciprocal of the times of ask-answer activities between the two nodes (users). Thus, a node with high closeness value (i.e., short distances to all other reachable nodes) should be the one which has many directed links to other nodes (i.e., the user should ask or answer a great number of questions).

Formally, let v denote one node, $CC(v)$ is the connected component reachable from v, the closeness value of v is defined as the reciprocal of the average distance to all other nodes in $CC(v)$:

$$Close(v) = \frac{|CC(v)|}{\sum_{w \in CC(v) \setminus v} dis(v, w)}. \quad (1)$$

Figure 3 reports the users with the top 10 highest closeness values (red ones) and their neighbors in Yahoo! Answer and Baidu Zhidao.

Observation 2 [Activity range of top contributors]. The top 10 contributors in Baidu Zhidao are not connected and the separated structure becomes more obvious from 2008 to 2010. However, the top 10 users and their neighbors in Yahoo! Answers are connected with each other through ask-answer links. When we further investigate the profile of the top 10 users with the highest closeness values in BZ10, we find that they mostly ask or answer questions in few categories while most top contributors in YA10 post answers and questions in relatively more categories. Thus, our observation shows that: with the development of Baidu Zhidao, more and more top contributors are only focus on a few particular categories. But top contributors in Yahoo! Answers still keep active in wide topics.

2) Degree: Degree centrality is defined as the number of links incident upon a node. It describes the importance of one node in a macro level. If the network is directed, we usually define two separate measures of degree centrality: indegree and outdegree. Indegree is the count of links directed to the node, and outdegree is the number of links the node directs to others.

TABLE III
THE NUMBER OF OVERLAPPING USERS ACROSS FOUR DATASETS

	YA08	BZ08	YA10	BZ10
Indegree & Outdegree	7	1	13	0
Indegree & Degree	63	18	78	49
Outdegree & Degree	39	81	29	50
All three	7	1	13	0

In our experiments, we construct weighted directed graphs to calculate the indegree centrality and outdegree centrality for YA08, BZ08, YA10 and BZ10. The weight on each edge is the times of asking-answering happens between the two nodes (users). Thus, in our setting the indegree centrality of node v is

$$Indegree(v) = \frac{indeg(v)}{indeg(V)}, \quad (2)$$

where $indeg(v)$ is the sum of weights for all links directed to the node. The outdegree centrality is calculated in the similar way. In addition, we calculate each node's degree centrality from weighted undirected graphs.

The user with high indegree centrality is a top asker while the user with high outdegree is a top answerer. To explore the relationships among top contributors, we rank the users according to the indegree, outdegree, and degree centrality respectively. Then we calculate the number of overlapping users in top 100 users between each two of them and among all the three measures. Table III presents the results.

Observation 3 [Type of top contributors]. In Yahoo! Answers there are more top contributors who both ask and answer a large number of questions. However, in Baidu Zhidao, there are very few users who contribute to a large amount of questions and answers at the same time. In addition, in Yahoo! Answers most top contributors prefer to ask questions rather than answer and this tendency remains from 2008 to 2010. For Baidu Zhidao, the proportion of top answerers is much larger than the ratio of top askers in 2008, but in 2010 the ratios of top answerers and top askers are nearly the same.

V. CATEGORY ANALYSIS

In this section we go deep into some categories as we find that sub-communities exist in CQA portals and they are naturally formed by a few categories (Section V-A). Then we discover the correspondent categories in Yahoo! Answers and Baidu Zhidao by category mapping (Section V-B). Finally, in Section V-C we conduct statistical analyses on four typical category pairs in Yahoo! Answers and Baidu Zhidao to expose the similarities and differences between the sub-communities of them.

Observation 4 [Activity range of users]. In CQA services most users only ask or answer questions on a very few categories. Figure 4 shows the pie chart which reports the proportion of users versus the number of categories in which they have asker questions or posted answers. We can observe that for all datasets, most users' behaviors are only limited to one category. For instance, in Yahoo! Answers, more than 68% of users ask or answer questions in one categories

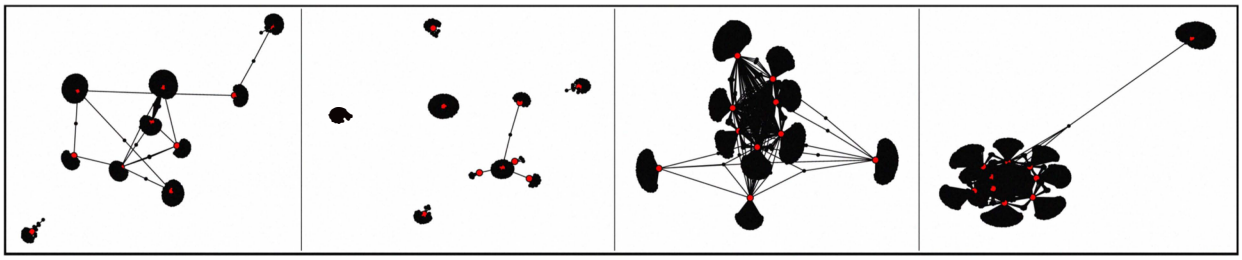

Fig. 3. The top 10 users with highest closeness values (red ones) and their neighbors in Yahoo! Answer and Baidu Zhidao (from left to right: BZ08, BZ10, YA08 and YA10).

and less than 16% of users are active in more than two categories. In Baidu Zhidao, more than 84% of users merely ask or answer questions in one category and less than 5% of users are active in more than two categories. In addition, this phenomenon aggravates in Baidu Zhidao while alleviates in Yahoo! Answers from 2008 to 2010.

A. Community Detection

Observation 4 motivates us to explore whether users can be naturally separated to sub-communities by the topic of questions and answers. Thus, we conduct community detection on YA08, BZ08, YA10 and BZ10. For each dataset, we construct a weighted undirected graph, in which each node represents an user and each edge means there exists ask-answer relationship between the two. The edge is weighted by the times of ask-answer activities in the two nodes (users). We do not consider the direction of the edges because we aim to investigate whether the community is established based on the topic (category). We apply the Louvain method [31] to detect sub-communities (with minimum number of users $\geq$ 10) in the above four graphs.

We calculate the weighted entropy for all detected sub-communities. Let S denote the number of sub-communities detected, SC_s represents the s_th of sub-community, c_{ij} represents the number of questions u_i have asked (answered, asked or answered) in category j, the weighted mean entropy $E_{wm}$is defined as follows:

$$E_{wm} = -\frac{|SC_s|}{\sum_{s=1}^{S}|SC_s|}\sum_{j=1}^{T}P_{sj}\ln(P_{sj}), \quad (3)$$

$$P_{sj} = \frac{C_{sj}}{|SC_s|}, \quad (4)$$

$$C_{sj} = \sum_{u_i \in SC_s}\frac{c_{ij}}{\sum_{k=1}^{T}c_{ik}}, \quad (5)$$

where T is the total number of categories. In addition, we calculate the entropy for the largest sub-communities $SC_\triangle$.

Table IV reports the results, from which we find that all the values of entropy in the largest detected communities and weighted entropy are much smaller compared with the case that users are equally distributed in each category (in this case the values of $E_{SC_\triangle}$ or E_{wm} are 1.792 for YA08, 2.463 for BZ08 and BZ10, 3.258 for YA10). Thus, the results demonstrate that users within one community prefer to ask and answer questions in the same few categories. Furthermore, the entropy values in question (Q), answer (A), question and answer (Q&A) are nearly equal except in $E_{SC_\triangle}$ of BZ08, which demonstrate that the sub-community is established based on topics rather than asking or answering.

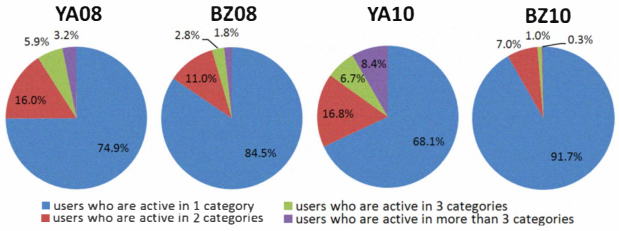

Fig. 4. Distribution of users across four datasets.

TABLE IV
THE WEIGHTED ENTROPY AND THE ENTROPY FOR THE LARGEST DETECTED SUB-COMMUNITIES ACROSS FOUR DATASETS

	K	$\lvert SC_\triangle\rvert$	$E_{SC_\triangle}$			E_{wm}		
			Q	**A**	**Q&A**	**Q**	**A**	**Q&A**
YA08	40	15467	0.66	0.76	0.72	1.04	1.15	1.12
BZ08	120	12843	2.06	1.68	1.80	1.71	1.69	1.69
YA10	32	33671	2.23	2.14	2.19	2.16	2.08	2.09
BZ10	338	11735	1.95	1.88	1.97	1.38	1.37	1.37

Observation 5 [Formation of sub-communities]. In CQA services users are naturally separately by topics (categories) and different sub-communities are formed based on different categories. Furthermore, this observation maintains in Yahoo! Answers and Baidu Zhidao from 2008 and 2010. The differences between Yahoo! Answers and Baidu Zhidao lie in that users in Yahoo! Answers prefer to be active in more categories (since the entropy is increasing) while users in Baidu Zhidao are more concentrated on fewer categories. It is reasonable to believe that if these tendencies keep going for Yahoo! Answers and Baidu Zhidao, users in Yahoo! Answers will become homogeneous while users in Baidu Zhidao will become heterogeneous.

B. Category Mapping

To facilitate the comparisons of sub-community structure in category level, we invite around 50 frequent internet to take part in a mapping survey through e-mails. These participants are all well educated with either having a Bachelor's or

Master's degrees or currently studying in their PhD programs. Moreover, their background cover a wide range of disciplines from natural sciences to social sciences, as well as engineering. All the participants are briefly informed of the research background and thereafter asked to map each category of Yahoo! Answers to that of Baidu Zhidao. They are given sufficient and flexible time, and as such they may visit Yahoo! Answers and Baidu Zhidao websites if necessary when they map the two systems to complete the e-mail survey. This mapping allows many-for-one and one-for-many cases, that is, there may be multiple categories (one category) of Yahoo! Answers correspond to one category (multiple categories) of Baidu Zhidao. Besides the original 14 categories in Baidu Zhidao, we also set another virtual one "Others". If the volunteer cannot map one category in Yahoo! Answers to any categories in Baidu Zhidao, he/she will map it to "Others".

Eventually, we received 32 responses and used their mapping results to construct a 26 by 15 matrix M[9]. Each row of M represents one category of Baidu Zhidao (plus "Others") and each column represents one category of Yahoo! Answers. The value of $m(i,j)(i \in [1,26], j \in [1,15])$ is the number of volunteers who map the two categories together. Note that this value can be fractional number in the case of one-for-many mapping. For instance, if one volunteer maps the 3rd category of Yahoo! Answers to both the 7th and 9th categories of Baidu Zhidao, we will add the value of $m(3,7)$ by 0.5 and $m(3,9)$ by 0.5. We apply Fleiss' kappa analysis [32] to measure the inter-rater reliability among these volunteers and obtain $\kappa = 0.54$, which means these 32 volunteers have moderate agreement on the category mapping from Yahoo! Answers to Baidu Zhidao according to [33] and the results can be used in our further study.

Observation 6 [Similar category settings]. We observe that in most cases, we can find a perfect one-to-one mapping. For example, *Sports* in Yahoo! Answers corresponds to *体育/运动* in Baidu Zhidao. However, both Yahoo! Answers and Baidu Zhidao have some special categories. For instance, Yahoo! Product, Environment and Cars & Transportation of Yahoo! Answers cannot be well mapped to any correspond categories in Baidu Zhidao. In turn, we cannot find the correspond category in Yahoo! Answers for the *烦恼* (*Trouble*) in Baidu Zhidao.

We design a measurement called *Fitness* to represent the degree of matching for two categories. Let $Fit(i,j)$ denote the *Fitness* of mapping the j-th category of Yahoo! Answers to the i-th category of Baidu Zhidao,

$$Fit(i,j) = \frac{2 \cdot m(i,j)}{\sum_{i=1}^{X} m(i,j) + \sum_{j=1}^{Y} m(i,j)}. \quad (6)$$

Table V reports the top 10 correspondent categories between Yahoo! Answers and Baidu Zhidao ranked by the value of *Fitness*. We find that 5 of all 6 top categories in YA08 can be well mapped to 4 categories in BZ08 (Both *Education & Reference* and *Science & Mathematics* in Yahoo! Answers can be mapped to *教育科学* (*Education & Science*) in Baidu Zhidao). In the next section, we compare the two portals' questions, answers, and users across the correspondent four category pairs (shown in Table V, in bold type).

[9]The result is shown at: http://appsrv.cse.cuhk.edu.hk/~bcli/doku.php?id=yabz

TABLE V
THE TOP 10 CORRESPONDENT CATEGORIES BETWEEN YAHOO! ANSWERS AND BAIDU ZHIDAO.

Yahoo! Answers	Baidu Zhidao	Fitness
Consumer Electronics	*电子/数码*	0.940
Sports	***体育/运动***	0.932
Games & Recreation	*游戏*	0.893
Computers & Internet	*电脑/网络*	0.891
Business & Finance	*商业/理财*	0.773
Arts & Humanities	***文化/艺术***	0.769
Health	***医疗/健康***	0.767
Education & Reference	***教育/科学***	0.550
Science & Mathematics	***教育/科学***	0.541
Entertainment & Music	*娱乐休闲*	0.538

C. Category Statistics

Tables VI and VII give a summary of the questions, answers and user structure across the four correspondent category pairs between Yahoo! Answers and Baidu Zhidao. We compare the properties of sub-communities in Yahoo! Answers and Baidu Zhidao from the following two perspectives: users' activeness in answering questions and the rate of asker-answerers among all users.

Observation 7 [Users' activeness]. We judge the community's activeness in answering questions from the number of answers per question. In Table VI, one question in Yahoo! Answers get 6 answers in average while one question in Baidu Zhidao only get 3 answers. It seems that users of Yahoo! Answers are more likely to provide answers. Comparing Table VI with Table VII, the average number of answers per question decreases to 5 in YA10 while it keeps same in BZ10. Specially, in *Health* category, this number decreases from 7 in YA08 to 3 in YA10, which is very interesting. Maybe the questions about the health are much harder to answer. For the four categories in Baidu Zhidao, the average number of answers per question have a slight decrease, too. Thus, the above results show that in CQA services, the number of answers for each question is decreasing from 2008 to 2010, which reflects the degeneration of the community.

Observation 8 [Usage of CQA]. In Table VI, the rate of asker-answerers in Yahoo! Answers (11.9%) is much higher than that of Baidu Zhidao (4.47%). This result shows that users in Yahoo! Answers prefer to help others who may meet the similar questions they have asked before. In addition, users in Yahoo! Answers are willing to help others when get others' help. Comparing Table VI with Table VII, in two years the rate of asker-answerers in Yahoo! Answers increases to 18.9%. However, in Baidu Zhidao, the rate decreases to 2.19%. In category view, the rate increases much in the *Art & Humanities*, *Health* and *Sports* categories for Yahoo! Answers, while decreases in all listed four categories. This change

TABLE VI
SUMMARY OF CORRESPOND CATEGORY PAIRS ACROSS YA08 AND BZ08.

Item	YA08					BZ08				
	A&H	**Edu**	**Health**	**Sports**	**All**	**A&H**	**Edu**	**Health**	**Sports**	**All**
# of questions	9,947	19,404	10,471	9,559	49,438	1,683	9,076	2,888	674	50,566
# of answers	54,154	82,216	70,069	93,726	300,575	5,890	25,437	9,101	2,956	158,388
# of answers per question	5.44	4.24	6.69	9.81	6.08	3.50	2.80	3.15	4.39	3.13
# of pure askers	5,810	10,926	5,590	2,674	20,080	1,544	7,215	2511	602	38,289
# of pure answerer	20,621	31,366	30,345	19,049	70,713	4,047	13,145	5,404	2,164	68,196
# of asker-answerer	1,890	2,900	2,520	2,885	12,293	55	447	73	19	4,984
Rate of asker-answerers(%)	6.67	6.42	6.55	11.72	11.93	0.97	2.15	0.91	0.68	4.47
Rate of users in all(%)	27.47	43.84	37.30	23.87	100.00	5.07	18.67	7.17	2.50	100.00

TABLE VII
SUMMARY OF CORRESPOND CATEGORY PAIRS ACROSS YA10 AND BZ10.

Item	YA10					BZ10				
	A&H	**Edu**	**Health**	**Sports**	**All**	**A&H**	**Edu**	**Health**	**Sports**	**All**
# of questions	6,608	13,597	9,405	7,609	162,175	2,351	7,071	2,187	552	43,001
# of answers	26,719	36,991	32,509	43,494	838,807	6,051	17,710	5,505	2,010	120,801
# of answers per question	4.04	2.72	3.46	5.72	5.17	2.57	2.51	2.52	3.64	2.81
# of pure askers	3,699	8,683	5,775	2,237	47,593	2,159	6,056	1,866	495	35,358
# of pure answerer	9,232	13,301	14,820	8,772	107,375	4,305	11,642	3,942	1,715	76,001
# of asker-answerer	1,281	1,261	1,599	1,815	36,181	53	149	29	11	2,495
Rate of asker-answerers(%)	9.01	5.43	7.21	14.15	18.93	0.82	0.84	0.50	0.50	2.19
Rate of users in all(%)	7.44	12.16	11.61	6.71	100.000	5.72	15.68	5.13	1.95	100.00

reveals the great differences between the two portals: users of Yahoo! Answers prefer to help others when receiving others' help but users in Baidu Zhidao just want to ask questions or provide answers.

VI. COMMUNITY EFFICIENCY

Conclusions drawn from Section IV and Section V are presented succinctly below:

1) Sub-communities are formed in Yahoo! Answers and Baidu Zhidao based on a few categories.
2) Yahoo! Answers has a certain amount of asker-answerers while in Baidu Zhidao askers and answerers are dominant and there are much fewer asker-answerers.
3) In Yahoo! Answers more people prefer to ask or answer questions in more categories while in Baidu Zhidao, most people (especially the top contributors) prefer to ask or answer questions in a few categories which they are interested in.
4) The phenomena of 2 and 3 are more evident in Yahoo! Answers and Baidu Zhidao from 2008 to 2010.

Since there are many differences in the structures of communities (and sub-communities) in Yahoo! Answers and Baidu Zhidao, one interesting question is whether these differences affect the question solving. To find the answer to this question, we track 3,000 newly posted questions in these portals respectively to check whether the states of these questions change within two days. The 6,000 questions compose the datasets YA-Q and BZ-Q.

Table VIII reports the tracking results, from which we find that in Yahoo! Answers 17.6% of questions receive satisfied answers within 48 hours. For those unresolved questions, nearly 1/5 of them receive no response (211 questions' states are missed since they are deleted by the system). For Baidu Zhidao, 22.7% of questions are well resolved, which is 28.97% higher than that of Yahoo! Answers. However, 42.8% of unresolved questions receive no response and the rate is much higher than that of Yahoo! Answers. We observe that users in Baidu Zhidao tend to give professional answers whereas users in Yahoo! Answers offer quick answers. These results are probably attributed to the structures and characteristics of communities (i.e., the composition and behavior of users). In Yahoo! Answers, more users who are active in visiting various categories and both asking and answering questions (as asker-answerers do) greatly enhance the efficiency of solving questions. However, more users in Baidu Zhidao visit fewer categories and only prefer to ask or answer questions (as askers and answerers do) and thus questions are not handled quickly, but if the questions are answered, the answers are probably of hight quality.

TABLE VIII
STATUS OF TRACKED QUESTIONS TWO DAYS AFTER BEING POSTED (YA=YAHOO! ANSWERS, BZ=BAIDU ZHIDAO).

Dataset	# of resolved Qs	# of unresolved Qs with at least one answer	# of unresolved Qs with no answer
YA-Q	527	1,820	442
BZ-Q	682	1,325	993

VII. CONCLUSIONS

In this paper we reveal the communities in two leading CQA portals, namely Yahoo! Answers and Baidu Zhidao. We conduct detailed and comparative analyses of the community (and sub-community) structures and user behaviors of the two

portals through social network and statistical analyses. We find that users in CQA portals form sub-communities within one or a few categories, and there are great differences in the communities (and sub-communities) structures and users' characteristics between Yahoo! Answers and Baidu Zhidao. In particular, in Yahoo! Answers there are a certain amount of asker-answers but in Baidu Zhidao askers and answers are dominant and there are much fewer asker-answerers. Furthermore, in Yahoo! Answers, more people prefer to ask or answer questions in more categories while in Baidu Zhidao, most people (especially the top contributors) prefer to ask or answer questions in a fewer categories which they are interested in. In addition, our temporal analysis reveals that the above two phenomena are aggravating from 2008 to 2010. What's more, users of Yahoo! Answers are more active in providing answers, although the enthusiasm about answering questions drops a little from 2008 to 2010.

The findings also confirm the influence of community (and sub-community) structure and characteristics on question solving. We find that more questions in Yahoo! Answers obtain answers efficiently but more questions in Baidu Zhidao receive high-quality answers and thus be resolved effectively. As we have mentioned in Section VI, community structures of Yahoo! Answers and Baidu Zhidao complement with each other in efficiency and effectiveness of question solving.

The observations and findings in this paper help us to better understand the community and design CQA portals to improve user experience and question solving. For instance, we may encourage more askers of Baidu Zhidao to provide answers through reward systems. In addition, we may identify the experts in Yahoo! Answers and route new questions to them to reduce wait time. In the future, we plan to further investigate the impact of community on question and answer qualities, which is another essential research issue in CQA research.

ACKNOWLEDGMENT

The work described in this paper was fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 413210 and CUHK 415311) and two grants from Google Inc. (one for Focused Grant Project "Mobile 2014" and one for Google Research Awards). The authors also would like to thank all participants in the category mapping and the reviewers for their helpful comments.

REFERENCES

[1] J. Jeon, W. B. Croft, and J. H. Lee, "Finding semantically similar questions based on their answers," in *Proc. of SIGIR '05*, 2005.

[2] 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 '09*, 2009.

[3] J. Jeon, W. B. Croft, and J. H. Lee, "Finding similar questions in large question and answer archives," in *Prof. of CIKM '05*, 2005, pp. 84–90.

[4] G. Zhou, L. Cai, J. Zhao, and K. Liu, "Phrase-based translation model for question retrieval in community question answer archives." in *ACL'11*, 2011, pp. 653–662.

[5] X. Xue, J. Jeon, and W. B. Croft, "Retrieval models for question and answer archives," in *Proc. of SIGIR '08*, 2008, pp. 475–482.

[6] K. Wang, Z. Ming, and T.-S. Chua, "A syntactic tree matching approach to finding similar questions in community-based qa services," in *Proc. of SIGIR '09*, 2009.

[7] X. Cao, G. Cong, B. Cui, and C. S. Jensen, "A generalized framework of exploring category information for question retrieval in community question answer archives," in *Proc. of WWW '10*, 2010, pp. 201–210.

[8] W. Wang, B. Li, and I. King, "Improving question retrieval in community question answering with label ranking," in *Prof. of IJCNN*, 2011.

[9] Y.-I. Song, C.-Y. Lin, Y. Cao, and H.-C. Rim, "Question utility: a novel static ranking of question search," in *Proc. of AAAI '08*, 2008.

[10] Z.-Y. Ming, T.-S. Chua, and G. Cong, "Exploring domain-specific term weight in archived question search," in *Proc. of CIKM '10*, 2010.

[11] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, "A framework to predict the quality of answers with non-textual features," in *Proc. of SIGIR '06*, 2006.

[12] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proc. of WSDM '08*, 2008.

[13] X.-J. Wang, X. Tu, D. Feng, and L. Zhang, "Ranking community answers by modeling question-answer relationships via analogical reasoning," in *Proc. of SIGIR '09*, 2009.

[14] M. A. Suryanto, E. P. Lim, A. Sun, and R. H. L. Chiang, "Quality-aware collaborative question answering: methods and evaluation," in *Proc. of WSDM '09*, 2009.

[15] P. Jurczyk and E. Agichtein, "Discovering authorities in question answer communities by using link analysis," in *Proc. of CIKM '07*, 2007.

[16] P. Jurczyk and E. Agichtein, "Hits on question answer portals: exploration of link analysis for author ranking," in *Proc. of SIGIR '07*, 2007.

[17] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," in *Proc. of WWW '07*, 2007.

[18] M. Bouguessa, B. Dumoulin, and S. Wang, "Identifying authoritative actors in question-answering forums: the case of yahoo! answers," in *Prof. of KDD '08*, 2008.

[19] Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao, "Routing questions to the right users in online communities," in *Proc. of ICDE '09*, 2009, pp. 700–711.

[20] B. Li and I. King, "Routing questions to appropriate answerers in community question answering services," in *Proc. of CIKM '10*, 2010.

[21] B. Li, I. King, and M. R. Lyu, "Question routing in community question answering: putting category in its place," in *Proc. of CIKM '11*, 2011, pp. 2041–2044.

[22] J. Guo, S. Xu, S. Bao, and Y. Yu, "Tapping on the potential of q&a community by recommending answer providers," in *Proc. of CIKM' 08*, 2008, pp. 921–930.

[23] H. Ma, T. C. Zhou, M. R. Lyu, and I. King, "Improving recommender systems by incorporating social contextual information," *ACM Trans. Inf. Syst.*, vol. 29, pp. 9:1–9:23, Apr. 2011.

[24] W.-Y. Chen, D. Zhang, and E. Y. Chang, "Combinational collaborative filtering for personalized community recommendation," in *Proc. of KDD '08*, 2008, pp. 115–123.

[25] M. Mo and I. King, "Exploit of online social networks with community-based graph semi-supervised learning," in *Proc. of ICONIP '10*, 2010, pp. 669–678.

[26] M. Mo, D. Wang, B. Li, D. Hong, and I. King, "Exploit of online social networks with semi-supervised learning," in *Proc. of IJCNN '10*, 2010, pp. 1 –8.

[27] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, "Knowledge sharing and yahoo answers: everyone knows something," in *Proc. of WWW '08*, 2008.

[28] E. Mendes Rodrigues and N. Milic-Frayling, "Socializing or knowledge sharing? characterizing social intent in community question answering," in *Proc. of CIKM '09*, 2009.

[29] Y. Liu, J. Bian, and E. Agichtein, "Predicting information seeker satisfaction in community question answering," in *Proc. of SIGIR '08*, 2008.

[30] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," *Computer Networks*, 2000.

[31] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[32] J. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, 1971.

[33] K. L. Gwet, *Handbook of Inter-Rater Reliability (Second Edition)*. Advanced Analytics, LLC, 2010.