Questioning Yahoo! Answers

Zoltán Gyöngyi
Stanford University
zoltan@cs.stanford.edu

Georgia Koutrika
Stanford University
koutrika@stanford.edu

Jan Pedersen
Yahoo! Inc., CA 94089, USA
jpederse@yahoo-inc.com

Hector Garcia-Molina
Stanford University
hector@cs.stanford.edu

ABSTRACT

Yahoo! Answers represents a new type of community portal that allows users to post questions and/or answer questions asked by other members of the community, already featuring a very large number of questions and several million users. Other recently launched services, like Microsoft's Live QnA and Amazon's Askville, follow the same basic interaction model. The popularity and the particular characteristics of this model call for a closer study that can help a deeper understanding of the entities involved, their interactions, and the implications of the model. Such understanding is a crucial step in social and algorithmic research that could yield improvements to various components of the service, for instance, personalizing the interaction with the system based on user interest. In this paper, we perform an analysis of 10 months worth of Yahoo! Answers data that provides insights into user behavior and impact as well as into various aspects of the service and its possible evolution.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous

General Terms

Human factors, experimentation

Keywords

question answering systems, user behavior

1. INTRODUCTION

The last couple of years marked a rapid increase in the number of users on the internet and the average time spent online, along with the proliferation and improvement of the underlying communication technologies. These changes enabled the emergence of extremely popular web sites that support user communities centered on shared content. Typical examples of such *community portals* include multimedia sharing sites, such as Flickr or YouTube, social networking sites, such as Facebook or MySpace, social bookmarking sites, such as del.icio.us or StumbleUpon, and collaboratively built knowledge bases, such as Wikipedia. Yahoo! Answers [12], which was launched internationally at the end of 2005, represents a new type of community portal. It is "the place to ask questions and get real answers from real people,"[1] that is, a service that allows users to post questions and/or answer questions asked by other members of the community. In addition, Yahoo! Answers facilitates the preservation and retrieval of answered questions aimed at building an online knowledge base. The site is meant to encompass any topic, from travel and dining out to science and mathematics to parenting. Typical questions include:

- "*If I send a CD using USPS media mail will the charge include a box or envelope to send the item?*"
- "*Has the PS3 lived up to the hype?*"
- "*Can anyone come up with a poem for me?*"

Despite its relative novelty, Yahoo! Answers already features more than 10 million questions and a community of several million users as of February 2007. In addition, other recently launched services, like Microsoft's Live QnA [7] and Amazon's Askville [2], seem to follow the same basic interaction model. The popularity and the particular characteristics of this model call for a closer study that can help a deeper understanding of the entities involved, such as users, questions, answers, their interactions, as well as the implications of this model. Such understanding is a crucial step in social and algorithmic research that could ultimately yield improvements to various components of the service, for instance, searching and ranking answered questions, measuring user reputation or personalizing the interaction with the system based on user interest.

In this paper, we perform an analysis of 10 months worth of Yahoo! Answers data focusing on the user base. We study several aspects of user behavior in a question answering system, such as activity levels, roles, interests, connectedness and reputation, and we discuss various aspects of the service and its possible evolution. We start with a historical background in question answering systems (Section 3) and proceed with a description (Section 4) and modeling (Section 5) of Yahoo! Answers. We provide our approach for measuring user reputation based on the HITS randomized algorithm (Section 6) and we present our experimental findings (Section 7). On the basis of these results and anecdotal evidence, we discuss possible research directions for the service's evolution (Section 8).

2. RELATED WORK

Whether it is by surfing the web, posting on blogs, searching in search engines, or participating in community portals,

WWW2008, April 21–25, 2008, Beijing, China.

[1] `http://help.yahoo.com/l/us/yahoo/answers/overview/overview-55778.html`

users leave their traces all over the web. Understanding user behavior is a crucial step in building more effective systems and this fact has motivated a large amount of user-centered research on different web-based systems [1, 3, 6, 9, 11, 10]. [1] presents a large scale study correlating the behaviors of Internet users on multiple systems. [10] examines user trails in web searches. A number of studies focus on user tagging behavior in social systems, such as del.icio.us [3, 6, 9, 11]. For instance, an experimental study of tag usage in My Web 2.0 shows that people naturally select some popular and generic tags to label Web objects of interest [11], while other studies identify factors that influence personal tagging behavior, such as *people's personal tendency* to apply tags based on their past tagging behaviors and *community influence* of the tagging behavior of other members [3, 6, 9].

In this paper, we study the user base of Yahoo! Answers. To the best of our knowledge, this is the first study of question answering systems of this type, and the first that analyzes many aspects of the user community, ranging from user activity and interests to user connectivity and reputation.

3. HISTORICAL BACKGROUND

Yahoo! Answers is certainly not the first or only way for users connected to computer networks to ask and answer questions. In this section, we provide a short overview of different communication technologies that determined the evolution of online question answering services and we discuss their major differences from Yahoo! Answers.

- **Email.** Arguably, the first and simplest way of asking a question through a computer network is by sending an email. This assumes that the asker knows exactly who to turn to for an answer, or that the addressee could forward the question over the right channel.
- **Bulletin board systems, newsgroups.** Before the web era, bulletin board systems (BBS) were among the most popular computer network applications, along email and file transfer. The largest BBS, Usenet, is a global, Internet-based, distributed discussion service developed in 1979 that allows users to post and read messages organized into threads belonging to some topical category (newsgroup). Clearly, a discussion thread could represent a question asked by a user and the answers or comments by others, so in this sense Usenet acts as a question answering service. However, it has a broader scope (such as sharing multimedia content) and lacks some features that enable a more efficient interaction, e.g., limited time window for answering or user ratings. Also, in recent years Usenet has lost ground to the Web. While it requires a certain level of technological expertise to configure a newsgroup reader, Yahoo! Answers benefits from a simple, web-based interface.
- **Web forums and discussion boards.** Online discussion boards are essentially web-based equivalents of newsgroups that emerged in the early days of the Web. Most of them focused on specific topics of interest, as opposed to Yahoo! Answers, which is general purpose. Often web forums feature a sophisticated system of judging content quality, for instance, indirectly by rating users based on the number of posts they made, or directly by accepting ratings of specific discussion threads. However, forums are not focused on question answering. A large fraction of the interactions starts with (unsolicited) information (including multimedia) sharing.
- **Online chat rooms.** Chat rooms enable users with shared interests to interact real-time. While a reasonable place to ask questions, one's chances of receiving the desired answer are limited by the real-time nature of the interaction: if none of the users capable of answering are online, the question remains unanswered. As typically the discussions are not logged, the communication gets lost and there is no knowledge base for future retrieval.
- **Web search.** Web search engines enable the identification and retrieval of web documents that might contain the information the user is looking for. The simplicity of the keyword-based query model and the amazing amount of information available on the Web helped search engines become extremely popular. Still, the simplicity of the querying model limits the spectrum of questions that can be expressed. Also, often the answer to a question might be a combination of several pieces of information lying in different locations, and thus cannot be retrieved easily through a single or small number of searches.

Category Name	Abbreviation
Arts & Humanities	Ar
Business & Finance	Bu
Cars & Transportation	Ca
Computers & Internet	Co
Consumer Electronics	Cn
Dining Out	Di
Education & Reference	Ed
Entertainment & Music	En
Food & Drink	Fo
Games & Recreation	Ga
Health & Beauty	He
Home & Garden	Ho
Local Businesses	Lo
Love & Romance	Lv
News & Events	Ne
Other	Ot
Pets	Pe
Politics & Government	Po
Pregnancy & Parenting	Pr
Science & Mathematics	Sc
Society & Culture	So
Sports	Sp
Travel	Tr
Yahoo! Products	Y!

Table 1: Top-level categories (TLC).

4. SYSTEM DESCRIPTION

The principal objects in Yahoo! Answers are:

- *Users* registered with the system;
- *Questions* asked by users with an information need;
- *Answers* provided by fellow users;
- *Opinions* expressed in the form of votes and comments.

One way of describing the service is by tracing the life cycle of a question. The description provided here is based on current service policies that may change in the future.

Users can register free of charge using their Yahoo! accounts. In order to post a question, a registered user has

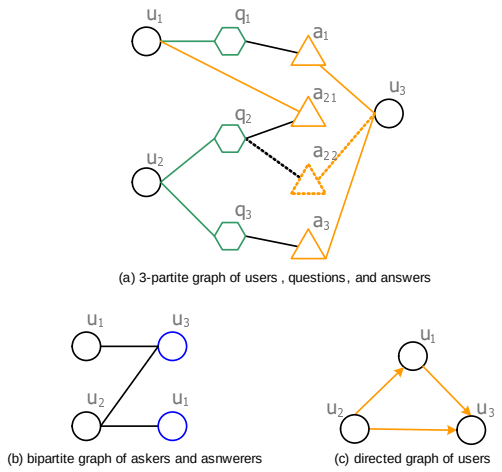

Figure 1: Example graphs for Y! Answers.

to provide a short question statement, an optional longer description, and has to select a category to which the question belongs (we discuss categories later in this section). A question may be answered over a 7-day *open* period. During this period, other registered users can post answers to the question by providing an answer statement and optional references (typically in the form of URLs). If no answers are submitted during the open period, the asker has the one-time option to extend this period by 7 additional days.

Once the question has been open for at least 24 hours and one or more answers are available, the asker may pick the best answer. Alternatively, the asker may put the answers up for a community vote any time after the initial 24 hours if at least 2 answers are available (the question becomes *undecided*). In lack of an earlier intervention by the asker, the answers to a question go to a vote automatically at the end of the open period. When a question expires with only one answer, the question still goes to a vote, but with "No Best Answer" as the second option. The voting period is 7 days and the best answer is chosen based on a simple majority rule. If there is a tie at the end, the voting period is extended by periods of 12 hours until the tie is broken. Questions that remain unanswered or end up with the majority vote on "No Best Answer" are deleted from the system.

When a best answer is selected, the question becomes *resolved*. Resolved questions are stored permanently in the Yahoo! Answers knowledge base. Users can voice their opinions about resolved questions in two ways: (i) they can submit a *comment* to a question-best answer pair, or (ii) they can cast a thumb-up or thumb-down vote on any question or answer. Comments and thumb-up/down counts are displayed along with resolved questions and their answers.

Users can retrieve open, undecided, and resolved questions by browsing or searching. Both browsing and searching can be focused on a particular category. To organize questions by their topic, the system features a shallow (2-3 levels deep) category hierarchy (provided by Yahoo!) that contains 728 nodes. The 24 top-level categories (TLC) are shown in Table 1, along with the name abbreviations used in the rest of the paper. We have introduced the category *Other* that contains dangling questions in our dataset. These dangling questions are due to inconsistences produced during earlier changes to the hierarchy structure. While we were able to manually map some of them into consistent categories, for some of them we could not identify an appropriate category.

Finally, Yahoo! Answers provides incentives for users based on a scoring scheme. Users accumulate *points* through their interactions with the system. Users with the most points make it into the "*leaderboard*," possibly gaining respect within the community. Scores capture different qualities of the user, combining a measure of authority with a measure of how active the individual is. For instance, providing a best answer is rewarded by 10 points, while logging into the site yields 1 point a day. Voting on an answer that becomes the best answer increases the voter's score by 1 point as well. Asking a question results in the deduction of 5 points; user scores cannot drop below zero. Therefore users are encouraged to not only ask questions but also provide answers or interact with the system in other ways. Upon registration, each user receives an initial credit of 100 points. The 10 currently highest ranked users have more than 150,000 points each.

5. MODEL

We can represent the interaction model in Yahoo! Answers as a graph, where nodes map to objects and edges to their interconnections in the system. There are several different graphs that can be defined, each one focusing on certain objects and their interactions in the system. Each of these graphs can possibly provide a different perspective on the system and allows studying different properties and phenomena in it. Next, we discuss some possible graphs that we have adopted for our analysis.

We consider an undirected tripartite graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. Nodes in $\mathcal{V}$ correspond to: (a) users, $\mathcal{U}$, (b) questions, $\mathcal{Q}$, and (c) answers, $\mathcal{A}$; hence, $\mathcal{V}$=$\mathcal{U} \cup \mathcal{Q} \cup \mathcal{A}$. Edges in $\mathcal{E}$ connect users to the questions they asked and to the answers they posted and questions to their answers, including a best answer.

Figure 1(a) presents a simple tripartite graph capturing the interactions of 3 users (circles) through the questions (hexagons) they pose and the corresponding best answers (triangles) and non-best answers (dashed triangle) they provide to questions of their fellow users.

In order to understand the user base of Yahoo! Answers, we consider different graphs capturing the connections between users asking and answering questions. There are several ways in which one could connect askers and answerers through the content they generate. Accordingly, there are different user-connection graphs that can be derived. We consider two different graphs described below.

We consider an undirected bipartite graph $\mathcal{G}'(\mathcal{V}', \mathcal{E}')$. Nodes in $\mathcal{V}'$ correspond to: (a) askers, $\mathcal{U}_1$, and (b) answerers, $\mathcal{U}_2$; hence, $\mathcal{V}$=$\mathcal{U}_1 \cup \mathcal{U}_2$. Each undirected edge in $\mathcal{E}'$ corresponds to the connection between a user posting a question and some other user providing an answer to that question. Edges are unweighted and could represent a single question-answer pair between the users, or several ones.

Figure 1(b) depicts the bipartite graph $\mathcal{G}'$ for the askers (black) and the answerers (blue) present in the example of Figure 1(a). To illustrate, the edge between user u_1 and u_3 indicates that there is at least one question asked by u_1 that got answered by u_3. Also, note how each user may appear in either or both of two distinct roles in the graph, independently as an asker and as an answerer. For instance, u_1 appears twice in the graph, both on the left and the right sides, acting as an asker in one case and as an answerer in the other case, respectively.

Finally, we focus on the most important relationship be-

tween users, i.e., that of asker-best answerer, and we consider the directed graph $\mathcal{G}''(\mathcal{V}'', \mathcal{E}'')$. Nodes in $\mathcal{V}''$ map to users. Each directed edge in $\mathcal{E}''$ from a user node u_i to node u_j indicates that user u_i received a best answer to a particular question by user u_j. Figure 1(c) depicts the directed graph $\mathcal{G}''$ for the users present in the example of Figure 1(a). For instance, the edge from user u_1 to u_2 indicates that u_2 has provided a best answer for u_1.

6. MEASURING REPUTATION

Community portals, such as Yahoo! Answers, are built on user-provided content. Naturally, some of the users will contribute more to the community than others. It is desirable to find ways of measuring a user's *reputation* within the community and identify the most important individuals. One way to measure reputation could be via the points accumulated by a user, on the assumption that a "good citizen" with many points would write higher quality questions and answers. However, as we will see in Section 7, points may not capture the quality of users, so it is important to consider other approaches to user reputation. For instance, possible measures may be the number of best answers, or the ratio between best answers and all answers.

Yet another option is to use reputation schemes developed for the Web, adapted to question answering systems. In particular, here we consider using a scheme like HITS [5]. The idea is to compute:

- *A Hub Score* (ρ) that captures whether a user asked many interest-generating questions that have best answers by knowledgeable answerers, and
- *An Authority Score* (α) that captures whether a user is a knowledgeable answerer, having many best answers to interest-generating questions.

Formally, we consider the directed graph $\mathcal{G}''$ indicating the connection between askers and best answerers (as depicted in Figure 1(c).) We then use the randomized HITS scores proposed in [8], and we compute the vectors α and ρ iteratively:

$$\alpha^{(k+1)} = \epsilon\,\bar{1} + (1-\epsilon)A_{\text{row}}^{T}\rho^{(k)} \text{ and}$$

$$\rho^{(k+1)} = \epsilon\,\bar{1} + (1-\epsilon)A_{\text{col}}\alpha^{(k+1)},$$

where A is the adjacency matrix (row or column normalized) and ϵ is a reset probability. These equations yield a stable ranking algorithm for which the segmentation of the graph does not represent a problem.

Once we obtain the α and ρ scores for users, we can also generate scores for questions and answers. For instance, the best answer score might be the α score of the answerer. The question score could be a linear combination of the asker's ρ score and the best answerer's α score,

$$c\rho_i + (1-c)\alpha_j,$$

These scores could then be used for ranking of search results.

7. EXPERIMENTS

Our data set consists of all resolved questions asked over a period of 10 months between August 2005 and May 2006. In what follows, we give an overview of the service's evolution in the 10 months' period described by our data set. Then, we dive into the user base to answer questions regarding user activity, roles, interests, interactions, and reputation.

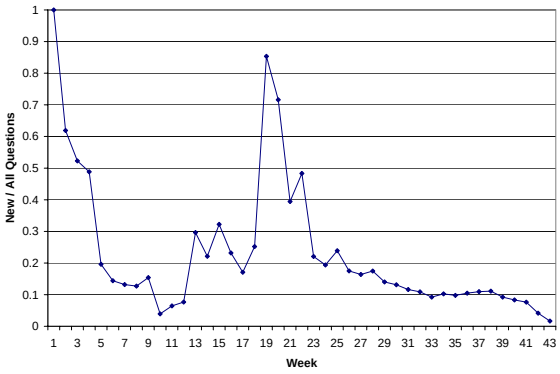

Figure 2: Ratio of new to all questions per week.

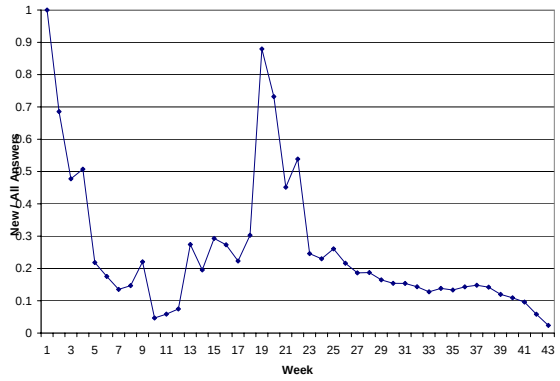

Figure 3: Ratio of new to all answers per week.

7.1 Service Evolution

Interestingly, the dataset contains information for the system while still in its testing phase and when it officially launched. Hence, it allows observing the evolution of the service through its early stages.

Table 2 shows several statistics about the size of the data set, broken down by month. The first column is the month in question and the second and fourth are the number of questions and answers posted that month. The third and fifth columns show the monthly increase in the total number of questions and answers, respectively. Finally, the last column shows the average number of answers per question for each month. We observe that while the system is in its testing phase, i.e., between August 2005 and November 2005, a small number of questions and answers is recorded. Then, in December 2005, the traffic level explodes soon after the service's official launch. In following months, the number of new questions and answers per month keeps increasing but the increase rate slows down. These trends are also visualized in Figures 2 and 3.

Figure 2 shows the ratio of new questions to all questions in the database on a weekly basis. Similarly, Figure 3 shows the ratio of new answers to all answers corresponding to the questions asked over these weeks. We observe the spikes between weeks 17 and 23 corresponding to a rapid adoption phase. Interestingly, about 50% of the questions and answers in the database were recorded during the 23^{rd} week. Then, the growth rate slowly decreases to $2-7\%$ after week 40. Despite the decreased growth rate, Table 2 shows that the average number of answers per question increased from 4 to 8.25 over the same period of 43 weeks. One explanation is

Month	Questions	Incr. Rate Q	Answers	Incr. Rate A	Average A / Q
August 2005	101	—	162	—	1.6
September 2005	93	92%	203	125%	2.18
October 2005	146	75%	256	70%	1.75
November 2005	881	259%	1,844	297%	2.09
December 2005	74,855	6130%	233,548	9475%	3.12
January 2006	157,604	207%	631,351	268%	4.00
February 2006	231,354	99%	1,058,889	122%	4.58
March 2006	308,168	66%	1,894,366	98%	6.14
April 2006	467,799	61%	3,468,376	91%	7.41
May 2006	449,458	36%	3,706,107	51%	8.25

Table 2: Number of resolved questions and answers per month.

that as more and more users join the system, each question is exposed to a larger number of potential answerers.

Overall, question answering community portals seem to be very popular among web users: in the case of Yahoo! Answers, users were attracted to the service as soon as it was launched, contributing increasingly more questions and answers. For Yahoo! Answers, there is also a hidden factor influencing its popularity: it is a free service in contrast to other systems, such as the obsolete Google Answers, that were based on a pay-per-answer model [4].

Altogether, 996,887 users interacted through the questions, answers, and votes in our data set. In what ways and to what extent each user contributes to the system? Do users tend to ask more questions than provide answers? Are there people whose role in the system is to provide answers? These questions are investigated in the following subsection.

7.2 User Behavior

First, we investigate the activity of each user through a series of figures. Figure 4 shows the distribution of the number of questions asked by a user. The horizontal axis corresponds to the number of questions on a logarithmic scale, while the vertical axis (also with logarithmic scale) gives the number of users who asked a given number of questions. For instance, 9077 users asked 6 questions over our 10 month period. A large number of users have asked a question only once, in many cases out of their initial curiosity about the service. The linear shape of the plot on a log-log scale indicated a *power law* distribution. In a similar fashion, Figure 5 shows the distribution of answers per user and the distribution of best answers per user. The horizontal and vertical axes are logarithmic and correspond to number of (best) answers and number of users, respectively. Again, the distributions obey a *power law*.

Consequently, askers can be divided into two broad classes: a small class of highly active ones, and a distinctively larger class of users with limited activity, who have contributed less than 10 questions. Similarly, answerers can be divided into very active and less active. In fact, these user behavior patterns are a commonplace in online community systems.

The above raise a number of possible questions: What is the role of a typical user in the system? For instance, could it be that a reserved asker is a keen answerer? What is the overlap between asker and answerer populations?

Figure 6 shows the percentage of users over the whole population that asks questions, answers, and votes. It also shows the percentage of users who perform two activities, e.g., they ask but also answer questions. We observe that

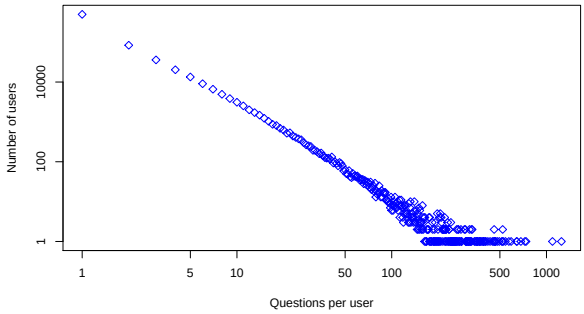

Figure 4: Distribution of questions per user.

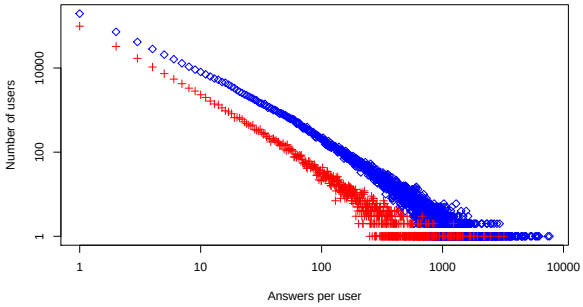

Figure 5: Distribution of answers (diamonds) and best answers (crosses) per user.

most users (over 70%) ask questions and only a small percentage of them helps the user community through answering (27%) or voting (18%). On the other hand, almost half of the population has answered at least one question while only about a quarter of the users cast votes.

Overall, we observe three interesting phenomena in user behavior. First, the majority of users are askers, a phenomenon partly explained by the role of the system: Yahoo! Answers is a place to "ask questions". Second, only a small fraction of these users gets involved in answering questions or providing feedback (votes). Hence, it seems that a large number of users are keen only on "receiving" from rather than "giving" to others. Third, there is a small portion of the user population that provides answers or votes.

The above observations trigger a further examination of questions, answers, and votes. Figure 7 shows the number of answers per questions as a log-log distribution plot. Note that the curve is super-linear: this shows that, in general, the number of answers given to questions is less than what would be the case with a power-law distribution. A possible

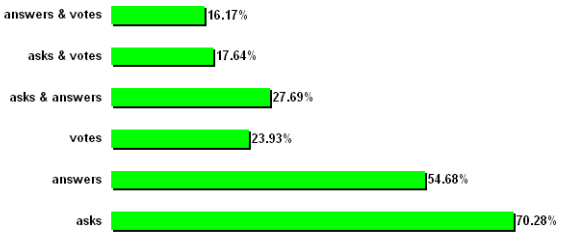

Figure 6: User behavior.

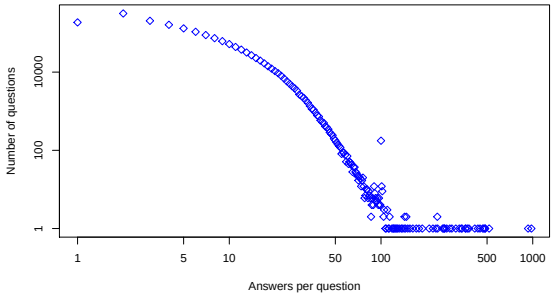

Figure 7: Distribution of answers per questions.

explanation of this phenomenon is that questions are open for answering only for a limited amount of time. Note the outlier at 100 answers—for several months, the system limited the maximum number of answers per question to 100. The figure reveals that there are some questions that have a huge number of answers. One explanation may be that many questions are meant to trigger discussions, encourage the users to express their opinions, etc.

Furthermore, additional statistics on votes show that out of the 1,690,459 questions 895,137 (53%) were resolved by community vote while for the remaining 795,322 (47%) questions the best answer was selected by the asker. The total number of 10,995,265 answers in the data set yielded an average answer per question ratio of 6.5.

The number of votes cast was 2,971,000. More than a quarter of the votes, 782,583 (26.34%), were self votes, that is, votes cast for an answer by the user who provided that answer. Accordingly, the choice of the best answer was influenced by self votes for 536,877 questions, that is 31.8% of all questions. In the extreme, 395,965 questions were resolved through a single vote, which was a self vote.

Thus, since votes may not be objective, in practice, a large fraction of best answers are of questionable value.

7.3 User Connectedness

In order to understand the user base of Yahoo! Answers, it is essential to know whether they interact with each other as part of one large community, or form smaller discussion groups with little or no overlap among them. To provide a measure of how *fragmented* the user base is, we look into the connections between users asking and answering questions.

First, we generated the undirected bipartite graph $\mathcal{G}'$ of askers and answerers (see Figure 1(b) for illustration). The graph contained 10,433,060 edges connecting 700,634 askers and 545,104 answerers. Each edge corresponded to the connection between a user posting a question and some other user providing an answer to that question. The described Yahoo! Answers graph contained a single large connected

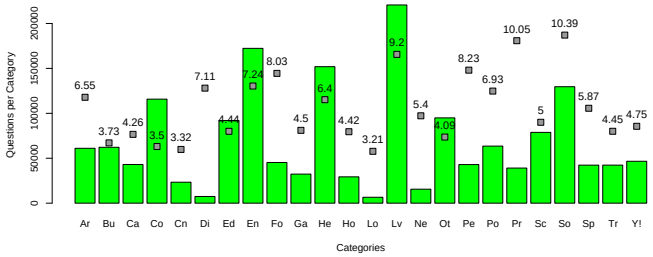

Figure 8: Questions and answer-to-question ratio per TLC.

component of 1,242,074 nodes and 1,697 small components of 2-3 nodes. This result indicates that most users are connected through some questions and answers, that is, starting from a given asker or answerer, it is possible to reach most other users following an alternating sequence of questions and answers. Second, we generated the directed graph $\mathcal{G}''$, which represents best-answer connections (see an example graph in Figure 1(c)). This graph contained the same number of nodes as the previous one and 1,624,937 edges. Still, the graph remained quite connected, with 843,633 nodes belonging to a single giant connected component and the rest participating in one of 30,925 small components of 2-3 nodes.

To test whether such strong connection persists even within more focused subsets of the data, we considered the questions and answers listed under some of the top-level categories. In particular, we examined *Arts & Humanities*, *Business & Finance*, *Cars & Transportation*, and *Local Businesses*. The corresponding undirected bipartite graphs had 385,243, 228,726, 180,713, and 20,755 edges, respectively. In all the cases, the connection patterns were similar as before. For instance, the *Business & Finance* graph contained one large connected component of 107,465 nodes and another 1,111 components of size 2-5. Interestingly, even for *Local Businesses* the graph was strongly connected, despite that one might expect fragmentation due to geographic locality: the large component had 15,544 nodes, while the other 914 components varied in size between 2 and 10.

Our connected-component experiments revealed the cohesion of the user base under different connection models. The vast majority of users is connected in a single large community while a very small fraction belongs to cliques of 2-3.

The above analysis gave us insights into user connections in the Yahoo! Answers's community but what are users interested in? Is a typical user always asking questions on the same topic? Is a typical answerer only posting answers to questions of a particular category, possibly being a domain expert? How diverse the users' interests are? The next subsection investigates these issues.

7.4 User Interests

One first thing to assess is whether user behavior is similar for all categories, or different topics correspond to varying asking and answering patterns.

In Figure 8, the horizontal axis corresponds to the various TLCs, while vertical bars indicate the number of questions asked in each particular category; the numbers above the gray squares stand for the average number of answers per question within the category. First, we observe that certain topics are much more popular than others: *Love & Romance*

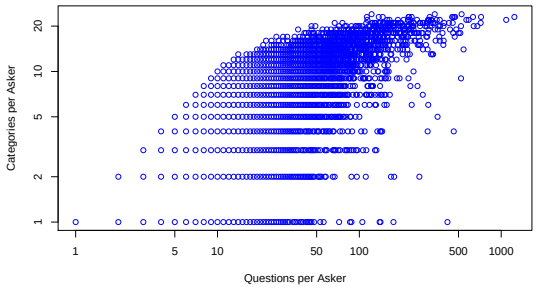

Figure 9: Questions vs. TLCs per asker.

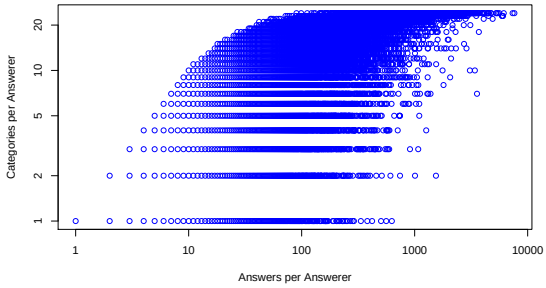

Figure 10: Answers vs. TLCs per answerer.

is the most discussed subject (220,598 questions and slightly over 2 million answers) with *Entertainment* following next (172,315 questions and around 1.25 million answers). The least popular topics are the ones that have a local nature: *Dining Out* (7462 questions and 53078 answers) and *Local Businesses* (6574 questions and 21075 answers). Interestingly, the latter two TLCs have the deepest subtree under them. The reason why topics of local interest might be less popular is that the corresponding expert answerer bases are probably sparser (a question about restaurants in Palo Alto, California, can only receive reasonable answers if there are enough other users from the same area who see the question). The perceived lack of expert answerers might discourage users even from asking their questions. On the other hand, there may be better places on the Web to obtain the pertinent information.

Focusing on the number of answers per question in each category, we observe how the average number of answers per question is the largest for the TLCs *Pregnancy & Parenting* and *Social Science*. The corresponding user base for pregnancy and parenting issues is expected to consist of adults spending significant time at home (in front of the computer), and willing to share their experience. The *Social Science* TLC really corresponds to an online forum where users conduct and participate in polls. A typical question under this category might be "Are egos spawned out of insecurity?". As it does not require any level of expertise for most people to express their opinions on such issues, this TLC seems to attract discussions, which yields a higher answer-to-question ratio. The category with the fewest answers per question is once again *Local Businesses*. Interestingly, the next to the last is *Consumer Electronics* with an answer-to-question ratio similar for the TLCs *Computers & Internet*, *Consumer Electronics*, and *Games & Recreation*, which are arguably related, even though the absolute numbers vary.

In conclusion, users of Yahoo! Answers are more interested in certain topics than others, and various issues attract different amounts of feedback (answers). General, lighter, topics, which interest a large population, such as Love & Romance, attract many inquiries. On the other hand, topics that do not require expert opinions, such as Social Science, or that can be answered based on personal experience, such as Pregnancy & Parenting, attract many answers.

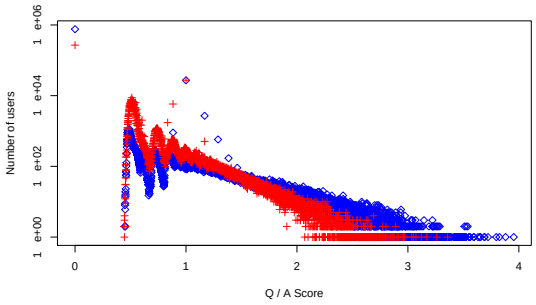

Figure 11: Randomized HITS authority (A, diamonds) and hub (Q, crosses) score distributions.

Moving down to the level of individual users, the average number of TLCs in which a user asks questions is 1.42. Figure 9 provides a distribution of TLCs per user as a function of the total number of questions asked by the user. The horizontal axis corresponds to the number of questions per user, while the vertical axis indicates the number of TLCs in which the questions were asked. The plot reveals that users cover the entire spectrum: 66% of askers (620,036) ask questions in a single TLC. This is expected as out of these, 80% (498,595) asked a single question ever. 29,343 users asked questions in two TLCs. Interestingly, there are 136 users who asked questions in 20 or more TLCs. The Pearson correlation coefficient between categories per asker and questions per asker is 0.666, indicating that there is little correspondence between curiosity and diversity of interest.

Figure 10 shows the distribution of TLCs per answerer as a function of the total number of answers provided by the user. The plot is similar to the previous one, covering the entire spectrum. Interestingly, users answer under more categories than ask. The average number of TLCs in which a user answers is 3.50, significantly higher than 1.42 for questions. The median is 2, so more than half of the answerers provide answers in two or more categories. Another interesting fact is that the number of users answering in most or all TLCs is significantly higher than for askers: 1,237 answerers cover all 24 TLCs, while 7,731 cover 20-23 TLCs. The Pearson correlation coefficient between categories per answerer and answers per answerer is 0.537, indicating an even weaker connection than for questions.

These results reveal that most answerers are not domain experts focusing on a particular topic, but rather individuals with diverse interests quite eager to explore questions of different nature.

Irrespective of their role, as askers and/or answerers, users of community-driven services, such as Yahoo! Answers, have some role in influencing the content in the system. In the following subsection, we discuss the particular features of the randomized HITS scores, and how they compare to other, possible measures of reputation discussed in Section 6.

7.5 User Reputation

In our computation of the HITS scores, we used a random

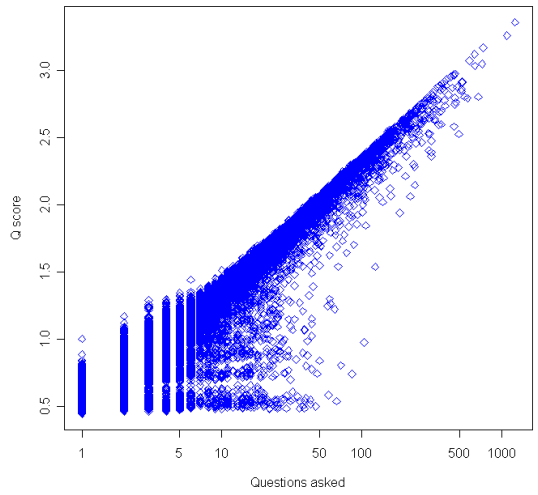

Figure 12: Correlation between hub (Q) score and number of questions asked.

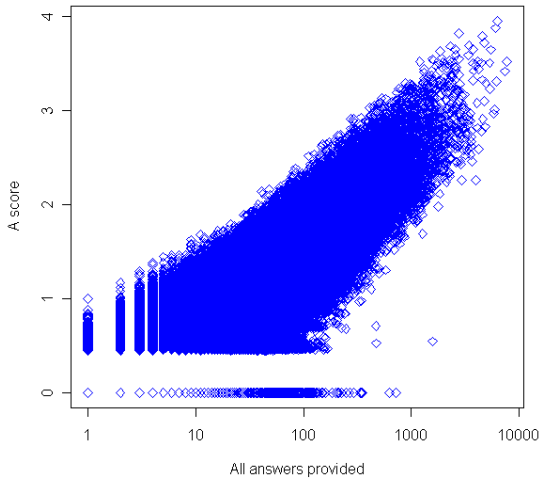

Figure 13: Correlation between authority (A) score and number of all answers provided.

jump of $\epsilon = 0.1$ and ran the computation for 100 iterations. We transformed the scores for legibility by taking the base-10 logarithm of the original numbers. Figure 11 shows the distribution of hub and authority scores. The horizontal axis shows the (transformed) hubs (light crosses) and authorities (dark diamonds) scores. The vertical axis corresponds to the number of users with particular hub or authority score. Note that the two points in the upper left corner of the graph correspond to the isolated users, who did not ask a single question, or never provided a best answer. The majority of the users have hub or authority scores below 2, while a small fraction obtained scores as high as 4.

The connection between the hub score of users and the number of questions they asked is shown in Figure 12. The horizontal axis corresponds to the number of questions asked by a user, while the vertical axis to the hub score of that user. Data points represent individual users with particular question counts and hub scores. Overall, the linear correlation is fairly strong (the Pearson correlation coefficient is 0.971), but the induced rankings are quite different (the Kendall τ distance between induced rankings is 0.632). However, notice the group of users with a relatively large number of questions but low hub score. For instance, 7,406 users asked more than 10 questions, but obtained only a hub score of at most 1.5. The questions of these users did not attract attention from authoritative answerers.

Similarly, Figures 13 and 14 show the correlation between the number of answers given by a user (and best answers, respectively) and the authority score of the user. When it comes to the correlation between all answers and authority score, the Pearson correlation coefficient is 0.727, while the Kendall τ measure is only 0.527. This indicates that when it comes to ranking users based on the number of answers given and their authority, the corresponding orderings are quite different – the users giving the most answers are not necessarily the most knowledgeable ones, as measured by the authority score. When it comes to the correlation between best answers given and authority score, the Pearson correlation coefficient is 0.99, but the Kendall τ only 0.698. Again, as in the case with questions and hub scores, users who provide the most best answers are not necessarily the most authoritative ones.

Overall, using HITS scores, we capture two significant intuitions: (a) quality is more important than quantity, e.g., users giving the most answers are not necessarily the most knowledgeable ones; and (b) the community reinforces user reputation, e.g., askers that did not attract attention from authoritative answerers do not have high hub scores.

One interesting question is whether the current point system, used for ranking the users in Yahoo! Answers, matches the notion of authority captured by our HITS scores. Figure 15 shows that there is a certain correlation between authority (vertical axis) and points (horizontal axis) for the majority of users. However, we can identify two user groups where the connection disappears. First, there are some users with zero authority, but with a large number of accumulated points. These users did not ever provide a best answer, though they were active, presumably providing many non-best answers. The lack of best answers within this group indicates the low quality of their contribution. In this respect, it would be inappropriate to conclude that a user with many points is one providing good answers. Second, there are many users with a decent authority score (between 0.5 and 1.5), but with very few points (around or less than 100). Remember that when users sign up with the system, they receive 100 points; they can only lose points by asking questions. Thus, our second group corresponds to users who occasionally provide best answers to questions, but they also ask a disproportionately large number of questions. Accordingly, their reputation as a good answerer is overshadowed by their curiosity. Hence, it would be hasty to assume that users with few points contribute content of questionable quality.

Finally, we wish to determine whether authority scores can be used to identify experts who consistently produce only high quality answers. For this, we consider the ratio of best answers to all answers for each user. Intuitively, the closer this ratio is to 1, the higher the overall quality of the user's answers is. Figure 16 presents results for all users, while Figure 17 focuses on users with at least 10 an-

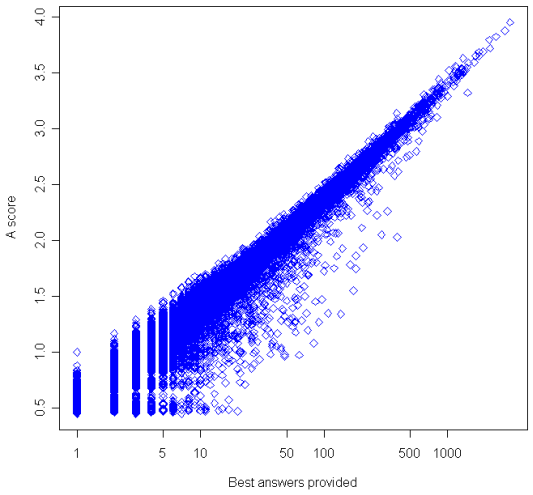

Figure 14: Correlation between authority (A) score and number of best answers provided.

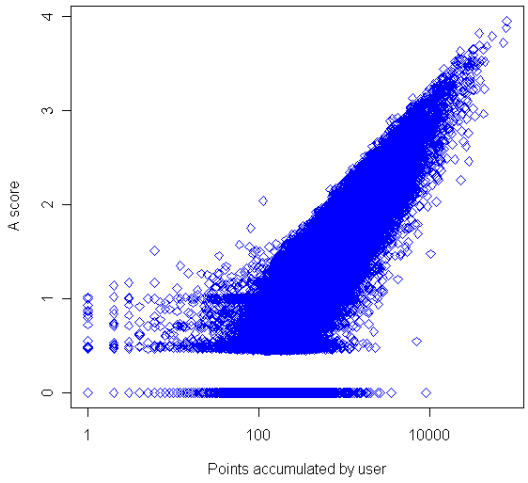

Figure 15: Correlation between authority (A) score and points accumulated by user.

swers. The horizontal axis of the figures corresponds to the ratio between best answers and all answers; the vertical axis shows the authority score. One may notice that there is no immediately recognizable pattern in the plots. In fact, the authority score is not a good indicator of the ratio: the authority score of users who only provided best answers can be as little as 0.5 and as large as 2.2. This variation is expected, because the authority score depends on the hub scores of the people who asked the corresponding questions. Accordingly, authority scores represent more than an intrinsic measure of how good the answerer is; they also capture the importance of the questions answered. Similarly, some users have a best-to-all ratio close to zero, but still achieve a high authority score: most of their answers may be of lower quality, as long as some are selected best.

8. DISCUSSION AND CONCLUSIONS

In this paper, we have presented our findings from an initial analysis of 10 months worth of Yahoo! Answers data shedding light on several aspects of its user community, such as user interests and impact. Here, we complement these results with anecdotal evidence that we have collected during our analysis. In particular, we have discovered three fundamental ways in which the system is being used.

(1) Some users ask *focused questions*, triggered by concrete information needs, such as,

"*If I send a CD using USPS media mail will the charge include a box or envelope to send the item?*",

The askers of such questions typically expect similarly focused, factual, answers, hopefully given by an expert. Arguably, the World Wide Web contains the answers to most of the focused questions in one way or another. Still, Yahoo! users may prefer to use this system over web search engines. A reason may be that the answer is not available in a single place and would be tedious to collect the bits and pieces from different sources, or that it is hard to phrase the question as a keyword-based search query. It may also be the case that the asker prefers direct human feedback, is interested in multiple opinions, etc.

It is quite possible that supporting such focused question answering was the original intention behind launching Yahoo! Answers. In practice, focused questions might not always receive an adequate answer, possibly because they never get exposed to a competent answerer (no such user exists in the system, or the question does not get routed to him/her). It is also common that some of the answers are low quality, uninformative, unrelated, or qualify more as comments than answers (e.g., a comment disguised as an answer might be "Who cares? You shouldn't waste your time and that of others.").

(2) Many questions are meant to *trigger discussions*, encourage the users to express their opinions, etc. The expected answers are subjective. Thus, in a traditional sense, there is no bad or good (or best) answer to such "questions." Consider, for example,

"*What do you think of Red Hot Chili Peppers?*"

At best, one may claim that the collection of all answers provided represent what the questioner wanted to find out through initiating a poll. In case of questions with subjective answers, it is quite common that a particular posted answer is not addressing the original question, but rather acts as a comment to a previous answer. Accordingly, discussions emerge. Note, however, that Yahoo! Answers was not designed with discussions in mind, and so it is an imperfect medium for this purpose. Users cannot answer their own questions, thus cannot participate in a discussion. Similarly, a user may only post at most one answer to a question. Yahoo! Answers has no support for threads that would be essential for discussions to diverge.

(3) Much of the interaction on Yahoo! Answers is just *noise*. People post random thoughts as questions, perhaps requests for instant messaging. With respect to the latter, the medium is once again inappropriate because it does not support real-time interactions.

This evidence and our results point to the following conclusion: *A question answering system, such as Yahoo! An-*

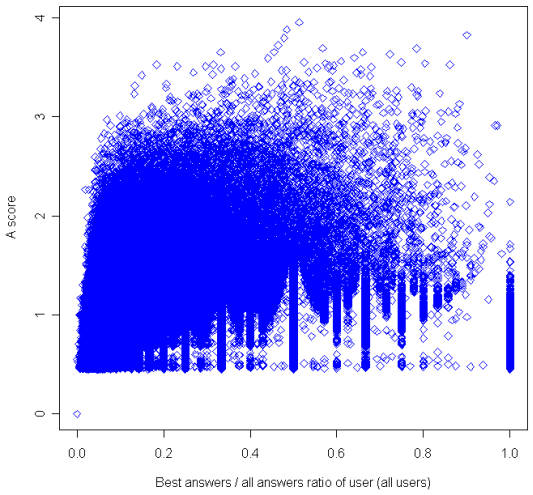

Figure 16: Correlation between authority (A) score and best answer per all answers ratio by user.

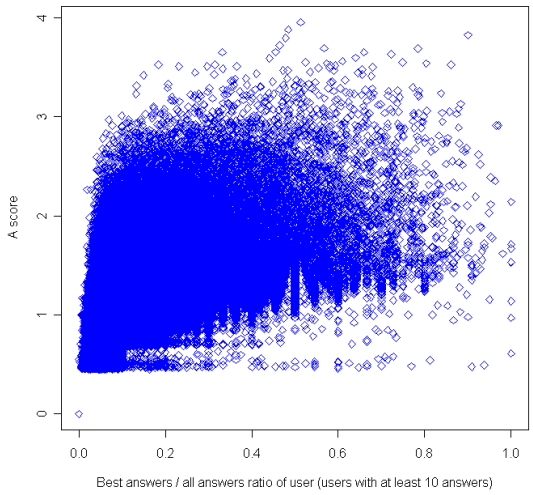

Figure 17: Correlation between authority (A) score and best answer per all answers ratio by user for users who provided at least 10 answers.

swers, needs appropriate mechanisms and strategies that support and improve the question-answering process per se. Ideally, it should facilitate users finding answers to their information needs and experts providing answers. That means easily finding an answer that is already in the system as an answer to a similar question, and making it possible for users to ask (focused) questions and get (useful) answers, minimizing noise. These requirements point to some interesting research directions.

For instance, we have seen that users are typically interested in very few categories (Section 7.4.) Taking this fact into account, personalization mechanisms could help route useful answers to an asker and interesting questions to a potential answerer. For instance, a user could be notified about new questions that are pertinent to his/her interests and expertise. This could minimize the percentage of the questions that are poorly or not answered at all. Furthermore, we have seen that we can build user reputation schemes that can capture a user's impact and significance in the system (Section 7.5.) Such schemes could help provide the right incentives to users in order to be fruitfully and meaningfully active reducing noise and low-quality questions and answers. Also, searching and ranking answered questions based on reputation/quality or user interests would help finding more easily answers and avoiding posting similar questions.

Overall, understanding the user base in a community-driven service is important both from a social and a research perspective. Our analysis revealed many interesting issues and phenomena. One could go even deeper into studying the user population in such a system. Interesting questions, which could be answered based on graph analysis, include: what is the nature of user connections, whether user-connection patterns are uniform, or there are certain *central* users representing connectivity hubs. Along similar lines, one could study various hypotheses. For example, one could hypothesize that most interaction between users still happen in (small) groups and that the sense of strong connectivity could be triggered by users participating in several different groups at the same time, creating thus points of overlap. This theory could be investigated with graph clustering algorithms.

9. REFERENCES

[1] E. Adar, D. Weld, B. Bershad, and S. Gribble. Why we search: Visualizing and predicting user behavior. In *Proc. of the* 16^{th} *Int'l WWW Conf.*, 2007.

[2] Amazon Askville. http://askville.amazon.com/.

[3] S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. of Information Science*, 32(2):198–208, 2006.

[4] Google Operating System. The failure of Google Answers. *http://google system.blogspot.com/2006/11/failure-of-google-answers.html.*

[5] J. Kleinberg. Authoritative sources in a hyperlinked environment. *J. of the ACM*, 46(5):604–632, 1999.

[6] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Position paper, tagging, taxonomy, flickr, article, toread. In *Proc. of the HT Conf.*, pages 31–40, 2006.

[7] Microsoft Live QnA. http://qna.live.com/.

[8] A. Ng, A. Zheng, and M. Jordan. Stable algorithms for link analysis. In *Proc. of the* 24^{th} *Annual Int'l ACM SIGIR Conf.*, 2001.

[9] S. Sen, S. Lam, A. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. Tagging, communities, vocabulary, evolution. In *Proc. of the CSCW*, 2006.

[10] R. White and S. Drucker. Investigating behavioral variability in web search. In *Proc. of the* 16^{th} *Int'l WWW Conf.*, 2007.

[11] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *Proc. of the Collaborative Web Tagging Workshop in conj. with the* 15^{th} *WWW Conf.*, 2006.

[12] Yahoo Answers. http://answers.yahoo.com/.