Experiments in Community Q&A

Introduction

Community question-and-answer (CQA) sites such as Stack Overflow or Yahoo! Answers represent a fundamentally compelling mix of information and conversation; of help-seeking, help-providing, and self-expression; of content space and social space, of achievement-seeking and altruism; of meeting ephemeral, immediate needs and creating a resource for posterity. They also represent an interesting point in the design space of social information spaces. Their primary structure is the asking of questions and the provision of answers. Hence they are a venue where people seek help and provide help, only incidentally providing a great information resource. They occupy a middle ground between collective authoring sites such as Wikipedia (where contributors explicitly build a resource for others and for posterity), and conversation venues such as Usenet News and the myriad discussion forums, where contributors are primarily engaged in dialog, and only incidentally may produce something that has lasting value. We have been studying CQA sites since 2006, and in the process have learned a great deal about the potential of these sites for helping people (Harper 2008), the variety of types of questions posed and ways to identify them (Harper 2009, 2010) the role of different users (notably so-called experts and super-users) and ways to identify and nurture them (Pal 2011, 2012a, 2012b; Farzan 2012), and more. This proposal is motivated by the questions we could not answer. Questions we could not answer because our prior research was grounded in the study of CQA systems from the outside. Mostly our work resulted from analysis of data visible at the system interface, or released to us by the CQA operator. Specifically, we studied Yahoo! Answers and StackOverflow through their web interfaces and released datasets, and we worked with Intuit Corp. to study TurboTax Online Community with a mix of public and private data. At times we conducted experiments by posting content. At times we were given access to internal data. And at times we even contacted users directly to survey them or invite them to experiments. But the core structure of the CQA sites was beyond our reach.

The frustration over the questions we could not answer led us to build our own CQA sites. We first built a small prototype site—MovieLens Q&A—which we have operated successfully for nearly three years. But MovieLens Q&A is small and narrowly focused. The experience led us to design and prototype GopherAnswers, a CQA site designed to support the nearly 100,000 members of the University of Minnesota community. Over the course of the past year we have built a prototype site and also built the needed relationships to integrate the site within the university. Details are described below, but the current status of the site is such that we could launch it to the community within 6-8 weeks. We plan to launch it as the 2013-2014 academic year starts—a time that would provide the greatest opportunity to promote it to users and that would coincide with the commencement of this proposed project. By building our own site, we open up the power of online field experiments (Konstan 2007) -- the ability to provide different experiences to different users and compare their resulting behaviors. And control of our own site opens up the ability to monitor factors normally hidden in public sites, including both building deep logging into the system to capture user browsing and drafting patterns and linking user data to study the role of external social connections to behavior inside the CQA site.

We present examples of the six types of questions posed in these sites (Harper 2010), with examples drawn from the expected domain of GopherAnswers:

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Description of Type</th>
<th>Example Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual</td>
<td>Seeks an objective answer that! emp_</td>
<td>Q1. How late is the bowling alley open at the student union?</td>
</tr>
<tr>
<td>Prescriptive</td>
<td>Seeks known steps or strategies to solve the questioner’s problem.</td>
<td>Q2. My roommate has season tickets for football and is going out of town this weekend. The tickets are loaded onto his ID. How can I use his ticket?</td>
</tr>
</tbody>
</table>
Experiments in Community Q&A

<table>
<thead>
<tr>
<th>Quality</th>
<th>Seeks to weigh the relative merits of alternatives or to find “best” and “worst” examples.</th>
<th>Q3. I have to take Calculus 2 next semester, preferably on MWF mornings. Who’s the best instructor to get?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Dis)Approval</td>
<td>Seeks to get subjective opinions on favorites, likes, dislikes, etc.</td>
<td>Q4. If I could date anyone on our football team, it would be Marqueis Gray. How about you?</td>
</tr>
<tr>
<td>Identification</td>
<td>Seeks to start discussion, build relationships, self-expression.</td>
<td>Q5. Hey folks, what do you think of the U’s new tuition model?</td>
</tr>
<tr>
<td>Advice</td>
<td>Seeks new or tailored solutions or approaches to addressing a question and inform future action.</td>
<td>Q6. What should I do? I’m a first-year EE student and at mid-term I’m getting Cs in both Physics and Calc. Should I drop them? Forget EE? What?</td>
</tr>
</tbody>
</table>

In our prior work we have observed and tracked these question as they occur on existing CQA sites, but we have not been able to actively or dynamically intervene in sites’ management of these question types.

This proposal at its core is both curiosity- and opportunity-driven. We do not claim that the diverse set of questions we propose to investigate fit a simple coherent structure. Rather, they emerge from an interdisciplinary collaboration between social computing computer scientists and specialists on online communication. And they emerge from the questions each of us felt we needed to ask, but could not answer without the power of experimental control.

In this proposal, we detail five specific research questions:

RQ1. What kinds of front-page recommendations most improve community outcome?

   This question addresses our longest-standing frustration. As experts in recommender systems, we have a long history of guiding users towards content—content they would find interesting or content that would help them serve the community. In this question, we apply a variety of computational techniques to explore presenting different mixes of questions based on users’ past selections, actions, explicit preferences, and the current state of open questions.

RQ2. How should intelligent task routing be deployed to manage workload distribution?

   This question addresses a challenge we discovered in quantitative and qualitative observation: too often CQA sites have a handful of “super users” who end up doing the vast majority of the work. Many tasks, including question-answering, could be performed by less active users. A variety of activities - including easy ones like voting or tagging - can deepen commitment to the site, helping users to develop into more reliable resources for the community. There is currently no infrastructure for engaging these users, or for triaging which contributions should be made by whom.

RQ3. How should referrals function within CQA?

   For many questions there is someone who really does know the answer, perhaps even someone officially charged with answering. That person may not spend time monitoring the site. How can we make referrals (both within the site and to outside) a meaningful contribution so that users are motivated to help each other get the best answers, even if that involves reaching out to a third party?

RQ4. How are user profiles used in CQA, and how can they help the community?

   Many CQA systems maintain profiles of their contributors, whether simple aggregates of activity or more elaborate information about qualifications and interests. Whether and how these profiles are used is not visible through public interfaces. We look forward to experimenting to learn how profiles can influence trust, behavior, and perceptions.

RQ5. To what extent can emphasizing altruism improve participation?

   Our studies of top contributors found that being helpful—altruism—was a top motivator for contributing effort to CQA. By contrast nearly all reward systems and communications emphasize
Experiments in Community Q&A

correctness, quantity, and competition over basic helpfulness. We will experiment with varying rhetoric and reward bases.

This set of questions is not exhaustive. A major reason for operating a live system is to figure out the right questions to ask. Our work plan includes spots for pure exploratory study—for example, studying data from the start-up phase of the community—and for addressing questions we have not yet learned enough about to formalize. In the end, we hope to develop and disseminate a deep understanding of the user dynamics of CQA.

Finally, we emphasize that understanding CQA sites is not simply an interesting curiosity. CQA has become an essential tool in both professional self-support (e.g., sites like Stack Overflow that support software developers) and in commercial “level 0” customer support (e.g., sites like TurboTax Live Community where volunteers help people with tax preparation questions). As businesses and other organizations seek ways to reduce the costs of customer support, online community self-support in the form of CQA has the potential to be an immense cost-saver while helping build community. The same demand is coming to exist in education, where companies are providing CQA sites to support instructors in both traditional courses and large scale open courses (e.g., MOOCs). Understanding how to make these communities function effectively will not only help business and professional development, but also in education and other areas of human endeavor.

Related Work

Understanding Question Asking and Answering

Describing Patterns of Use. People have used digitally networked technology—email, Usenet, and other discussion forums, for example—to ask questions long before the invention of CQA services. However, several social and technical advancements have come from more modern user interface designs that specialize in a question and answer format. In particular, modern CQA sites are search engine optimized, incentivized (with money or points), and categorized (with tags or site-specific taxonomies). These advancements, along with the accompanying increases in user-base and question/answer volume, have caused researchers across several fields to investigate the mechanics of these sites and the underlying motivations and actions of the users.

Perhaps the most foundational work in this research space aggregates large datasets to understand user behavior. (Adamic 2008) present a highly-detailed descriptive study of Yahoo Answers, a very diverse and active CQA site. This study leverages the categorical nature of Yahoo Answers to learn about topic-level differences, and uses social networking techniques to discover user-level differences. One of the key insights in this paper is that different categories of question-asking and answering function remarkably differently in terms of user behavior. Subsequent work by Mamykina et al. (2011) analyzes a data dump from the programming-oriented CQA site Stack Overflow, finding that 92% of questions are answered multiple times, and the median time to an “accepted answer” is 21 minutes. They explore site usage in terms of user stereotypes, when and how often questions are viewed, and success rates. To make sense of these findings, they survey the site’s developers and key members of the community, finding that developer participation in the community was an important factor, as was the site’s interaction design.

Much of the research interest in CQA technology is related to its position as the “social” cousin of search engines like Google. Morris et al. (2010b) directly compare the results of searching and asking in a lab study, finding that subjects preferred searching, but asking provided greater personalization and increased confidence that the answers were correct. Hecht et al. (2012) acknowledge the complementary strengths of the two question-asking modalities, proposing an interface that posts search results directly into the stream of answers.
Research has also focused on discovering insights into key features or modalities of CQA interaction. For instance, moderation or reputation systems are prevalent in CQA designs. Paul et al. (2012) used data analysis and interview techniques to learn about users’ perceptions of the reputation system in Quora, suggesting that the de facto points models could be improved by incorporating additional features, such as tying real identity to screen names. Lee et al. (2012) study the differences in questions and answers when the interaction is moved to a mobile environment, finding shorter content and more pervasive use.

Other non-specialized social networking technologies also support CQA interactions. Morris (2010a) examined why users choose to ask questions on Facebook, and what types of questions they ask there; Lampe (2012) examined users’ perceptions of the informational value they receive there. Paul et al. (2011) investigate how often different types of questions are asked on Twitter (rhetorical and factual) and Twitter’s success in answering those questions (most questions get no responses).

**Categorizing Content and Users.** Most descriptive work on user behavior in CQA platforms relies on some form of categorization for questions and answers (e.g., Adamic 2008) or users (e.g., Mamykina 2011). Such categorization helps divide the spaces of questions and answers into distinct components that behave differently (e.g., questions about professional wrestling differ from those on theoretical physics; questions seeking opinions differ from those seeking facts). Some work has focused on building rigorous and generalizable taxonomies. For instance, Gazan (2007) applies the theoretical perspective of “specialists” and “synthesists” to question asking and answering, finding empirical differences in the quality of answers provided by the synthesists. Harper et al. (2010) build a theoretically-grounded taxonomy of questions types, based on Aristotelian rhetorical theory along with more modern theoretical rhetoric work. Efron and Winget (2010) develop and evaluate a taxonomy of question types asked on Twitter, arguing that questions there are fundamentally different from those asked on other platforms, perhaps due to Twitter’s mechanisms for sharing/consuming content.

**Understanding the Effect of Economic Incentives.** Google Answers is a CQA site that launched in 2001. The most notable aspects of Google Answers are that (a) users paid between $2 and $200 for answers, and (b) there was a limited pool of “expert” researchers who locked questions for answer. Though Google Answers closed in 2006, it served as inspiration for a body of research into the monetary incentives underlying an online “information market”, including empirical work (e.g., Edelman 2011), theoretical economics work (e.g., Jain 2009), and applied systems-building work (e.g., Hseih 2009).

One of the most interesting conversations to emerge from the research analysis of market-based CQA is whether paying more for information results in higher quality information. In fact, there is disagreement in the field (Jeon 2010) concerning the answer to this question. Chen et al. (2010b) conducted a field experiment in Google Answers, controlling the questions while manipulating the prices, finding that higher prices led to longer, but not better answers. Harper et al. (2008) found contradictory results, finding that paying $30 resulted in qualitatively better answers than paying $3.

**Understanding Cultural and Contextual Effects.** Question-asking online is a global phenomenon. One would expect cultural differences to exist that fundamentally influence how sites are designed, and how users interact. Nam’s (2009) study found similarities between Korean site KnowledgeIn and US-based Yahoo Answers, but underscored how the cultural emphasis on altruism shaped user behavior. Yang’s (2011) survey study of question-asking across four countries found strong cultural influences in question types and topics, the relative importance of questions, and user motivations for participation.

Perhaps more subtle differences may be found when comparing a professional context for question asking (e.g., using a corporate intranet) with a public context. Classic work in the field (Ackerman 1990, Ackerman 1996) proposed design strategies optimized to the particular problem of question asking and answering in a corporate environment. More recently, Thom et al. (2011) examine use of question asking in an enterprise social networking tool, characterizing question types and topics in this setting.
Experiments in Community Q&A

Simulation, Modeling, and Algorithm Design for Q&A Communities

Predicting Quality. Perhaps the research questions that have attracted the most attention in recent years relate to measuring and predicting the quality of user contributions in CQA systems (e.g., Janes 2001, Teevan 2011). One reason for this interest is the obvious presence of a large amount of low-quality or off-topic content, especially in large, general-purpose sites like Yahoo Answers.

Our work (Harper 2008) presents a cross-site perspective on the predictors of answer quality. We conducted a controlled field experiment where we injected a set of questions into a variety of CQA sites, measuring how factors such as question type and topic related to quality-centric outcomes. Among other results, we found interaction effects between question attributes and the site where the question was asked, and we found that community size was critical in producing consistently good answers.

Other research has experimented with a variety of computational models for predicting the quality of answers. Adamic et al. (2008) found strong predictive power in simple answer attributes such as length and the user’s reputation score, with evidence that the quality of resulting predictive models varied by the category of the question. Agichtein et al. (2008) add more complex factors to their machine learning classifier, finding a variety of features that added predictive power. In particular, they found that using text alone was a poor predictor of quality, and that no one feature stood out as dominant.

Subsequent work has focused on building statistical models of question-asker satisfaction, a quality that may be defined by particular actions, such as marking an answer in the user interface as accepted, and assigning some minimum rating value to that answer (Liu 2008). One interesting result of this research is that the most predictive features relate to past activity by the question-asker, rather than attributes of the answerers. In follow-up work, the researchers apply a similar model to predict web searcher satisfaction when they land on a CQA page with existing answers (Liu 2011a).

Little work has investigated the problem of predicting question quality, though the quality of questions certainly impacts the eventual archival quality of question-answer sets. We ran a preliminary experiment in this vein (Harper 2009), describing an ensemble machine learning algorithm that could accurately separate conversational and informational questions, which we found to be empirically different in terms of their potential archival value.

Discovering Experts. Related to the problem of automatically measuring the quality of the content in a CQA site is the problem of automatically measuring the quality of users. In particular, research has focused on methods for detecting the most elite users, sometimes called “experts”. These users often contribute much more than their share of the high quality content in a Q&A community, and have been shown empirically to behave differently from other users (Pal 2010).

Several studies focused on the development of features to assist in this prediction task. Though CQA sites usually do not offer explicit social network relationship features, it is possible to build a network structure by drawing edges between a question asker and the users who answer that user’s question. For instance, (Jurczyk 2007a) and (Jurczyk 2007b) experiment with using the HITS algorithm on this structure, finding success in predicting outcomes such as “best answer” votes. (Zhang 2007) developed a new metric - the “z-score” - that combines a user’s frequency of asking and answering into a single number. This CQA-specific metric also contains predictive power in models that attempt to predict which users will provide the highest quality answers. Subsequently, researchers have leveraged these metrics and others to build accurate predictive models of expert users from early in their lifecycle (Pal 2011, Pal 2012a), and to automatically infer the distinctions between experts and non-experts (Bouguessa 2008).

Question Routing and Recommendation. CQA sites are dominated by power laws, where a small number of users perform most of the work. They are also subject to problems of information overload, where users are unable to effectively locate information or opportunities for contribution. The CQA research community has addressed these problems through the development and evaluation of question
routing techniques that attempt to engage a greater user base (Guo 2008), or to help power users more efficiently use their time (Zhou 2009; Li 2011). The most common techniques involve building statistical models of users’ interest in a set of topics (e.g., Li 2011). Liu et al. (Liu 2011b) extended this topic-oriented technique by building a model of when and how users choose questions to answer, showing, for example, that users post answers in relatively short bursts, and that these bursts tend to be focused on a single category. Zhou et al. (2009) describe another effective question-routing framework based on multi-component model built from user profiles, thread structure, and clustering techniques. In some CQA contexts - such as Quora or Facebook - there is an opportunity to route questions using social networking techniques. Researchers have investigated these techniques, finding that so-called “strong ties” behave better than expected (and better than weak ties) in Facebook questions (Panovich 2012), but that strangers willingly answer questions that are routed to them in Twitter (Nichols 2012).

The problem of question routing has also been explored in synchronous CQA systems like Aardvark (Horowitz 2010), where the system is designed to deliver nearly immediate answers using communications technologies like instant messaging. Aardvark routed questions based on a statistical model based on the probability of each user answering a given question, along with the social connectivity and profile similarity between the asker and the answerer. White et al. (2011) designed a longitudinal study to investigate the effects of question routing in a synchronous system. They found that larger communities of answerers led to better outcomes across a variety of measures, but that increased routing frequency correlated with lower user satisfaction. Researchers have worked to mitigate these interruption costs using market mechanisms (Hsieh 2009).

Research Questions

As discussed above, the following questions bridge the frustrations inherent in questions we could not answer without direct access to CQA site internals and the opportunities that arise with access to our own research site. We reiterate that they are not meant to be exhaustive. Other questions are omitted both due to space and due to the limitations on our ability to anticipate the interesting research that will emerge as we accumulate data and experience. For each question, we briefly review the motivation and specific related work not already discussed above, present specific research hypotheses or sub-questions, and briefly review any research methods not common across all CQA research. We discuss the research infrastructure, general metrics, and online field studies together after all of the research questions.

**RQ1.** What kinds of “front page” recommendations most improve community outcomes?

**Introduction.** Personalization of user interfaces demonstrably affects user behavior and community outcomes. For instance, one study found that personalized messages encouraged use of an online discussion forum, but that some forms of personalization were more effective than others (Harper 2007b).

Another study found that recommendation algorithms for suggesting friends had a strong impact in shaping the formation of relationships in an online social network (Chen 2009). These studies are examples of “mixed-initiative” user interface design (Horvitz 1999) that seek to balance user customization with intelligent filtering and recommendation.

The front page of a CQA site is a prime example of an interface where such mixed-initiative design principles may be employed to influence behavior and outcomes. On the front page, system developers have an opportunity to encourage users to perform actions of value to the community by displaying personalized information. But such a display should not restrict users from exploring the question and answer space in the way that they wish. Popular CQA sites use a range of approaches to this design challenge – Yahoo Answers displays most recent questions with support to refine by category, Quora shows a personalized feed of questions based on the activities in the logged-in user’s social network, and Stack Overflow allows the user to browse recent activity in a variety of ways and provides an interface for choosing and browsing to tags of interest.
Experiments in Community Q&A

This research seeks to understand the algorithmic design space of recommending questions (and answers) to users on the front page, and to experimentally determine how different design choices impact user experience and community outcomes. Some of the algorithmic considerations for finding content to display include:

- recency: CQA site users are more interested in new activity than old activity
- answer quality: unanswered questions, and questions with low quality answers may need more attention than questions that have been well-answered
- social proximity: users may be more interested in activity from contacts in their social network
- topic proximity: users may be more interested in activity relating to familiar topics

Specific Hypotheses.

H1.1 Emphasizing recent questions decreases the average time to first answer
H1.2 Emphasizing questions with no answers or low-quality answers leads to higher average answer quality
H1.3 Emphasizing the activity of social connections leads to (a) higher activity levels per user, (b) greater use of social networking features, including profile pages
H1.4 Emphasizing topics of interest leads to (a) higher average answer quality, (b) higher rates of unanswered questions

Experimental Methods and Metrics.

To explore the differential impact of different methods for recommendation we will extend our site with a home page module that allows us to drive content using a variety of algorithms. We will divide users into discrete experimental groups, where each group will have a different algorithm driving home page content. We will also provide users an option in their "settings" page to change their default home page to any of the other algorithms at any time - we will measure the rates at which users switch away from the different algorithms as one measure of subjective satisfaction with the method used. We will measure user actions, such as the number and quality of contributions, time on site, and login events, to determine objective high-level differences between the groups. In addition, after we have collected data over the course of several months, we will survey users to discover their subjective preferences.

RQ2. How should intelligent task routing be deployed to manage workload distribution?

Introduction. CQA sites are not perfect. Questions go unanswered, or answers take too long to arrive. Users are unable to find opportunities to contribute, or miss the questions and answers they would enjoy the most. Too often, a small group of super users end up doing the vast majority of the work.

One approach to mitigating these problems is to incorporate intelligent task routing algorithms (Cosley 2006) to help users find personalized opportunities to contribute to the site. Prior work has investigated opportunities for routing in Q&A communities. Several researchers (e.g., Guo 2008; Qu 2009; Nichols 2012) have looked at developing models of user topic interests. Liu et al. (2011b) extend this form of recommendation by building deeper models of how users choose questions to answer.

In this research, we are interested in three concepts that have not been deeply explored in other work, but that the GopherAnswers experimental infrastructure will support. First, we aim to explore the possibilities for routing a variety of CQA-related tasks, not just question answering. CQA systems benefit from a variety of types of work, such as rating, commenting, tagging, and marking duplicate questions. Synergistically, some users will be able to contribute these lower-effort tasks, but will have a more difficult time providing new answers to questions. Second, we aim to explore the effects of explaining the recommendations provided by our routing algorithms. Prior work in recommender systems has shown that explaining recommendations can improve user acceptance of the algorithmic recommendations (Herlocker 2000). We wish to learn if there are similar techniques to increase adoption of routing interfaces. Third, we aim to study the impact of when and where routing recommendations are
displayed. When are users looking for opportunities to contribute, and how can we most effectively communicate what we’d like them to accomplish?

**Specific Hypotheses.**

- **H2.1** intelligent task routing will increase the overall level of contribution per user, site-wide.
- **H2.2** recommending “easier” tasks, such as evaluating whether questions are duplicates, increases the number of those contributions in the system, and does not lower the overall level of contribution per user.
- **H2.3** explaining routing decisions causes users to view recommendations more frequently, and to evaluate them more highly.
- **H2.4** routing options that mirror recent actions increase user contributions more than routing options that are independent of recent actions.
- **H2.5** routing algorithms that push recommendations throughout the site’s pages generate more actions than routing algorithms that are only displayed on the home page.

**Experimental Methods and Metrics.** We will evaluate the effectiveness of our routing interventions by conducting A/B testing - different users will see different recommendation interfaces when they use the site. We will launch separate experiments to cover the three research aims discussed above. We will also employ survey methodology to first gather information about what users would like to see and the types of contributions they are most interested in pursuing. After we have deployed a routing interface for several weeks, we will follow up with an additional survey to gauge user satisfaction.

**RQ3. How should referrals function within CQA?**

**Introduction.** Organizational knowledge management research (Alavi 2001) cites the importance of infomediaries and network hubs (Hagel 1999) as critical components in an organizational information structure. These individuals are not experts or even question-answerers, per se, rather, they understand the information resources in an organization, and help information seekers identify the right resources. In the context of CQA, these are people who may not know the answer to a question, but know the people who do know those answers. They can provide referrals, thereby bringing the question to the attention of those who can answer it. This research question explores the concept of encouraging, rewarding, and capturing referrals as first-class contributions to a CQA system.

**Example.** Recall Q2 above—a student (call him Bob) wanted to know how to use his roommate’s football season ticket. Let’s assume Toni reads this question, perhaps seeing a couple of speculative answers. She doesn’t know the right answer, but she does know that Jill is director of student ticketing in the athletic department. She can help Bob get the answer he wants by referring his question to Jill. Jill can then post an authoritative answer. This example illustrates the variety of design variables that can be explored. Is the referral in-band (and visible in the question stream) or out-of-band (visible only to Jill)? How, if at all, is Toni credited with helping get the answer?

**Specific Hypotheses.** RQ3 has four primary hypotheses:

- **H3.1** Adding a referral mechanism increases overall answer rate and quality.
- **H3.2** Visible referrals yield more referrals and higher-quality outcomes than hidden referrals.
- **H3.3** Referrals lead to participation from a group of users who otherwise tend not to participate (e.g., authorities).
- **H3.4** Incentives for referral lead to more referrals and higher quality outcomes.

RQ3 also has two cross-over questions.

- **RQ3a** How do human referrals compare with automated question routing? Are there domains or circumstances where either is consistently better?
- **RQ3b** How do referrals relate to the social networks of the referrers? Do people refer primarily to friends? Are friend referrals more or less useful than out-of-network referrals?
Experimental Methods and Metrics. To explore referrals, we will add a contribution type (a new class of contribution) for referral, with a referral contribution screen that allows users to browse the directory to find the appropriate reference(s). Configuration options will include whether the referrals are displayed or not, and whether to allow anonymous referrals. Referrals will generate a contact (e-mail by default, other types configurable) with a link to the specific question and any referrer’s message. We will log both referrals themselves and subsequent activity (including link-following, answering, and reading/rating the generated answers). We will also experiment with points and other rewards for successful referrals (i.e., those that result in answers; those that result in high-rated answers), measuring the degree to which these are impacted by different visibility and incentives. In addition to frequency of referral and quality of resulting answers, we will employ our standard suite of satisfaction and engagement metrics, for referrers, referees, and those who ask/read those answers.

RQ4. How are user profiles used in CQA, and how can they help the community?

Introduction. A majority of CQA sites provide user interfaces to edit and view user profiles. Despite this, there has been little research that explores the rhetorical effects profiles have on CQA exchanges. Raban (2009) explores how users employ a combination of implicit and explicit cues to affect a CQA exchange, but does not focus her study around user profiles. From a different perspective, Shah et al. (2008) research the user profiles of Yahoo Answers and Google Answers, but limit their analysis to the quantitative data linked to a user’s number of questions, answers, and points/rewards.

We intend to extend this work by looking at additional features of a profile including the information users voluntarily disclose about themselves. Within the context of CQA exchanges, profiles have the potential to give users background information about fellow users and potentially shape how CQA interactions occur. This research question would explore two key functions of user profiles. First, it would determine whether or not viewing a user’s profile affects a CQA exchange. Specifically, we will perform textual analysis of CQA exchanges in relation to the content viewed in a user profile to determine whether it leads to better ratings, additional comments, adding contacts, and/or other improvements to the site’s ecology. Second, our research would explore what components of a profile are most important to other users and — more importantly — work to determine how that affects CQA exchanges. GopherAnswers will provide a rich environment to study this question due to our ability to precisely measure what components of a profile are selected (clicked on) by users, how long users view profiles, and addresses whether CQA sites truly function as “online communities,” and how profile-related community reinforcing behaviors can affect the content on the site.

Example. Consider Q3 above, where a student (let’s call her Jane) asks about Calculus 2 instructors. Tran is a graduate student in Math and has been a TA for several of the instructors. He considers posting a response, but wants to be sure this isn’t just an attempt to bait people into bad-mouthing professors. He also thinks the answer to who is “best” depends on the student. He looks at Jane’s profile and sees that her previous questions and answers have all been sincere. He also sees that she’s linked to a homepage that indicates she’s a first-year student planning to major in biology. Armed with that information, Tran posts a reply. He even decides that it is worth the time to post some advice on how to get the most out of Calculus 2. Jane is both impressed and grateful. She sends a note of thanks, and also adds Tran as a “favorite” to follow his later comments. This example explores the potential that user profiles can have on a CQA site. Does viewing profiles affect the quantity/quality of responses a user receives to a question? Do user profiles encourage “community” like behavior and additional user interactions?

Specific Hypotheses. RQ 4 has three primary hypotheses:

H4.1 Viewing a user’s profile changes the likelihood of answering that individual’s question.
H4.2. Viewing a user’s profile changes the likelihood for users to have repeat interactions.
H4.3 Viewing a user’s profile changes how an individual responds to her/his questions.
H4.4 For each of the above, the direction of change is related to the quantity and quality of past contributions and the affinity/similarity of viewer/user profile demographics and attributes.

RQ4 also has the following sub-questions:
- RQ4a What information do users tend to add to their profiles?
- RQ4b What links/anchors within the profile are clicked on most by other users?
- RQ4c What profile content is most associated with answering? With not answering?

**Experimental Methods and Metrics.** To explore this research question, we will customize the GopherAnswers user profile system with an experimental module to support A/B field testing. Our methods will include a mix of qualitative exploration, quantitative modeling and hypothesis testing, and explicit experimentation. We will build models of profile-viewing and subsequent actions to understand both profile-viewing behavior and its consequences. We will carry out supervised laboratory studies to understand profile usability and usage at a qualitative level. And we will use textual analysis of profiles and CQA exchanges, coding for behavior indicators (e.g., humor, trust, shared values), and overall quality of answer.

**RQ5.** To what extent can reward systems and rhetoric emphasizing altruism improve participation?

**Introduction.** Most of the major online CQA services have coalesced around questioners and readers voting answers up and down and/or designating specific answers as “best” answers. Users receive points and recognition from accumulating best answer credits or simply from total numbers of answers posted. We know from our recent work surveying top contributors to Yahoo! Answers that these individuals overwhelmingly cite altruism as the primary reason for their high participation on that site—they want to help others. This mismatch between helpful and most/best is rhetorical in a deep sense—the most/best framing focuses on individual recognition while the helpful framing emphasizes the other and the community. An example of this gap occurs in cases such where multiple answers together provide the greatest value—not any particular one. A competitive framing often disproportionately rewards one response and may discourage those who have something to add, but not as much to say as the earlier posters. We will experiment with varying rhetorics and reward structures, with variations including recognizing all contributions equally, recognizing the best, recognizing all helpful contributions, and separating out helpful to the asker and helpful to the rest of the community.

**Example.** Q6 above seeks advice on how to handle unexpectedly low grades. This question may elicit a dozen different answers, ranging from personal anecdotes to statements of policy about dropping courses and information about grade distributions and later success. We are interested in how interactions change when the asker (or readers) are asked to designate answers as “best” vs. “helpful”, and how feedback on quality, quantity, or helpfulness affects subsequent participation in later exchanges. Consider the personal anecdotes, in a “best” system, would contributors feel unrewarded?

**Specific Hypotheses.** RQ5 has three primary hypotheses:
- H5.1 There are a significant number of helpful answers that would not be acknowledged as valuable in a system based on voting for “best” or “correct” answers.
- H5.2 Emphasizing helpfulness within both the language and the structure of the acknowledgments within the site will increase overall participation throughout the site.
- H5.3 Emphasizing helpfulness over “best, correctness, quantity” will lead to greater community identity within the site.

**Experimental Methods and Metrics.** We have long wished to pursue this question but existing CQA sites are understandably reluctant to tweak formulas that they perceive as successful, thus underscoring the value of this project’s opportunities to work “behind the curtains.” We have two models for testing variant feedback and rewards. First, we can roll out variant feedback in different sections of the site (i.e., divided by topic). This will allow us to explore different alternatives before dividing our population. For
Experiments in Community Q&A

a complete study, however, we will divide the site population such that different sets of users see different voting, feedback, etc. Since not all questions get votes, we can hide the other-system votes from the view of the users of each system. After a suitable period (probably 3 months) we will analyze the effects, conduct surveys of the two groups of users, and then either deploy the “winning” strategy or deploy the strategies varyingly in domains where they are most effective. We will also monitor in-site responses to the acknowledgment systems as they occur, and observe whether different types of users rise to the top of the two parallel experimental settings. We believe this experiment has significant implications for site design in the more general CQA ecology.

Research Infrastructure, Methods, and Metrics

In this section we discuss the GopherAnswers research infrastructure (both technology and the flexibility it offers), the basic research methods we will use across the study, and a set of general metrics we will be collecting. This section is not meant to be exhaustive, but reflects the most important points in each.

GopherAnswers. As discussed above, GopherAnswers is our second community CQA site, one developed specifically for the purpose of experimental study. GopherAnswers is integrated into the University of Minnesota’s user management environment, which allows us access to user authentication

Figure 1. Screenshots of the GopherAnswers prototype: (a) home page; (b) ask a question; (c) answer a question; (d) view a user profile.
without the need for separate accounts. We also provide public access (for those not currently affiliated with the University) and the option of anonymous postings (subject to moderation). Like most CQA infrastructure, GopherAnswers provides a web-based interface to browse, post, and answer questions, profile facilities, a feedback system, and moderation and management tools (see figure 1). We will be extending these tools to provide special infrastructure for controlled experiments and field trials, including cross-section A/B testing, and to add the special features needed to implement the experiments.

The data model underlying GopherAnswers at the time of this writing is shown in figure 2, illustrating the flexibility to customize user data, question attributes and tags, and voting structures. We will supplement this basic data model with a fine-grained logging database that will capture a detailed set of metrics, including page views, clicks, search text, browser information, and localization information. To offer browser-level logging hooks, we will initiate logging actions from an asynchronous JavaScript snippet installed on each page (modeled on the logging infrastructure used in Google Analytics).

The GopherAnswers software platform is a fork of the open source project OSQA. It is built in Python, using the Django web framework, and depends on MySQL, an open source relational datastore. The user login framework is extensible; we will tie in with the University of Minnesota’s SAML-based Single Sign-On provider. Our research group has considerable experience hosting public web sites and tools, including movielens.org, which has seen more than 150,000 user registrations over its lifetime, and cyclopath.org. Our strategy for hosting GopherAnswers will follow a similar pattern - we will host the service internally in the University of Minnesota Department of Computer Science data center, employing the same rigorous practices to ensure data backup and security.

**Research Methods.** This work depends on a number of standard research methods that we have employed extensively in our other work:

*Logging and Analysis.* Online field experiments (Konstan 2007; Kraut 2012) are valued because of their high potential for experimental validity, and resultantly have a history in fields as wide-ranging as medicine (e.g., Lohr 1986) and economics (e.g., Harrison 2004). Online field experiments require either direct experimental control of a “real” resource (e.g., Ling 2006; Gilbert 2007), or experimenter participation under assumed identities (e.g., Burke 2007; Resnick 2006). Our research group has published a substantial number of results across several fields by conducting field experiments in settings like CQA sites (Harper 2008), movielens.org (e.g., Chen 2010a; Drenner 2008; Harper 2007a; Cosley 2006; Harper 2005), and cyclopath.org (e.g., Priedhorsky 2010; Panciera 2010).
**Qualitative Research.** Surveys and in-person interviews provide mechanisms for collecting qualitative data on user preferences and experiences (Cozby 2000). These methods add color and detail to quantitative methods, and can be used, for example, to better understand user affective and aesthetic preferences (Wright 2008). Our research group has experience using survey data as a primary, quantitative research mechanism (e.g., Harper 2008), as a coding tool (e.g., Harper 2009), and as a mechanism for adding detail to primarily quantitative work (e.g., Chen 2009; Cosley 2006). We have also conducted interviews to better inform experimental details and outcomes (e.g., Panciera 2010; Harper 2005).

**Metrics Collected.** In addition to specific metrics collected for individual research questions or hypotheses, we will collect a broad set of process and outcome measures to assess the health of the community and impact of interventions on individual users. The table below summarizes the most important metrics, broken down in two dimensions. The first grouping is user-level vs. community-level: some metrics relate to the health and activity of the community as a whole, while others measure the contribution of a particular user (user-level metrics can be aggregated into community-level metrics, but community-level metrics cannot be split out to measure individuals). Also, we cannot compute user-level metrics for not-logged-in users (though, we can compute session-level metrics for these users as appropriate). The second grouping is quality vs. quantity. Quantity is necessary to ensure a sustainable (or growing) community, while quality is a desirable aspect for informational systems.

<table>
<thead>
<tr>
<th></th>
<th>Community-level</th>
<th>User-level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quantity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>user retention rate; daily visitors</td>
<td></td>
<td>login count; login rate over time</td>
</tr>
<tr>
<td>bounce rate (visits with 1 page view)</td>
<td></td>
<td>page views; page views by type</td>
</tr>
<tr>
<td>% anonymous users</td>
<td></td>
<td>contributions (e.g., questions, tags)</td>
</tr>
<tr>
<td>distribution of page views / visit</td>
<td></td>
<td>contributions per sessions</td>
</tr>
<tr>
<td>distribution of time on site</td>
<td></td>
<td>distribution of activity by topic/tag</td>
</tr>
<tr>
<td>questions / day; answers / question</td>
<td></td>
<td>z-score (Zhang 2007)</td>
</tr>
<tr>
<td># active topics</td>
<td></td>
<td>response to out-of-band contact</td>
</tr>
<tr>
<td><strong>Quality</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>distribution of answers, tags, points</td>
<td></td>
<td>points received / question, answer</td>
</tr>
<tr>
<td># duplicate questions</td>
<td></td>
<td>has custom profile and avatar image</td>
</tr>
<tr>
<td>avg. archival quality of questions</td>
<td></td>
<td>answers: topic diversity</td>
</tr>
<tr>
<td>% askers who are answerers</td>
<td></td>
<td>% own questions with accepted answer</td>
</tr>
</tbody>
</table>

In addition, we expect to use a variety of non-computed metrics to assess the more qualitative measures of community health. We will survey users to collect measures of their satisfaction with the system overall, and with particular features or experimental conditions (e.g., see Harper 2005).

**Work Plan**

The work plan for this project involves three parallel sets of activities coordinated over a period of four to five “work thrusts” which will be outlined below. Staffing for the project includes a research associate who will serve as project coordinator and as a co-lead researcher, a 50% research assistant focusing on computational issues and research infrastructure and tools, and a 25% research assistant focusing on communications and community issues. In practice, following our past models, we will work closely together. We anticipate weekly formal research meetings and more frequent informal ones. We also anticipate involving several MS and undergraduate students in these projects.

<table>
<thead>
<tr>
<th>Thrust 1 (Months 1 – 9)</th>
<th>Launch and collection/analysis of start-up data; experiments on front-page recommendations (RQ1); development of infrastructure to support RQ3, RQ4.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thrust 2 (Months 10-17)</td>
<td>Studies of profile use (RQ4; long term results during Thrust 3); release and initial study of referral system (RQ3); development of infrastructure for RQ1.</td>
</tr>
</tbody>
</table>
Experiments in Community Q&A

<table>
<thead>
<tr>
<th>Thrust 3 (Months 18-25)</th>
<th>Deployment and experiments with intelligent task routing (RQ2), experiments with rhetoric of helpfulness (RQ5); selection of followups and addl. questions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thrust 4 (&amp;5) (through end of project)</td>
<td>Follow-up studies of RQ1-5 as warranted plus additional questions; preparation of GopherAnswers for transition to long-term status.</td>
</tr>
</tbody>
</table>

**Conclusion**

Community Q&A systems have become widely deployed for social, professional, and commercial support. Detailed research into these systems has been hampered by a lack of experimental control over successful deployed systems. In this proposal we address a variety of questions whose study is enabled by controlled research infrastructure. In particular, we seek to generate new understanding of how CQA can best be fostered through experimental study of interfaces and algorithms that affect what users see, how tasks are directed at them, and how reward systems work. We will also generate new software tools to support CQA through recommendation and task routing; and prototypes of new functionality and interfaces such as referral and improved feedback and reward systems.

Given the broad use of CQA, we look forward to helping disseminate these findings—both through publication and through broader outreach efforts to leading practitioners.

**Results of Prior Support**

This work follows on a stream of NSF support in this area, including the following grants:

<table>
<thead>
<tr>
<th>Grant</th>
<th>Amount</th>
<th>Title</th>
<th>Konstan</th>
</tr>
</thead>
<tbody>
<tr>
<td>IIS-0324851</td>
<td>$1,246,017</td>
<td>ITR: Collab. Research: Designing On-Line Communities to Enhance Participation -- Bridging Theory and Practice</td>
<td>PI</td>
</tr>
<tr>
<td>IIS-0534420</td>
<td>$630,133</td>
<td>Helping Hands: Computer Support for Community-Maintained Artifacts of Lasting Value</td>
<td>co-PI</td>
</tr>
<tr>
<td>IIS-0808692</td>
<td>$2,439,875</td>
<td>HCC: Large: Collab Research: Understanding Online Volunteer Communities: Toward Theory-Based Design</td>
<td>co-PI</td>
</tr>
<tr>
<td>IIS-0812148</td>
<td>$461,519</td>
<td>HCC-Small: Understanding and Supporting Online Question-Answering Sites</td>
<td>PI Logie co-PI</td>
</tr>
</tbody>
</table>

This project follows most directly from the HCC Question-Answering grant that is now finishing, and can be logically thought of as a continuation of that work (though with a significantly different focus). We focus on the specific results of that project as of this writing:

**Contributions to Science.** We have made several significant contributions to the scientific understanding of online CQA sites and the behavior of the individuals using those sites—particularly the top-level answerers who provide so much of the value those sites contain. Here is a brief summary of the work:

- This work was grounded in two foundational studies of online CQA. (Harper et al. 2008) reported on an experimental evaluation of CQA sites’ ability to provide quality answers to complex questions. It found that paying helped, but delivering the question to a large community was the most consistent important factor in obtaining high quality responses. (Harper 2009) focused on the differences between conversational and informational questions, finding that most of the archival value in CQA sites comes from responses to informational questions, and that a composite classifier could classify questions as information vs. conversational with nearly 90% success.

- Based on the recognition that conversational/information was too simple, obscuring the real nature of questions, we developed a new, theoretically grounded taxonomy of the six types of questions users ask on online systems (Harper 2010). We also used that taxonomy in much of our subsequent work as we explored both site differences and expert behaviors (Logie 2011). We are also delighted that
Experiments in Community Q&A

this taxonomy is catching on, with increasing citations and use by others in studies of question-asking and –answering.

- In collaboration with colleagues at CMU, and working with Intuit Corp., we studied the question-answering behavior of users on Intuit’s TurboTax Live Community, building a predictive model that can detect potential “superusers” within their first weeks of contribution (Pal 2011) and then experimented with a socialization and training program designed to retain, motivate, and boost the knowledge of these high-potential users, with mixed results (Farzan 2012).

- Based on the hypothesis that experts tend to direct their work where it is needed, rather than simply where it is easiest to post or where there are points available, we build and tested a model of expert question selection bias. We confirmed that this bias is real and substantial, and that it is possible to identify actual and potential high-value users by looking for users who direct their answers towards questions with low existing value. (Pal 2010; Pal 2012a)

- We also had a set of technical advances needed to process and experiment with online CQA. These include a detailed study of entity-linking interfaces (Dong 2011) and a study of identification of question temporality (Pal 2012c).

- Finally, we have completed substantial qualitative work including surveys and deep studies of the question-answering behavior of top answerers on the Yahoo! Answers site. This work (under review) shows different roles assumed by top answerers, highlights the importance of the community as a whole, and provides insights on reputation systems.

Software Development and Research Artifacts. As part of our research, we developed and deployed a CQA system as part of the MovieLens recommender system (www.movielens.org). This deployment was successful technically – it helped us refine the needed software architecture for a research CQA system. But its limited scope did not allow us to explore many of the experimental questions that required a larger community. Hence, we have started the development of GopherAnswers discussed above.

Contributions to Education. We are proud of the educational success we have had in this project. Over the course of a single small grant we’ve provided education for four doctoral students (Aditya Pal who earned his Ph.D. in Computer Science and is now a research scientist at IBM Research; Anuradha Uduwage, a current Ph.D. student in Computer Science; Joseph Weinberg who is now writing his dissertation in Rhetoric; and Andrew Virtue who is starting on his dissertation research related to online communication). We also had three students complete MS projects in Computer Science, and engaged ten undergraduates through REU and UROP research projects. Our work with undergraduates is one of the areas we’re most proud of; we’ve been able to attract a diverse and excellent group of students, many of whom did not previously consider research as a possible career direction. It is a point of pride to be able to write letters of recommendation for them as they apply to graduate school or technically advanced jobs.

Relationship of past research to proposed research. This project is grounded in the work we’ve done before, but it explores a very new direction. It builds on the experimental method of online field experiments (Konstan 2007), developing the infrastructure to apply that method for the first time to substantial online CQA systems. Our prior work reached the limits of what we could understand through data mining and observational study, and we now pose questions that can only be explored through interventional experiments.