Socializing or Knowledge Sharing? Characterizing Social Intent in Community Question Answering

Eduarda Mendes Rodrigues
Microsoft Research
7 JJ Thomson Avenue
Cambridge, CB3 0FB, UK
+44 (0) 1223 479 700
eduardamr@acm.org

Natasa Milic-Frayling
Microsoft Research
7 JJ Thomson Avenue
Cambridge, CB3 0FB, UK
+44 (0) 1223 479 700
natasamf@microsoft.com

ABSTRACT

Knowledge sharing communities, such as Wikipedia or Yahoo! Answers, add greatly to the wealth of information available on the Web. They represent complex social ecosystems that rely on user participation and the quality of users' contributions to prosper. However, quality is harder to achieve when knowledge sharing is facilitated through a high degree of personal interactions. The individuals' objectives may change from knowledge sharing to socializing, with a profound impact on the community and the value it delivers to the broader population of Web users. In this paper we provide new insights into the types of content that is shared through Community Question Answering (CQA) services. We demonstrate an approach that combines in-depth content analysis with social network analysis techniques. We adapted the Undirected Inductive Coding method to analyze samples of user questions and arrive at a comprehensive typology of the user intent. In our analysis we focused on two types of intent, *social* vs. *non-social*, and define measures of social engagement to characterize the users' participation and content contributions. Our approach is applicable to a broad class of online communities and can be used to monitor the dynamics of community eco-systems.

Categories and Subject Descriptors

H.1.2 **[Information Systems]**: User/Machine Systems – *human factors*; H.3.5 **[Information Systems]**: On-line Information Services.

General Terms

Experimentation, Human Factors, Measurement

Keywords

Q&A Community, Question Typology, User Intent, Social Scores.

1. INTRODUCTION

Community Question Answering (CQA) services have gained wide adoption over recent years, providing a community approach to question answering. Unlike automatic question answering systems (e.g., TREC Q&A Track [17]) or expert networks (e.g., allexperts.com or justanswer.com), answers are provided by a large community of users who can actively engage in answering any question, irrespective of their level of expertise. Still, the answer quality can sometimes reach, or even surpass, the quality of answers given by library reference services and experts [11]. Furthermore, the promptness of the answers is very attractive to users, even more so when users seek advice or opinions, which are unlikely to be obtained through standard Web search.

This type of social media generates a rich and evolving knowledge base, valuable to the broader population of Web users. Search engines have already started to surface CQA content, when deemed relevant to users' queries. However, given the diversity in content quality [14], it is important to develop methods for identifying relevant questions and answers [1, 2, 8] and ranking them accordingly [3, 15]. At the same time, it is important to understand how affordances of the CQA services reflect upon the interaction among individuals [13] and what can be done to encourage the exchange of desirable content. Inevitably, the CQA eco-system depends on the expertise of community members it attracts, the community responsiveness to questions, and the nature of users' interactions. Thus, it is important to gain a good understanding of the community behavior and individuals' contributions, e.g., by analyzing properties of the network representations that capture question response patterns [1, 2].

In our research we begin with the hypothesis that each question reflects a particular intent and therefore instigates a specific type of engagement by the community. Thus, characterizing the user intent is an important step towards understanding the community as a whole. For that reason we devised a method to analyze questions and arrive at a comprehensive typology of user intent. In this paper, we primarily focus on the impact that social behavior has on CQA communities and thus consider two broad classes of user intent, *social* vs. *non-social*. More precisely, all the questions that are intended for purely social engagement are considered *social*, and those that seek information, advice or opinion are considered *non-social* but instigating a knowledge sharing engagement. In order to characterize the communication patterns around *social* and *non-social* intent, we combine content analysis with social network analysis and develop measures of social engagement that quantify the users' participation and content contributions.

In the following section we discuss related work and, in section 3, we provide the relevant background about CQA communities. In section 4 we describe the methodology for deriving a typology of user intent and, in section 5, we use the typology to analyze the types of questions asked and answered by the top content contri-

CIKM'09, November 2–6, 2009, Hong Kong, China.

butors in two communities, Yahoo! Answers and MSN QnA. In section 6, we present the results of automatic classification of questions and we classify about half million questions from MSN QnA to characterize the level of *social* vs. *non-social* engagement in that community. We conclude with the summary of our work and directions for future work.

2. RELATED WORK

CQA services include a social networking component, whereby users may create ties to other users through asking and answering questions and establish a reputation in the community for their content contributions. At the same time, CQA services offer a vast and evolving knowledge base of questions and answers. They are a valuable resource for users with similar information needs. Recently, the research community has been actively looking into various problems associated with this type of communities. That includes research on identifying and predicting the quality of answers [1, 2, 8] and user satisfaction [10], learning to rank answers [3, 15], modeling network evolution [9], and modeling the user authority and level of expertise [4, 12].

Agichtein et al. [2] proposed a classification model for estimating the quality of answers in Yahoo! Answers based on features that are derived from the content and the answer-to social network, e.g., the authority measures. They classified question-answer pairs along several aspects, including how well-formed, readable, useful, and interesting they are. Further work on characterizing CQA content was done by Adamic et al. [1] by focusing on three specific topic categories from Yahoo! Answers. The three categories were selected among the 189 most active Yahoo! Answers categories, as the best representatives of the 3 clusters obtained by k-means clustering. The clustering considers three primary dimensions: *thread length*, given by the average number of answers, *text length*, given by the average number of characters in the answers, and *asker/replier overlap*, obtained by the cosine similarity between the asking and replying frequency of users. From the patterns in user interactions, they identified users with behavior similar to the 'answer-person' roles found in newsgroup communities [18].

Bian et al. [3] presented a ranking framework for retrieving factual information from social media that utilizes data about user interaction to retrieve high quality content. Harper et al. [11] probed various Q&A services, including Yahoo! Answers and MSN QnA, to assess the value of answers provided by each of these communities. They defined a set of probing questions according to three generic categories, *factual*, *opinion* and *advice*, posted them to each service, and subsequently analyzed the quality of received answers.

Our work complements the existing research by focusing on the user intent, as reflected in the user's questions, and the social engagement, rather than the quality of the content exchanged.

3. PRELIMINARIES

CQA services provide a platform for users to engage through asking and responding to each others' questions. They include features that facilitate posing questions and providing answers, and incentives to encourage user participation and self-regulation of the content quality, e.g., through content ratings, abuse reports, user reputation scores, and user rankings. Most of the existing research has focused primarily on Yahoo! Answers. We study and contrast two CQA communities: Yahoo! Answers and MSN QnA.

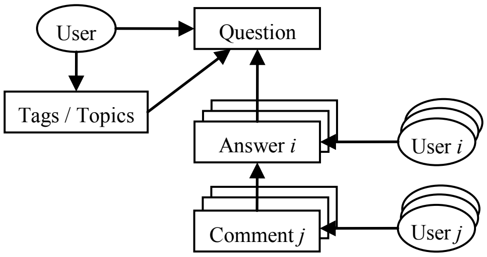


Figure 1. Main entities involved in a CQA thread.

We begin by providing a brief description of the Yahoo! Answers and MSN QnA services and discuss their social network structure.

3.1 Yahoo! Answers and MSN QnA

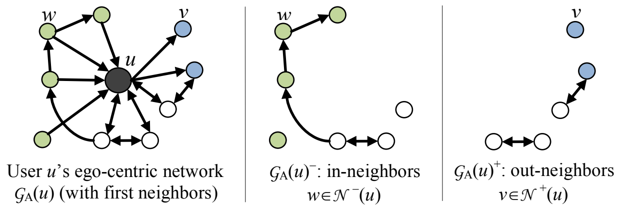
CQA services offer a fairly common set of features. Figure 1 represents a typical QA thread with the main entities involved in the question answering process. A question can receive several answers from multiple users during an *answering phase*. The community is then given an opportunity to provide comments and to vote on the quality of the received answers. In Yahoo! Answers users can cast both positive and negative votes (thumbs-up and thumbs-down) during the *voting phase*. They can also provide comments once the *best answer* has been selected. At the time of posting the question, the user is asked to select one of the topics from the set of predefined Yahoo! categories. The categories are used by other members of the community to find content on topics of their interest.

MSN QnA users can perform the same functions: ask, answer questions, comment, vote, and categorize questions. However, these are supported in slightly different ways. The user's vote is interpreted as approval and candidacy for the best answer. The users can comment on any answer at any time during the question lifecycle. This flexibility opens up opportunities for richer user interactions during the answering process, which results in a different structure for the QA threads. Finally, at the time of posting the question, the user is asked to assign a set of tags that best describe the question. The service suggests a set of candidate tags from a pool of community generated tags but the user can choose to create new tags as appropriate.

CQA services also provide search and browsing facilities so that users can explore recently posted questions and answers. In Yahoo! Answers browsing is supported through a topic taxonomy and each question is associated with a single topic. In MSN QnA, content is described through tagging which leads to ever-growing non-hierarchical, community generated categorization scheme. The tagging approach enables browsing through questions by topic but is also conducive to creating tags that can cover questions of social nature [13]. Such are questions with social intent, e.g., announcements, greetings, celebrations and news, personal questions, and similar.

3.2 CQA Social Networks

CQA includes three main types of user interactions that lead to implicit social networks: (1) *answer to* other users' questions, (2) *comment on* other users' answers, and (3) *vote on* other users' answers. The *answer-to* network, in particular, represents the primary form of user interaction on CQA services. We analyze this social network to identify structural features that are useful for predicting question types and user intent. We represent the



User *u*'s ego-centric network $\mathcal{G}_A(u)$ (with first neighbors)

$\mathcal{G}_A(u)^-$: in-neighbors $w \in \mathcal{N}^-(u)$

$\mathcal{G}_A(u)^+$: out-neighbors $v \in \mathcal{N}^+(u)$

Figure 2. Ego-centric network of user *u* and the sub-graphs of its in- and out-neighborhoods.

network as a directed graph, $\mathcal{G}_A$=($\mathcal{V}$,$\mathcal{E}$), where a node $u \in \mathcal{V}(\mathcal{G}_A)$ denotes a community member and a directed edge $(u,v) \in \mathcal{E}(\mathcal{G}_A)$ indicates that user *u* answered a question from user *v*. We use $\mathcal{N}^-(u)$ to denote the set of *in-neighbors* of *u* and $\mathcal{N}^+(u)$ the set of *out-neighbors* of *u* (see Figure 2). The in-neighbors of *u* are users who have responded to questions from *u*, and the out-neighbors are users whom *u* has responded to. More precisely, $\mathcal{N}^-(u)$= $\{w \in \mathcal{V}(\mathcal{G}_A)$: $(w,u) \in \mathcal{E}(\mathcal{G}_A)\}$ and $\mathcal{N}^+(u)$=$\{v \in \mathcal{V}(\mathcal{G}_A)$: $(u,v) \in \mathcal{E}(\mathcal{G}_A)\}$. Furthermore, we use $\mathrm{d}^-(\mathrm{u})$=$|\mathcal{N}^-(\mathrm{u})|$ and $\mathrm{d}^+(\mathrm{u})$=$|\mathcal{N}^+(\mathrm{u})|$ to denote the in-degree and out-degree of *u*, respectively. Later, in section 6.2, we use these notions to characterize the social engagement of individual users. We calculate separately the edge density for the in-neighbor and out-neighbor sub-graphs. More precisely, we compute the clustering coefficient of the in-neighborhood C^-(u) and out-neighborhood C^+(u):

$$C^-(u) = \frac{|\{(w_i,w_j)\}|}{d^-(u)\cdot(d^-(u)-1)} : w_i, w_j \in \mathcal{N}^-(u), (w_i,w_j) \in \mathcal{E}(\mathcal{G}_A) \quad (1)$$

$$C^+(u) = \frac{|\{(v_i,v_j)\}|}{d^+(u)\cdot(d^+(u)-1)} : v_i, v_j \in \mathcal{N}^+(u), (v_i,v_j) \in \mathcal{E}(\mathcal{G}_A) \quad (2)$$

3.3 Datasets

We collected and analyzed data from both Yahoo! Answers and MSN QnA services.

YA. The Yahoo! Answers dataset was obtained by seeding a crawler with pages linked to the top level categories. Each category page lists recent questions assigned to that category and its sub-categories. Over 95% of the content included in our dataset was created during a three month period, from March to May 2008. It comprises 309,599 questions, posted by 217,615 distinct users and 1,151,453 answers, provided by 195,869 distinct users. On average 72.5% (±17.1%) of users active in a specified day answered questions, 55.2% of users only answered questions and 14.9% of users both asked and answered questions. Our crawl does not include all the questions and all the answers posted within those three months. Thus, we are working with a partial representation of the *answer-to* social network (see network statistics in Table 1).

QNA. The MSN QnA dataset spans the first year of the service, starting with its release in September 2006. The complete dataset consists of 488,760 questions, 1,330,819 answers, and 901,752 comments. The questions were posted by 241,616 distinct users, while the answers and comments were contributed by 42,941 and 34,068 distinct users, respectively. On average, 45.5% (±16.2%) of users active in a day answered questions, 8.2% of users only answered questions and just 9.6% of users both asked and answered questions. In this case, the dataset is a full snapshot of the service and thus, the *answer-to* social network is complete (see Table 1).

Table 1. *Answer-to* network statistics: $\mathcal{T}_D$ no. of days present in the dataset, $\mathcal{V}$ no. of nodes (i.e. users), $\mathcal{V}_A$ and $\mathcal{V}_Q$ no. of nodes without incident and outgoing edges, respectively, $\mathcal{E}$ no. of edges and $\mathcal{E}_{Bi}$ no. of bidirectional edges.

Dataset	$\mathcal{T}_D$	$\mathcal{V}$	$\mathcal{V}_A$	$\mathcal{V}_Q$	$\mathcal{E}$	$\mathcal{E}_{Bi}$
YA	143	217,615	195,869	124,932	723,955	190,718
QNA	397	241,616	42,941	190,399	879,706	416,741

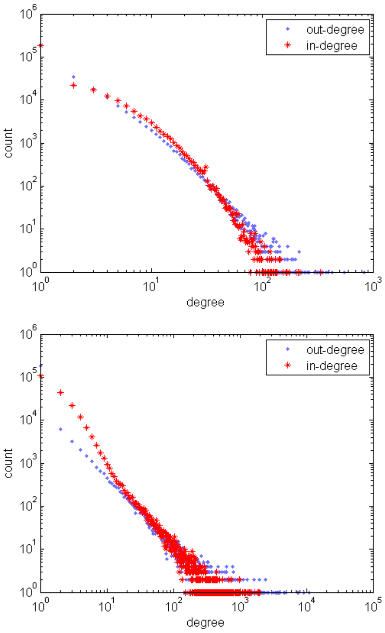


Figure 3. Distributions of in-degree and out-degree for the YA (top) and QNA (bottom) *answer-to* social network datasets.

Figure 3 shows in- and out-degree distributions for the *answer- to* social networks from both datasets. We see that most users provide answers to very few other users and, similarly, receive answers from very few users. This indicates a low level of involvement in answering questions.

4. TYPOLOGY OF QUESTIONS

A question posed by a CQA user reflects a specific intent. The user may ask for advice or wish to instigate a debate, or learn about other members of the community. We manually inspected over five thousand questions from Yahoo! Answers and MSN QnA. From the preliminary analysis it was apparent that some questions requested factual information, some sought advice, and others requested opinions (section 5 provides further details about the questions analyzed). These three broad types were also acknowledged in [1, 2] and used to develop questions for the study in [11]. However, it was also apparent that a fair number of questions, especially on MSN QnA, were posted to engage with other community users through informal conversations, as typically occurs in online forums and chat rooms. In this section, we describe the method we used to characterize question types and we summarize the main outcomes.

4.1 Method

In order to achieve a systematic characterization of question types, we used an *Undirected Inductive Coding* (UIC) method [16] to capture the intent behind each question. We performed a qualitative content analysis of questions sampled from both communities, defined the codes that described the intent, and developed a detailed typology[1] of intents that emerged organically from the data.

The UIC method involves identifying and assigning to each question a set of codes that characterize the question. New codes are generated every time a data item cannot be covered by the existing ones. Eventually the coding scheme stabilizes. The second stage involves the *reduction of codes* by identifying commonality and thus, grouping the codes and corresponding data examples. There are two advantages of this approach. First, we can evolve the number of dimensions by simply adding new code types and refining the taxonomy as needed. The overhead is in making sure that the previously processed questions are tagged with new codes. However, it is assumed that codes are generated exhaustively when processing the questions. Thus, the need for retrospective work is expected to be minimal. Second, since the coding is done at the *atomic level*, i.e., starting with basic concepts, we can flexibly define higher level categories by combining the appropriate codes. This is in contrast with the common approach where higher-level categories are defined first and the labeling of data needs to be repeated whenever a different perspective is taken.

4.2 Coding Reduction

There are many different perspectives that one can take when coding the content. Similarly, there are different ways in which one can synthesize the codes into higher level categories and arrive at the typology with reduced dimensions. We focus on the *users' objectives* and the *type of information request* conveyed by the question. For example, for the question "*Name all the presidents of the United States?*" we capture (1) the intent of satisfying the questioner's information need and (2) the particular type of content, i.e., obtaining an objective response, such as facts that are verifiable or commonly accepted as knowledge by the society. More specifically, we group codes along the following dimensions:

- ***Personal vs. General perspective***—the answer is expected to provide a personal perspective, experience, preference, or offer generally adopted truth and knowledge.
- ***Community vs. Individual issue***—the question is directed to the community on a community matter or to an individual, on a personal matter, such as habits, preferences, etc.
- ***Social vs. Non-social intent***—the objective of the question is to engage in conversion with the community as a whole or at a personal level, or is a request for information.

These groupings lead us to eight main question types on which we decided to focus our analysis of user intent (see also Table 2):

- ***Factual Information*** **(FI)**—the question is a request for factual information or a source of information. It may require more or less expertise to answer: "*Who was the last president of Zambia?*"

Table 2. Definition of question types along 3 dimensions: General vs. Personal perspective, Community vs. Individual issue, and Social vs. Non-social intent.

Question Type	General	Personal	Community	Individual	Social	Non-social
Factual Information (FI)	•		•	•		•
General Advice (GA)	•		•			•
Personal Advice (PA)		•	•	•		•
General Opinion (GO)	•		•	•	•	•
Personal Opinion (PO)		•	•	•	•	
Chatting (C)		•		•	•	
Entertainment (E)	•			•	•	
Other (O)						

- ***General Advice*** **(GA)**—request for advice or recommendation about general or personal issues, but provided from an objective stance, involving verifiable facts. The user intent is to solve a problem, make a decision or carry out an action: "*Can anyone tell me how to adjust the carb on a 49cc 2 stroke dirtbike?*"
- ***Personal Advice*** **(PA)**—request for advice that involves personal experience, preferences, and insights of the person who answers the question. The user intent is to gain support or help with a problem that does not have an objective solution and thus requires a personal perspective: "*Has anyone got any ideas for my little sisters 16th birthday* […] *extra special! Ideas for pressies and what to do*?"
- ***General Opinion*** **(GO)**—request of opinion about a general issue, a community stance, or similar. The user seeks opinions on a general matter and may phrase it as a hypothetical question, an open question, a poll about a general topic: "*Why does mankind keep recreating the same reality*?"
- ***Personal Opinion*** **(PO)**—request for opinion about a stance on a personal issue. The user seeks opinions which reflect the preference or personal experience of the answerer: "*Do you like LA?*" or "*What is your favorite baking scent?*"
- ***Chatting*** **(C)**—the question is a vehicle for the user to chat with other users, through light informal conversation with the community as a whole or at a personal level. It can be a request for personal experiences, a direct communication with specific users, or a way of sharing one's feelings about oneself, personal achievements or the moment: "*I'm eating a slice of a home-made pie. Anyone want some?*"
- ***Entertainment*** **(E)**—the user intent is to entertain the community by asking riddles, posting trivia or puzzles for the community to solve: "*(Besides 'War' and 'Bush') _______ is just getting out of hand.*"
- ***Other*** **(O)**—the question is not formulated as a question, e.g. it is nonsensical, an advertisement, etc.: "*Oops excuse the spelling. It was a fish with no tail.*"

[1] The question typology specification is available at: http://research.microsoft.com/pubs/79403/QCodes.pdf.

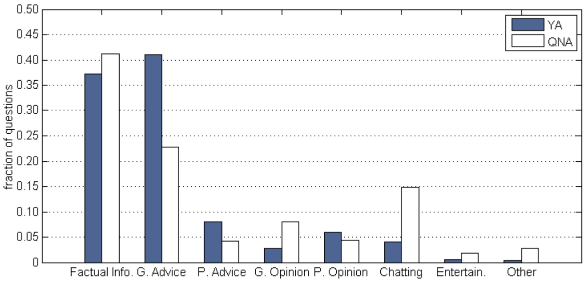


Figure 4. Distribution of question types across the samples of questions from YA and QNA.

5. ANALYSIS OF QUESTION TYPES

5.1.1 Question Samples

The question typology that we developed emerged from manual annotation of random samples of questions from both datasets (YA and QNA). The questions were coded and labeled by 2 annotators—the inter-annotator agreement was measured using the Kappa statistic [5], which indicated substantial agreement: κ=0.761. The labels that were assigned consist of the 8 main question types described in the previous section.

Figure 4 shows the overall distribution of questions per type after labeling 2000 questions from each dataset. We found that in both samples, *factual information* and *general advice* were the predominant question types. However, we also observed a considerable number of *chatting* questions in the QNA sample (~15%), indicating that users of this community often contribute content with *social intent*. Such questions are often personal or useless outside of their original context (e.g., time sensitive: "*How's the weather in your home town today?*"), which makes them poor contributions to the question answering knowledge base. Nonetheless, they strengthen the ties among core users who regularly visit the service and are likely to be central to establishing a sense of community. The challenge lies on striking the right balance between *social* and *non-social* interactions so that all the users can benefit from the service, including the majority of one-time users who come with a question and expect quick and quality responses.

Based on our analysis of the sample questions we raised the hypothesis that users' activity levels (i.e., number of content contributions) and social network ties might be reflected in the types of questions they ask and answer. To investigate this hypothesis, we ranked users based on the overall number of questions and answers contributed to the service, and selected the top 5 contributors from each dataset for further analysis, as we present next.

5.1.2 Top Contributors

The level of users' activity in CQA can be measured by the number of individual content contributions (i.e. questions, answers and comments posted). We focus this analysis on the top 5 contributors of questions and top 5 contributors of answers in each dataset. From the whole dataset we randomly sampled 50 questions and 50 answers (including the respective questions), from each of those users. For users who posted less than 50 questions we took all that were present in the dataset. This resulted in a total of 1,446 questions, 537 from YA and 909 from QNA. The questions were then manually labeled with respect to their type by the same annotators as before. In QNA, two of the top answerers were also among the top questioners, while on YA some of the top answerers did not post any questions within the time span of our sample. Table 3 and Table 4 present the distribution of manually assigned labels to: questions answered (A) and questions asked (Q) by the top answerers and the top questioners.

Table 3. Top 5 users from YA with the most answers (u_1 to u_5) and with the most questions (u_5 to u_{10})

Question Label		FI	GA	PA	GO	PO	C	E	O
Top Answerer									
User u_1	A	**20**	**29**	0	0	1	0	0	0
	Q	1	0	0	0	0	0	0	0
User u_2	A	**19**	**23**	0	5	3	0	0	0
	Q	-	-	-	-	-	-	-	-
User u_3	A	**26**	**12**	4	2	3	3	0	0
	Q	-	-	-	-	-	-	-	-
User u_4	A	**31**	9	3	2	4	1	0	0
	Q	-	-	-	-	-	-	-	-
User u_5	A	**11**	**16**	**18**	2	3	0	0	0
	Q	0	0	0	1	2	**0**	0	0
Top Questioner									
User u_6	A	-	-	-	-	-	-	-	-
	Q	**10**	**10**	3	**22**	4	0	0	0
User u_7	A	5	4	0	0	2	1	0	0
	Q	**11**	**20**	0	3	7	0	0	0
User u_8	A	-	-	-	-	-	-	-	-
	Q	**35**	6	0	0	0	0	0	0
User u_9	A	**11**	**28**	2	1	2	3	0	0
	Q	**12**	**9**	1	0	7	2	0	1
User u_{10}	**A**	1	0	0	0	0	0	0	0
	Q	1	**17**	0	2	6	0	0	0

Table 4. Top 5 users from QNA with the most answers (u_1 to u_5) and with the most questions (u_4 to u_8)

Question Label		FI	GA	PA	GO	PO	C	E	O
Top Answerer									
User u_1	A	**14**	1	2	5	3	**22**	3	0
	Q	4	5	4	3	**10**	**23**	1	0
User u_2	A	6	2	1	1	4	**33**	1	2
	Q	6	2	0	1	3	**36**	2	0
User u_3	A	7	1	4	2	8	**28**	0	0
	Q	0	3	1	1	3	**41**	1	0
Top Answerer = Top Questioner									
User u_4	A	**8**	2	4	1	7	**28**	0	0
	Q	0	1	0	0	0	**32**	**17**	0
User u_5	A	**14**	2	5	1	9	**18**	1	0
	Q	5	2	1	0	0	**40**	2	0
Top Questioner									
User u_6	A	4	5	4	3	**10**	**23**	1	0
	Q	1	1	2	2	9	**32**	3	0
User u_7	A	7	3	2	1	6	**28**	3	0
	Q	**12**	3	4	0	3	**28**	0	0
User u_8	A	**10**	4	1	2	3	**28**	2	0
	Q	9	1	1	3	5	**11**	**19**	0

Type of Content. We note that YA top contributors engage mostly in questions that seek *factual information* and *general advice*. For example, most of the contributions from YA user u_1 provide advice on digital cameras, with many of the answers being referrals to camera-related websites, not offering direct solutions to the problem as such. In contrast, the predominant question type in which the top QNA contributors engage in is *chatting*, indicating strong participation with a *social intent*. However, they

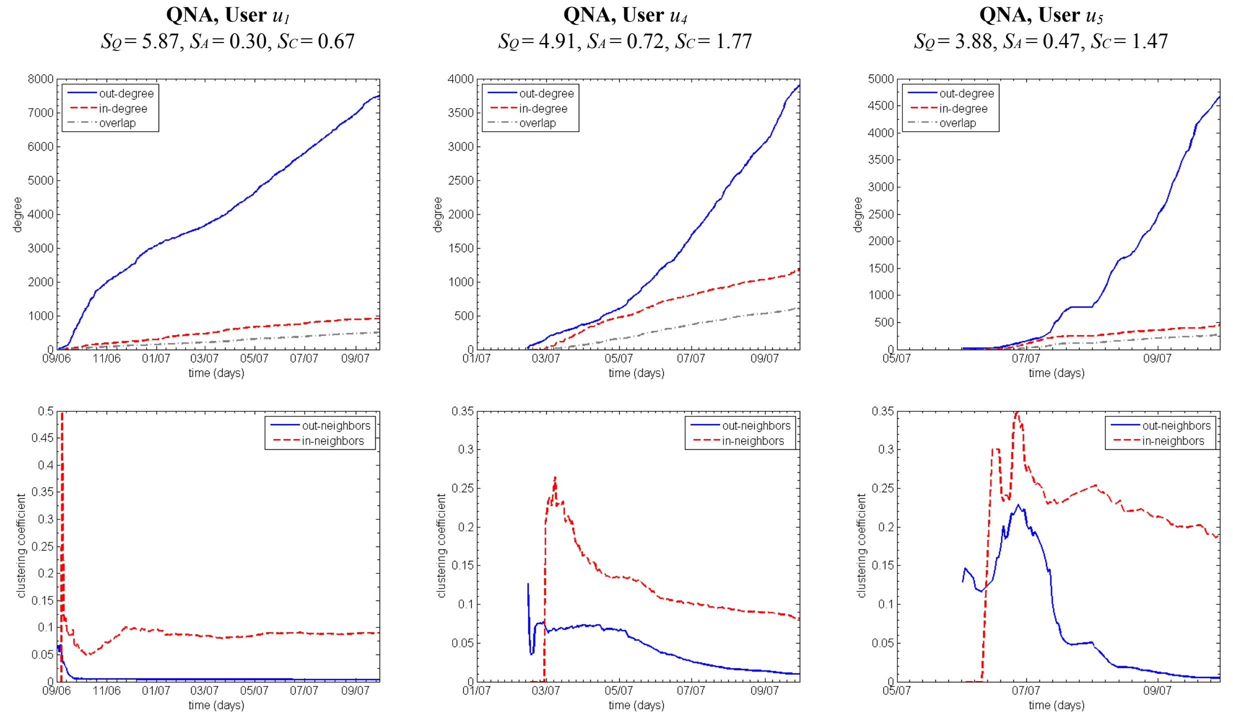


Figure 5. For 3 QNA top contributors the plots show the evolution over time of—Top: in-degree $d^-(u)$, out-degree $d^+(u)$, and overlap between in- and out-neighbors, $|\mathcal{N}^-(u)\cap\mathcal{N}^+(u)|$; Bottom: in- and out- clustering coefficient, $\mathcal{C}^-(u)$ and $\mathcal{C}^+(u)$.

also make contributions on other question types. Interestingly, some QNA users (e.g. users u_1, u_4 and u_5), primarily ask *chatting* questions but answer to various question types, including *factual information* ones. Thus, the QNA dataset presents rich data for gaining new insights into *social* vs. *non-social* interactions in CQA.

Social Engagement. Figure 5 shows the evolution over time of ego-centric network structure for QNA users u_1, u_4 and u_5, in terms of: (1) in-degree, out-degree, neighborhoods overlap, and (2) in- and out-clustering coefficients, for the full time span of our dataset ($\mathcal{T}_D$=397 days). In all cases we observe a high overlap of in- and out-neighbors, which indicates high reciprocity in question answering. For example, about half of the users who responded to user u_1 also received answers from u_1. Such high reciprocity indicates that users are acquainted with many of the users who respond to their questions. Furthermore, their acquaintances are likely to be connected among themselves, considering the relatively high values and stabilization of the in-clustering coefficients. In real networks the clustering coefficient of a node tends to decrease with its degree [9] – that is also the case here. However, we see periods when there was increase of $\mathcal{C}^-(u)$ or relatively steady densification of the in-neighborhoods, despite the increase of $d^-(u)$. For example, about one month after u_1 and u_5 joined the community there was an increase in their $\mathcal{C}^-(u)$, followed by steadiness or a slow decrease. $\mathcal{C}^-(u_4)$ was also relatively stable two months after user u_4 joined the community. Note that the initial sharp oscillations of the clustering coefficients are due to the low degree values – e.g. for a degree of 2, a single edge between the 2 neighbors results in a clustering coefficient $\mathcal{C}$ = 0.5; the arrival of a new neighbor can decrease it to $\mathcal{C}$ = 0.16.

In summary, the social network properties of users who predominantly ask *Chatting* questions indicate reciprocal ties and connectedness among acquaintances. These findings prompted us to analyze the whole community with respect to the types of questions posted to the CQA services. For that we first trained automatic classifiers for selected classes of questions using manually labeled data. We then used them to classify the full set of questions and analyze the social network structure for all the users in the QNA community. These experiments are described in the next section.

6. COMMUNITY ANALYSIS

CQA services can only thrive if individuals who join the community have their information needs satisfied and are willing to contribute by helping others to meet their needs. They cultivate and rely upon the users' sense of community. As suggested by the findings in the previous section, the users build social ties and it is important to understand how different question types affect the social network structure. To that effect we investigate the influence of questions with *social* vs. *non-social* intent (section 6.1) on the community social structure (section 6.2).

6.1 Classification of Questions

We took the manually labeled samples of questions from YA and QNA and grouped them based on the intent: questions labeled as FI, GA and PA types are part of the *non-social* training data set, and PO, C, and E are part of the *social* training set. Since the general opinion type (GO) included both *social* and *non-social* questions and accounts for a small percentage of questions in the training dataset, we did not attempt to train a classifier to predict such type of questions.

Table 5. Description of feature sets.

Feature	Description	F
Question Features		
Q_{TEXT}	Normalized tf.idf scores for single terms and n-grams	**1**
Q_{TLEN}	Length of question title in number of characters	**2**
Q_{DLEN}	Length of question description in number of characters	**2**
Q_{URL}	URL present in question (binary value:1=true, 0=false)	**2**
Thread Features		
TP_A	Number of answers in thread	**3**
TP_{AC}	Number of answers with comments	**3**
TP_C	Number of comments in thread	**3**
TP_{CUQ}	Number of comments by the user who asked the question	**3**
TP_{CUA}	Number of comments by users who answered question	**3**
TP_U	Number of users involved in the thread	**3**
TP_{UA}	Number of users who answered question	**3**
TP_{UC}	Number of users who commented on the answers	**3**
TP_{ALEN}	Average answer length in number of characters	**3**
TP_{ATIME}	Time elapsed until first answer was received (in sec.)	**3**
TP_{AURL}	URL present in answers (binary value: 1=true, 0=false)	**3**
Tag & Topic Features		
T_{TXT}	Tag / topic text	**4**
T_Q	Number of questions with tag	**5**
T_A	Average number of answers by tag	**5**
T_C	Average number of comments by tag	**5**
T_{QTLEN}	Average length of question title by tag	**5**
T_{QDLEN}	Average length of question description by tag	**5**
T_{ALEN}	Average answer length by tag	**5**
T_{CLEN}	Average comment length by tag	**5**
T_{TLEN}	Average thread length by tag	**5**
Social Network Features		
SN_d^-	User in-degree	**6**
SN_d^+	User out-degree	**6**
SN_{OV}	Overlap between in- and out-neighbors	**6**
SN_C^-	Clustering coefficient for in-neighbors	**6**
SN_C^+	Clustering coefficient for out-neighbors	**6**

Table 6. QNA classification results: *non-social* type

Feature Set	P		R		F1		BEP	
	avg	*std*	*avg*	*std*	*avg*	*std*	*avg*	*std*
F1 (baseline)	0.809	±0.028	**0.950**	**±0.012**	0.873	±0.016	0.861	±0.015
F1,F3	0.850	±0.021	0.898	±0.011	0.873	±0.010	0.865	±0.019
F1,F2,F3	0.809	±0.024	0.932	±0.016	0.866	±0.017	0.856	±0.014
$F4_{TAG}$	0.785	±0.014	0.931	±0.015	0.852	±0.009	0.814	±0.015
F5	0.781	±0.018	0.933	±0.014	0.850	±0.011	0.811	±0.014
$F4_{TOPIC}$	0.732	±0.019	0.983	±0.005	0.839	±0.011	0.733	±0.020
$F4_{TAG}$,F5	0.786	±0.017	0.926	±0.012	0.850	±0.011	0.810	±0.017
F1, …,F5	0.810	±0.013	0.930	±0.007	0.866	±0.008	0.866	±0.008
$F6_Q$	0.835	±0.032	0.865	±0.045	0.849	±0.034	0.848	±0.027
$F6_Q$-	0.768	±0.020	0.949	±0.018	0.849	±0.016	0.859	±0.023
$F6_Q$+	**0.895**	**±0.020**	0.785	±0.034	0.836	±0.022	0.851	±0.023
$F6_Q$,$F6_A$	0.789	±0.028	0.896	±0.026	0.839	±0.020	0.823	±0.027
F1,$F6_Q$,$F6_A$	0.795	±0.030	0.921	±0.013	0.853	±0.020	0.832	±0.026
F1,$F6_Q$	0.810	±0.022	0.896	±0.013	0.851	±0.012	0.852	±0.022
F1,$F6_A$	0.795	±0.030	0.921	±0.013	0.853	±0.020	0.832	±0.026
F1,F3,$F6_Q$	0.847	±0.030	0.894	±0.017	0.870	±0.019	0.865	±0.019
F1,F3,$F6_Q$+	0.856	±0.011	0.904	±0.022	**0.879**	**±0.010**	**0.871**	**±0.010**

Table 7. QNA classification results: *social* type

Feature Set	P		R		F1		BEP	
	avg	*std*	*avg*	*std*	*avg*	*std*	*avg*	*std*
F1 (baseline)	0.766	±0.033	0.313	±0.072	0.439	±0.077	0.612	±0.011
F1,F3	0.759	±0.057	0.259	±0.037	0.385	±0.043	0.598	±0.036
F1,F2,F3	0.755	±0.068	0.234	±0.049	0.353	±0.059	0.593	±0.055
$F4_{TAG}$	0.712	±0.059	0.306	±0.021	0.427	±0.017	0.474	±0.034
$F4_{TOPIC}$	**0.809**	**±0.091**	0.145	±0.031	0.244	±0.047	0.307	±0.037
F5	0.709	±0.081	0.311	±0.031	0.430	±0.029	0.493	±0.030
$F4_{TAG}$,F5	0.692	±0.072	0.309	±0.033	0.424	±0.026	0.474	±0.029
F1,…,F5	0.723	±0.087	0.330	±0.040	0.450	±0.039	0.598	±0.024
$F6_Q$	0.626	±0.073	0.251	±0.023	0.357	±0.033	0.544	±0.023
$F6_Q$-	0.621	±0.076	0.236	±0.022	0.342	±0.032	0.535	±0.022
$F6_Q$+	0.550	±0.047	0.334	±0.021	0.415	±0.027	0.534	±0.040
$F6_Q$,$F6_A$	0.648	±0.055	0.342	±0.054	0.446	±0.054	0.540	±0.015
F1,$F6_Q$,$F6_A$	0.670	±0.035	0.394	±0.049	0.494	±0.042	0.582	±0.024
F1,$F6_Q$	0.670	±0.035	0.394	±0.049	0.494	±0.042	0.582	±0.024
F1,$F6_A$	0.670	±0.035	0.356	±0.040	0.463	±0.033	0.578	±0.031
F1,F3,$F6_Q$	0.694	±0.043	**0.443**	**±0.039**	**0.539**	**±0.031**	0.618	±0.031
F1,F3,$F6_Q$+	0.723	±0.039	0.346	±0.033	0.468	±0.035	**0.618**	**±0.015**

6.1.1 Features

As part of the training process, we represented each question by a vector of representative features. We defined the features by considering various entities associated with CQA threads (see Figure 1): the question itself, information about user tags assigned to the question, characteristics of the answers and comments, properties of the users involved in the thread, and similar. The complete list of features we used in our classification experiments is presented in Table 5.

The first set of features refers to *Question* properties. We model the question text as a bag-of-words feature vector, and include information about its length and the presence of URLs in the text. We take a simple classifier with the Q_{TEXT} features as the baseline for other experiments. The second set of features describes the properties of the CQA thread. The *Thread Features* include the number of answers and comments given, the number of users involved in the thread, and similar.

We also include aggregate statistics about the usage of tags in the QNA dataset and the usage of topic labels in YA (see *Tag & Topic Features*). Building on previous work [13], that classified questions from QNA onto the Yahoo! Answers topic hierarchy, we consider both the QNA tags and the automatically assigned Yahoo! topics.

Finally, we take into account properties of the *answer-to* network that characterize interactions of a given user with the rest of the CQA community. Given that the network is directed and that a user may exhibit distinct questioning and answering behavior, we consider these two types of interactions separately. We use similar network metrics as in the previous section: in- and out-degree, overlap between in- and out-neighbors, and clustering coefficients of in- and out-neighborhoods. To capture the dynamic properties of the social network, we calculate these metrics at the time the interaction occurred. Consequently, the same user will have different scores depending on the timestamp of the question.

Table 8. Questions, answers and comments per question class.

Type of Post	Total	Posts per Question Thread	
Non-social questions	412,910	*avg.*(±*std*)	*median*
Answers	874,615	2.12 (±2.08)	1.00
Comments	396,552	0.96 (±1.97)	0.00
Social questions	31,653	*avg.* (±*std*)	*median*
Answers	218,918	6.92 (±4.33)	6.00
Comments	274,687	8.68 (±10.63)	6.00

Table 9. Results of the full dataset classification.

SVM Classifier	P	R	F1	Accuracy
non-social	**0.405**	0.608	0.486	0.711
social	**0.857**	0.545	0.667	0.591

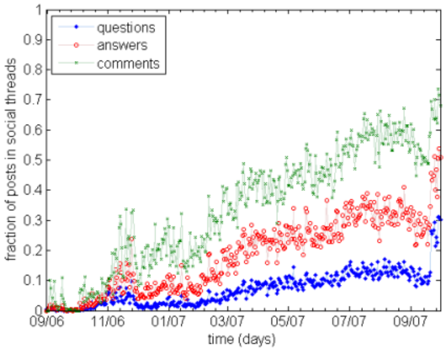


Figure 6. Ratio of social questions, answers and comments, over the total contributions of each type, over time.

6.1.2 *Automatic Classification*

We conducted a comprehensive set of experiments with linear SVM classifiers [6, 7] to investigate feature sets that are effective in predicting the two classes: *social* and *non*-social. We applied one vs. all approach for multi-class classification using SVM as the binary classifier. In order to account for imbalance between the positive and negative class in our data samples, we modified the SVM cost function to increase the penalty for misclassifying social questions. For each class, we ranked the classified questions based on the SVM score and calculated the break-even-point (BEP) for the ranked list (i.e., the rank at which the p*recision* and *recall* are equal). To assess the performance of the classifiers, we performed 5-fold cross-validation on the training dataset.

In Table 6 and Table 7 we detail the classification results with individual feature sets and several combinations of feature sets for the QNA dataset. We also performed classification experiments on the YA dataset but, as expected, the small number of representative questions of the *social* type was not sufficient to train a classifier for this class. The tables show *P, R, F1* calculated as *set* precision, recall, and F1 measures, and *BEP* scores determined from the ranked lists of classified questions. We note that the baseline performance of the classifier for the *non-social* class is quite high. This is partially due to the relatively small data sample and the imbalance in the number of questions in the *social* and *non-social* classes. Nevertheless, we can observe the relative contribution of different feature types and use that as a guide for further work in question classification.

Classification of non-social questions

- The use of content features in addition to thread properties (F1, F3), leads to improved performance over the baseline comprising bag-of-words (F1), with more precise classification (increase from 0.809 to 0.850) at the expense of recall.
- The use of simple social network features considering the out-neighborhood of the question author (F6Q+), leads to even more precise classification. That can be attributed to the fact that novice users whose social network is distinctive (e.g., sparser neighborhood connections), typically ask *non-social* questions.
- Combining network features with the content and thread features (F1, F3, F6Q+) resulted in the best BEP.

Classification of social questions

- The baseline classifier (F1) performs well in comparison with other types of features sets. However, better recall was achieved when combining content, thread and network features (F1, F3, F6Q+).
- It is interesting to observe that the classifier F4TOPIC that uses topic labels from Yahoo! categories for QNA questions causes the linear SVM classifier to make a significantly different trade-off between precision and recall. By the SVM determined threshold, the classifier leads to 0.809 precision in identifying socially intended questions, with much lower recall.

It is worth noting that the Yahoo! topics associated with the *social* questions in QNA span 93 categories, among which the most frequent are *Entertainment and Music*, *Games and Recreation*, *Education and References*, and *Polls and Surveys.* This insight may be helpful in understanding where social engagements may happen within Yahoo! Answers.

6.1.3 *Classification of the Full QNA Dataset*

We applied the binary SVM classifiers with the best performing feature set (F1, F3, $F6_Q$+) to the full set of questions from QNA comprising almost 0.5 million questions. This resulted in 84.5% questions assigned to the *non-social* class and 6.5% assigned to the *social* class*.* The remaining 9.0% of questions were not assigned to either class. Table 8 lists the total number of questions and respective answers and comments for assigned to each class.

A semi-supervised learning approach could have been used to leverage the large proportion of unlabelled data in our dataset and possibly yield increased classification accuracy. However, we do not expect that would significantly impact our analysis. We were able to perform a small scale validation of our results by using the manually labeled questions as test data. The classifiers performance is given in Table 9, showing high precision for questions of the social type.

Next, we present analysis of the social network structure considering the breakdown of the full set of QNA questions into the two classes.

6.2 Social Network Structure

Analysis of the classified questions reveals that 16.4% of the total answers and 30.5% of the total comments in the full dataset belong to *social* question threads. This is quite significant considering that such questions only represent 6.5% of the total in the dataset. Furthermore, questions assigned to the *social* class receive on average more answers and more comments than the

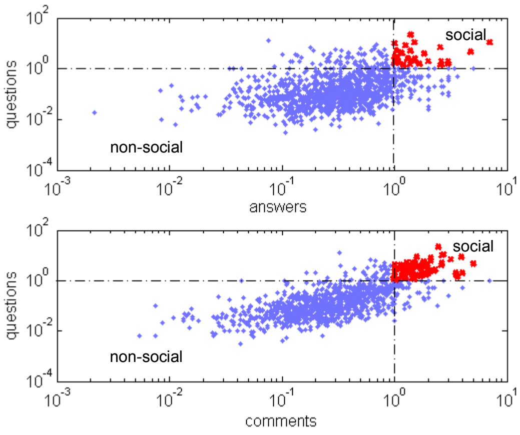
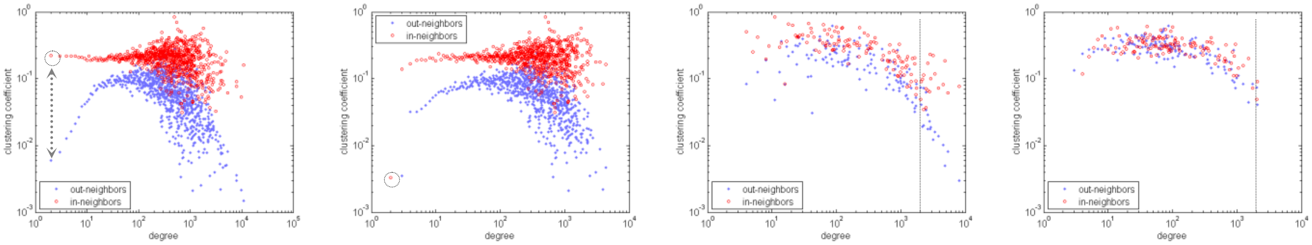


Figure 7. Scatter plot of the *social scores* of the whole QNA community (241,616 users). Top: S_A vs. S_Q; Bottom: S_C vs. S_Q.

Table 10. Average in- and out- degrees, neighbors overlap, and clustering coefficients for users with a given *social score*, for question (S_Q), answer (S_A) and comment (S_C) contributions.

Questions & Answers	$d^-(u)$	$d^+(u)$	$\lvert\mathcal{N}^-\cap\mathcal{N}^+\rvert$	$C^-(u)$	$C^+(u)$
S_Q≤1; S_A≤1 (*n*=21701)	19.4	29.6	2.1	0.162	0.046
S_Q≤1; S_A>1 (*n*=258)	39.5	32.0	5.6	**0.317**	0.298
S_Q>1; S_A≤1 (*n*=89)	**582.0**	**1637.2**	**196.0**	0.168	0.133
S_Q>1; S_A>1 (*n*=32)	335.4	312.2	90.6	0.274	**0.312**
Questions & Comments	$d^-(u)$	$d^+(u)$	$\lvert\mathcal{N}^-\cap\mathcal{N}^+\rvert$	$C^-(u)$	$C^+(u)$
S_Q≤1; S_C≤1 (*n*=11558)	33.1	52.0	3.7	0.196	0.067
S_Q≤1; S_C>1 (*n*=96)	74.1	149.6	18.9	**0.366**	**0.190**
S_Q>1; S_C≤1 (*n*=51)	**464.7**	**1722.5**	164.4	0.182	0.157
S_Q>1; S_C>1 (*n*=60)	641.3	1114.5	**199.0**	0.181	0.147



(a) **(b)** **(c)** **(d)**

Figure 8. Average clustering coefficients of in- and out-neighborhoods, $C^-(u)$ and $C^+(u)$, for users with a given degree for: (a) all users in the QNA dataset; (b) users with *social scores* $S_Q \le 1, S_A \le 1$ *and* $S_C \le 1$; (c) and users with S_Q>1, (d) users with S_A>1.

ones assigned to the *non-social* class. Over time, the fraction of social questions posted to the service has increased noticeably, as did the number of respective answers and comments (Figure 6). This indicates that the MSN QnA community ecosystem is evolving in a way that encourages interactions of a *social* nature.

In order to quantify the users level of social engagement we define a *social score*, $S(u)$, as the ratio of *social* vs. *non-social* content contributions to the CQA service. $S(u)$>1 indicates that user u predominantly contributes content with a *social* intent. We analyze the *social score* of the all the users in the community from questioning (S_Q), answering (S_A) and commenting (S_C) perspectives, i.e. considering each type of contribution (no. questions, no. answers and no. comments) separately. Figure 7 shows scatter plots of these *social scores* for the whole QNA community (i.e., the 241,616 users).

Points above the horizontal dotted lines correspond to users who predominantly ask *social* questions, and points to the right of the vertical dotted line correspond to users whose answers (top plot) or comments (bottom plot) were predominantly given to *social* questions. We observe that there is a minority of users, shown in red, who contribute *social* content (less than 100). However, among these users are some of the top contributors (see *social scores* of users shown in Figure 5).

Table 10 contains the number of users, n, on each quadrant, and the average in-degree, out-degree, neighborhood overlap, and the in- and out-clustering coefficients. We can see that users who generate more *social* questions than *non-social* ones (S_Q>1) have on average a fairly high in- and out-degree. We also note a relatively large overlap of in- and out-neighbors, which indicates that these users establish reciprocal ties with other users, who answered their questions at least once. Furthermore, we observe a large $C^-(u)$ for users who answered and commented primarily on *social* questions, and posted primarily *non-social* questions. In Figure 8, we take a closer look at the two clustering coefficients metrics, $C^-(u)$ and $C^+(u)$. In plot (a) we show the average clustering coefficients for users of a given degree, considering the whole community. We observe a significant difference in these two metrics for users of low degree (see distance between $C^-(u)$, in red, and $C^+(u)$, in blue). This is due to the fact that more than 82.2% of users post very few answers or even none, resulting in very low out-degree and clustering coefficient score close to zero, which brings the average down quite significantly.

In plots (b)-(d) we show the same two metrics for subsets of users, which we clustered based on the *social scores*. Plot (b) refers to users who primarily contribute *non-social* questions, answers, and comments. It is interesting to observe that the exclusion of users with high *social score* from this plot leads to a sharp decrease of the average $C^-(u)$ for users of degree equal to 2 (points on the left of the chart indicated with a circle). This implies that these users primarily receive responses from users who were removed from this view, i.e., users who engage in social interactions.

Plot (c) refers to users who predominantly ask questions with a *social* intent and plot (d) to users who primarily answer *social* questions. We observe that in (c) users with very high degree have on average a relatively high $C^-(u)$ and a much lower $C^+(u)$ (see points with degree >11K). Since users with such degrees do not appear in plot (d) this implies that the top contributors who ask *social* questions are also those who respond to many users who ask *non-social* questions. These findings suggest that the core participants of the MSN QnA community have social interactions with one another, while at the same time provide answers to *non-social* questions generated by the rest of the community.

Our findings suggest that the approach of combining the analysis of social network structure with the understanding of the user intent offers a promising framework for studying online knowledge sharing communities.

7. CONCLUSIONS

Modeling the entire social ecosystem of a CQA service or any other online community, is challenged by the scale, the complexity of the user interactions, and the dynamic nature of these services. Thus, it is essential to diversify the analysis and develop methods that enable us to probe into different aspects of such communities.

In this paper we present a method for analyzing the nature of the interactions among individuals in CQA communities, focusing on the user intent that is reflected in their questions. Our approach consists first of developing a typology of questions with respect to the user intent, through an iterative manual method. The typology can be used both for higher level analysis (e.g., for distinguishing between *social* and *non-social* questions) and for focusing on a very specific (lower-level) user intent. This represents a new contribution to the research community that can support future studies of similar knowledge-sharing communities.

We demonstrate the use this typology to investigate features that would be useful to identify socially intended questions in CQA. Using a fine level characterization of question types into factual information, general advice, personal advice, general opinion, personal opinion, chatting, and entertainment, we analyzed samples of question from Yahoo! Answers and MSN QnA, to gain insights about the community as a whole and its most active community members. We developed effective SVM classifiers to distinguish between *social* and *non-social* questions and use them to analyze the entire social network of the MSN QnA community. The same classifiers could be used in search applications to ensure that the service primarily surfaces the *non-social* content.

Second, we define new metrics, the *social scores* of individual users, to complement the social network metrics and analyze the community behavior. The social scores and the social network metrics consider questioning and answering activities separately, which enables us to investigate the presence of social ties among the most active users.

The answer to our research question: “Socializing or knowledge sharing?” in CQA communities is thus “both”. Indeed, the CQA users do engage in information sharing and they do socialize. Our methodology can be extended to include the level of users’ social engagement and can offer new insights that can be exploited by the online services in various ways. Examples include monitoring the community, devising incentive mechanisms to promote quality contributions, defining user reputation, or recommending content to individuals, so that users interested in socializing could easily find social questions and those who are topic experts could be guided to respond to information or advice seeking questions.

8. REFERENCES

[1] Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S., Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of* WWW '08, pp. 665-674, 2008.

[2] Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G., Finding high-quality content in social media. In *Proceedings of* WSDM '08, pp. 183-193, 2008

[3] Bian, J., Liu, Y., Agichtein, E. and Zha, H. Finding the right facts in the crowd: Factoid question answering over social media. In *Proceedings of* WWW '08, 2008.

[4] Bouguessa, M., Dumoulin, B., and Wang, S., Identifying authoritative actors in question-answering forums: the case of Yahoo! answers. In *Proceedings of* KDD '08, pp. 866-874, 2008.

[5] Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, vol. 20, no. 1. pp. 37-46, April 1960.

[6] Cortes, C. and Vapnik, V. Support vector networks. Machine Learning, vol. 20, pp. 273–297, 1995.

[7] *Text Garden*, Grobelnik, M. and Mladenic, D., available at: http://www.textmining.net/.

[8] Gyongyi, Z., Koutrika, G., Pedersen, J., Garcia-Molina, H., Questioning Yahoo! Answers. *First Workshop on Question Answering on the Web*, held at WWW '08, 2008.

[9] Leskovec, J., Backstrom, L., Kumar, R., and Tomkins, A. 2008. Microscopic evolution of social networks. In *Proceedings of* KDD '08, pp. 462-470. 2008.

[10] Liu, Y., Bian, J., and Agichtein, E. 2008. Predicting information seeker satisfaction in community question answering. In *Proceedings of* SIGIR '08, pp. 483-490, 2008.

[11] Harper, F. M., Raban, D., Rafaeli, S., and Konstan, J. A., Predictors of answer quality in online Q&A sites. *In Proceedings of* CHI '08, 2008.

[12] Jurczyk, P. and Agichtein, E., Discovering authorities in question answer communities by using link analysis. In *Proceedings of* CIKM '07, pp. 919-922, 2007.

[13] Mendes Rodrigues, E., Milic-Frayling, N., and Fortuna, B. Social Tagging Behaviour in Community-driven Question Answering. In *Proceedings of* IEEE/WIC/ACM WI 2008, 2008.

[14] Su, Q., Pavlov, D., Chow, J., and Baker, W. Internet scale collection of human-reviewed data. In *Proceedings of* WWW '07, 2007.

[15] Surdeanu, M., Ciaramita, M. and Zaragoza, H. Learning to Rank Answers on Large Online QA Collections. In *Proceedings of* ACL '08, 2008.

[16] Thomas, R. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, vol. 27, no. 2, pp. 237-246, 2006.

[17] Voorhees, E. M. 2001. The TREC question answering track. *Natural Language Engineering,* n. 7, v. 4, pp. 361-378, 2001.

[18] Welser, H.T., Gleave, E., Fisher, D., Smith, M. Visualizing the Signatures of Social Roles in Online Discussion Groups. Journal of Social Structure, no. 8, vol. 2.