The Impact of Social Design on User Contributions to Online Communities

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Franklin Maxwell Harper

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Joseph A. Konstan

May, 2009

Acknowledgements

This thesis was made possible (and more fun) thanks to the contributions of many:

- *Mentoring.* I have been lucky to find such supportive and kind mentors. Joe Konstan has been a sagacious, supportive, and often humorous advisor. Loren Terveen, John Riedl, and John Carlis have provided direction, wisdom, and encouragement. Dan Frankowski, Dan Cosley, Sean McNee, and Pam Ludford were influential senior lab-mates who taught me the ropes and helped me through sharing their own experiences.

- *Collaboration.* This thesis is multidisciplinary, and has taught me about much more than just computer science. Thanks to Yan Chen and Sherry Li for their contributions to Chapter 2, Bob Kraut and Sara Kiesler for their contributions to Chapter 3, and Sheizaf Rafaeli and Daphne Raban for their contributions to Chapter 4. All of these people broadened my perspective on research, which has kept things interesting.

- *Research Support.* Dan Frankowski and Rich Davies provided me with friendly and skilled technical help throughout this thesis. Shilad Sen helped me to discover my interest in data mining and was always valuable as a sounding board. Sara Drenner provided programming support that made the work in Chapter 3 possible, and Daniel Moy made the machine learning algorithms better in Chapter 5.

- *Friendship.* GroupLens has been a friendly and intellectually rich environment to conduct my graduate studies. I'm glad to have landed in such a great lab.

- *Funding.* I gratefully acknowledge the financial support of the National Science Foundation through grants IIS 03-24851 and IIS 08-12148.

I also wish to thank my family – especially my wife Cindy Harper – for strongly supporting me and caring for me through this process. You are awesome. Thank you!

Abstract

The World Wide Web has become increasingly participatory through the widespread adoption of interfaces that facilitate user-generated content. These interfaces can be made more social by allowing users to view and respond to the actions of others. For example, Flickr (http://flickr.com) encourages photo sharing by allowing users to view and comment on others' photos, and Amazon (http://amazon.com) encourages purchases through the use of book reviews, discussion forums, and recommendations. In this thesis, we explore the utility of social designs for improving the quality and quantity of user contributions to online communities.

We investigate the use of social design at several levels. First, in a series of online field experiments in MovieLens (http://movielens.org), we examine the potential for increasing the quantity of user contributions through the display of personalized, social information. Second, in a comparative, controlled field study across a variety of popular question and answer (Q&A) sites, we compare different models of participation and their impact on the quality of user contributions. Finally, in a study of hand-coded questions from several Q&A sites, we use machine learning techniques to understand the characteristics of users' requests that are predictive of informational quality.

We find evidence that appropriate use of social information can increase the quantity of user contributions: social comparisons led MovieLens members to rate more movies, and persuasive messages to visit the MovieLens discussion forum were most effective when they compared the user viewing the message to another member. We also find evidence that the unstructured participation models characteristic of Web 2.0 sites increase the quantity, diversity, and responsiveness of user contributions, with no apparent overall cost to information quality. However, we find that unstructured participation leads many users to treat these sites as purely social resources. Therefore, to better support the utility of Q&A sites as informational resources, we contribute a computational framework that can reliably characterize user interactions as informational or conversational.

Table of Contents

List of Tables

List of Figures

Chapter 1
Moving Toward a Social Web

What is "Web 2.0"? It is a term, popularized in 2004, that has come to mean different things to different people. It means the rise of start-up companies with glossy logos, two employees, and an espresso machine. It means Web pages that use asynchronous Javascript calls to dynamically change the contents of a page without a reload. It means light-weight development environments like Ruby on Rails and software engineering practices like "agile" software development.

Importantly, the use of the term Web 2.0 recognizes that a significant shift has occurred toward an interactive, participatory, and engaging Web. The term "browsing the Web" is already becoming an anachronism, as the simple act of passively viewing Web pages is becoming subsumed by more interactive experiences. We update our Facebook status, comment on a friend's Wordpress blog entry, rate a video on YouTube, and ask a question at Yahoo Answers. Even traditional news sources like The New York Times now include features that encourage users to submit pictures, share articles on external Web sites, and leave comments. Our online actions are becoming increasingly public, and the opportunities for contribution are expanding. And this trend appears to be gaining momentum.

One of the most pervasive and interesting aspects of the move to Web 2.0 is how the shift towards interactivity has increased the opportunities for *social interaction*. We are not just interacting more with Web sites, but we are interacting more with each other. Through comment boxes, profile pages, and even product reviews, we can see the actions and preferences of other people, comment on them, and interact directly with them.

To better understand what these social designs are like in practice, let us take the music-oriented Web 2.0 site last.fm as an example. Last.fm incorporates user participation in nearly every aspect of its site design. For example, every user owns a personal profile page that provides

information about their listening interests. Users can join groups and become friends with one another, guided by algorithms that show users their "neighbors" in taste. Users can leave comments for one another and listen to other users' playlists, and they are also given the ability to edit artists' (wiki-based) profiles and vote for the photos that should represent artists.

Like other Web 2.0 sites that incorporate social design elements, Last.fm capitalizes on the intuitive notion that social design is engaging to users. In fact, there is some empirical evidence in support of this intuition. One study of EBay Germany found that new users who chose to participate in the online community were much more prolific users of the e-commerce features of the site (Algesheimer 2006). Also, social psychologists argue that social interactions and the sharing of personal profile information leads to increased participation and a reduced likelihood of leaving the community (Ren 2007).

In this thesis, we explore the impact and usefulness of social design from several perspectives. We are interested in understanding how social design affects communities, users, and the quality and quantity of user-generated content. Our goal in exploring these themes is to help Web site designers better understand the design space, and so help them to create richer experiences for users and better outcomes for their community or business.

1.1 Our Perspective: Designing for Online Communities

In this thesis, we are concerned with better understanding design in the context of an online community. It is worth defining these terms, as they are widely used in different ways.

Figure 1-1 sketches the components and interactions in an online community that are of greatest interest to this work. It contains three primary components: the *actors* who participate in the community, the *Web site implementation* that mediates that participation, and the *data repository* that stores the results of that participation. The arrows shown in the diagram represent interactions between the components.

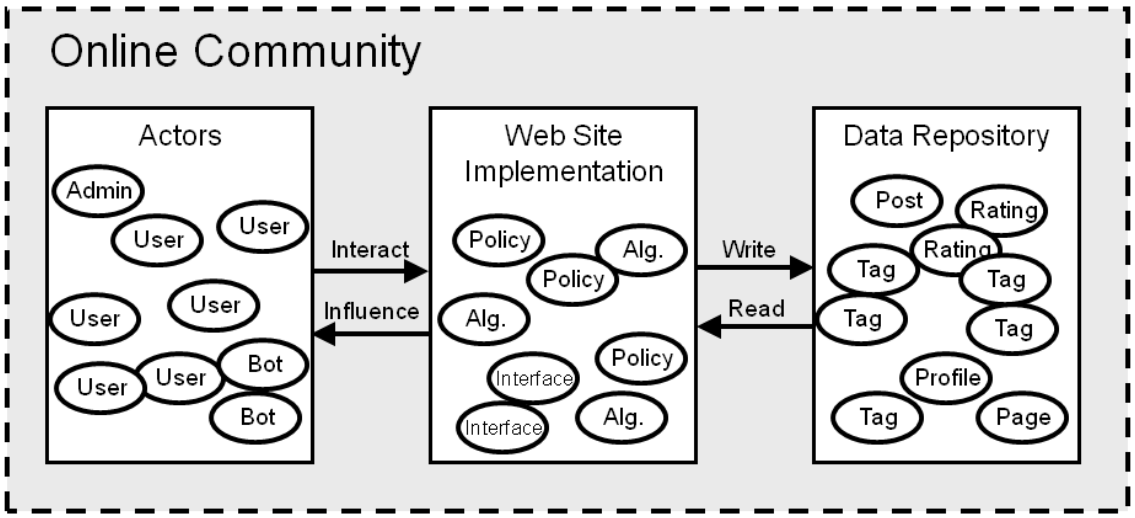


Figure 1-1. A conceptual diagram that summarizes our perspective on online communities. We identify three primary components: the actors who participate in the community, the Web site implementation that mediates that participation, and the repository that stores the results of that participation.

Let us step through this conceptual model using last.fm as our example. We start with the Web site implementation, a collection of algorithms, user interfaces, and policies that mediate the user experience and control the data repository. In last.fm, users can see music-oriented interfaces – in this case, dynamic Web pages – about artists and online radio stations, as well as socially-oriented interfaces about other users and interest groups. Users are also subject to a set of policies – for example, users are not allowed to upload music files that are copyrighted unless they have permission. Finally, algorithms that operate behind the scenes determine what information to display. For example, last.fm has algorithms that determine which artists, events, and users are recommended to each user on the site's home page.

Last.fm's Web site implementation requires a data repository filled with users' contributions such as posts, wiki edits, tags, and profile information. The Web site implementation is able to read this content as input to algorithms or interfaces, and it is able to write new content to the repository for future use or further processing. These data are interrelated. For example, a user might apply the tag "electronic" to the musical group Underworld; this tag can then be turned into a link by the Web site implementation and clicked by other users to locate all the groups that have been tagged with "electronic".

Finally, the last.fm online community is brought to life by the actors who use the Web site. Most common among these actors are regular users of the site, who might browse the site

anonymously, or become active members who contribute regularly. Users – and other actors, such as administrators and automatic bots – interact with one another as mediated by the Web site implementation. For example, when a last.fm user posts a message, last.fm's algorithms first process that message to automatically hyperlink any references to artists or albums; the post is then written to the data repository and may be subsequently be read as input to other algorithms or interfaces. Users are in turn influenced by what they see as they use the Web site; a last.fm user who sees a particularly poor description of their favorite band might be influenced to edit that description to make it better.

Web site administrators can hope to influence user behavior – and therefore community outcomes – through design. We refer to design broadly as any decision that can shape a user's experiences when interacting with the Web site. Last.fm has made many such decisions – from the links they show on their navigation bar, to the algorithms that recommend artists.

1.2 Research Goals: The Impact of Social Design

This thesis is organized into a series of chapters, each of which pursue a goal. We choose these goals because we believe they are both commonly held and considered important by community users or administrators. Each goal is linked to our overarching theme of understanding and innovating social designs that foster participation and encourage quality contributions. Table 1-1 presents a summary of these goals.

	Chapter 2	Chapter 3	Chapter 4	Chapter 5
Goal	↑ participation	↑ awareness/ consumption	good answers	archival content
Technique	social comparisons	social invitation	community structure	machine learning
Study Type	1-shot intervention	sustained intervention	online field experiment	artifact study
Level of Analysis	individual	individual	community	community
Social Context	no	little	some	more
Quality	yes	no	yes	yes
Quantity	yes	yes	yes	no
Platform	MovieLens	MovieLens	Q&A	Q&A
Original Publication	(Harper 2007a)	(Harper 2007b)	(Harper 2008)	(Harper 2009)

Table 1-1. A summary of the research in this thesis by chapter.

> ***Goal 1: Increase User Participation.*** *Communities may wish to encourage existing members to contribute additional "work" for the good of the community. We study a social mechanism for encouraging members to rate more movies and moderate a movie information database in a movie-oriented online community.*

In chapter 2, our goal is to encourage members of MovieLens (http://movielens.org), a movie recommendation site, to rate more movies and to help moderate the database of movie information. The MovieLens community relies on these contributions in several ways. The movie recommendation algorithm's performance increases when more users rate movies – especially movies that few people have seen. Therefore, encouraging members to rate more movies increases the predictive power of the recommendation engine for all members. The movie information database requires member maintenance in order for the library of recommended movies to be kept up-to-date and accurate. All members benefit from a cleaner, more complete database of movies to browse and rate, though the job of editing the database may not seem like much fun.

We address this goal by emailing MovieLens members a personalized message with information about their ratings history. The email messages contain a statement that compares users to one another in terms of their performance: users are told whether they are above, below, or about average, compared with other users like them. Thus, we are issuing a "social

comparison" (Festinger 1954), which social psychologists speculate causes people to evaluate their own performance, and set baselines for success. The email message also contains links to several potential activities in MovieLens, including "rate popular movies", "rate rare movies", and "help us update the MovieLens database". We subsequently examine the effect of comparisons of different "directions" on users' propensity to contribute in different ways.

Goal 2: Increase User Awareness of a New Feature. *Communities may wish to make their members aware of a new site feature. We study an automated, personalized mechanism for encouraging members to visit and post in a new discussion forum in a movie-oriented online community.*

In chapter 3, our goal is to promote users' awareness of a newly launched discussion forum in MovieLens. The addition of a discussion forum has been argued to have benefits to the community, such as increased member attachment (Algesheimer 2006). However, after MovieLens added a discussion forum, a very small percentage of users tried the new feature, and an even smaller percentage actively posted. Without member visits, a feature such as a discussion forum cannot succeed.

We address this goal by developing "personalized invitations" to visit the forums that appear at the top of the MovieLens home page. A personalized invitation is a small text box that contains a suggestion for action. For instance, the invitation might say that someone recently wrote a post about Star Trek II: The Wrath of Khan (a movie of great interest to the user), and suggest that the user go read that post. The invitations are *personalized* because they contain different messages for each user, based on their history of activity in MovieLens. We investigate the overall effectiveness of invitations in increasing member awareness of the discussion forums as well as the relative effectiveness of different strategies for personalization. We report on the differences between social vs. non-social personalization strategies and the inclusion of socially-oriented vs. data-oriented entities in the message.

Goal 3: Encourage High Quality Responses. *Communities may wish to provide an environment where users' questions are met with high-quality responses. We study the impact of high-level design decisions on response quality in several question and answer sites.*

In chapter 4, our goal is to understand how high-level site design affects response quality in several question and answer (Q&A) sites. Q&A sites are places where people ask questions, and others answer those questions – with varying degrees of quality. We believe that the quality of answers is strongly influenced by design decisions that determine who can answer questions and whether the site costs money to use. Designing for "open participation" is a key element of many Web 2.0 sites; we report on the impact of this particular design choice on response quality, diversity, and timeliness.

Our method is to ask questions of our own design in a variety of Q&A sites. We study five different sites, including Yahoo Answers (where everyone can contribute equally), AllExperts (where only select participants answer questions), and Google Answers (where there are articulated roles and fee-based asking and answering). We control for several aspects of the questions we ask, such as the difficulty, the topic, and the type of the question. We collect quality scores by asking a group of judges to rate the answers to questions. From these data, we report a set of quantitative and qualitative observations regarding the relative performance of different community designs.

> ***Goal 4: Identify Archival-Quality User-Generated Content.*** *Communities may wish to differentiate between user contributions that have informational content and those that do not. We study the indicators of this distinction and develop computational techniques to perform the differentiation in several question and answer sites.*

In chapter 5, our goal is to understand the structural differences between conversational and informational questions in a social Q&A site, and to determine the implications of this difference on potential archival quality. While conversational questions have a role in social Q&A sites – they engage users – they may not lead to information that's useful to people searching the Q&A site for answers to questions. However, manually labeling questions as informational or conversational is difficult and potentially inaccurate. We address this problem by developing techniques for algorithmically distinguishing between these question types.

To achieve this goal, we hand-code a dataset of questions collected from three Q&A sites, then apply data mining techniques to the coded dataset. We develop three feature sets that are useful in discriminating between informational and conversational questions: categorical features, textual features, and social network features. We use these feature sets to analyze the structural

differences between the question types, as well as to build machine learning models capable of distinguishing between the types with high accuracy.

1.3 What's to Come

The next four chapters are research-oriented and focused – each chapter represents one study (or group of related studies) that corresponds with one of the four high-level research goals discussed above. Each of these chapters may stand on its own – each contains focused discussion of any related work, contributions, and opportunities for future work.

In Chapter 6, we examine the context and contributions of this thesis. Where Chapters 2-5 tackle specific research goals, Chapter 6 attempts to synthesize these studies by presenting them in a larger context. Specifically, we discuss the *theoretical*, *methodological*, *computational*, and *empirical* context in which this work takes place. Once this context is established, we return to our research contributions with an eye to the broader implications of this work. We conclude with a discussion of the most promising areas for future work.

Chapter 2

Social Comparisons to Motivate Contributions to an Online Community*

In December, 2006, Time Magazine awarded its annual Person of the Year award to "You" (Grossman 2006) in a nod to the changing nature of the Internet. No longer are Web sites exclusively created by editors and read by everyone else; increasingly, they allow content to be contributed by anyone who so wishes. Wikipedia, MySpace, and YouTube have become some of the top-visited sites on the Web, based entirely on content contributed by their members. As a case in point, the Web page displaying the Person of the Year article contains several buttons that make it easy for readers to recommend the article to others via Web sites such as Facebook.

What motivates people to edit encyclopedia entries at Wikipedia, write movie reviews at Rotten Tomatoes, or share Time Magazine articles at Facebook? On the surface, many of these types of contributions have little personal benefit – editing an article in Wikipedia may help other users, but takes one's own time. Therefore, people must be motivated by intrinsic factors – for example, a desire to achieve status within a community (Bernheim 1994), or a desire to reciprocate the efforts of other users (Rabin 1993).

We may think of Web sites built on member contributions as public goods, subject to the problems of free-riding. We know from economics research that the environment in which decisions are made affects contributions (Ledyard 1994). Thus, designers of Web sites can hope to affect the volume of user contributions through design. They might take action to change the

---

* This chapter extends the work originally published as (Harper 2007a), co-written with Xin Li, Yan Chen, and Joseph Konstan. Also see (Chen 2009) for a deeper treatment of this experiment from an economic perspective. The author of this thesis built, managed, and analyzed the experiment. All authors contributed to the design of both surveys, as well as to the overarching experimental design.

costs of the contribution by making contributions easier to make. For example, social networking sites such as LinkedIn provide tools for members to import their contact lists, to save them the effort of entering contact information manually. Other sites attempt to increase the benefit to contributors. For example, the technology news-oriented site Slashdot unlocks extra features for members after they have provided high-quality contributions to the site.

Previous research on the voluntary provision of public goods has shown that information about social norms can affect contributions. For example, people recycled more materials when they were provided with information about how much other people had recycled (Schultz 1999). Can a similar comparison make a Wikipedia member edit more articles or a Rotten Tomatoes member write more movie reviews?

2.1 Research Questions

In this research, we use email to deliver a feedback intervention to make the norms of an online community of users salient. We extend prior work in several ways. First, we investigate the effect of leveraging social influence in an anonymous online system. Second, we investigate the effect of upwards, downwards, and no-difference comparisons. Our goal is to determine methods for eliciting additional contributions from these members. We investigate the following research questions:

> ***RQ Activity***. *How does social comparison in an online community affect members' propensity to visit and contribute?*
>
> ***RQ Perception***. *How does social comparison in an online community affect members' self-reported motivations to visit and contribute?*

In subsequent sections, we describe an online field experiment designed to answer these research questions. In this study, we find: (1) that messages containing comparison information focus members' energy to improve their relative standing, but do not increase overall interest in the community, and (2) that men and women believe themselves to be motivated by comparison information in very different ways.

2.2 Background: Social Influence and Comparison

To evaluate our abilities, actions, and opinions, we compare ourselves to others (Suls 2002). In some cases, we make these comparisons because we are presented with information about others' actions or information revealing social norms. Social influence and comparison has been the subject of much study in the social sciences; we use this work to inform our research on comparisons in an online system.

It matters who we compare ourselves to. Festinger (1954), in his classic work on social comparison, theorized that we compare ourselves to others who are better off for guidance, while we compare ourselves to others who are worse off to increase our self-esteem. Subsequent research, however, has found conflicting results regarding so-called upwards and downwards comparisons (Suls 2002). Wheeler and Miyake found that upward comparison decreased subjects' feelings of well-being, while downward comparison increased feelings of well-being (Wheeler 1992). However, Lockwood et al. found that upward comparisons can inspire people if success seems attainable (Lockwood 1997), and Buunk et al. found that downward comparisons actually make individuals feel worse about themselves in some contexts (Buunk 1990). Thus, we are left with little guidance about how comparisons made in an online system will make users feel – it is apparently highly dependent on the context and the individual.

We may be more hopeful that social comparisons can motivate individuals to increase contributions to a public good. Several studies have shown that making social norms visible can increase pro-social behavior. Frey and Meier conducted a study in which subjects were given information on the percentage of people donating to a social fund. They found that showing a percentage reflecting greater participation led subjects to participate more themselves (Frey 2004), but only for those subjects who had not already participated in the past. Croson and Shang found a similar result in testing social influence on donations to a public radio station. In this study, first-time donors who were told that another member had contributed $300 gave 29% more than first-time donors who were not given that information (Croson 2005). However, a meta-analysis of studies such as these shows that so-called feedback interventions can lead to negative effects on performance (Kluger 1996). This work proposes that interventions providing positive feedback to a subject in the absence of further opportunities to improve lead to

decreased effort. On the other hand, interventions providing negative feedback tended to increase performance, as long as increased effort from the subjects perceivably improved their standing.

2.3 Research Platform: MovieLens

To evaluate the effects of comparative messages, we ran a field experiment in MovieLens, an online movie recommendation Web site (http://movielens.org) where members rate movies and receive personalized movie recommendations (see Figure 2-1 for a screenshot). MovieLens uses a collaborative filtering algorithm (Resnick 1994) to predict how well members will like movies in its database. Because collaborative filtering works based on finding statistical correlations between users or items in the database, MovieLens relies on member-contributed ratings data. Newly-released movies and rarely-viewed movies are especially difficult to recommend due to a scarcity of ratings. 6.8% of the movies in MovieLens's database have fewer than 10 ratings, below the threshold required by the collaborative filtering algorithm to make predictions. Thus, one of the goals of this study is to find ways to encourage members to rate more of the movies they have seen.

Figure 2-1. Screenshot of the MovieLens home page.

At the time of this study, MovieLens did not contain exhortations to rate movies. Also, members had no way to see one another's ratings, activity, or opinions. When members joined the system, they were told that by rating more movies, they would receive more accurate recommendations. In addition to this information, a number at the top of each page reminded members of how many movies they had rated; this was hyperlinked to a page with statistics about those ratings.

2.3.1 MovieLens User Motivations

Why do MovieLens members rate? A survey of 357 MovieLens members (Harper 2005) showed that different members were motivated to rate movies in different ways. The most common reason was to improve the quality of the movie recommendations. Next most popular were the reasons of rating for fun, and rating to keep a list of movies that the user had seen. This survey also revealed that MovieLens members did not often think about one another (at the time, MovieLens did not have online community features). Few members claimed to rate movies to voice their opinion or to influence others. See Table 2-1.

	Top Reason	Second Reason (If Any)	Third Reason (If Any)	Total
Improve my Recommendations	211	73	18	302
Rating is Fun	50	66	77	193
Keep List of Movies I've Seen	50	75	48	173
Help MovieLens System	9	47	84	140
Help Others' Recommendations	16	39	32	87
Remove Movies from Screen	1	15	13	29
Voice my Opinion	3	7	18	28
Influence Others	3	9	14	26
Other	4	2	6	12

Table 2-1. Users' responses to the survey question: "please rank your top 3 reasons to rate movies". 357 users took this survey.

We also asked MovieLens users about their motivations to use the system, in general. Unsurprisingly, most users responded that they like to view movie recommendations. Other

users used MovieLens *because it was a place where they could rate movies* – revealing that the act of rating is not simply viewed as a cost, but as a benefit of the system as well! At the time of this survey (2004), relatively few users claimed to use the system for social reasons such as seeing other people's opinions or sharing their own opinions with others. See Table 2-2.

	Top Reason	Second Reason (If Any)	Third Reason (If Any)	Total
I like to view movie recommendations	192	102	31	325
I like to rate movies	91	97	61	249
I like to search for movies	39	78	70	187
I like to see what other people think of movies I like	4	28	67	99
I like to share my opinions	5	21	47	73
Other	18	7	15	40

Table 2-2. Users' responses to the survey question: "please rank your top 3 reasons for using MovieLens". 357 users took this survey.

2.4 Methods

In this section we first describe the mechanism used to inject social comparisons into MovieLens: personalized email newsletters. We then describe the methods used for subject recruitment and the timeline of the study.

2.4.1 Injecting Social Comparisons: Personalized Email Newsletters

To deliver our social comparison intervention, we designed two personalized email newsletters to send to MovieLens members. The experimental version contained a message about how many movies the recipient of the email had rated compared with other members in the system. The control version contained information about the member's ratings without comparison to other members.

The experimental and the control newsletter were similar in design. Each was formatted in html, although members with text-only email clients received a text-only version. Each contained a header with the MovieLens logo and some non-personalized statistics about the site. Below the header was a section with personalized information according to the subject's experimental group, as described below. Following this was a section containing a short news item about recent feature additions to MovieLens, and finally a section containing a reminder that this newsletter was sent as part of an experiment. Figure 2-2 and Figure 2-3 show screenshots of the control and experimental newsletters.

To deliver the social comparison, the experimental newsletter contained the following text at the top of the message:

> *Ever wondered how many movies you've rated compared with other users like you? You have rated [num_ratings] movies. Compared with other users who joined MovieLens around the same time as you, you've rated [more, fewer, about as many] movies than the median (the median number of ratings is [median_ratings]).*

In contrast, the control newsletter contained a personalized message about members' participation in MovieLens without any comparison to other members:

> *Here are some statistics about your ratings behavior for one popular movie genre. About [percent] of the movies that you've rated are comedies. Your average rating in this genre is [average_rating].*

Values for items in brackets were personalized based on the member's usage history, as described below.

movielens

helping you find the *right* movies

Now with 8,715 movies, 96,940 users, and 11,931,422 ratings!

MovieLens Experimental Newsletter v.1

Your Profile

Here are some statistics about your ratings behavior for one popular movie genre.

About **38.6%** of the movies that you've rated are comedies. Your average rating in this genre is **3.5**.

Interested in getting more out of MovieLens? Here are some options:

- help us update the MovieLens database - updating the MovieLens database will improve the quality of information in the system.
- invite a buddy to use MovieLens - having a buddy in MovieLens will give you personalized group recommendations.
- rate popular movies - rating more popular movies will link you with other users and improve the quality of your recommendations.
- rate rare movies - rating rare movies will help others get more movie recommendations.

Or, you can just visit MovieLens.

Figure 2-2. The control version of the email newsletter.

movielens

helping you find the *right* movies

Now with 8,715 movies, 96,940 users, and 11,931,422 ratings!

MovieLens Experimental Newsletter v.1

Your Profile

Ever wondered how many movies you've rated compared with other users like you?

You have rated **287** movies. Compared with other users who joined MovieLens around the same time as you, **you've rated more movies** than the median (the median number of ratings is 100).

If you'd like to **rate more movies**, here are some options:

- rate popular movies - rating more popular movies will link you with other users and improve the quality of your recommendations.
- rate rare movies - rating rare movies will help others get more movie recommendations.

If you'd like to **try new features**, you may want to:

- invite a buddy to use MovieLens - having a buddy in MovieLens will give you personalized group recommendations.
- help us update the MovieLens database - updating the MovieLens database will improve the quality of information in the system.

Or, you can just visit MovieLens.

Figure 2-3. The experimental version of the email newsletter, personalized for an above average member.

The newsletter followed this personalized message with five links (underlined below) along with neighboring text that explained the benefit of these actions:

- rate popular movies – rating more popular movies will link you with other users and improve the quality of your recommendations.

- rate rare movies – rating rare movies will help others get more movie recommendations.

- invite a buddy to use MovieLens – having a buddy in MovieLens will give you personalized group recommendations.

- help us update the MovieLens database – updating the MovieLens database will improve the quality of information in the system.

- Or, you can just visit MovieLens.

Because our results rely on members understanding and acting on the email newsletter that we sent, we pre-tested the usability of the newsletter via 14 phone interviews with MovieLens members. We found that, in general, members were able to understand the contents. 10 of the 14 subjects understood the concept of a median, while the remaining 4 interpreted the word as "average". 11 out of 14 subjects, after being asked to look away from the newsletter, were able to recall whether the newsletter had said they were above, below, or about average.

2.4.2 Recruitment, Participation, and Timeline

To solicit volunteers for the study, we emailed 1,966 MovieLens members, chosen randomly from those who had logged in during the past year, who had rated at least 30 movies, and who had given us permission to send them email. This email contained a link to a MovieLens Web site with a consent form describing the study. 629 members clicked on the email link, of whom 268 consented to participate and were included in the study.

We randomly assigned half of the 268 subjects to an experimental group and half to a control group. Subjects in the experimental group would receive an email newsletter with ratings comparison information, while subjects in the control group would receive a newsletter without comparisons, as described above. Since we were comparing members based on how many movies they had rated, we wished to ensure that new members to the system were not being (unfairly) compared with long-time members. Thus, we further divided subjects into three equal-sized groups based on their seniority in MovieLens (see Table 2-3). Within each of these

seniority-based groups, we call the one-third of subjects with the most ratings "above average", the one-third with the fewest ratings "below average", and the final one-third "average". These labels correspond to whether the subject was told that he or she had rated more, fewer, or about the same number of movies as the median member in their age group.[2] See Table 2-4 for an overview of the distribution of ratings across each experimental condition.

Treatment	Seniority	N	Avg # Weeks Member	Avg # Logins	Avg # Ratings
Control	New	45	12.5	8.6	287.0
	Mid	45	50.3	42.8	431.0
	Old	44	214.6	153.9	747.5
Comparison	New	45	15.2	12.9	399.1
	Mid	45	63.5	63.2	502.4
	Old	44	233.0	225.8	898.5

Table 2-3. Number of subjects and average activity prior to the study by treatment. By definition, members with more seniority had belonged to the site longer on average. As expected, members with more seniority had rated and logged in more often on average.

Seniority	Comparison	N	Min # Ratings	Median # Ratings	Max # Ratings
New	Rated Fewer	15	58	120	210
	Same	15	211	291	372
	Rated More	15	391	632	1408
Mid	Rated Fewer	15	88	224	306
	Same	15	310	391	510
	Rated More	15	525	680	1962
Old	Rated Fewer	15	177	340	520
	Same	14	567	784	1040
	Rated More	15	1082	1479	2165

Table 2-4. Number of subjects and statistics summarizing ratings activity prior to the experiment for subjects receiving the Comparison Treatment. The number of ratings a subject had provided prior to the experiment determined their placement in a comparison group.

[2] Unfortunately, the division of users into experimental groups has a bias: users in the comparison group are, on average, longer tenured users with more ratings. We first used a randomizer to divide users into the comparison and control groups, then later sub-divided these two groups into seniority groups. The initial randomization into two groups is where the bias was introduced (randomly!).

Members who consented to participate in this study were immediately redirected to an online survey. This survey was designed to collect subjects' perceptions of the benefits and costs of using MovieLens, using questions drawn from our earlier study (Harper 2005), as well as to discover how they believed they compared with other members in the study in terms of ratings. Two weeks after sending the initial invitation to participate in the study, we personalized and sent the email newsletter manipulation. We logged when subjects clicked on links in the newsletter as well as their actions in MovieLens following the email. Finally, one month after sending the email newsletter, we emailed subjects one final time asking them to take another survey. This survey asked members how well they liked the newsletter, and which links they thought were valuable. Subjects in the experimental condition were reminded of the comparison they saw in the newsletter and were asked how it made them feel. See Figure 2-4 for an overview of the timeline of the study for the 268 subjects.

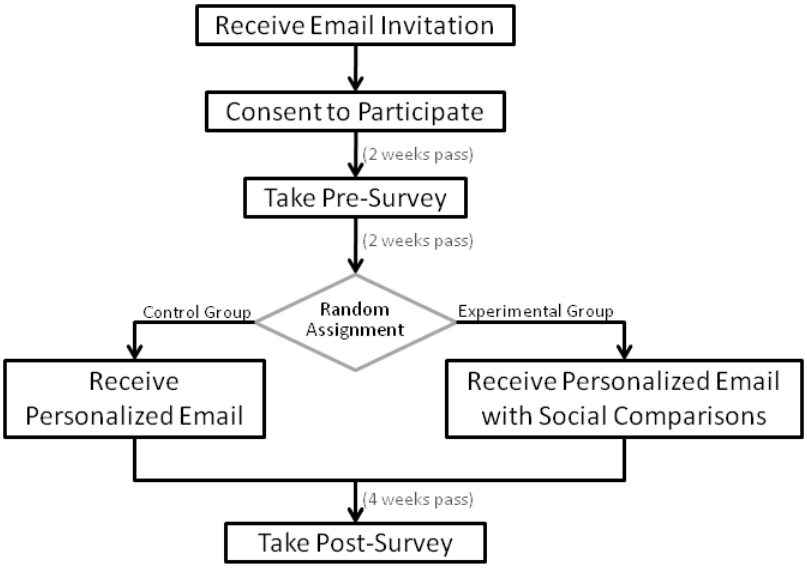


Figure 2-4. An overview of the experiment from the perspective of a user. Periods of experimental inactivity are noted in parentheses.

2.5 Results

Upon sending the email newsletter manipulation, subjects immediately began to visit MovieLens and rate movies. In the week following the manipulation, 49.3% (132/268) of

subjects clicked one or more links in the email message, 60.4% (162/268) of subjects logged in, and 48.5% (130/268) of subjects rated one or more movies. The five links displayed in the email newsletter were not clicked or acted on with equal likelihood; see Table 2-5 for a summary.

Suggested Action	# Users to Click	# Users to Act
rate popular movies	54	120
rate rare movies	79	78
invite a buddy to use MovieLens	7	2
help us update the MovieLens database	23	22
just visit MovieLens	19	162

Table 2-5. Response to the five suggested actions in the email newsletter across all experimental conditions, including the number of users who clicked each link in the newsletter, and the number of users who performed the suggested action in the week following the manipulation.

2.5.1 Effect of Social Comparisons on User Activity

Propensity to Click. Subjects who received the social comparison manipulation were no more or less likely to click on a link in the email newsletter. 48.5% (65/134) of subjects in the control condition clicked on one or more links, as compared with 50% (67/134) of subjects in the experimental condition (ChiSquare=0.06, df=1, p=0.81). See Table 2-6 for a summary.

Link Clicked	Control Group	Experimental Group	P-value
Rate Pop	22	32	0.13
Rate Rare	45	34	0.14
Invite Buddy	1	6	0.06
Maintain DB	13	10	0.51
Just Visit	8	11	0.48
(any link)	65	67	0.81

Table 2-6. The number of users (out of 134 in each group) to click on links in the experimental email, and the p value of the test of statistical significance (Chi-square). Users were able to click on more than one link at different times.

The direction of the comparison did not significantly affect subjects' propensity to click (ChiSquare=0.91, df=3, p=0.82), though subjects told they had rated fewer movies than other users clicked the least (44.4%), while subjects told they had rated more movies than others clicked the most (53.3%). However, there was some variation in the links that subjects chose to click, as summarized in Table 2-7. Subjects told they had rated more movies than other members were most likely to click the two links under the heading "try new features": invite a

buddy to use MovieLens and help us update the MovieLens database (ChiSquare=7.26, df=1, p<0.01). We also observed some interesting, but not statistically significant effects. First, subjects told they had rated about the same number of movies as other members were nearly twice as likely to click on the link just visit MovieLens as other members (ChiSquare=1.29, df=1, p=0.26). Second, subjects told they had rated fewer movies than other members were most likely to click rate popular movies (ChiSquare=2.39, df=1, p=0.12).

	Click Target				
Comparison	**Rate Pop.**	**Rate Rare**	**Invite Buddy**	**Maintain DB**	**Just Visit**
No Comparison	16.4%	33.6%	0.7%	9.7%	6.0%
Rated Fewer	28.9%	15.6%	2.2%	6.7%	6.7%
Same	25.0%	34.1%	0.0%	0.0%	11.4%
Rated More	17.8%	26.7%	11.1%	15.6%	6.7%

Table 2-7. Percentage of subjects clicking on each of the five links in the email newsletter by social comparison condition. Although there were no significant differences between overall click rates based on the direction of the comparison, there were differences in which links subjects chose to click.

Propensity to Act. Subjects receiving the social comparison manipulation rated significantly more movies the week after the email than subjects in the control group (means 13.17 vs. 6.51, p=0.03). However, the experimental manipulation did not have a statistically significant effect on other behaviors. See Table 2-8 for a summary.

Metric	**Control Group**	**Experimental Group**	**P-value**
Ratings	6.51	**13.17**	**0.04**
Logins	1.36	1.44	0.86
Buddies Invited	0.00	0.01	0.16
DB Edits	0.88	0.36	0.60

Table 2-8. The average number of actions taken by subjects in the control group or the experimental group, and the p value of the test of statistical significance (Wilcoxon non-parametric). Only the difference in ratings is statistically significant.

See Table 2-9 for an overview of the effect of the direction of the comparison on subjects' behavior. Subjects that were told they had rated fewer movies than other members rated significantly more movies in the week following the manipulation than other subjects (means

19.1 vs. 8.0, F=7.68, p<0.01). This group was also the only one to rate more movies in the week following the manipulation than their lifetime per week average. There was no significant difference in number of logins in the week following the manipulation between the control group (mean 1.36) and the experimental group (mean 1.44) (F=0.07, p=0.79). Subjects in all conditions averaged more logins in the week following the email newsletter than their lifetime per week login average. Subjects told they had rated fewer movies than other members logged in the fewest times (mean 0.8) in the week following the manipulation (F=3.43, p=0.07), although their rate of logging in increased at approximately the same rate as subjects in the control group.

Comparison	Ratings	Ratings/Week Change	Logins	Logins/Week Change
No Comparison	6.51	-7.08	1.36	0.45
Rated Fewer	19.09	11.06	0.78	0.46
Same	8.20	-1.71	1.52	0.87
Rated More	12.04	-17.95	2.02	0.32

Table 2-9. Average activity in the week after the email newsletter, and the average difference between this activity and members' lifetime per week activity. Members told they had rated fewer movies than others saw the largest increase in ratings, while members told they had rated about the same number of movies as others saw the largest increase in login activity.

To better understand how the social comparison affected users of different levels of activity, we conduct an exploratory analysis based on pairing users across experimental conditions. To construct these pairings, we first use the experimental criteria to assign users from the control group to seniority (new, mid, old) and comparison (rated fewer, same, rated more) conditions. Thus, our "inferred group assignment" reveals the type of newsletter that control subjects hypothetically would have received, had they been assigned to the experimental condition. Ordering within these groups by ratings prior to the experiment, we pair users from the experimental group and the control group. In Table 2-10 below, we summarize pairwise comparisons of these users across two metrics: number of ratings and number of logins the week after the email newsletter. The only significant difference appears for users who were told they were below average: users in the experimental group rated nearly 5 times as many as users in the control group the week after the newsletter ($p < 0.01$).

Comparison	Ratings (control)	Ratings (expt)	P-value (2 tailed)
Rated Fewer	3.91	**19.09**	**< 0.01**
Same	9.23	8.20	0.41
Rated More	6.89	12.04	0.36

Comparison	Logins (control)	Logins (expt)	P-value (2 tailed)
Rated Fewer	0.89	0.78	0.64
Same	1.36	1.52	0.78
Rated More	1.82	2.02	0.77

Table 2-10. Paired analysis of (above) the mean number of ratings the week after the newsletter and (below) the mean number of logins the week after the newsletter. Subjects in the control condition were not assigned to groups; this analysis uses post-hoc inferred group assignment as described above in the text. Significance testing is conducted with two-tailed paired t tests; these results should be treated as exploratory.

2.5.2 User Perceptions of the Social Comparisons

78.7% of subjects (211/268) participated in the survey that we launched one month after the email newsletter, including 104/134 subjects in the control group and 107/134 subjects in the experimental group. 50 women and 152 men took the survey (9 participants declined to identify their gender).

When asked if they liked receiving the email newsletter, subjects averaged 3.8 on a 5 point Likert scale ranging from strongly disagree (1) to strongly agree (5). There was no difference in responding to this question between the control and experimental subjects or between subjects with different comparison directions.

Subjects were asked to agree or disagree that "I didn't care" about the comparison in the newsletter. Overall, 48.1% of subjects agreed; there were no significant differences between the experimental groups or the directions of comparison. However, men were less likely to agree than women (40.3% vs. 68.2%, ChiSquare=5.33, df=1, p=0.02).

Subjects were asked if they agreed that "I wanted to do something to help increase my score". Subjects told they had rated fewer movies than others were the most likely to agree (53.8%), followed by those told they had rated about the same number as others (48.5%), or more than others (28.9%).

Those members who agreed that the newsletter made them want to do something to increase their score actually used the system more than those who disagreed. They rated more movies in the week after the manipulation (means 19.9 vs. 8.8, F=4.30, p=0.04). They also logged in slightly more in the week after the manipulation (means 1.80 vs. 1.50), but that difference is not statistically significant (F=0.27, p=0.60).

There were differences between men and women in how much they agreed that they wanted to do something to increase their score. Women were most motivated to agree when they were told they were the same as others (71.4%), while men were most motivated to agree when they were told they had rated fewer movies than others (68.4%). In a logistic regression model to predict whether a subject wanted to do something to increase his or her score, both the experimental group (p=0.05) and the interaction between experimental group and gender (p=0.02) were significant. Gender was not significant (p=0.13), although there is a trend that men (43.1%) were more motivated to agree than women (28.9%). See Figure 2-5 for a graph of the interaction between comparison condition and gender.

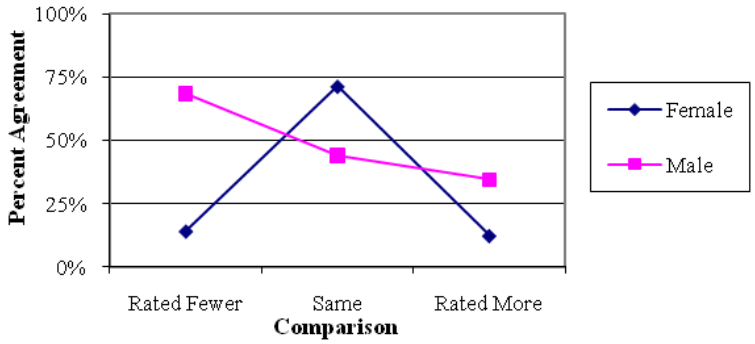


Figure 2-5. Percent of subjects agreeing that "I wanted to do something to help increase my score" by comparison condition and gender. While gender is not a statistically significant predictor of response, comparison condition and the interaction between condition and gender are both significant.

2.6 Discussion

RQ Activity. How does social comparison in an online community affect members' propensity to visit and contribute? While subjects who received an email message with the comparison manipulation were no more likely to click on one of the links or log in to the system, they were more likely to rate movies. Thus, while we do not find conclusive evidence that a comparison influences a member's overall interest in using the system, we do find evidence that it changes their focus within the system.

One important question this raises is whether or not shifting members' attention towards rating might cause them to do less in other areas of the system. We cannot answer this question definitively in our study, but we can give some preliminary data. Subjects receiving the comparison manipulation contributed fewer edits to the MovieLens database (editing 48 entries) compared to the control group (editing 118 movies). This is, however, not a statistically significant difference ($F=2.37$, $p=0.12$). Future work should look at whether the effects of social comparisons or other non-monetary incentives are inherently zero-sum, or if these features can instead boost overall levels of member activity.

We also found that subjects who were told they had rated fewer movies than others rated the most movies and changed their rating behavior the most in the week following the newsletter. One potential caveat to this result is that the marginal cost of providing ratings increases over time, as members find it increasingly difficult to find seen but unrated movies in the system (Harper 2005). However, we note that members who were told they had contributed fewer ratings didn't just rate popular movies. In fact, in the week following the manipulation, this group rated more rarely-rated movies per member (1.27) than any other group (the other three groups averaged 1.11). This difference is not statistically significant ($F=0.06$, $p=0.81$), but it does underscore the fact that these members were contributing ratings of value to the system.

RQ Perception. How does social comparison in an online community affect members' self-reported motivations to visit and contribute? We see from the behavioral data that subjects from all conditions were approximately equally likely to click on a newsletter link and visit MovieLens the week after the manipulation was made. Also, there was no difference across conditions in how well members claimed to like the newsletter. The interesting aspect of these

data is that there was no apparent negative side-effect of telling below-average members how they compare. In fact, 44% of these subjects agreed that "I didn't care" about the comparison, while only 8% agreed that they felt envious about other members. However, we remain cautious recommending our particular design for use in real systems; in a telephone interview before the study, one subject professed to feeling slighted that the newsletter said he was below average.

Subjects who were told they had rated fewer movies than others agreed more often than other subjects that the newsletter made them want to do something to increase their score. Interestingly, subjects who agreed that they wished to increase their score actually rated more movies. Thus, upwards comparisons might be seen as explicitly motivational in this case: subjects were aware of their motivation, and acted on it.

Men and women had statistically significant differences in how they perceived the comparison information. In general, men were more likely to say that they wanted to take action, and less likely to agree that "I didn't care" about the comparison. Just as interesting, women appeared to be most motivated by a message that their contributions were average, a result that would not have been predicted by any theories we know of. In fact, conformity theory (Bernheim 1994) would predict quite the opposite. We are unsure of the generality of this result, and we are hopeful that other researchers will investigate it further.

2.7 This Work in Context

The study reported in this chapter is part of a group of related work that is not in the scope of this thesis. This chapter builds on a prior study of user rating behavior in MovieLens (Harper 2005), where the goal was to better understand user motivations to contribute information to an online community. We found that users rated for both selfish reasons (e.g., increasing the quality of their recommendations) and altruistic reasons (e.g., helping other members' recommendations). This work also led us to the conclusion that different users have different motivations for acting, and that personalization technology could potentially be leveraged to coax additional effort from members.

We followed this study of user motivations with a larger study investigating ways of eliciting increased participation from members. The full study is presented in (Chen 2009). In this work, we expand the experimental mechanism reported here to also include a third newsletter variant that compares members on a "net benefit" score, derived from an economic notion of "utility". In this additional treatment, we found that telling members they were getting more "benefit" from the system than others were most motivated to contribute work of benefit to the community (such as updating the movie database).

2.8 Conclusion

In this study, we used email newsletters to tell members of an online movie recommendation site how they compared with other members in terms of movie ratings. In so doing, we established a social norm in a community where such a norm had been absent. We found that this type of comparison is potentially a powerful way to redirect members' attention – while members who received a comparison message rated more movies than members in a control condition, they were no more likely to click on links in the email newsletter or visit the site.

Online communities wishing to promote contributions of a certain kind may wish to display information that leads members to evaluate their level of contribution. While many Web sites display information about superstar users (such as with Amazon's "Top Reviewers" list), it is also possible to compare users with their peers in the system. In this way, users may be motivated by the presence of more attainable goals (Lockwood 1997). However, since our results also provide support for the notion that upward comparisons are the most motivational, systems may wish to adopt a "carrot on a stick" approach to keep goals just out of reach.

Our study has limitations. We have only presented short-term data regarding the effect of social comparison. Additional work is needed to determine whether the continuing presence of such a feature can lead to long-term behavioral changes. Also, while we presented survey data that shows significant differences between men and women in terms of their perceptions of online social comparisons, further work is needed to translate this result into useful design principles. Finally, our methods employed natural group selection – we honestly told members how they compared with the median user in their cohort – rather than random group selection. Thus, we

can present only correlational evidence concerning the varying impact of the different "directions" of a comparison, while a different methodology could reveal whether it is actually the *state* of being below average that leads to increased activity, or if it is comparison itself. It would also be interesting to investigate if there is an interaction effect between the subjects' actual standing and the direction of the comparison.

In future work, we hope to continue to investigate the use of non-monetary incentives in online communities. We are especially interested in two common design features which facilitate social comparison: leaderboards and contribution-based status levels. We are also interested in developing and evaluating personalization algorithms that find especially compelling comparisons for display by leveraging the system's knowledge of users' relationships, interests, and behavior. We hope that this research will lead to the development of tools that will help online communities improve, focus, or diversify contributions from their members.

Chapter 3

Talk Amongst Yourselves: Inviting Users To Participate In Online Conversations*

Vibrant online communities offer ways for people with common interests to connect and organize their contributions for common purpose. Many online communities, especially those providing member-contributed content, contain both an archive of domain information and a social space where members exchange information and interact (Ridings 2004). Typically, visitors first "lurk" at the periphery of the information space; later some of them become active participants in the social space (Nonnecke 2000).

Not all communities are equally successful. Some communities have a sufficient or excessive volume of posts, and may wish to encourage lurking over posting (Nonnecke 2000). Other communities die from lack of participation. Butler found that over 50% of a large and diverse sample of email-based groups failed to receive a single message over the course of a four month study (Butler 1999). These communities might be helped by the presence of additional posters, who will contribute content, or even by additional lurkers, whose visible presence (demonstrated by read counts, for example) can help encourage contributions from other users (Preece 2004).

MovieLens (http://movielens.org), the movie recommendation community that is the site of the research reported in this chapter, is one such community that would benefit from increased

---

* This chapter extends the work originally published as (Harper 2007b), co-written with Dan Frankowski, Sara Drenner, Yuqing Ren, Sara Kiesler, Loren Terveen, Robert Kraut, and John Riedl. The author of this thesis built, managed, and analyzed these experiments with technical and analytical contributions from Dan Frankowski. Yuqing Ren and Robert Kraut contributed some statistical analysis for both experiments. All authors contributed to the overarching experimental design.

posting and lurking. At the time of this research (Winter, 2005-2006), MovieLens had just launched discussion forums as a new feature. Members had been slow to adopt this feature, with only 19% of members having visited and 2% having posted. Because the system used data from the forums to generate social movie recommendations, the community would benefit from an increased volume and diversity of forum posts.

3.1 Personalized Invitations

How might MovieLens promote awareness of this new social feature, and subsequently encourage its use? While the design space for potential solutions is large, in this chapter we examine the use of text-based messages inserted into the site's interface to drive traffic.

Specifically, we explore the design and effectiveness of *personalized invitations*. Personalized invitations encourage members to visit or post in a discussion forum by augmenting the user interface to emphasize the presence of interesting content. The success of personalized invitations relies on intelligent computation to generate appropriate content for display, as well as appropriate presentation to maximize the visibility and potential effectiveness of the appeal.

There is reason to believe that displaying invitations to users may lead to action. A fundamental principle in human behavior is that people do things to minimize their behavioral costs or effort. Zipf (1949) identified this principle in the 1940s, using it to account for humans' tendencies to develop shorter words as the words become more frequent in a language (e.g., "television" compressing to "TV"), to communicate most with people who are close by, and to select the pie closest to the front of the freezer. This principle helps explain why people make decisions heuristically, rather than through a rational analysis of costs and benefits (Simon 1971), and why they use heuristic processing of persuasive messages rather the more systematic analysis of the evidence that a message presents (Chaiken 1989).

We believe that the use of personalization to tailor the content of an invitation to a particular user will improve on a non-personalized call to action. Prior work in recommender systems has shown that personalization can help users make decisions when faced with uncertainty (Hill 1995). E-commerce Web sites have used this knowledge to build personalized interfaces to

increase sales (Schafer 2001). In addition, Internet users claim to prefer personalized content to non-personalized content (Greenspan 2004).

3.2 Research Questions

There are three research questions that we address in this chapter:

> *RESEARCH QUESTION 1: PARTICIPATION. Do personalized invitations lead to increased participation?*

Our primary design goal is to increase members' awareness and use of the MovieLens forums. We measure the impact of personalized invitations on rates of viewing and creating forum posts.

> *RESEARCH QUESTION 2: ALGORITHMS. Do different algorithms for choosing the content of the invitation affect users' response rates?*

Though we believe that personalizing an invitation to a user's particular interests will improve its appeal, it is less clear how to design these personalization algorithms. Entities such as forum posts and users contain many dimensions along which they may be recommended. In this chapter, we describe and evaluate several algorithms inspired by social psychology theories that may be used to recommend entities for inclusion in an invitation.

> *RESEARCH QUESTION 3: SUGGESTION. How does the suggestion made in the invitation affect users' willingness to act?*

Personalized invitations might suggest that users act in different ways. In the case of MovieLens, they might ask that users read posts, reply to posts, or start new conversational threads. They might be designed to disclose more or less information about the recommended content. They might use different wordings. We explore these fundamental design decisions in this third research question.

3.3 Research Context

We have chosen to pursue this research through a field experiment in MovieLens.[3] At the time of this research, MovieLens had more than 100,000 registered users, and averaged over 2,000 unique visitors each month. In June 2005, MovieLens was augmented to include a discussion forum. MovieLens's forum (based on the open source mvnForum project) is a non-hierarchical, threaded conversation space with two main areas of conversation: one for talking about movies, and the other for talking about MovieLens. These forums are publicly visible, but posting requires (free) registration.

While the discussion forum has attracted dedicated users, the number of regular forum users is a small fraction of the total number of MovieLens users. Of the approximately 12,000 unique members who have visited MovieLens since the launch of the forums, only 19% have visited the forums, and only 2% have posted one or more messages. Of the people who read the forums, 88% are lurkers (people who read but do not post). Lurkers may help increase participation, by motivating posters (Preece 2004). MovieLens shows the presence of lurkers by displaying read counts next to threads, and by showing a list of online users on the front page of the forum.

One distinguishing feature of the MovieLens forums is its ability to recognize and understand references to movie titles in posts (Drenner 2006). Recognized movies are hyperlinked and recommendation information is presented alongside the forums interface. Figure 3-1 shows a post with two linked references. Movie recognition enables several of the personalization algorithms described in this research below.

[3] See chapter 2 for more information about MovieLens.

Figure 3-1. An example forum post in MovieLens with two movie references. Each reference to a movie title is hyperlinked in the text of the post; widgets on the side panel allow users to rate and bookmark these movies.

Movie recognition is enabled by two custom tools, the movie-linker and the movie-finder. The movie-linker allows members to manually insert a movie reference into the text of their post, using an AJAX-based completion interface for quick movie searches in-place. The movie-finder automatically inserts linked references when titles are found in the content of the post. Our approach to automatically discovering movie entities in post text is related to prior work done in the field of natural language processing (e.g. (McDonald 1996)), although our focus has been on improving the usability of named entity recognition systems. The movie recognition interface and architecture used in MovieLens are described in more detail in (Drenner 2006).

3.4 The Design Space

While the potential design space for personalized invitations is large, we consider two main aspects: the algorithms used to select content for display, and the nature of the suggestion made to the user. We do not investigate other potentially interesting aspects, such as visual design or the timing of the presentation.

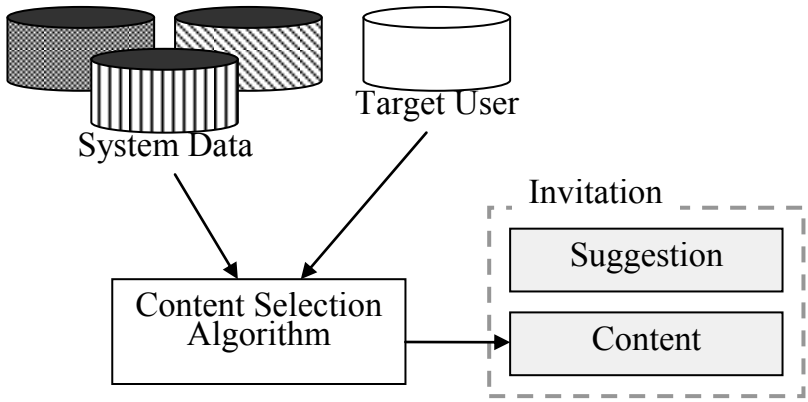


Figure 3-2. A conceptual diagram of the design space for personalized invitations. We construct an invitation by combining the output of a content selection algorithm with a suggestion for user activity.

Figure 3-2 gives a conceptual overview of the design space we consider. An invitation, shown in the lower-right corner, has two components. *Content* refers to a system entity (e.g., a forum post, a movie, or a user name) that comprises the focus of the invitation. *Suggestion* refers to the request made of the user, and the presentation of the request. While a suggestion is hard-coded into the invitation's design, content is dynamically generated. To find content, a *content selection algorithm* searches through system data for entities to display, possibly based on the target user's preferences or usage history.

Figure 3-3 shows an example of a personalized invitation on the MovieLens home page. The content selection algorithm (*Rare Rated*, see below) has found a movie entity (*Nicholas Nickleby (2002)*) that the target user has rated, but that few other users have rated. The invitation suggests that the user start a new thread about this movie, stating that the user is "one of only a few" members who is able to take this action.

Figure 3-3. MovieLens home page with one of the invitation variants from experiment 1 (magnified).

In the following section, we describe the first of two experiments on the effect of personalized invitations. We propose and evaluate several algorithms for intelligently choosing content for display, and examine the effectiveness of various suggestions for action.

3.5 Experiment 1: Invitations to Post

Our first experiment tested the overall effectiveness of invitations that ask users to post messages in the MovieLens discussion forum. Invitations designed for this experiment asked a user to either start a new thread about a specific movie or to reply to an existing post. We developed several algorithms for recommending posts or movies to supply the topic of the invitation.

People universally seem curious to learn about themselves and how they are similar to and different from others (Festinger 1954), suggesting invitations that emphasize ways that a person differs from other members of the community. Prior work by Ludford et al. (2004) showed that participants in a discussion forum were motivated by knowing what unique perspective they can

bring to a group, and that knowledge increases their participation. Thus, a target user's uniqueness is something that designers might choose to emphasize in an invitation. For example, in MovieLens, we might point out that a user can write about a movie that few other members have seen, or that a user's opinion about a movie is unique. We extend the work of Ludford et al. by building algorithms that automate the selection of content that emphasizes uniqueness, and by evaluating the algorithm's potential for enhancing the effectiveness of invitations to participate.

Extrapolating from the law of least effort (Zipf 1949), we can expect to elicit more contributions if we minimize the effort to participate. One way to vary the effort involved in writing a post is to either ask users to reply to an existing post, or to ask them to start a new topic of conversation. We hypothesize that replying to a post is a lower cost action than starting a new thread. By offering a specific post to reply to, we constrain the space of things to talk about. We test this idea by comparing the effectiveness of invitations asking users to start a new thread with invitations asking users to reply to an existing post.

3.5.1 Content Selection Algorithms

We recommended two types of content that have the potential to make an invitation compelling in MovieLens – movies and forum posts. The trick is choosing the right movie or the right post.[4]

Recommending Movies. The MovieLens movie recommendation system offers many plausible ways for choosing movies to include in an invitation, such as: recently-rated movies, highly-rated movies, poorly-rated movies, rarely-rated movies, highly-recommended movies, and movies that are influential to a user's recommendation model. Cosley et al. (2006) investigated several of these algorithms to find movies for users to edit for accuracy in MovieLens. They found that the choice of algorithm had a large effect on users' willingness to help.

---

[4] While there is a variety of personalized and non-personalized methods for selecting content to display, we have chosen to explore only personalized methods. While non-personalized methods – for example, choosing a random movie, or the most recently written post – are plausible candidates for inclusion, we believe these methods are less interesting and probably less effective than their personalized counterparts. While we will not explore non-personalized methods in this work, future would could better inform our understanding of the trade-offs involved.

Based on the best algorithm from (Cosley 2006), we developed *Rare Rated*, an algorithm designed to choose a movie the user will be able to write about in the forums. This algorithm searches the target user's ratings history for rarely-rated movies – defined as movies with fewer than 250 system-wide ratings – and picks one to display at random. As a baseline algorithm, we also developed *Rated*, an algorithm that chooses a movie for display at random from the target user's rating history.

Recommending Posts. There are a variety of ways to choose forum posts for inclusion in an invitation. One can imagine many plausible algorithms for this task, ranging from collaborative filtering to content-based algorithms.

One approach to choosing forum posts for display relies on what we call *indirect recommendation*: the use of knowledge in one domain to recommend items in another domain. In MovieLens, we can choose posts for display in an invitation by combining our knowledge of users' movie tastes with our knowledge of which movies have been mentioned in the forum.

There are several ways to use indirect recommendation to choose posts for display. For example, we might recommend posts that mention movies a user is *familiar* with, that mention movies a user *likes* or *dislikes*, that mention movies the poster and the recipient of the invitation *agree* or *disagree* about, or that are written by authors who are *similar* or *dissimilar* to the recipient of the invitation.

For this experiment, we developed *Disagree*, an algorithm which recommends forum posts for users to reply to. This algorithm searches for posts referring to movies the target user has rated. Each post is scored on the maximum difference between the user's rating and the post author's rating of a movie referenced in the post. The system chooses the post with the largest disagreement score for display in the invitation. We developed the baseline algorithm *Random Rated Post*, which randomly chooses among all posts which refer to a movie the target user has rated.

3.5.2 Invitation Variants

We designed four variants of a personalized invitation in this experiment. These variants shared the same basic visual design: each contained a link to a movie the user had rated and a link to a

page for writing posts. Figure 3-3 (above) shows a screenshot of the MovieLens home page, with an example invitation highlighted. The four invitation types were:

1. *New Thread + Random Rated*. This invitation asks users to start a new thread about a movie. The system randomly chooses a movie the user has rated for display in the invitation. This invitation is worded as follows:
   *"Tell Others About [MovieTitle] – **You have rated the movie** [MovieTitle]. Post and share your thoughts."*

2. *New Thread + Rare Rated*. This invitation asks MovieLens users to start a new thread about a movie, emphasizing their ability to make a unique contribution by selecting content with the Rare Rated algorithm. We modify the bolded portion of the previous wording to state "you're one of only a few MovieLens members who have rated...".

3. *Reply + Random Rated Post*. This invitation asks users to reply to a post in the forums. The system searches for posts that refer to movies the user has rated, and chooses one at random. This invitation is worded as follows:
   "Tell Others About [MovieTitle] – A recent post mentions the movie [MovieTitle], which you rated [YourRating]. Post and share your thoughts."

4. *Reply + Disagree*. This invitation asks users to reply to a post in the forums, emphasizing the potential of the user to provide a new perspective to the discussion by selecting content with the Disagree algorithm. We modified the wording above to include the sentence "We think you disagree with the poster about this movie."

3.5.3 Methods

Our subjects were drawn from the pool of new and returning MovieLens members during 17 days in December 2005. We randomly chose 1/5 of the MovieLens members as the control group; the remaining 4/5 were assigned to the experimental group.

The control group continued to use an unmodified MovieLens interface for the duration of the study. The unmodified interface includes several paths to the forums: every page contains a header with a link to the forums, the front page contains a list of links to three recent forum

posts, recently mentioned movies in search results link to forum posts, and each page devoted to a movie links to recent relevant forum posts.

The experimental group used the same MovieLens interface as the control group, with the addition of an invitation at the top of three pages: the MovieLens home page, the forums front page, and the page listing the threads in each forum. Each time a user in the experimental groups viewed one of these pages, one of the four invitation types was chosen at random, personalized, and displayed.

3.5.4 Results

Overall Participation. The experimental group consisted of 1,611 users who logged in a total of 6,392 times. The control group consisted of 410 users who logged in a total of 1,552 times. 19% of the experimental group viewed at least one post, as compared with 20% of the control group. 3% of the experimental group posted at least once, as compared with 1% of the control group. 9% of the experimental group clicked on an invitation at least once. See Table 3-1 for a summary of the differences between the control group and the experimental group.

Metric	Control Group	Experimental Group	P-value
# Logins	3.79	3.97	0.85
# Posts Viewed Per User	11.87	15.24	0.40
Fraction of Users: Viewed Post	0.20	0.19	0.46
# Posts Written Per User	0.04	0.17	0.07
Fraction of Users: Created Post	0.01	0.03	0.07

Table 3-1. A comparison between users in the control group and users in the experimental group across a variety of metrics. Statistical significance is tested using the non-parametric Wilcoxon test.

Though users in the experimental group averaged more post views than users in the control group, the difference is not statistically significant. This is explained by the presence of a few users in the experimental group who viewed thousands of posts. However, users in the experimental group were substantially more active in their posting behavior, a difference that trends towards statistical significance (p=0.07).

Finding 1. Invitations asking users to post increased posting behavior, but not viewing behavior.

New Thread vs. Reply. Table 3-2 summarizes the total number of invitation views, clicks, and posts over the duration of the experiment, across all users.

Reply invitations were clicked 237 times, while *New Thread* invitations were clicked 65 times. This difference is statistically significant, based on a logistic regression model built to predict whether or not a user clicked on the invitation ($z=4.84$, $p<.01$).

	Views	Clicks	Posts
New Thread	2397	17	9
New Thread + Rare Rated	2500	48	5
New Thread Subtotal	4897	65	14
Reply	2480	65	1
Reply + Disagree	2446	172	5
Reply Subtotal	4926	237	6

Table 3-2. Total views, clicks, and posts by invitation type. Only posts directly caused by an invitation are counted as posts. It is impossible to know exactly how many posts were indirectly caused by the presence of the invitations.

However, *New Thread* invitations directly led to 14 posts, while Reply invitations only led to 6 posts. There was not enough data to detect a statistically significant difference.

Finding 2. Users were more likely to click on invitations asking them to reply to a post than on invitations asking them to start a thread.

Table 3-3 summarizes the number of clicks per user, grouping invitations by algorithm type. Across both the *New Thread* and *Reply* invitation types, users clicked on invitations with content emphasizing their uniqueness (*Rare Rated* and *Disagree*) more frequently than on invitations using the baseline algorithms (*Random Rated* and *Random Rated Post*). The effect of these algorithms on users' click rates is positive and statistically significant, tested with a logistic regression model built to predict whether or not a user will click on an invitation ($z = 3.52$, $p < .01$).

	Low Uniqueness	High Uniqueness
New Thread	.0036	.0099
Reply	.0137	.0401

Table 3-3. Invitation clicks per user by invitation type. High Uniqueness counts the *Rare Rated* and *Disagree* algorithms together, while Low Uniqueness counts the two baseline algorithms together.

There was no observable difference in this experiment between the high uniqueness and the low uniqueness algorithms in terms of posting behavior – both algorithms led to 10 posts.

Finding 3. Invitations emphasizing uniqueness led to more clicks, but not more posts.

3.5.5 Discussion

Research Question 1: Participation. Compared with the control group, users who received invitations posted more, but did not read more, as shown in Finding 1. It also appears that invitations increased the overall activity level in the forum. There were on average more posts per day during the experimental period (19.5) than during the two weeks immediately before (13.7) and after (12.5) the experiment. However, this difference is not statistically significant.

Did invitations spark valuable content, or did they lead to lower quality posts? Although quality is subjective and difficult to measure, we propose that one sign of a post's quality is whether it leads to further conversation. Of the 293 posts written during this experiment, the 20 posts directly caused by invitations have been replied to an average of 1.7 times, as compared with an average of 0.8 replies for the remaining 273. 75% of the posts directly caused by invitations received at least one reply, as compared with 64% of the other posts. Of the 23 threads started during the experiment, there is not a statistically significant difference between the threads that were started from invitations and threads that started "naturally" in terms of subsequent views and replies (views: $p = 0.91$; replies: $p = 0.84$). Thus, invitations led to posts that were read and replied to.

The MovieLens pages that contained invitations in the experimental condition were viewed 27% more often by subjects in the experimental condition than by subjects in the control condition. Recall that in this study, each time a user views a page with an invitation, the content of the invitation is changed. We hypothesize that users viewed these pages more often because

they were intrigued by the presence of the invitations, and acted to explore the different invitation types.

There were also undesirable outcomes. In one instance, a new post author created a thread about a movie that had already been discussed elsewhere – an action which more senior members noticed and corrected. Lampe and Johnston suggest that old and new members alike benefit when new users spend time learning the customs of the community by reading posts before they first post themselves (Lampe 2005). Thus, in our second experiment, we adjusted our suggestions to ask users to read posts, in order to give new users a chance to acclimate to the community before being thrust into the position of posting.

Research Question 2: Algorithms. Finding 3 shows that the algorithm for choosing content does matter. For both invitation types, personalization that emphasized uniqueness made a positive difference to response rates. One user commented on the personalization:

> *It's kind of interesting to see what movies pop up... and that I am "only one of a few" to have rated the movie.*

The *Reply + Disagree* invitation variant was especially effective at generating clicks. It led to nearly three times as many clicks as the *Reply + Random Rated Post* variant.

Research Question 3: Suggestion. Finding 2 presents conflicting results concerning fundamental issues of presentation. While users were more likely to click on an invitation asking them to reply to a post, they were more likely to write a post when asked to start a new thread. Although the post-writing result was not significant, it remains a surprise. Replying to a message intuitively feels like a lower-cost action than writing a new thread, because the topic of conversation is more constrained. However, it might actually be the case that new users to the forums found it easier to start a new topic: they could simply write their thoughts without having to read lots of posts to understand the context of their contribution.

Alternately, based upon feedback from MovieLens users, this data may be explained by a shortcoming in the algorithms used to find posts for the Reply invitation variants: we asked users to reply to potentially very old posts (up to six months old). One user wrote:

> *Don't ask me to reply to a post that's more than a month old. (Better still, two weeks.)*

In the next experiment, we address this concern by filtering recommended content for recency.

Next Steps. One of the most significant effects from experiment 1 was the increased activity generated by the *Disagree* algorithm. This invitation variant led to nearly three times as many clicks as the baseline invitation asking users to reply to a post that mentioned a movie. It may have been successful by creating a sense of curiosity about other users: "who do I disagree with about this movie, and why do our opinions differ?" We hypothesize that the *Disagree* algorithm made more visible the social nature of the discussion forum. Therefore, in experiment 2, we explore other invitations which emphasize the social nature of the forum.

3.6 Experiment 2: Invitations to Read

We revisited our design of personalized invitations in our second experiment, based on the experimental findings from the first experiment and feedback from users. We wished to study other aspects of invitation design. First, we wished to design invitations that bring users into the forums as readers, to give new users the chance to explore before they post. Second, we wished to understand other social dimensions of algorithms for choosing content.

People use familiarity to reduce cognitive effort. Habit enables people to make repeated decisions without having to think through the alternatives each time. People reduce the possibility of incorrect or suboptimal decisions by seeking familiar sources. Studies show that "mere exposure" (Zajonc 1968) can explain people's attraction not just to other people, but also to music, art, and food. Repeated contact over time causes a person to like other people and objects more. These observations suggest that invitations presenting familiar items or people will be more effective than invitations that present unfamiliar items or people. Thus, in this experiment, we either chose familiar or unfamiliar content for display in the invitation.

How much should we reveal about the content of the invitation? The invitation might be more or less specific about the recommended content. For example, we might choose to display an entire post, or we might simply show the subject line. The study of the psychology of curiosity (e.g., Loewenstein 1994) posits that motives often stem from incongruities or information gaps in the world. If we display less information about a post, will that enhance users' curiosity? Or,

will displaying less information cause us to omit details that are especially compelling to the user, such as a the title of a favorite movie, or an especially intriguing username. We test these ideas by varying the specificity of the information displayed in the invitation.

3.6.1 Content Selection Algorithms

Because we wished to emphasize social aspects of the discussion forum, we exclusively chose post content for inclusion in invitations. We especially were interested in examining the power of emphasizing the social nature of the forums. Thus, we developed two algorithms, one that is intended to emphasize social content, and one that is intended to emphasize non-social content.

Our social approach to choosing posts relies on users' history of viewing the discussion forum. By tracking which users have viewed which conversation threads, we can compute a familiarity score from any user to any other user. The *Relaviz* system (Webster 2006) used a graphical display of the asymmetrical familiarity between pairs of users to encourage participation by connecting lurkers and posters. We use a similar computation to discover relationships between users, but use the output to recommend posts written by *familiar* or *unfamiliar* users. The recommendation of familiar users is only possible for MovieLens members who have previously viewed forum posts.

We call our algorithm that implements this idea *Familiar Poster*. This algorithm chooses among posts written in the last week based on whether or not the target user has previously viewed posts by that author five or more times. Requiring fewer views weakens users' familiarity with post authors, while requiring more views further restricts the set of posts to recommend; we looked to balance these two factors. We developed a corresponding baseline algorithm, *Unfamiliar Poster*, to recommend posts written by authors the target user has seen fewer than five times.

To test whether any effects of familiarity were due to the social effects of making other users visible, we also developed a non-social approach to recommending posts with familiar or unfamiliar content. Just as we may believe that users are attracted to the forum by the presence of other users, we may also believe that users are attracted to the forum by the movies that are discussed.

Our algorithm for recommending posts on the basis of their movie content is called *Familiar Movie*. It returns the set of posts written in the last week that reference one or more movies the target user has rated. We developed a corresponding algorithm, *Unfamiliar Movie*, that returns the set of posts that mention movies the target user has not rated. While a single post may mention both rated and unrated movies, in the invitation we only show the movie chosen by the algorithm.

3.6.2 Invitation Variants

We made several changes to the overall design of invitations in MovieLens in preparation for our second experiment. The biggest change was that the invitations no longer asked users to post, but instead recommended that users visit the forums to read a post.

Some subjects thought that getting many different invitations per session (the time between login and logout) was too demanding in the first experiment. Thus, we designed these invitations so that users would view the same invitation throughout their session. Clicking removed the invitation from the user's interface for the duration of the session. Users could also explicitly hide the invitations for the remainder of the session using a "hide" link. Users who had clicked "hide" in three or more sessions were given a "hide forever" link that permanently removed invitations from their view when clicked. 65 out 1,917 users clicked the "hide" link at least once; 11 users chose to permanently hide invitations, of whom 6 are repeat forum posters.

Figure 3-4 shows a sample invitation. The most prominent visual change to the invitation design was the inclusion of the subject line of the recommended post, as well as up to 125 characters of preview post text. To avoid confounding our experimental manipulation, we stripped references to movies from the preview text. To do this in a natural way, we searched the post for the first phrase beginning with 125 characters without a movie reference. Failing this, we used the first 125 characters of the post, replacing movie titles with the string "...".

Figure 3-4. A sample invitation from the MovieLens home page during experiment 2. The post preview text is used with the author's permission.

There were a number of variants of this basic invitation, as described below.

Specificity Variations. We call invitations that contain more information about the recommended post more specific. Users were either shown the name of the author of the recommended post, or they were not. Similarly, users were either shown the name of a movie referenced in the recommended post, or they were not. We randomized the order of the name and the movie in invitations that displayed both. Named entities were shown in bold, green text to draw attention to their presence. We added small icons next to these named entities to further distinguish their presence. We adjusted the two icons to be the same size and approximately the same level of luminosity.

Familiarity Variations. We also varied whether or not the invitation's named entities were familiar to the user. In the case of a movie, we varied whether we used the Familiar Movie or the Unfamiliar Movie algorithm to select the content. In the case of a post author, we either used the Familiar Poster or the Unfamiliar Poster algorithm. We randomly chose among the intersection of these sets of posts for display.

Credibility Variations. We designed four different wordings to use as the opening phrase in the invitation. Two of the wordings were designed to exploit the credibility of the MovieLens recommender system by implying that "our system" is making a recommendation for the user.

Our system predicts you'll enjoy the following new post

Our system recommends the following new post for you

The other two wordings did not imply that it is the system making the recommendation, but left the source of the recommendation unspecified:

We think you'll like the following new post

Check out the following new post

3.6.3 Methods

The second study took place for 17 days in February and March, 2006. All MovieLens users who logged in during this period and who met our entrance criteria were exposed to the experimental manipulation. We did not include a control group because we had established the efficacy of the invitations in the first experiment, and because we were primarily interested in measuring the effect of experimental manipulations.

For the purposes of analysis, we split user sessions in this experiment into two groups:

1. *ForumHistory*: users who have visited the forums in a previous session

2. *NoForumHistory*: users who have never visited the forums

Users in *ForumHistory* received the full set of invitation variations, while users in *NoForumHistory* received all variations except for those that require a familiar post author. Users were moved from group *NoForumHistory* to group *ForumHistory* after their first visit to the discussion forum. Users who could not receive the full set of invitation variations did not receive any invitation, and these sessions are not included in the analysis. For example, users who had not rated any of the movies that had been referenced in the past week of forum posts could not receive the familiar movie variation, and were excluded.

Due to the exploratory nature of our study and a limited pool of users, we used a half fractional factorial design, a design that gives us main effects and lower-order interactions, but sacrifices higher-order interactions. We chose 8 out of 16 runs of a full, four-factor, two-level, factorial design (Box 1978). The factors are: showing a movie title, familiarity with the movie title, showing a user name, and familiarity with the user name. The levels are: *ForumHistory* and *NoForumHistory*.

Our data are nested by nature, since each user can have multiple sessions. We analyzed the data using hierarchical linear modeling (HLM) techniques (Bryk 1992) to control for random effects at the user level. We then determined the significance of our results using HLM analyses.

3.6.4 Results

2,415 users logged in to MovieLens during experiment 2. 1,917 of these users participated in the experiment by viewing an invitation. 10.5% of the participating users clicked on an invitation at least once. Table 3-4 summarizes the number of users, sessions, and invitation clicks that took place during the study.

	Unique Users	Sessions	Clicks
ForumHistory	704	3225	193
NoForumHistory	1213	3012	74
Total	1917	6237	267

Table 3-4. Number of users, sessions, and invitation clicks in experiment 2 by user group.

Overall, users in *ForumHistory* clicked on an invitation in 6.0% of their sessions, as compared with users in *NoForumHistory*, who clicked in 2.5% of their sessions. This effect is significant in our model ($p < .01$). This effect might be expected, given that users in *ForumHistory* had previously expressed interest in the forum while users in *NoForumHistory* had not.

Credibility. Table 3-5 summarizes the percentage of invitations that were clicked by wording. The two wordings that contained the phrase "our system" led to users clicking on the invitation 50% more often than with the other two, a significant effect in the regression model ($p < .01$). Overall, the most effective wording was "Our system predicts you'll enjoy the following new post," with 99 clicks. The least effective wording was "Check out the following new post," with 51 clicks. The message "We think you'll like" appears to represent a middle ground (74 clicks) by making implicit reference to the system.

Finding 4. The wording of the invitations mattered. Invitations were more effective when they were worded to emphasize the credibility of the recommendation.

Wording	% Clicked
Our system predicts you'll enjoy the following new post	5.4
Our system recommends the following new post for you	5.0
We think you'll like the following new post	4.1
Check out the following new post	2.6

Table 3-5. Percentage of invitations clicked for each wording, across all groups. The first two wordings emphasize the credibility of the source of the recommended post, while the second two wordings leave the source of the recommendation ambiguous.

Specificity. Table 3-6 summarizes the number of invitations clicked in each specificity condition. Simply showing the name of a forum poster in the invitation has a positive and significant statistical effect on click rates ($p < .01$). Importantly, though, this effect only applies to users in *ForumHistory*. Consequently, the interaction effect between user group and showing the poster's name is also statistically significant in our model ($p < .01$).

Finding 5. Invitations were more effective when they showed the name of a post author, but only for users who had previously visited the forum.

Showing the title of a movie had ambiguous effects and is not statistically significant in our model. It actually slightly depressed the click rate for users in *NoForumHistory* (from 2.5% to 2.4%), although it increased the click rate for users in *ForumHistory* (from 5.6% to 6.3%).

Finding 6. Invitations were not improved or worsened by displaying the name of a movie.

	Movie Shown?		Name Shown?	
	No	Yes	No	Yes
ForumHistory	5.6% 89/1582	6.3% 104/1643	4.4% 71/1615	7.6% 122/1610
NoForumHistory	2.5% 37/1480	2.4% 37/1532	2.5% 37/1505	2.5% 37/1507
Total	4.1%	4.4%	3.5%	5.1%

Table 3-6. Specificity Results. Percentage of invitations clicked per session (and raw numbers) across specificity conditions and user groups. More specific invitations are more effective, but only for users that have previously visited the forums.

Familiarity. Table 3-7 summarizes the number of invitations clicked in each familiarity condition. We failed to find evidence that our personalization algorithms for detecting familiar

movies and forum authors improved the success of our invitations. While we initially saw evidence that suggested showing familiar movie titles increased click rates (from 4.0% to 5.4% across groups), further analysis showed that our findings were confounded. Whether or not a user was familiar with a movie is correlated with the total number of times the movie has been rated in MovieLens (correlation = .448). Thus, it is impossible to know if the increased click rate is due to the user's familiarity with the movie or its overall popularity. When both factors are included in our model, neither is significant.

Likewise, we do not find evidence that showing familiar user names affects click rates. While it is the case that click numbers were slightly higher when a familiar name was shown (8.3% click rate when the name was familiar vs. 6.9% when the name was unfamiliar), the effect is not statistically significant in our model.

Finding 7. Invitations were not improved by familiarity-based personalization.

	Familiar Movie?		Familiar Name?	
	No	Yes	No	Yes
ForumHistory	5.4% 44/813	7.2% 60/830	6.9% 55/798	8.3% 67/812
NoForumHistory	1.6% 12/753	3.2% 25/779	*N.A.*	*N.A.*
Total	3.6%	5.3%	6.9%	8.3%

Table 3-7. Familiarity Results. Percentage of invitations clicked per session (and raw numbers) across familiarity conditions. While click-through rates are slightly higher for invitations containing familiar entities, the differences are not statistically significant.

3.6.5 Discussion

Research Question 2: Algorithms. We failed to show statistically that our algorithms for choosing posts on the basis of entity familiarity had an effect on users' rates of clicking through to the discussion forum. Perhaps this finding (Finding 7) points to a fundamental tension between familiarity and curiosity: while familiar people and items may be more comfortable and liked, unfamiliar people and items may heighten a target user's curiosity. For example, we might hypothesize that some forum users would be more inclined to click on an invitation with an unfamiliar user name ("who is this?") while other users might be more inclined to click on an invitation with a familiar name ("I remember you, you were interesting").

We are not sure the degree to which this finding might generalize. It is possible that there are types of content or other domains for which familiarity algorithms are more useful. For example, in domains where the volume of traffic is much higher than ours, choosing familiar users could make the community feel more intimate, encouraging users to be social.

Research Question 3: Suggestion. Finding 5 and Finding 6 taken together show that in MovieLens, showing user names helped increase clicks, while showing movie titles did not. Why is this the case? One possible explanation is that showing the name of a poster emphasizes the social nature of the discussion forums. The perceived value of visiting the forums is enhanced by emphasizing features that are not available from the movie recommendation interface: the written opinions of other users.

3.7 Conclusion

We have investigated the usefulness of *personalized invitations*, an intervention designed to increase participation in an online discussion forum. The results from this research can be used by designers who wish to increase participation in a discussion forum. Intelligent algorithms can relate users in the community with one another as a way of breaking the ice and encouraging new relationships. And new interfaces can emphasize social features such as the names of other users and recent post text to emphasize the presence of interesting people and discussions.

Our two experiments were exploratory in nature: we evaluated many design options over a short time period, possibly at the cost of detecting statistically significant differences in some cases. In these explorations, we found that invitations had an immediate impact in MovieLens over the short-term, causing users to write and view more posts. It would be interesting to explore the effect of invitations in longer experiments. We also found that invitations had a low average click-through rate. Follow-up work might explore ways of timing the delivery of invitations: for example, invitations might be particularly effective at moments when the user has just completed a task, and is deciding on the next activity.

We experimented with several algorithms for choosing the content of invitations, with varying success. The design of these algorithms was influenced by theories from social psychology:

uniqueness and familiarity. While uniqueness turned out to be a useful principle in our algorithm design, familiarity did not, perhaps confounded by users' curiosity to learn about new things. It is possible that familiarity-based personalization algorithms would be more effective in other domains, especially larger communities where it is more difficult to locate content written by familiar users.

Overall, we found that invitations emphasizing the social nature of the discussion forum increased user activity, while invitations emphasizing other details of the forum were less successful. While showing the name of a movie did not greatly increase user interest, showing the name of a post author did. *Disagree*, the algorithm we consider the most successful in this work, emphasized the presence of social interaction and related the target user directly to a post author. The forums are a social space; users were more drawn to them when the social aspects were emphasized.

Chapter 4

Predictors of Answer Quality in Online Q&A Sites*

User contributions continue to generate increasing amounts of rich online content. One relatively recent manifestation of this trend is question and answer (Q&A) sites – places where users ask questions and others answer them.

South Korean Internet portal Naver (http://naver.com), illustrates the potential of Q&A sites. As of July, 2007, Naver handled 77% of internet searches originating in South Korea, dwarfing worldwide leaders Yahoo (4.4%) and Google (1.7%) (Sang-hun 2007). One of the reasons for this disparity is the relatively small Korean language corpus available for crawling. To address this shortcoming, Naver built a Q&A site called Knowledge iN that encourages users to type questions for others to answer, rather than relying on search results (Chae 2005). Since their 2002 launch of Knowledge iN, Naver has accumulated 70 million questions and answers, and continues to receive over 40,000 questions and 110,000 answers per day (Sang-hun 2007).

Similar sites are now common worldwide. Yahoo now offers Q&A sites localized to 26 countries.[5] As of December, 2007, Yahoo Answers has attracted 120 million users worldwide, and has 400 million answers to questions (Leibenluft 2007). Yahoo incorporates their Q&A data into their search results. Although Google closed their U.S. Q&A site in 2006, they rejoined the trend with new services in Russia and China in 2007.

---

* This chapter extends the work originally published as (Harper 2008), co-written with Daphne Raban, Sheizaf Rafaeli, and Joseph Konstan. The author of this thesis managed and analyzed the experiment; he and Joseph Konstan each wrote half of the questions used in the experiment. All authors contributed to the overarching experimental design.

[5] According to http://answers.yahoo.com, as of 1/7/2008

Given the increasing competition for users' questions and answers, different designs have emerged. Some sites allow anyone in the community to answer questions, while others have individual "experts" filling that role; some charge askers and pay answerers, while others use leaderboards, points, or stars to encourage answering. Design decisions such as these are likely to have a large impact on the type and volume of questions asked, as well as the quality and responsiveness of the answers (and corresponding value to "social search" system providers) (Raban 2008), yet we still know little about the specific effects of these decisions.

From a user's perspective, it may be unclear which sites to turn to for high quality, friendly, or responsive answers. For users who have not yet joined a Q&A community, it is not obvious which questions to ask rather than search for. In Q&A sites that require payment, it is not obvious whether spending more money will earn a better answer. And it is possible that different types of questions and different rhetorical strategies will receive different responses, depending on the community. In general, it is unclear what approach to selecting and using a site will yield the most useful results.

In this chapter we seek to answer some of these questions through a comparative, controlled field study of responses provided across several online Q&A sites. We report both on quantitative data regarding site performance, and qualitative observations that illustrate several interesting and nuanced characteristics of these sites.

4.1 Question and Answer Sites

For the purposes of this study, a question and answer (Q&A) Web site is purposefully designed to allow people to ask and respond to questions on a broad range of topics. Fundamentally, all Q&A sites offer some interface designed for asking and answering questions. Commonly, users will be asked to categorize their question in some way, to route the question to answerers. Also, most Q&A sites offer an interface for searching and browsing, often organized by question status (e.g. "open", or "closed"). Yahoo Answers presents an example of such an interface (see Figure 4-1).

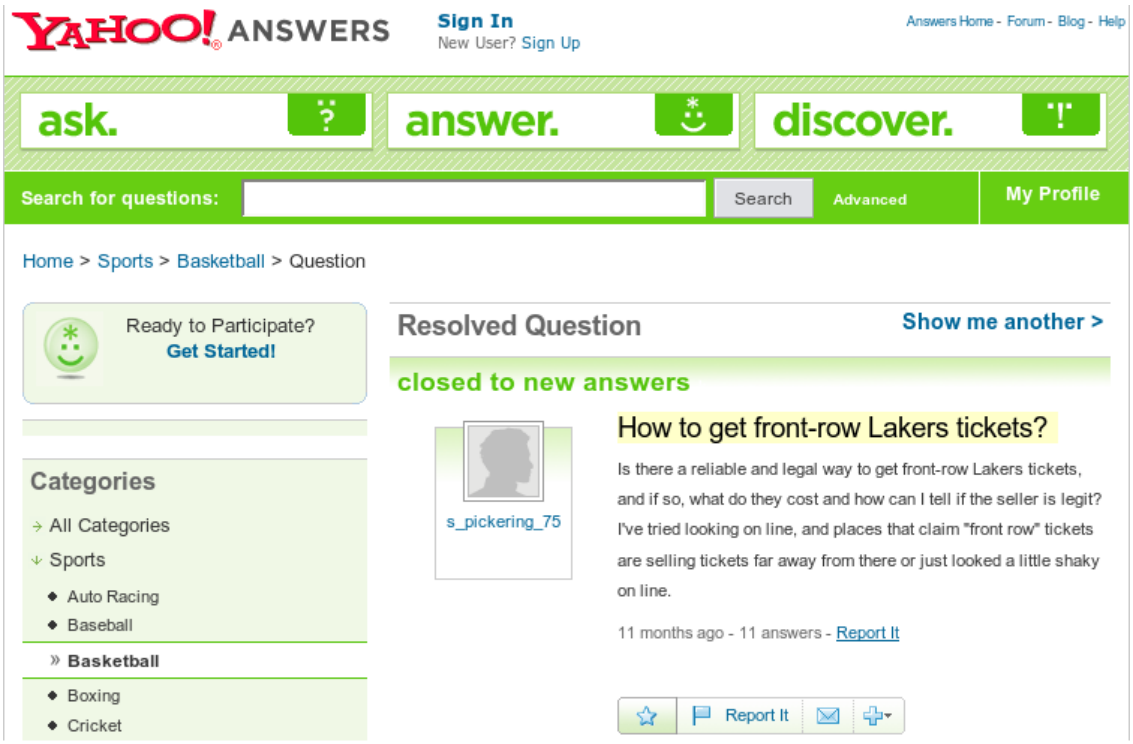


Figure 4-1. A question in Yahoo Answers.

4.1.1 Three Types of Q&A Sites

From the research literature and our observations and use, we identify three types of Q&A sites: "digital reference services", "ask an expert services", and "community Q&A sites." All three of these site types are in common use today. While they are all built to help users ask and answer questions, they achieve this goal in very different ways:

Digital reference services represent the online analogue to library reference services (Pomerantz 2004). Traditional library reference services employ expert researchers[6] to help people find useful information. Today, many public libraries have added digital reference services, such as the New York Public Library's "Ask Librarians Online" (http://www.nypl.org/questions/). Digital reference services typically use basic tools for online communication: in one survey, 71.4% elicited questions via a web form, and 80% respond to questions via email (White 2001).

[6] In the U.S., reference services are typically staffed by librarians with a master's degree in library science.

Digital reference services rely on specific people performing specific tasks (such as question routing, researching, and answering) as well as highly-constructed workflow (e.g., by using issue tracking software for tracking open/closed questions). Thus, this type of reference service maintains the library's organized and structured model of question answering; some researchers have argued that it is these "clearly defined policies and procedures that are well understood by all the participants" that are a model for success (McClennen 2001).

Ask an expert services[7] represent a first step, technologically and socially, away from the structure and formality of digital reference, while retaining the overall goal of providing quality question answering service. These services are staffed by "experts" (of varying credentials), often in a relatively circumscribed topic area, such as science (e.g. at "MadSci Network", http://www.madsci.org) or oceanography (e.g. at "Ask Jake, the SeaDog", http://www.whaletimes.org). Ask an expert sites tend to have some organizational and procedural structure, though less so than digital reference services. For example, in some sites the category of the question asked may determine which expert will respond; other systems allow experts to declare which questions they will answer by locking the question, or by other means. Ask an expert services may be thought of as online communities, though member interactions tend to be very topic-oriented, "discussions" tend to read like FAQs, and askers and answerers do not interact as peers.

Community Q&A sites[8] leverage the time and effort of everyday users to answer questions – they represent Web 2.0's answer to more traditional online reference services. Established examples of community Q&A sites are Yahoo Answers and Knowledge iN. Community Q&A sites have little structural or role-based organization, typically, although some sites have moderators, or users who have earned elevated privileges based on past contributions. These sites manifest strong online community features, including off-topic discussions and discussions where participants reply to one another, and the presence of repeat/regular users. Community Q&A sites also tend to embrace newer interaction designs than the other types of Q&A sites, by

[7] Also known as: expert services, knowledge networks, or information exchanges.

[8] Also known as: social Q&A or knowledge search; not to be confused with Question Answering (QA), a research area in natural language processing.

providing features like tagging and ratings interfaces, RSS feeds, and highly interactive browsing and searching capabilities.

4.1.2 Related Work on Q&A Sites

Q&A sites have been the subject of related work from researchers in such diverse communities as information science, economics, and information retrieval.

Recent work has examined question answering roles and motivations in community Q&A sites. Gazan (2006), in a study of AnswerBag (http://www.answerbag.com), distinguishes between two roles: *specialists* and *synthesists*. While specialists claim expertise in a given topic and answer questions based on their own knowledge, synthesists gather pointers to outside resources. The researchers found that answers from synthesists were slightly more useful than answers from specialists. Nam et al. (2009) study user motivations and roles in the Korean site Knowledge iN, finding that levels of activity correlate with contribution quality.

Google Answers (http://answers.google.com) has attracted quite a bit of research attention, primarily because it created a monetized "information market", where users choose how much to pay for a question, and optionally tip answerers for work well done. Edelman found that answerers provide better answers as they gain experience, and that topic-specialists provide higher quality answers than generalists (Edelman 2004). Rafaeli et al. found in observational study that levels of payment alone were not sufficient to explain variations in answer quality, but that non-monetary incentives such as ratings influenced answer quality (Rafaeli 2005, 2007). Hseih and Counts (2009) conducted an experimental study of a real-time Q&A system, finding that the presence of money (where askers pay for answers) made users more selective in the questions they chose to ask and answer, reducing the prevalence of "less important" questions, but also reducing the perceived sociality of the system.

Other researchers have investigated differences among Q&A sites. A cross-site comparison of ask an expert services was conducted across 20 such sites (Janes 2001). In this study, the researchers injected questions into the sites and measured outcomes such as response rate, response time, and verifiable answers. This research led to modest findings: ask an expert services in general responded to 70% of all questions, and commercial sites were more likely to provide one or more answers than noncommercial sites. Perhaps more importantly, this study

presents a useful methodology for studying these sites that includes "developing" questions by revising existing questions from the Internet Public Library (based on a small set of criteria) with verifiable answers. This methodology is less useful for studying community Q&A sites, as answers from one site quickly show up in search results, cross-contaminating conditions; our methodology addresses this problem.

More recently, Rousch conducted an informal comparison of six popular community Q&A sites by searching for and asking a small set of questions at each site (Rousch 2006). Based on this informal study, the author concluded that Yahoo Answers was the best performer, and ventured that the reason was their large base of users. This speculation led us to study the influence of a community of users in a larger and more formal study.

As Q&A technologies continue to evolve, qualitative observation has the potential to greatly increase our understanding of how questions are asked and answered online. One interesting example of qualitative observations is presented in (Lee 2005), where the researchers (who are studying music information retrieval) present data on attributes of the questions asked in Knowledge iN and Google Answers. The researchers describe the questions asked on these sites as frequently "vague, incorrect, and incomplete". They also provide a breakdown of questions by the type of "information need", finding that in the domain of pop music queries, a large majority of users wish to identify a particular song or artist, or receive music recommendations. In this research, we report on qualitative observations to better understand the characteristics that emerge in different types of Q&A sites.

4.2 Research Questions

Both site designers and question askers would benefit from a better understanding of the predictors of Q&A quality, responsiveness, and effort. In this context, we present two main research questions that summarize the high level goals of this research.

Research Question 1: How do Q&A sites differ in the quality and characteristics of answers to questions?

We hypothesize that Q&A sites differ in how well they answer questions, and that there are a number of dimensions along which sites differ that influence the quality and characteristics of answers. Specifically, we investigate the following themes:

- How do different Q&A sites differ in answer quality, answerer effort, and responsiveness?
- How do community Q&A sites compare with sites that rely on individuals for answers?
- Do for-fee Q&A sites outperform free sites?
- How do Q&A sites with topic experts compare with those with research experts?

Research Question 2: What can question askers do to receive better answers from a Q&A site?

A first time visitor to a Q&A site might try any number of strategies to get useful answers to a question. We investigate some fundamental strategies to question asking:

- Does paying more in a for-fee Q&A site result in better answers?
- Do simple rhetorical strategies such as thanking the answerer or indicating prior effort affect answer quality?
- Does the topic of the question or the type of question affect answer quality?

4.3 Methods

To determine how Q&A sites differ, and to determine strategies for question askers to receive better responses, we conducted a six week field study, using five sources of online answers. In this study, we asked real questions using made-up identities, then used a panel of blinded judges to rate the questions as well as the answers. In this section, we describe the sites and methodology used in this research.

4.3.1 Q&A Sites Used in This Research

We selected the following Q&A sites for our study:

Library Reference Services represent a traditional form of digital reference service. We consider these sites as a baseline for question answer quality and responsiveness. We divided our questions among eight brick-and-mortar libraries from the United States and Canada[9] that offer Web-based digital reference, plus the Internet Public Library's "Ask a Question". These services all operate using the same Q&A model and site interface – questions are submitted via a web form, responses come via email.

Google Answers is a hybrid service that combines a digital reference service with some community features. Google Answers employed approximately 500 paid researchers, who would answer questions and ask for question clarifications when necessary. To ask a question, users would declare how much they would pay for an answer, between $2 and $200. Researchers would earn 75% of the price of each question they answered. Optionally, users could "tip" researchers after they returned an answer. Both researchers and regular users could comment on questions; these comments often contained valuable information. Because Google Answers was our only paid site, we studied it at three difference price-points – $3, $10, and $30.[10] Google Answers closed for undisclosed reasons after our study was complete, on December 1, 2006.

AllExperts was among the largest and broadest ask an expert sites operating at the time of our study. To ask a question, users must first locate an appropriate "expert" by navigating a taxonomy of question categories such as "Style→Fashion→Hairstyling". If there are multiple

---

[9] Public libraries located in: Alberta, Kansas City, Memphis, Minneapolis, New York, Saskatchewan, Seattle, and Florida

[10] To determine monetary values to use, we observed one week of data (the week of August 20, 2006) to determine the price distribution of questions asked. During this time period, question prices ranged between the minimum and maximum payments possible ($2-$200); the median and mode prices were both $10. Using this distribution as a guide, we chose the 20th percentile ($3) as a "low" payment, the 50th percentile ($10) as a "medium" payment, and the 80th percentile ($30) as a "high" payment.

experts for a given category of questions, users can look at experts' personal profiles and the ratings their past answers have received from other users. Questions are not publicly viewable until the "expert" has responded, and optionally stay private. After an answer has been provided, other users of the site may add comments. This design of this site is strongly oriented towards asking questions; it is a challenge to browse and comment on previously asked questions.

Yahoo Answers is the most visited community Q&A site in the United States. As of December, 2007, Yahoo Answers has provided over 400 million answers to questions (Leibenluft 2007). It also represents what is quickly becoming the de facto standard community Q&A interface: questions are categorized and broadcast to the community, any user can answer any question, and users can rate questions and vote on "best answers". The design of this site encourages browsing questions by category, and emphasizes the newest content. Due to the heavy traffic of this site, questions often receive many replies very quickly after the question is asked. However, questions more than a few hours old often stop receiving further answers as they are lost in the flood of new information.

Live QnA is Microsoft's community Q&A site. While this site features a very similar interface to Yahoo Answers, it does not have the same level of usage. Microsoft QnA launched on August 1, 2006, about two months before we began asking questions (Oct 10, 2006). We chose this site because it serves as an interesting contrast with Yahoo Answers, and because it is interesting to study emerging communities. Today, Microsoft QnA has almost 100x less traffic than Yahoo Answers.

4.3.2 Methodology Overview

We employed a comparative, controlled field study of responses provided across several online Q&A sites. Figure 4-2 summarizes our methodology for each of the 126 questions that we developed and asked – which includes a process for vetting questions according to our independent variables (described below), as well as a process for evaluating the "output" from the Q&A site. In the following sections we describe these steps in more detail.

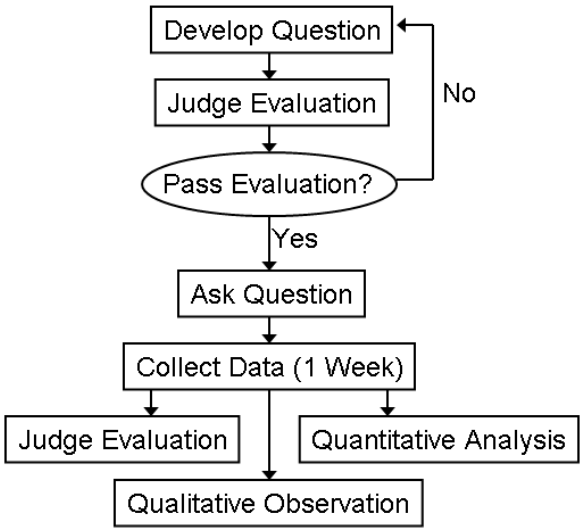


Figure 4-2. Methods: from question development to analysis.

4.3.3 Experimental Design

At the heart of our study is a set of questions that we developed and asked in Q&A sites. As mentioned earlier, there is a precedent in the Q&A research literature for the custom development of questions to enhance experimental control (Janes 2001). Given full control over the questions asked, we designed an experiment with several controlled and balanced independent variables, which we describe below. We chose these variables because they represent some of the high-level characteristics of questions that are commonly asked, and because they plausibly would have an effect on answer quality or responsiveness.

Topic represents the information domain of the question. We examined three topics commonly seen in Q&A sites: technology, business, and entertainment.

Type represents the style of answer the question requires. This variable has three categories: *factual*, *opinion*, and *personal advice*. Factual questions seek objective data or pointers to content; these questions are geared towards researchers. Opinion questions seek others' thoughts on a topic of general interest; these questions do not have a "correct" answer and may be answered without reference to the question askers needs. Personal advice questions seek recommendations based on the asker's own situation; answerers must understand the question asker's situation to provide a good answer.

Prior Effort represents whether the question indicates explicitly that the asker had spent time trying to find the answer before turning to the Q&A site. This variable has two levels: prior effort and no prior effort. A question might indicate prior effort with statements such as "I did a Google search to figure out how to do this.." or "I've asked our office IT folks, and...".

Gratitude represents whether and how the question asker thanks the prospective answerer(s), and may be considered one component of a question's politeness. This variable has three levels: no thank you message, a short (3 words or less) thank you message, and a long (5 words or more) thank you message. We always placed the thank you statement (if any) as the last sentence of the question.

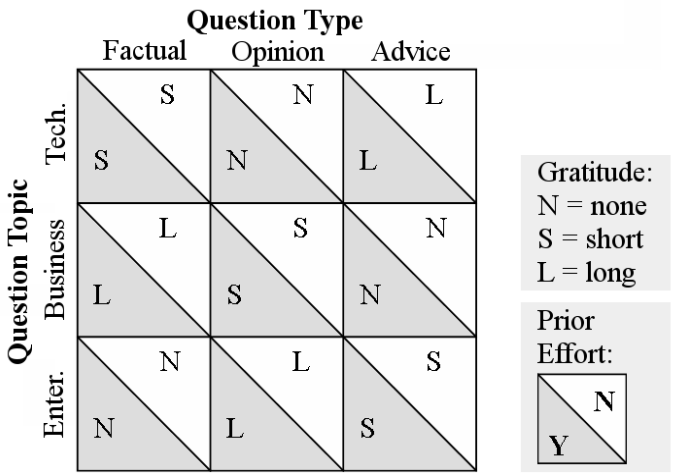


Figure 4-3. Our experimental design in a nutshell. We developed seven questions for each triangle.

In addition to these four independent variables, we distributed questions across different Q&A systems, as described below.

Destination collapses together Q&A site and cost. We make this combination because cost is only applicable to one of our chosen Q&A systems: Google Answers. This variable further collapses the nine library reference sites into a single conceptual destination to avoid overburdening any particular library, to avoid detection (these sites are very low volume), and to avoid the risk of using an outlier service. Our seven destinations are: Google Answers ($3), Google Answers ($10), Google Answers ($30), Library Reference, AllExperts, Yahoo Answers, and Live QnA.

To trim the number of experimental groups necessary while preserving main effects and low order interactions, we employed a fractional factorial experimental design (Box 1978) to choose 18 out of 54 runs of a full factorial design (see Figure 4-3). Our four factors were: topic, type, prior effort, and gratitude. We distributed these 18 runs in a balanced fashion across the seven levels of the destination variable. Thus, for all of these combinations of topic, type, prior effort, and gratitude, we developed seven questions. We randomly assigned each of these questions to one destination.

4.3.4 Developing Questions

Several criteria governed our process of question development. First, we required realistic questions that might be plausibly asked online. Second, we required that our questions would need research or expertise for high quality answers – for each question, we tried one or more intuitive Google searches to ensure there were no solutions (or existing answers on Q&A sites) in the immediate search results. Third, we required questions that adhered to the constraints of our experimental design.

To impartially evaluate questions (and later, answers), we employed a panel of six judges. These judges were all juniors and seniors in college, majoring in either English or Rhetoric. We selected these students because we believed their training in the use of language would enable them to understand and evaluate online questions and answers. The judges were each paid for their participation. They were blinded to the context of the question – they evaluated the unformatted text of the question and answer(s) – and they were not aware that we were the authors of the questions.

We wrote 18 question *templates*, representing the relevant combinations of the four independent variables in our fractional factorial design (as described above). From each of these templates, we wrote seven questions, one for each destination. The seven variants all contained the same number of sentences, and shared characteristics as governed by the experimental design. But they differed in the object of the question enough to avoid detection, and to avoid cross-contaminating other sites with search engine results. For example, two variants of the question defined by entertainment (topic), factual (type), no prior effort, and no thank you message were:

"Are there videos available of the Tonight show from the Steve Allen and Jack Paar years? For sale, online or even just in a museum, I'd like to find them."

"Where could I find unusual Gilligan's Island Memorabilia for a big fan's birthday? Not the usual videos and pictures, but preferably something like props from the original set."

To ensure that our questions reflected our desired independent variables, the judges evaluated each question across a series of criteria. For example, the judges were asked to classify each question as "business", "technology", or "entertainment", and to rate each question in its difficulty. We discarded or rewrote any question that did not get a majority of judge agreement.

See Appendix A for more examples of questions.

4.3.5 Pilot Study and Judge Training

About a month before the launch of our experiment, we conducted a pilot study to train our judges on pilot data and to refine our experimental procedure. We developed seven factual questions and asked one in each destination. We collected responses for a week, then brought the six judges to a training session where we set expectations and answered questions.

4.3.6 Asking Questions

Beginning Oct 10, 2006, we spent six weeks asking our questions online. To avoid the potentially confounding effects of time of day or day of week, we asked all of our questions at approximately 1:00pm local time, on Tuesday, Wednesday, and Thursday of each week. We asked seven questions per day (one per destination) – twenty-one questions per week. For each question, we created a pseudonym using a random name generator, which we methodically turned into a Q&A account sign-in name and email address from a free web-based email provider.

4.3.7 Outcome Measures

Once we had asked a question, we recorded all responses that arrived for up to one week. In our analysis of Google Answers, we consider both "comments" and formal answers as part of the answer set, as comments often contained valuable information to the asker. Google Answers also permitted answerers to request clarifications. We posted a total of four clarifications by writing a short, neutral reply that reiterated a relevant aspect of the original question.

Recall that all sites allow for multiple answers (or no answers at all). In our analysis, we analyze answers in aggregate unless otherwise noted (e.g. in reporting per answer metrics) – in this research study we are primarily interested in Q&A site usefulness from the perspective of the question asker, rather than in characteristics and variations between individual answerers. As such, our judges were trained to evaluate answers as a set, rather than individually, looking at the overall quality and feel of the answer(s) given. We presented questions in the order they were posted on the Web site. We did not vote for best answers or rate answerers in any case, though we did always pay Google researchers without additional tipping.

We consider several primary outcome measures in our analysis. First, we report on several simple metrics about answers, such as the number of answers received to a question, the length of these answers (in characters), and the number of links in these answers. Second, we report on two index variables constructed from the judges' evaluations of answers. To construct each of these variables, we normalized and summed a set of Likert scale survey questions, then renormalized to a 0-1 scale. The first of these variables, *judged answer quality*, is constructed from five Likert scale survey questions that we think reflect the overall goodness or value of the answer(s) provided. These five survey questions measure: (1) answer correctness, (2) asker confidence in answer, (3) helpfulness of answer, (4) progress towards receiving an answer, and (5) monetary worth of the answer. The second of these variables, *judged answerer effort*, is constructed from four Likert scale survey questions that we think reflect the (perceivable) time and energy answerers spent answering the question. These survey questions measure: (1) degree of personalization in answer, (2) perceived answerer effort, (3) answerer friendliness, and (4) ease of use of answer. Both measures appear to be internally consistent (answer quality: Cronbach's $\alpha = 0.94$; answerer effort: Cronbach's $\alpha = 0.87$).

4.3.8 Analysis Methods

In this analysis we focus primarily on main effects. Except where noted, we use regression analysis to build predictive models of four main dependent measures: rated answer quality, rated answerer effort, number of answers, and answer length. We used a mixed model to measure the significance of the judged quality and effort metrics. As each question/answer was rated across many attributes by several different judges, we cannot assume that each judge-score is an independent observation. Thus, to control for judges' biases, we analyzed the data treating

judge as a random effect. In our analysis we used a Restricted Maximum Likelihood method for fitting mixed models, as this method does not depend on balanced data.

4.4 Quantitative Results

Table 4-1 summarizes several outcome measures. Across all destinations, 84.1% (106/126) of our questions received at least one answer. We received a total of 276 answers. On average, each question received 1.73 answers in the first day and 2.19 in the first week. The average judged answer quality was 0.48, and the average judged answerer effort was 0.51. Notably, the top 17 answers in terms of judged quality came from Google Answers, 8 of which were $30 questions, and 7 of which were $10 questions. The top overall answer scored 0.88 on our judged quality scale. As ranked by judged quality, the top AllExperts answer ranked #18, the top Yahoo Answers answer ranked #21, the top library reference answer ranked #26, and the top Live QnA answer ranked #38.

Destination	% Questions with >0 Responses	# Answers/ Question	Length (characters)	Quality	Effort
AllExperts	61%	0.61	629.45	0.33	0.37
Google A. ($3)	78%	2.39	571.47	0.41	0.44
Google A. ($10)	89%	2.78	815.60	0.59	0.60
Google A. ($30)	100%	2.83	1393.92	0.68	0.71
Library Reference	78%	0.83	802.13	0.41	0.47
Live QnA	89%	1.89	257.76	0.40	0.43
Yahoo Answers	94%	4.00	319.24	0.51	0.53
Overall	84%	2.19	678.07	0.48	0.51

Table 4-1. A comparison of the seven destinations across three quantitative metrics (% questions receiving at least one response, average number of answers/question, average answer length in characters) and two index variables (average judged answer quality, average judged answerer effort).

In general, answers with high quality ratings were long and contained many links. We are able to explain 32% of the variance in judged answer quality using just two simple variables: the total length of the answers received, and the total number of hyperlinks in the answer. Our method here is to build a regression model predicting judged answer quality, controlling for the judge as a random effect, using answer length and number of links as our independent

measures. We find that both length and number of links are statistically significant (answer length: $p<0.01$; number of links: $p<0.01$; total model: $R^2=0.32$) and positively correlated with judged quality.

Because we randomly assigned questions to destinations, we expect no difference in question difficulty across destinations. Indeed, as gauged by our judges, there was no significant difference between the seven destinations in predicting question difficulty ($p=0.76$). In addition, there is no evidence that our questions were perceived as out of the ordinary in any site: none of our 126 questions received comments or replies indicating that they were out of place.

4.4.1 Research Question 1: How do Q&A sites differ in the quality and characteristics of answers to questions?

Destination Characteristics

As a starting point in our cross-site comparison, we ask which destinations perform the best across a variety of metrics. See Table 4-1 for an overview of our results. We find that that Google Answers at the $10 and $30 level outperformed all other destinations across most metrics, while Yahoo Answers provided the most answers per question. There is statistically significant evidence that the choice of Q&A site has strong effects on outcomes. According to our model, destination is a significant predictor of answer quality ($p<0.01$), answerer effort ($p<0.01$), total answer length ($p<0.01$), and number of answers ($p<0.01$).

Judged Quality and Effort. See Figure 4-4 for an overview of how each destination compares in terms of judged quality and effort. Pairwise Wilcoxon tests, pairing questions from the same templates, allows assessment of the statistical significance of the observed differences between destinations. We find that Google Answers at $30 provides significantly higher judged answer quality and effort than Google Answers at $3 ($p<0.01$ for both tests). There is trend data suggesting that Yahoo Answers has higher answer quality than library reference services ($p=0.08$), but no conclusive evidence that Yahoo Answers outperforms Google Answers in terms of quality at the $3 level ($p=0.16$). It is interesting to observe that Google Answers appears to exhibit lower variance of quality and effort as payment increases (σ quality: \$3=0.30, \$10=0.25, \$30=0.21; σ effort: \$3=0.28, \$10=0.24, \$30=0.14).

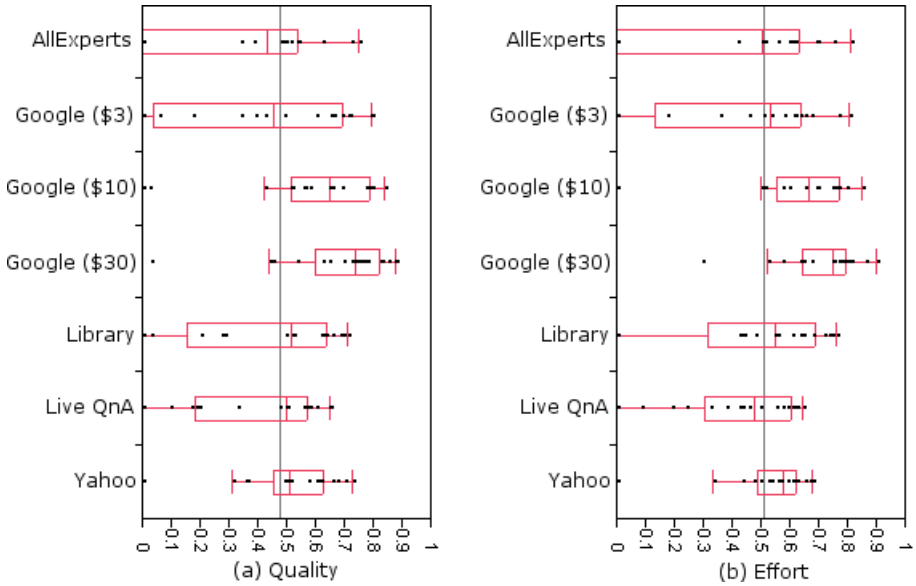


Figure 4-4. (*a*) A box plot of the aggregate judged *quality* metric per destination. The vertical line at 0.48 represents the overall mean judged quality for a question in our experiment. (*b*) A box plot of the aggregate judged *effort* metric per destination. The vertical line at 0.51 represents the overall mean judged effort for a question in our experiment. The ends of the boxes are at the 25th and 75th percentiles.

Diversity and Timeliness. Another way of looking at answer quality is to examine the diversity of answers, and the speed with which answers arrive. Figure 4-5 shows a summary of the number of answers per question by destination. Yahoo Answers provided the most answers per question, on average (4.0), while AllExperts provided the fewest (0.6). Using pairwise Wilcoxon tests, pairing questions from the same templates, we find that Yahoo has higher answer diversity than Live QnA ($p<0.01$). As might be expected, AllExperts and library reference provided lower diversity than all other sites ($p<0.01$ for all 10 pairwise comparisons).

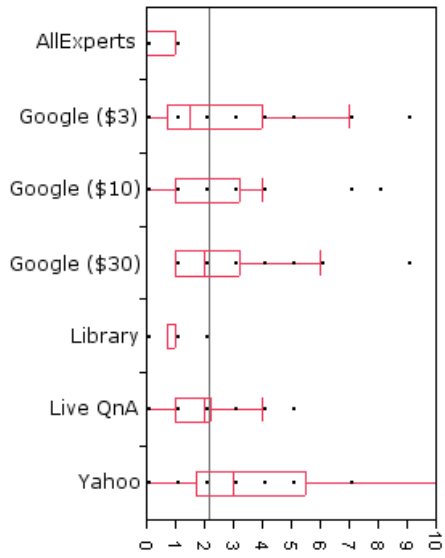


Figure 4-5. A box plot of the number of answers per question at each destination. The ends of the boxes are at the 25th and 75th percentiles. The vertical line at 2.2 represents the overall mean number of answers per question across destinations.

In general, most of the answers we received were timely, independent of destination. Overall, 79% of answers arrived within a day of asking the question. Figure 4-6 shows the total number of answers that we received to our questions, broken out by the time period in which we received the answer. Google provided the most answers after the first day had passed, indicating that while Google provides the highest quality answers (given sufficient payment), askers must have some patience when waiting for answers.

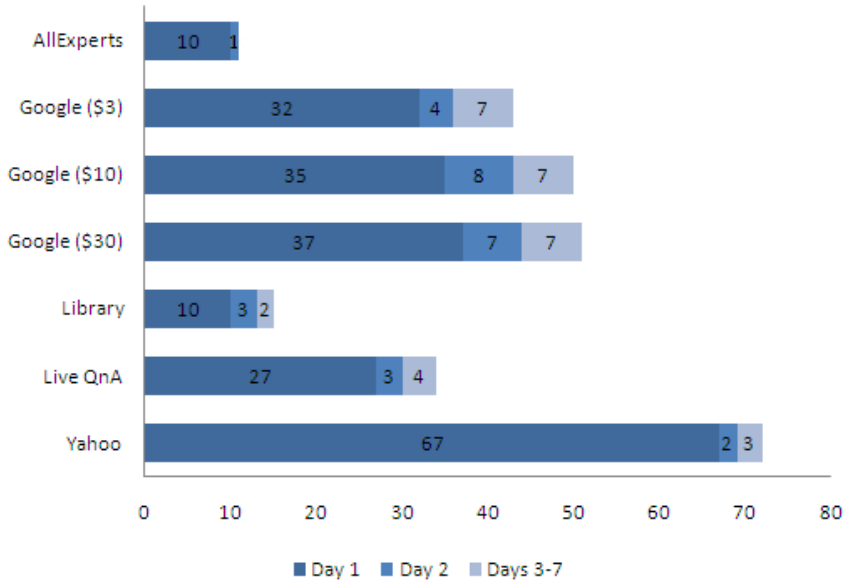


Figure 4-6. We asked 18 questions per destination. This chart shows the total number of replies by destination, stratified by when answers arrived. Overall, 79% of answers to our questions arrived in the first day.

Exploratory Analysis of Design Dimensions

We wish to make some claims regarding the effect of several design dimensions on answer quality in real Q&A sites. However, there are challenges in designing a field experiment that supports such an analysis, as there are many confounding factors inherent in all Q&A sites and it is difficult to isolate the effects of any single design feature. However, there is value in "real world" data, and this exploration may provide evidence to inspire future work in a different experimental setting.

For the following analysis, we rely on a similar statistical analysis as described in methods, but we treat Q&A site as a nested variable, as each site occurs at exactly one level of the dimensions that we test.

Community vs. Individual. We wish to understand whether answers from communities outperform answers from individuals. Of the five sites used in our study, two (Yahoo Answers and Live QnA) give community answers, and two (AllExperts and Library Reference) give answers from individuals. We omit Google Answers data from this analysis, as it is the only fee-based site (which could dominate other factors), and because we consider this a hybrid system: while individuals "answer" questions, others may comment and add to the answer.

As we might expect, we find that community sites provide more answers than individual sites (means: 2.94 community vs. 0.72 individual; $p<0.01$; $R^2=0.39$). We find trends, but not strong evidence, of other effects. We find that community has marginally significant, positive effects on judged answer quality (means: 0.47 community vs. 0.42 individual; $p=0.09$; $R^2=0.09$), and on the total length of answers received (means: 881.92 community vs. 526.56 individual; $p=0.06$; $R^2=0.17$). However, we find no effect on judged answerer effort ($p=0.60$).

Paid vs. Free. We have already presented data that Google Answers outperforms all other Q&A sites across our measures, except when we paid a small amount ($3). If we consider Google Answers to represent "paid" answers, and the other sites to represent "free" answers, we find – not surprisingly – that paid answers appear to outperform free answers. Paid answers had higher judged answer quality (means 0.61 vs. 0.45) and higher judged answerer effort (means 0.62 vs. 0.50); they received more answers (means 2.67 vs. 1.83 answers) and longer answers (means 2527 vs. 704 characters). However, these differences could simply be a result of Google's system outperforming the other systems on average, irrespective of cost. In the absence of free questions in Google Answers (and paid answers in other sites), we can make no claims about paid vs. free answers. However, we later return to the effect of amount paid in Google Answers, and provide evidence that in this case, paying more improves answers.

Specialists vs. Synthesists. As we discussed earlier, Gazan (2006) showed that in the Q&A site AnswerBag, answers from synthesists were judged better than answers from specialists. We are able to provide some corroborative evidence for Gazan's result, as we find number of links to be positively correlated with judged answer quality ($\rho=0.54$, $p<0.01$).

We can extend this analysis by making the assumption that Google Answers and library reference sites consist of a majority synthesists, given their reputation as employing professional or expert researchers. The other three sites we consider to consist of a majority specialists: AllExperts because they employ experts in particular subject areas, and the other sites because we speculate that users spend the majority of their time browsing and answering questions about favorite topics. There is some evidence that this assumption is plausible: taking number of links per answer to as a strength-of-synthesist metric, we find that Google Answers (3.23) and library reference sites (1.89) score highest, trailed by AllExperts (0.81), Yahoo Answers (0.52) and Live QnA (0.36). Given this separation, our model supports Gazan's

finding and shows a positive effect for synthesist sites in answer quality (means: 0.54 synthesist vs. 0.45 specialist, $p<0.01$).

4.4.2 Research Question 2: What can question askers do to receive better answers from a Q&A site?

Level of Payment

A question asker in Google Answers might wonder how much money to spend on a question to get the best answer. To help this person, we look at statistics that describe average answer characteristics across different payment levels. For this analysis we treat cost as a categorical variable; we are interested in whether cost is a significant factor in predicting outcomes, rather than modeling the relationship between actual price and outcome.

As gauged by our judges, when we paid more for an answer, we received higher quality answers and answerers spent more effort. $30 answers were rated 0.68 on judged answer quality, as compared with 0.62 for $10 answers and 0.47 for $3 answers; cost is a statistically significant factor in our model predicting quality ($p<0.01$). $30 answers were rated 0.71 on judged answerer effort, as compared with 0.63 for $10 answers and 0.49 for $3 answers. Cost is also a statistically significant predictor of effort in our regression analysis ($p<0.01$). Interestingly, paying more did not result in *more* answers, although it resulted in *longer* answers (means: $3, 1365 characters; $10, 2266 characters; $30, 3949 characters) ($p=0.01$).

These statistical results perhaps belie more nuanced community behavior which we revisit in later sections with qualitative observations.

Rhetorical Strategy

In our methodology, we varied our expression of gratitude and our indication of prior effort across questions. These are two simple rhetorical strategies that we speculated could be employed by question askers to affect Q&A outcomes.

Neither gratitude nor prior effort had a statistically significant effect in predicting answer outcomes. On average, using the longest thank you message led to the highest judged answerer effort (means: 0.53 long vs. 0.49 short vs. 0.51 none), and the absence of a thank you message led to the lowest judged answer quality (means: 0.48 long vs. 0.48 short vs. 0.47 none). In our

model, however, neither of these differences were statistically significant in predicting quality (p=0.56) or effort (p=0.31). See Table 4-2.

Thank You	% Questions with >0 Responses	# Answers/ Question	Length (characters)	Quality	Effort
None	86%	2.48	1467.62	0.47	0.51
Short	81%	1.88	1378.17	0.48	0.49
Long	86%	2.21	1610.12	0.48	0.53

Table 4-2. Mean outcomes based on the length of the thank you message in the text of the question.

On average, indicating prior effort actually decreased both judged answer quality (means: 0.49 no prior vs. 0.46 prior) and judged answerer effort (means: 0.52 no prior vs. 0.50 prior). In our model, prior effort was marginally significant in predicting quality (p=0.07) but not significant in predicting effort (p=0.32). See Table 4-3.

Prior Effort	% Questions with >0 Responses	# Answers/ Question	Length (characters)	Quality	Effort
No	86%	2.08	1364.00	0.49	0.52
Yes	82%	2.30	1606.60	0.46	0.50

Table 4-3. Mean outcomes based on whether or not we indicated "prior effort" in the text of the question.

We might hypothesize that these rhetorical strategies could interact with the Q&A destination. For example, perhaps being polite matters in one community with one set of norms, while it is considered strange in another community. We see some evidence of this: there is a significant interaction effect between gratitude and destination in predicting quality (p<0.01) and answerer effort (p<0.01). Every destination appears to respond to thank you messages differently. We can speculate that different Q&A sites have different cultures, a latent factor that is interacting with our different messages. Prior effort, on the other hand, does not interact with destination in a statistically significant way (p=0.28). See Table 4-4 and Table 4-5.

	Gratitude		
Destination	None	Short	Long
AllExperts	0.29	0.37	0.31
Google A. ($3)	0.47	0.57	0.21
Google A. ($10)	0.58	0.58	0.62
Google A. ($30)	0.62	0.70	0.71
Library Reference	0.43	0.25	0.57
Live QnA	0.47	0.38	0.34
Yahoo Answers	0.46	0.49	0.58
Overall	0.47	0.48	0.48

Table 4-4. Mean judged answer quality broken out by thank you message and site.

	Prior Effort	
Destination	No	Yes
AllExperts	0.30	0.35
Google A. ($3)	0.34	0.49
Google A. ($10)	0.67	0.52
Google A. ($30)	0.70	0.65
Library Reference	0.50	0.33
Live QnA	0.41	0.38
Yahoo Answers	0.54	0.48
Overall	0.49	0.46

Table 4-5. Mean judged answer quality broken out by prior effort and site.

Type and Topic of Question

By varying two independent variables – type and topic – we investigate whether the informational goal of a question affects its resulting answer quality in Q&A sites.

Our data suggest that topic has a potentially large effect on the number of answers received (means: 3.07 ent. vs. 1.90 tech. vs. 1.60 bus.; $p<0.01$), a small and marginally significant effect on answer quality (means: 0.49 tech. vs. 0.48 bus. vs. 0.46 ent.; $p=0.06$), and no effect on answerer effort ($p=0.74$) or the length of answers received ($p=0.94$). In particular, it seems that entertainment-oriented questions received many replies, but those replies were poor in judged quality relative to other topics. See Table 4-6.

Topic	% Questions with >0 Responses	# Answers/ Question	Length (characters)	Quality	Effort
Business	86%	1.60	1515.64	0.48	0.51
Entertainment	86%	3.07	1528.43	0.46	0.51
Technology	81%	1.90	1411.83	0.49	0.51

Table 4-6. Mean outcomes based on the topic of the question.

Destinations appear to vary in the quality of answers they provide to questions in different topics. See Table 4-7 for an overview. Library reference services provided higher quality answers in business- or technology-oriented questions than they did in entertainment-oriented questions. Live QnA provided their best answers to technology questions, and Yahoo Answers performed particularly well on business questions.

	Topic		
Destination	Business	Entertainment	Technology
AllExperts	0.35	0.31	0.31
Google A. ($3)	0.35	0.50	0.39
Google A. ($10)	0.65	0.55	0.59
Google A. ($30)	0.73	0.58	0.72
Library Reference	0.46	0.34	0.44
Live QnA	0.21	0.43	0.56
Yahoo Answers	0.59	0.48	0.46
Overall	0.48	0.46	0.49

Table 4-7. Mean judged answer quality broken out by question topic and site.

Asking different types of questions also appears to affect outcomes. Our data show that type has a statistically significant effect on quality (means: 0.55 advice vs. 0.46 opinion vs. 0.42 factual; $p<0.01$), effort (means: 0.57 advice vs. 0.51 opinion vs. 0.45 factual; $p<0.01$), and length (means: 2028 advice vs. 1408 opinion vs. 1020 factual; $p=0.02$); type has no effect on the number of answers received ($p=0.19$). Thus, requests for personal advice appear to receive the most – and the best – attention of any question type in our study. Conversely, factual questions appear to receive the fewest – and the lowest quality – responses. See Table 4-8.

Type	% Questions with >0 Responses	# Answers/ Question	Length (characters)	Quality	Effort
Advice	93%	2.43	2027.76	0.56	0.57
Factual	74%	1.71	1020.24	0.42	0.45
Opinion	86%	2.43	1407.90	0.46	0.51

Table 4-8. Mean outcomes based on the type of the question.

As with topic, destinations appear to vary in the quality of answers they provide to different types of questions. See Table 4-9 for an overview. Library reference services provided much higher quality responses to advice or factual questions than opinion questions. Yahoo Answers and AllExperts, on the other hand, provide their lowest quality answers for factual questions.

	Type		
Destination	Advice	Factual	Opinion
AllExperts	0.40	0.23	0.35
Google A. ($3)	0.43	0.45	0.37
Google A. ($10)	0.73	0.46	0.60
Google A. ($30)	0.74	0.60	0.69
Library Reference	0.48	0.44	0.33
Live QnA	0.51	0.35	0.33
Yahoo Answers	0.59	0.40	0.54
Overall	0.55	0.42	0.46

Table 4-9. Mean judged answer quality broken out by question type and site.

4.5 Qualitative Observations

Through the course of our six week study, we observed interactions that illustrate the strengths and weaknesses of the different Q&A sites. In this section, we share some highlights of our study, to deepen our understanding of the dynamics that govern the Q&A process in different sites.

The data presented above strongly make the case that Google Answers, on average, provided the best answers of any of the sites studied. We believe that the reasons for this success go deeper than the financial incentives. Rather, the community of researchers and regular users was passionate about answering questions, and appeared to enjoy the "game" of answering

challenging questions. In many cases, researchers and other users used the (unpaid) commenting feature to post lengthy replies to answers that other researchers had written. For example, we asked the following $30 question: "Which actress has the first female line in a talking movie? [...]" Within two hours, we received a long (3,800 character) answer that included information about the actress (*"Eugenie Besserer was the first female to speak in a full length talkie. She played Al Jolsen's mother, Sara Rabinowitz in the film the Jazz Singer"*), statistics about the first line, excerpts from the script, and six links to Web pages with further information. However, one community member disagreed with this answer, replying *"The actress with the first line in a talkie was Sarah Bernhardt in 'Le Duel d'Hamlet' around 1900 [...]"*. All this led to a five user, passionate discussion concerning the subtleties of the question. The discussion led to two formal answer clarifications and a congratulatory post: *"Well done everybody! A Great Question has brought a Great Answer and some interesting Comments."*

However, the pricing structure in Google Answers may lead to some awkwardness for new users. Google Answers researchers appear to have internalized a model of how much a question is "worth", while question askers (especially first-timers) may not understand how much money to offer for a question. One question we asked at the $3 level asked for "advice and pitfalls" concerning hiring a custodial service. In response, we received a single comment from a researcher: *"It'll cost you a lot more than $3 to hire a custodian and it would take a Google Answers Researcher a lot more time than $3 is worth to research this question."* On the flip side, we appeared to overvalue other questions, such as one $30 question with the subject line *"What e-mail system to use for mailing lists?"* This question received just two short comments with brief recommendations, rather than any "authoritative" answer that synthesized outside data, or gave an expert opinion with background information. In fact, only 11/18 of our $30 questions received an official answer – the seven unanswered questions spanned all three topics and types. However, all 18 received at least one response in the form of a comment, underscoring the value of the community features that Google added to their reference service.

We found that our questions in community Q&A sites were more likely to get some response (92%, vs. 81% in the other sites). However, the benefits of high responsiveness are potentially offset by other, qualitative shortcomings. For example, in one Yahoo Answers questions with the subject *"How to get front-row Lakers tickets?"*, we received two separate responses that read *"ebay"*, one that read *"buy them duh"*, one that read *"You could try giving Jack Nicholson*

a call [...]", and one that read *"Umm I can help ya... Sleep with Jack Nicholson."* On the other hand, we also received recommendations for two ticket resellers that both appear to be good options. Live QnA had a much less responsive feel, but this did not improve our perception of the signal-to-noise ratio. For example, we asked a question about reel film projectors, stating *"All my searches for movie projectors point me to digital projectors [...], not a reel film projector"*, our only response was a link to a site that sells only digital projectors, probably indicating that the answerer did not read the full question.

Although library reference services and AllExperts are fundamentally different types of Q&A services, they shared several of the same advantages and drawbacks, perhaps because they both depend on individuals for answers. In both sites, the biggest problem was getting any answer – in AllExperts, only 61% of our questions received a response, while across libraries, only 78% of our questions received a response.[11] In both services, we typically received exactly one response per question, so there were no opportunities for collecting diverse opinions without re-asking the question elsewhere. In one respect, however, the two services differed – AllExperts responses reflected the answerer's interest in the topic, while the librarians' answers reflected an interest in the research process, or a lack of interest. For example, we asked AllExperts for *"information on who might be the best baseball announcer of all-time"*, and the respondent enthusiastically wrote a lengthy response, stating *"The guy who I most enjoy listening to because he does all these things extremely well is Jon Miller [...] Listening to him makes the game much more enjoyable to me"*. In contrast, a question directed at library reference services about *"who is the most skilled celebrity chef?"* received the dry response: *"I do not have any reliable source for this information. I did find a Website with award information [...]"*

4.6 Discussion and Conclusion

In this study, we found that (1) you get what you pay for in Q&A sites, and (2) a Q&A site's community of users contributes to its success. Across our answer quality and responsiveness metrics, Google Answers (a fee-based Q&A site) was superior to each of the free Q&A sites we

[11] Two of the nine libraries we chose did not reply to any questions; all other libraries had a 100% response rate.

studied. Further, when we paid more for an answer at Google Answers, we typically received longer, better answers. Qualitatively, we found that the volunteer efforts of the Google Answers community helped make answers better, and gave the site more diverse opinions. Among the free sites, Yahoo Answers scored the best – its large community provided high answer diversity and responsiveness. Compared to Live QnA, a community with a very similar design but many fewer users, Yahoo Answers typically yielded better responses, further underscoring the importance of a large, active community.

There is an ongoing debate concerning the benefits and drawbacks of information derived from open community participation. For example, the community-edited Wikipedia has been favorably compared with the professionally-edited Encyclopedia Britannica in terms of science article quality (Giles 2005), but Wikipedia readers must cope with a small and increasing chance of viewing articles with intentional vandalism or misinformation (Priedhorsky 2007). In the Q&A domain, we find that the community in Yahoo Answers provides surprisingly high-quality (aggregate) answers compared with the professionally-staffed library reference services, but Yahoo Answers users must expect substantial variability in the quality of individual answers. On the other hand, the community discussion features in Google Answers appeared to add value to the system with no visible downside. Given the large body of literature that has shown extrinsic incentives "crowd out" intrinsic motivations (Deci 1999), we might not expect the unpaid members of the Google Answers community to contribute as much value as they did. Future work might leverage this finding to better understand the properties of online community design that allow paid and free contributions to coexist.

In general, community Q&A sites are fertile ground for future work. They have integrated many compelling features that encourage work from their users, such as points, ratings, voting, and leaderboards. Research exploring the impact of these features (e.g. following the examples of (Lampe 2004) and (Ling 2005)) on the quality and quantity of Q&A would help designers better understand when and where to deploy such features in their own online communities. Also, deep qualitative work holds much potential to better understand Q&A sites. There has been little research that seeks to understand what questions people ask, how they ask them, how they choose questions to answer, or how they respond to questions.

In conclusion, we leave you with a question: What change do you think will most improve Q&A sites of the future? $10 for the best answer.

Chapter 5

Facts or Friends? Distinguishing Informational and Conversational Questions in Social Q&A Sites*

Social question and answer Web sites (Q&A sites) leverage the wisdom of crowds to help us find answers to our questions. These are sites that we can turn to when our search terms fail to turn up useful results, or when we seek personal advice and the opinions of others. Social Q&A sites work on a simple premise: that any user can pose a question, and in turn other users – potentially many of them – will provide an answer.

This simple premise has turned out to be very popular. For example, Yahoo Answers – the leading U.S. Q&A site – fields approximately 90,000 new questions every day.[12] To put this number in perspective, Yahoo Answers generates as many new Web pages each month as are contained in the English language Wikipedia (2.8 million).[13]

One of the reasons why social Q&A sites have been launched by major search engine companies is due to their ability to simultaneously expand their searchable corpus and engage users. An early example of success in this area is Knowledge iN, a social Q&A site that was central to the rise of South Korean internet portal Naver (Chae 2005). However, outside of Asia,

---

* This chapter extends the work originally published as (Harper 2009), co-written with Daniel Moy and Joseph Konstan. The author of this thesis developed the data set and managed the data collection process; he developed the machine learning algorithms with technical contributions from Daniel Moy. Joseph Konstan contributed to the design of the coding tool, and to the overarching experimental design.

[12] Based on a sample collected between 2/20/2008 and 4/8/2008; the precise average was 88,121 questions per day.

[13] As of 4/14/2009, according to http://en.wikipedia.org/wiki/Special:Statistics.

social Q&A has failed to become a feasible alternative to search engines, or even a reliable source of high-quality searchable information.

Operators of social Q&A sites speculate that one of (potentially many) reasons for this failure is the interference of conversational question asking – questions like *"what are you doing right now?"* that are unlikely to lead to the creation of archival-quality information.

In this chapter, we seek to deepen our understanding of the differences between *conversational* questions and *informational* questions. Let us define these terms:

- *Informational questions* are asked with the intent of getting information that the asker hopes to learn or use via fact- or advice-oriented answers. An example: *What's the difference between Burma and Myanmar?*

- *Conversational questions* are asked with the intent of stimulating discussion. They may be aimed at getting opinions, or they may be acts of self-expression. An example: *Do you drink Coke or Pepsi?*

Thus, we focus our attention on questions rather than answers, with an attitude that some questions enable the creation of high-quality answers more than others. It is entirely possible that one cannot fairly measure the quality of an answer in the absence of information regarding the quality of the question being answered. Also, we focus on questions because recent research has focused primarily on detecting answer quality (see below), and investigating questions may reveal new insights into the Q&A process.

5.1 Research Questions

In this research, we explore the differences between conversational and informational Q&A using human coding, statistical analysis, and machine learning algorithms. We organize this work using the following research questions:

RQ1: Can humans reliably distinguish between conversational questions and informational questions?

We first seek to validate our division of questions into two primary types. While other researchers have observed social question asking (e.g., Adamic 2008, Agichtein 2008), we test the extent to which humans are able to agree when classifying random questions across several Q&A sites.

RQ2: How do informational and conversational questions differ in terms of writing quality and archival value?

Our second research question asks if conversational questions are perceived to have lower quality. We investigate two metrics in particular, one concerning the quality of the writing in the question, and the other concerning the long-term value of the question from the perspective of someone searching for information.

RQ3: What are the structural differences between conversational questions and informational questions?

In our third question, we explore the structure and nature of conversational and informational questions. We describe categorical, linguistic, and social differences between these two types of questions.

Research Challenge 1: Can we build algorithms that reliably categorize questions as informational or conversational?

We believe that developing automated techniques for separating informational and conversational content opens new possibilities for social media research. For instance, we can begin to develop information quality metrics that understand that not all user contributions are intended as archival-quality. Similarly, automated classification techniques open new possibilities for interaction designs, such as automated tagging of content and personalized content ranking systems that leverage user preferences for different content types.

5.2 Related Work

Recently, social Q&A sites have become the focus of much research attention. A substantial fraction of this attention has been spent understanding the properties of these new forums for online social discourse. In the previous chapter, we looked at the performance of social Q&A sites in comparison with more traditional online avenues for question asking such as online reference libraries, and found both benefits (more diverse opinions) and drawbacks (highly variable quality) (Harper 2008). Other researchers have reported on differences among categories in Q&A sites (Adamic 2008), user tendencies to specialize within topics in Q&A sites (Gyöngyi 2008), and the effect of user experience and rewards on performance (Shah 2008). Collectively, these studies have revealed that Q&A sites are inconsistent in terms of question and answer quality, and that there is a large and poorly understood social component to these systems.

The apparently extreme variability in answer quality in sites such as Yahoo Answers has led to recent work that attempts to quantify quality, and to leverage computational techniques to improve the user experience. For example, researchers have investigated the use of machine learning techniques to predict askers' satisfaction with answers (Liu 2008), build models of question and answer quality based on features derived from text and user relationships (Agichtein 2008), predict the occurrence of "best answers" (Adamic 2008), and compute metrics of expected user quality (Jurczyk 2007). These studies are encouraging: they all indicate that we can develop algorithms to infer the quality of Q&A discourse with some confidence. However, there are limitations to these studies. First, all this work has taken place on just one Q&A site – Yahoo Answers. While Yahoo is the largest social Q&A site in the United States, it is unclear the extent to which any of these methods generalize to the entire social Q&A space, rather than a single instance. Second, these studies treat all questions equally, but it seems likely that quality for a conversational task is a very different thing from quality for an informational task.

5.3 Data Collection And Coding Methods

We picked three social Q&A sites to study that offer similar Q&A interfaces, but that differ in the volume of contribution and membership: Yahoo Answers, Answerbag, and Ask Metafilter.[14] These sites each offer an opportunity for users to ask questions on any topic for the community to answer. While there are other types of online forums for online question asking and answering, such as digital reference services and "ask an expert" sites, we do not consider these sites in this analysis, as they have a more restrictive Q&A process – relying on single "experts" to answer questions – and experience empirically different quality problems than social Q&A sites (Harper 2008).

For each of the three sites, we collected information over a range of dates, including full text, user identifiers, category names and identifiers, and timestamps. See Table 5-1 for summary statistics of this dataset.

	Answerbag	Metafilter	Yahoo
# Days	180	808	49
# Users	51,357	11,060	1,575,633
# Users: Askers	26,399	7,452	1,200,413
# Users: Answerers	32,064	10,806	911,972
# Questions	142,704	45,567	4,317,966
# Questions per Day	793	56	88,122
# Answers	806,426	657,353	24,661,775
# Answers per Question	5.65	14.43	5.71
% Questions Answered	89.9%	99.7%	88.2%

Table 5-1. Properties of the three datasets used in this work.

Yahoo Answers (answers.yahoo.com) is the largest Q&A site in the United States, claiming 74% of U.S. Q&A traffic (Hitwise 2008). We downloaded data using the Yahoo Answers web API for a period of seven weeks, resulting in a data set of over 1 million users, 4 million questions,

[14] These are three of the four most frequently visited Q&A sites in the U.S. as of March, 2008 (Hitwise 2008). We did not study the second-ranked site, WikiAnswers.com, because its interface prevents us from unambiguously understanding when a question is asked, and who it is asked by.

and 24 million answers. Remarkably, the Yahoo Answers community asked an average of 88,122 questions per day over the period of our data collection. Notable features of the Yahoo interface include questions that accept new answers for just 4 days (or at the asker's request, 8 days), browsing organized around categories, and a prominent system of rewarding users with "points" for answering questions.

Answerbag (www.answerbag.com) is a much smaller Q&A site, with an estimated 4.5% market share (Hitwise 2008). We downloaded data using the Answerbag web API, for a data set that spans 6 months of activity. We chose to collect data on questions asked in the first half of 2007, as Answerbag questions remain "open" indefinitely, and many questions continue to receive answers long after they are asked (in contrast with the short lifespan of questions at Yahoo). Answerbag distinguishes itself by sorting questions either by a user-rated "interestingness" metric, or by the last answer received. Also, Answerbag questions do not have a separate subject field, and are limited to just 255 characters. Finally, Answerbag answers may themselves be commented on – we exclude this data from our analysis to ensure consistency across the data sets.

Ask Metafilter (ask.metafilter.com) is the smallest of our Q&A sites, with an estimated 1.8% share of U.S. Q&A traffic (Hitwise 2008), and an average of 56 questions per day over the course of our study. We collected over 2 years of data from Ask Metafilter using a custom scraper. Questions on this site may receive answers for a year after they are asked; we stopped scraping questions newer than May 2007 to ensure completeness. Ask Metafilter requires participants to pay $5 to join the community, which has led to a much lower volume of contributions. Also notable is the fact that nearly every question (99.7%) receives at least one answer.

5.3.1 Coding Methodology

Our research questions depend on human evaluations of question type and quality. We developed an online coding tool to better allow us to make use of available volunteers. To help ensure high-quality coding, the tool requires new users to complete a tutorial of instructions and quiz questions. The tutorial is designed to emphasize our own definitions of conversational and informational question asking, to ensure as much consistency across coders as possible. See Figure 5-1 for an example quiz question.

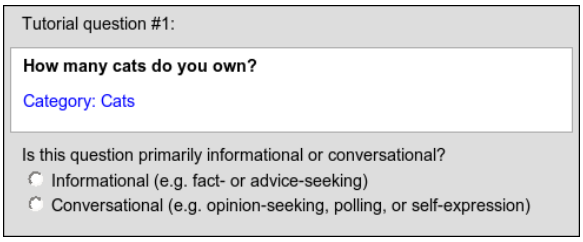


Figure 5-1. A sample quiz question from the coders' tutorial. This question is conversational, as the apparent intent of the question asker is to poll other users. Whether their answer is right or wrong, the tutorial shows an explanation that reinforces our definitions of the terms.

The online coding tool presents the text from all three Q&A sites in a common format, as shown in Figure 5-2. Other than the question text, the only information shown about each question is its category (which is site-specific), as some questions lack sufficient context without this information (e.g., the question *"How do I kick properly?"* does not make sense unless we understand it is part of the category "swimming"). We do not reveal the name of the Q&A site where the question was asked, and we do not provide a link to the original question or the answers – we wish to ensure that different questions are graded independent of any site-specific bias.

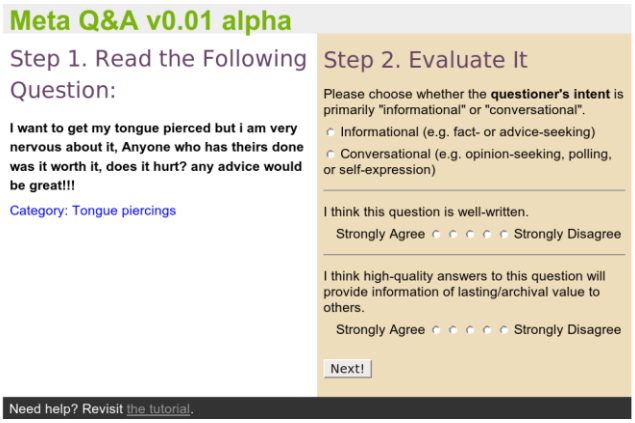


Figure 5-2. A screenshot from the online coding tool. Coders were asked (1) to determine if the question was asked primarily with informational or conversational intent, and (2) to assign subjective ratings of how well the question was written, and the potential archival value of the question.

The coding tool first asks users to determine if a question is asked with primarily informational or conversational intent. It then asks users to rate the question on two dimensions using Likert scales (5=strongly agree, 1=strongly disagree):

- WRITING QUALITY: I think this question is well-written.

- ARCHIVAL VALUE: I think high-quality answers to this question will provide information of lasting/archival value to others.

We ask coders to evaluate ARCHIVAL VALUE assuming that the question receives high-quality answers, rather than showing the actual answers, because our goal is to understand questions (not answers) and because we want consistent judgment across sites without the bias of different answer quality.

Two additional questions appear when a user codes a question as informational:

- APPROPRIATE RESPONSE: Which of the following types of responses would be appropriate for this question?

- PERSONALIZED: Do you think this question is highly personalized with regard to the asker's situation, completely independent of the asker's situation (generic), or somewhere in between?

APPROPRIATE RESPONSE asks users to choose one or both of (a) objective facts, (b) personalized advice. PERSONALIZED is evaluated on a 5 point scale, with the ends labeled "personalized" and "generic". See Figure 5-3 for a screenshot.

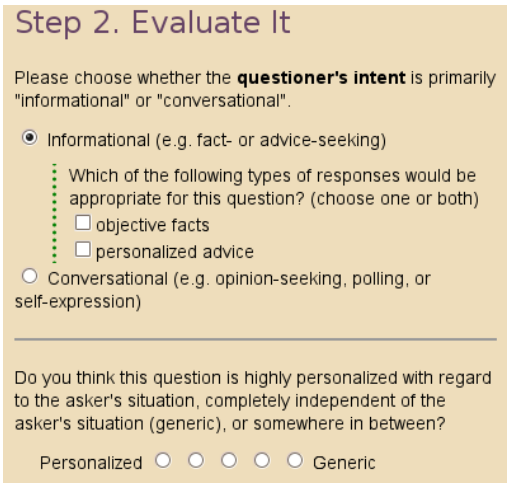


Figure 5-3. A cropped screenshot from the online coding tool. If the coder selected "informational", then two additional questions appeared. First, the coder was asked to evaluate whether responses containing objective facts and/or personalized advice are appropriate. Second, the coder was asked to evaluate the degree to which the question is personalized to the asker's situation.

The coding tool randomly samples questions from our data set, balanced across the three test sites. To ensure some consistency in our coding, each question was independently coded by at least two volunteers. In the case that the first two volunteers disagreed about the question type, we solicited two additional volunteers to code the question. For analysis that depends on question type, we did not consider questions where the four voters cannot come to a majority classification. While soliciting additional coders for each question would provide greater confidence in the results, coders' time and effort are limited resources. Thus, we selected this process to maximize the number of questions coded – necessary for machine learning techniques – while also providing some redundancy. See Figure 5-4 for an overview.

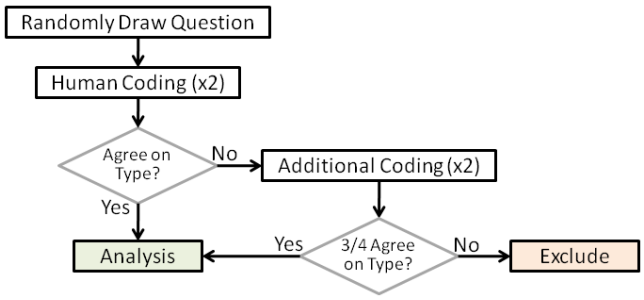


Figure 5-4. A flowchart showing how a randomly drawn question is classified by two or more coders.

5.4 Results Of Human Coding

In all, 30 people participated as coders, submitting a total of 1,106 evaluations across a set of 490 questions. In this section, we present descriptive statistics about this dataset, in the process addressing our first two research questions. For examples of questions from each site along with the coded question type, see Appendix B.

5.4.1 Human Coder Agreement

427/490 questions (87.1%) received agreement on question type from the first two coders. Of the remaining 63 questions that were coded by four people, 23 (4.7%) received a "split decision", where two coders voted conversational and two voted informational. Thus, for our machine learning sections below, we consider the set of 467 questions that received majority agreement among coders.

The APPROPRIATE RESPONSE variable can take three possible values, as the coder is asked to check one or both of "objective facts" or "personalized advice". Coders agreed about the "objective facts" choice 82.4% of the time, and about "personalized advice" 78.7% of the time; they agreed about both choices 64.5% of the time, and disagreed about both choices 3.5% of the time. To render a "consensus" coding in subsequent analysis, we average the coder's decisions, then round (in case of ties, we "round up" by declaring that the response type is appropriate). Then, we combine the two independent choices into a single, three level nominal variable.

The PERSONALIZED variable reflects coders' responses on a 5 point scale, ranging from "generic" (1) to "personalized" (5). Coders agreed within one point 66.9% of the time (mean disagreement = 1.24). Because we changed the wording of the PERSONALIZED question on the online form after the first day of coding, we discard 23 data points, leaving 285 of the 308 (92.5%) of the informational questions for subsequent analysis on this variable. The discarded rows are balanced across sites, and randomly distributed from the pool of informational questions.

The WRITING QUALITY and ARCHIVAL VALUE metrics reflect coders' responses to 5 point Likert scale questions. Coders agreed within one point 74.4% of the time on the WRITING QUALITY metric (mean disagreement = 1.02), and 70.4% of the time on the ARCHIVAL VALUE

metric (mean disagreement = 1.08). The distribution of coder disagreements does not change significantly across the different sites in our study.

5.4.2 Site Characteristics

Conversational questions are common in Q&A sites: we found that overall, 32.4% of the questions in our sample were conversational. However, it is clear from these data that different sites have different levels of conversational content ($p < 0.01$, $\chi^2 = 117.4$), ranging from Answerbag where 57% of the questions are coded conversational to Ask Metafilter where just 5% of the questions are coded conversational. All three pairwise comparisons are statistically significant (Answerbag (57%) > Yahoo (36%) > Metafilter (5%); $p<0.01$ for all pairwise tests). See Table 5-2 for more details.

	Total	Informational	Conversational	Disagreement
Answerbag	161	64 (40%)	92 (57%)	5 (3%)
Metafilter	166	151 (91%)	9 (5%)	6 (4%)
Yahoo	163	93 (57%)	58 (36%)	12 (7%)
Overall	490	308 (63%)	159 (32%)	23 (5%)

Table 5-2. Number of questions by coding result. "Disagreement" represents the case where four coders failed to reach a majority classification.

Informational questions are most often asked in such a way that either advice or objective facts are appropriate in response (52%); fewer informational questions are coded appropriate only for objective responses (41%), while relatively few informational questions are coded appropriate just for advice (7%). Different sites have different distributions of APPROPRIATE RESPONSE ($p<0.01$, $\chi^2 = 40.7$). For instance, 68% of informational questions in Ask Metafilter are appropriate for either advice or objective facts, compared with just 28% of informational questions in Answerbag. See Table 5-3 for more details.

	Total	Objective Only	Advice Only	Either/Both Ok
Answerbag	64	42 (66%)	4 (6%)	18 (28%)
Metafilter	151	36 (24%)	12 (8%)	103 (68%)
Yahoo	93	48 (52%)	5 (5%)	40 (43%)
Overall	308	126 (41%)	21 (7%)	161 (52%)

Table 5-3. Number of informational questions by APPROPRIATE RESPONSE category.

Informational questions may be highly personalized or completely generic, though the average degree of personalization varies by site (see Figure 5-5). Informational questions are most personalized to the asker's situation in Metafilter, and the most generic in Answerbag (means: Answerbag: 2.0, Metafilter: 3.2, Yahoo: 2.5). These differences are all statistically significant using a Tukey-Kramer HSD test (α=0.05) (Kramer 1956).

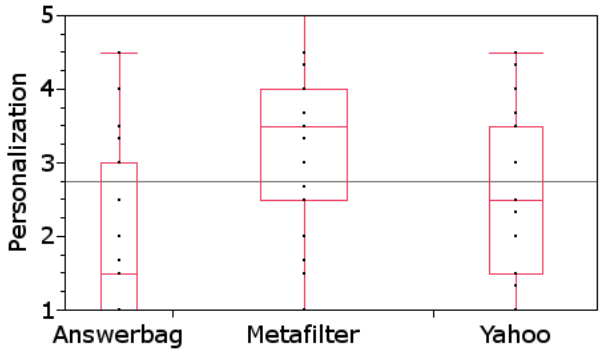


Figure 5-5. A box plot of the degree of personalization per question at each site, where higher scores mean questions that are more personalized to the asker's situation. The ends of the boxes are at the 25th and 75th percentiles. The horizontal line at 2.7 represents the overall mean personalization per question across destinations. The width of each box is proportional to the number of observations in this analysis.

In Figure 5-6, we compare the three Q&A sites in terms of coders' responses to our WRITING QUALITY and ARCHIVAL VALUE questions. Ask Metafilter was rated, on average, to have the highest WRITING QUALITY scores (means: Answerbag=3.2, Metafilter=3.9, Yahoo=2.7). Using a Tukey-Kramer HSD test, we find that all pairwise comparisons indicate statistically significant differences (α=0.05). Ask Metafilter was also rated, on average, to have the highest ARCHIVAL VALUE scores (means: Answerbag=2.3, Metafilter=3.69, Yahoo=2.4). An HSD test shows that the difference between Ask Metafilter and the other two sites is significant, but that the difference between Answerbag and Yahoo is not (α=0.05). More notably, the most common assessment for Yahoo and Answerbag is that the questions will not yield archival value (1 or 2 on a 5-point scale), while 1 and 2 are the least common assessments for Metafilter.

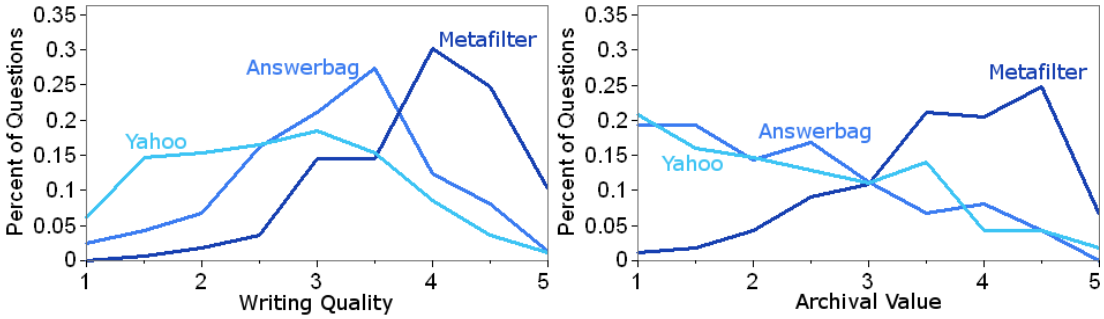


Figure 5-6. Distribution of aggregate scores assigned to questions (rounded to the nearest 0.5), split by Q&A site. Higher values are better.

5.4.3 Archival Value and Writing Quality by Question Type

On average, conversational questions received lower scores on both our WRITING QUALITY metric (means: Conversational: 2.9, Informational: 3.4) and our ARCHIVAL VALUE metric (means: Conversational: 1.7, Informational: 3.3). Figure 5-7 shows the full results. Notably, we find that fewer than 1% of conversational questions received ARCHIVAL VALUE scores of 4 (agree) or above.

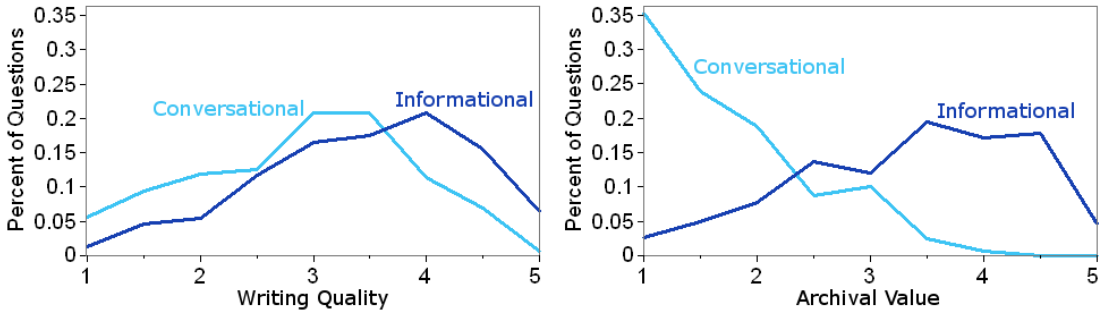


Figure 5-7. Distribution of aggregate scores assigned to questions (rounded to the nearest 0.5), split by question type. Higher values are better.

To isolate the effects of question type from the site in which it is asked, we build a regression model that includes Q&A site, question type, and the interaction of these two variables. After controlling for site, we still find that conversational question type is a significant negative predictor for both WRITING QUALITY ($p<0.01$, $F=10.7$) and ARCHIVAL VALUE ($p<0.01$, $F=125.7$). In the resulting prediction expressions, a conversational question lowers the WRITING QUALITY prediction by 0.36 points, and the ARCHIVAL VALUE prediction by 1.32 points.

Informational questions vary in quality based on the coded APPROPRIATE RESPONSE variable. Those that are receptive to *both* objective fact-based and advice-based answers have the highest ARCHIVAL VALUE scores (means: Objective:3.0, Advice:2.9, Both:3.6). See Figure 5-8. This difference is statistically significant using a Tukey-Kramer HSD test (α=0.05), and remains statistically significant after controlling for site in a regression model (F=9.1, p<0.01). APPROPRIATE RESPONSE is not a statistically significant predictor of WRITING QUALITY (F=1.9, p=0.15).

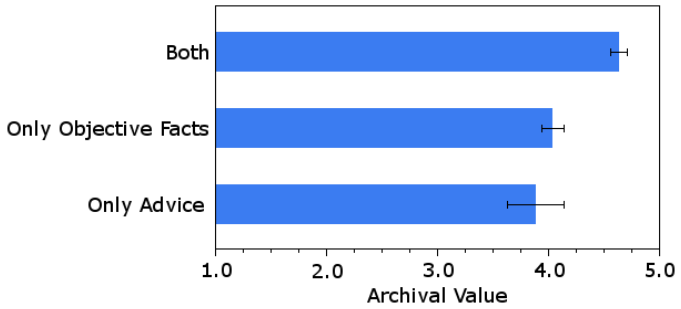


Figure 5-8. Mean archival value scores for informational questions (with standard error bars), by the type of answers that would be appropriate in answering the question: advice only, objective facts only, or both advice and objective facts.

There is not a strong relationship in this data set between PERSONALIZED in an informational question and the WRITING QUALITY or ARCHIVAL VALUE outcome measures. There is a trend towards lower WRITING QUALITY scores as PERSONALIZED increases (F=2.8, p=0.09) that is observable across all three sites. There is also a trend towards an interaction effect between site and personalization in this study (F=2.9, p=0.06): Ask Metafilter questions average slightly lower ARCHIVAL VALUE scores with increased personalization, whereas the other two sites average slightly higher scores as personalization increases.

5.4.4 Discussion

RQ1. Humans can reliably distinguish between conversational and informational questions in most cases. The first two coders agreed 87.1% of the time, while only 4.7% of questions failed to receive a majority classification.

However, the 12.9% of questions where the first two coders disagreed indicate that there is a class of questions that contain elements of both conversational and informational questions. Two example questions that received split decisions are *"how many people are on answer bag"*

and *"Is it me or do ipods vibrate?"* In each case, it is hard to determine whether the primary intent of the question asker is to learn information or to start a conversation. Because of these ambiguities, we cannot expect either humans or computers to achieve 100% accuracy in classifying questions.

RQ2. Conversational questions are associated with lower writing quality and lower potential archival value than informational questions. This effect is robust even after controlling for the site in which a question is asked. Few conversational questions appear to have the potential to provide long-term informational value, even if we assume the given answers are high quality. See Appendix C for an example of a high rated and a low rated question of each type.

Among informational questions, we find that questions that may be answered with *either* objective facts or personalized advice are judged to have the highest potential archival value. Thus, many of the "best" questions (from an informational standpoint) appear to have some objective basis, along with some unique aspect. Surprisingly, questions seeking only advice are rated to have similar potential archival value as questions seeking only objective facts – it is not correct to assume that advice-oriented questions have low informational potential. It is also interesting to note that questions seeking advice are much longer than those not seeking advice (median # characters: advice appropriate: 542, advice not appropriate: 156) and are also more personalized (mean PERSONALIZED: advice appropriate: 3.4, advice not appropriate: 1.9; t-test: $p<0.01$). There is no corresponding effect for questions seeking objective facts.

Surprisingly, the degree of the personalization of an informational question is poorly correlated with its potential for archival value ($\rho=0.12$). We do find, in predicting ARCHIVAL VALUE from PERSONALIZED, that a quadratic polynomial more closely fits the data than a linear model (quadratic: $R^2=0.06$, linear: $R^2=0.02$). This quadratic model is concave, with the peak coinciding with archival value of approximately 3 (the middle value in the rating scale). Thus, there is some evidence that there is a sweet spot for personalization: questions may be too specific to be widely useful, or too general to present opportunities for unique answers.

5.5 Structural Differences And Classifiers

We now turn our attention to the task of using machine learning algorithms to detect question type at the time a question is asked. These models help us to understand the structural properties of questions that are predictive of relationship-oriented or information-oriented intent. Specifically, we address RQ3 and begin to address Research Challenge 1 by learning about the categorical, linguistic, and social differences that are indicators of question type.

5.5.1 Machine Learning Methods and Metrics

In general terms, our task is to use the data available at the time a question is asked online to predict the conversational or informational intent of a question asker. To accomplish this task, we employ supervised machine learning algorithms (see (Mitchell 1997) for more information).

We use three primary metrics in reporting the performance of our machine learning classifiers: sensitivity, specificity, and area under the ROC curve (AUC). Because the output of our classification algorithm does not have a clear division between "positive" and "negative" values, we arbitrarily consider a positive to be a conversational question. For example, a "true-positive" is correctly predicting a conversational question, and a "false-negative" is incorrectly predicting an informational question. Our metrics may be interpreted as:

- **Sensitivity** - proportion of conversational questions that are correctly classified
- **Specificity** - proportion of informational questions that are correctly classified
- **AUC** - (area under ROC curve) a single scalar value representing the overall performance of the classifier.

For more information about these metrics, see (Fawcett 2004).

There is an important reason why we choose to use sensitivity and specificity (more common in the medical literature) over precision and recall (more common in the information retrieval literature): our data have different proportions of positives and negatives across the three Q&A sites. In general, precision and recall suffer from what is known as "class skew", where apparent classification performance is impacted by the overall ratio of positives to negatives in the data

set (Fawcett 2004). Thus, if we were to use precision and recall metrics, we could not fairly compare classifier performance across sites with different conversational/informational ratios: performance would appear worse for sites with few conversational questions.

Unless otherwise noted, we employ 5-fold cross validation to evaluate performance. Our data set excludes those "split decision" questions that received 2 votes each for informational and conversational by the coders. "Overall" performance of classifiers across all three sites represents a mathematical combination of the individual performance statistics rather than a fourth (i.e., unified) model. Finally, because the three site-specific ROC curves represent predictions over different datasets, in our overall performance we simply report the mean of AUC scores, rather than first averaging the individual ROC curves. Unless otherwise noted, we use the Weka data mining software package to build these predictive models (Witten 2005).

5.5.2 Baseline

We provide a baseline model as a frame of reference for interpreting our results: a 0-R algorithm that always predicts the most commonly occurring class. For example, in Yahoo Answers, 62% of the (non-excluded) questions were coded as informational, so the 0-R algorithm will always predict informational. See Table 5-4 for results. Note that our baseline outperforms random guessing, which would converge to an overall performance of sensitivity=0.5 and specificity=0.5.

	Sensitivity	Specificity	AUC
Answerbag	1.00	0.00	0.50
Metafilter	0.00	1.00	0.50
Yahoo	0.00	1.00	0.50
Overall	0.58	0.79	0.50

Table 5-4. Performance of the 0-R baseline classifier.

5.5.3 Predicting Type Using Category Data

Social Q&A sites typically organize questions with categories or tags. Prior work on Yahoo Answers showed that some categories resemble "expertise sharing forums", while others resemble "discussion forums" (Adamic 2008). In this section, we test the accuracy with which it is possible to classify a question as conversational or informational with just knowledge of that question's category.

All three sites in our study use categories, but they use them in different ways. At the time of our data collection, Metafilter had a flat set of 20 categories, while Answerbag had over 5,700 hierarchical categories and Yahoo had over 1,600 hierarchical categories. The problem with building a classifier across so many categories is that it is difficult to collect a training set that is populated with sufficient data for any given category. However, none of the three sites had more than 26 "top-level categories". Thus, to improve the coverage of our training data, we develop two features:

- "Top-level category" (TLC), a feature that maps each question to its most general classifier in the category hierarchy. For example, in Yahoo, the "Polls & Surveys" category is part of the TLC "Entertainment & Music".

- "Low-level category" (LLC), a feature that maps each question to its most specific category. However, we only populate this feature when we've seen a low-level category at least 3 times across our dataset. For example, we only have one coded example from Answerbag's "Kittens" category, and so leave that question's LLC unspecified in the feature set.

We implement this classifier with a Bayesian network algorithm. This classifier is able to improve on the baseline classifier (sensitivity=0.77, specificity=0.72, AUC=0.78). Compared to the 0-R baseline, the category-based classifier improves sensitivity 18%, but worsens specificity by 4%; see Table 5-5. In particular, we note that this classifier appears to work quite well on the Metafilter dataset – one where there are only 9 conversational instances – achieving AUC of 0.82.

	Sensitivity	**Specificity**	**AUC**
Answerbag	0.82	0.41	0.71
Metafilter	0.56	0.95	0.82
Yahoo	0.66	0.82	0.81
Overall	0.77	0.72	0.78

Table 5-5. Performance of the category-based classifier.

To better understand the relationship between category and question type, we may look at examples from across the three sites. Table 5-6 shows the three most popular categories in each Q&A site in our coded dataset. We may see from these examples that some categories provide a

strong signal concerning question type. For example, questions in Answerbag's "Outside the bag" category are unlikely to be informational. However, we find few TLCs that are unambiguously predictive.

The addition of low-level categories to the set of features does slightly improve performance as compared with a classifier trained only on TLCs. These categories appear to be especially important in the case of Yahoo, where some categories provide a very strong signal concerning the presence of conversational questions. For instance, Yahoo's Polls & Surveys category was rated 100% conversational (14/14), Singles & Dating was rated 70% conversational (7/10), and Religion & Spirituality was rated 100% conversational (6/6). See Table 5-7.

Answerbag	Outside the bag (19/20) Relationship advice (12/15) Entertainment (9/14)
Metafilter	computers & internet (0/34) media & arts (3/19) travel & transportation (1/16)
Yahoo	Entertainment & Music (20/26) Health (6/22) Science & Mathematics (2/12)

Table 5-6. Top 3 Top-Level Categories (TLCs) in the coded dataset and the fraction of conversational questions. Few TLCs provide an unambiguous signal regarding question type.

Answerbag	Outside the bag (12/13) Life & Society (4/4) Health & Fitness (2/4)
Yahoo	Polls & Surveys (14/14) Singles & Dating (7/10) Religion & Spirituality (6/6)

Table 5-7. Top 3 Low-Level Categories (LLCs) in the coded dataset and the fraction of conversational questions. Metafilter does not have LLCs, and is excluded from this table. Note that in Answerbag, it is possible a question to have the same LLC and TLC. E.g., the set of questions in the TLC "Outside the bag" is a superset of questions that actually belong to the category "Outside the bag".

5.5.4 Predicting Type Using Text Classification

One of the most effective tools in distinguishing between legitimate email and spam is the use of text categorization techniques (e.g., Sahami 1998). One of the insights of this line of work is

that the words used in a document can be used to categorize that document. For example, researchers have used text classification to algorithmically categorize Usenet posts based on the presence of requests or personal introductions (Burke 2007). In this section, we apply this approach to our prediction problem, in the process learning whether conversational questions contain different language from informational questions.

The text classification technique that we use depends on a "bag of words" as the input feature set. To generate this feature set, we parse the text of the question to generate a list of lower-case words and bigrams. To improve the classifier's accuracy, we only kept the 500 most-used words or bigrams in each type of question. We did not discard stopwords as they turned out to improve the performance of the classifier. We used Weka's sequential minimum optimization (SMO) algorithm.

The text classifier does outperform the baseline classifier (sensitivity=0.70, specificity=0.85, AUC=0.62) but, surprisingly, does not improve on the category-based classifier. However, there is reason for optimism, as this is the best-performing classifier so far on the Answerbag data set. See Table 5-8 for details.

	Sensitivity	Specificity	AUC
Answerbag	0.79	0.70	0.75
Metafilter	0.00	1.00	0.50
Yahoo	0.48	0.71	0.60
Overall	0.70	0.85	0.62

Table 5-8. Performance of the text-based classifier.

We now evaluate individual tokens, based on their impact on classification performance as measured by information gain, a metric that represents how cleanly the presence of an attribute splits the data into the desired categories (Kullback 1951). In this study, information gain can range from 0 (no information) to .93 (perfect information).

Question words. Questions in English often contain "interrogative words" such as "who", "what", "where", "when", "why", and "how". 75.4% of the questions in our coded dataset contain one or more of these words, the most common being the word "what", used in 40.3% of questions. While several of these words ("who", "what, "when") appear to be used in roughly equal proportion across conversational and informational questions, the words "how" and

"where" are used much more frequently in informational questions, while the word "why" is used much more frequently in conversational questions. See Table 5-9 for details.

	% Informational	% Conversational	Information Gain
where	**15.5%**	1.9%	0.033
how	**29.8%**	11.3%	0.029
why	5.6%	**17.0%**	0.012
what	30.2%	34.0%	< 0.001
who	13.9%	11.3%	< 0.001
when	17.0%	13.2%	< 0.001

Table 5-9. Six common interrogative words, and the percentage of questions that contain one or more instances of these words.

I vs. you. Conversational questions are more often directed at readers by using the word "you", while informational questions are more often focused on the asker by using the word "I". The word "I" is the highest-ranked token in our dataset based on information gain. See Table 5-10 for details.

	% Informational	% Conversational	Information Gain
I	**68.6%**	27.4%	0.124
you	25.8%	**54.7%**	0.050

Table 5-10. A higher percentage of questions with informational intent contain the word "I", while a higher percentage of questions with conversational intent contain the word "you".

Strong Predictors. To find textual features that possess a strong signal concerning question type, we sort all features by their information gain. In Table 5-11, we show the top features after filtering out common stopwords (they tended to be predictive of informational questions). Again, we find that questions directed at the reader (i.e., using the word "you") are conversational: phrases such as "do you", "would you" and "is your" have high information gain and are predictive of conversational intent. On the other hand, informational questions use words that reflect the need for the readers' assistance, such as "can", "is there", and "help".

	% Informational	% Conversational	Information Gain
can	**35.1%**	9.4%	0.069
is there	**10.4%**	0.6%	0.032
help	**19.8%**	5.7%	0.029
do I	**12.3%**	1.9%	0.028
do you	4.9%	**22.0%**	0.047
would you	1.3%	**8.8%**	0.023
you think	1.3%	**8.2%**	0.021
is your	0.0%	**3.7%**	0.020

Table 5-11. Tokens that are strong predictors of conversational or informational intent, sorted by information gain.

5.5.5 Predicting Type Using Social Network Metrics

Social network methods have emerged to help us visualize and quantify the differences among users in online communities. These methods are based on one key insight: to understand a user's role in a social system, we might look at who that user interacts with, and how often (Fisher 2006). In this way, each user can be represented by a social network "signature" (Welser 2007) – a mathematical representation of that user's *ego network*: the graph defined by a user, that user's neighbors, and the ties between these users (Hanneman 2005). Researchers have used these signatures to better understand canonical user roles in online communities (Welser 2007), including *answer people*, who answer many people's questions, and *discussion people*, who interact often with other discussion people. While these roles were developed in the context of Usenet, a more traditional online discussion forum, they are adaptable (and useful) in the Q&A domain (Adamic 2008).

In this section, we use social network signatures to predict question type. This analysis builds on features that describe each user's history of question asking and answering. We treat the Q&A dataset (summarized in Table 5-1) as a directed graph, where vertices represent users and directed edges represent the act of one user answering another user's question (see Figure 5-9 for an example).

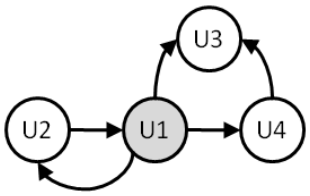


Figure 5-9. An example ego network for user U1. U1 has answered questions by U2, U3, and U4, while U2 has answered a question by U1. U1's metrics: *NUM_NEIGHBORS*=3, *PCT_ANSWERS*=0.75, *CLUST_COEFFICIENT*=0.17.

We allow multiple edges between pairs of vertices. For example, if user A has answered two of user B's questions, there are two edges directed from A to B. We model the three datasets collected from separate Q&A as three separate social networks. Because we wish to build models that can make predictions at the time a question is asked, we filter these structures by timestamp, ensuring that we have an accurate snapshot of a user's interactions up to a particular time.

To make use of these social networks in our machine learning framework, we construct the following features, with respect to the question asker (V):

NUM_NEIGHBORS. *The number of neighbors to the question asker.* This feature captures the number of users V has interacted with prior to this question. In our example in Figure 5-9, U1 has answered questions asked by U2, U3, and U4 and received an answer from U2. In total, U1 has interacted with three users (U2, U3, and U4); NUM_NEIGHBORS=3.

PCT_ANSWERS. *The question asker's number of outbound edges as a percentage of all edges connected to that user.* This feature captures the number of questions V has answered divided by the sum of that number and the number of responses to V's own questions. In our example in Figure 5-9, U1 has provided three answers (to questions asked by U2, U3, and U4) and received one answer (from U2); PCT_ANSWERS = 3/4 = 0.75.

CLUST_COEFFICIENT. *The clustering coefficient of the question asker's ego network*. This feature captures the degree to which V's neighbors have interacted with one another. We use the definition of clustering coefficient presented in (Watts 1998), modified to accommodate a

directed graph[15]. Suppose V has n neighbors. Then, at most n(n-1) directed edges can exist if each neighbor of V is connected to each other neighbor of V *in both directions*. CLUST_COEFFICIENT measures the fraction of these allowable edges that actually exist in the Q&A network. In our example in Figure 5-9, U1's three neighbors can potentially have 6 interconnecting links with each other. Only one of these potential links exists: U4 provided an answer to U3. Thus, CLUST_COEFFICIENT = 1/6 = 0.17.

The social network-based classifier, implemented with a Bayesian network algorithm, is able to correctly classify 69% of conversational questions and 89% of informational questions, overall. While the model appears successful across both Yahoo and Answerbag, performance is low at Metafilter (AUC=0.64). See Table 5-12 for details.

	Sensitivity	**Specificity**	**AUC**
Answerbag	0.71	0.87	0.81
Metafilter	0.87	0.61	0.72
Yahoo	0.00	1.00	0.64
Overall	0.69	0.89	0.72

Table 5-12. Performance of the social network-based classifier.

Analyzing a social network built from Q&A discourse reveals strong differences between users who ask questions with a conversational intent and users who ask questions with informational intent (see Figure 5-10 for an overview). The first metric, NUM_NEIGHBORS, shows that users who ask conversational questions tend to have many more neighbors than users who ask informational questions (means: Conversational=757, Informational= 252). This effect is significant across the three sites ($p<0.01$), as well as within each of the three sites ($p<0.01$ for all three).

[15] We discard superfluous parallel edges for this analysis.

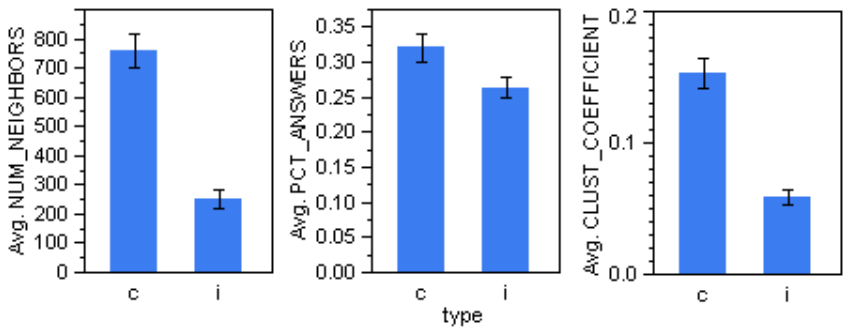


Figure 5-10. Differences in the three social network metrics across question type (c=conversational, i=informational), shown with standard error bars.

Conversational question askers tend to have higher PCT_ANSWERS scores than informational question askers (mean PCT_ANSWERS: Conversational=0.32 vs. Informational=0.26, $p=0.02$). This effect is robust across Yahoo ($p=0.02$) and Answerbag ($p<0.01$), though Metafilter exhibits the reverse effect ($p<0.01$) where informational askers have a higher PCT_ANSWERS score than conversational askers.

Finally, looking at CLUST_COEFFICIENT reveals that users asking conversational questions have more densely interconnected ego networks than users asking informational questions (mean CLUST_COEFFICIENT: conversational=0.15 vs. informational=0.06, $p<0.01$). This effect is robust across Yahoo ($p<0.01$) and Answerbag ($p<0.01$), but not significant in Metafilter ($p=0.35$).

5.5.6 Discussion

RQ3. There are several structural properties of conversational and informational questions that help us to understand the differences between these two question types. Though site-specific, some categories are strongly predictive of question type, such as the top-level category "computers & internet" in Metafilter (predictive of informational) and the low-level category "Polls & Surveys" in Yahoo Answers (predictive of conversational). Certain words are also strongly predictive of question type. For example, the word "you" is a strong indicator that a question is conversational. Finally, users that ask conversational questions tend to have larger, more tightly interconnected ego networks.

Research Challenge 1. All three of the models presented in this section outperform our baseline model. However, each individual model exhibits strengths and weaknesses. Moving forward, we attempt to address these limitations through the use of ensemble methods.

5.6 An Ensemble For Predicting Question Type

Intuitively, it would seem that our different feature sets complement one another, rather than simply presenting several different ways of arriving at the same predictions. In general, we assert a belief that in online content analysis, the use of multiple complementary feature sets is superior to the use of a single feature set.

5.6.1 Classifier Diversity

The key idea of ensemble learning methods is that we may build many classifiers over the same data, then combine their outputs in such a way that we capitalize on each classifier's strengths. In our case, we may intuitively note that a text classifier may perform better in the presence of more text, or that a category classifier will more accurately classify questions in some categories than in others. Thus, to assess the potential of our classifiers to collaborate, we first assess whether they are making errors on the same questions.

To measure our potential for improvement through ensemble methods, we may turn to diversity metrics (Polikar 2006). These metrics quantitatively measure the tendency for a pair of classifiers to make the same mistakes. In this analysis, we choose Yule's Q Statistic (Yule 1900). This metric varies from -1 to 1; classifiers that tend to categorize the same instances correctly take more positive values, while classifiers that tend to categorize different instances incorrectly take more negative values. Thus, lower values indicate greater diversity (Kuncheva 2000).

We find that the category-based and social network-based classifiers appear to pick up on much of the same signal (Q=0.72), perhaps showing that the same “types” of users tend to post in the same categories. On the other hand, we see better diversity scores when comparing the output of text-based and category-based classifiers (Q=0.31) as well as the social network-based and text-based classifiers (Q=0.58).

5.6.2 Algorithm Details and Results

We construct an ensemble classifier by running a meta-classifier on the output of three individual classifiers: a confidence score between 0 and 1 that a question is conversational. See Figure 5-11 for an overview of the architecture. Intuitively, we might believe that our meta-classifier picks up on signals concerning which classifier is strongest in each site and what different confidence scores mean in each classifier. We used JMP's neural network algorithm for this classifier, and report the results of 5-fold cross validation on reordered data.

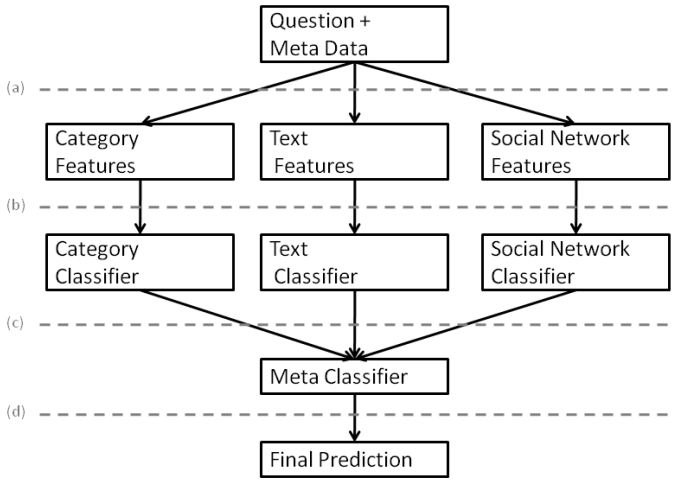


Figure 5-11. An overview of the architecture of the ensemble classifier. In stage (a), we extract features from the question to be classified, as described above. We read these features into the specialized classifiers in stage (b). In stage (c), each specialized classifier outputs a confidence score that serves as the input to the meta classifier. In stage (d), the meta classifier generates a final prediction: conversational or informational.

The ensemble method shows strong improvement over any of the individual classifiers for each site. This classifier returns AUC values > 0.9 for each site, correctly classifying 79% of conversational content and 95% of informational content (see Table 5-13 and Table 5-14), for an overall accuracy of 89.7%.

	Sensitivity	Specificity	AUC
Answerbag	0.85	0.84	0.91
Metafilter	0.33	1.00	0.91
Yahoo	0.78	0.95	0.95
Overall	0.79	0.95	0.92

Table 5-13. Performance of the ensemble classifier.

		Predicted	
		i	c
Actual	i	54	10
	c	14	78

Answerbag

		Predicted	
		i	c
Actual	i	151	0
	c	6	3

Metafilter

		Predicted	
		i	c
Actual	i	88	5
	c	13	45

Yahoo

Table 5-14. Confusion matrices for the ensemble classifier across the three datasets (i=informational, c=conversational).

All three individual classifiers mutually agreed on a classification 61.9% of the time – in these cases, the ensemble achieved 93.8% accuracy. In the remaining 38.1% of the instances where there was disagreement, the ensemble achieved just 83.2% accuracy. For example, the question *"What operating system do you prefer? Windows, Linux, Mac etc."* was correctly classified as conversational, despite a wrong prediction by the category-based classifier (the question is in Answerbag's Operating systems category). However, the question *"Which 1 do u need more??? Money or Love???"* was incorrectly classified as informational, as only the category-based classifier was correct (the question is in Yahoo's Polls & Surveys category). There is future work in using multiple learners as a means for generating confidence scores in the resulting classification.

5.6.3 Discussion

Research Challenge 1. Our classifier achieves 89.7% classification accuracy across the three Q&A sites, close to the rate at which the first two human coders agreed on a classification (91.4%). Unsurprisingly, the classifier was more accurate on these questions where the first two coders agreed than on questions where four coders were required (92.0% vs. 65.9% classification accuracy). Based on these results, we are optimistic about the potential for algorithms that are at least as reliable as humans for distinguishing conversational and informational questions.

What of the questions that received a split decision by coders? We trained a meta-classifier for each of the three sites with the full data set, then generated predictions for each of the questions that generated disagreement among the coders. See Table 5-15 for a summary. Perhaps unsurprisingly, the three individual classifiers agreed only 52.2% of the time (recall, they agree

61.9% of the time across questions in our primary dataset). Within Metafilter, the three classifiers always agree that the question is informational; within the other two sites there is substantially less agreement. Our optimal classifier is somewhat conservative in classifying questions as conversational: it classifies only 22% of these borderline questions as conversational.

	# Split Vote Questions	# Questions: All Classifiers Agree	# Questions: Coded Conversational
Answerbag	5	0	3 (60%)
Metafilter	6	6	0 (0%)
Yahoo	12	6	2 (17%)
Overall	23	12	5 (22%)

Table 5-15. Results of running the ensemble classifiers on the 23 questions where coders were split between conversational and informational, including the number of instances where all three constituent classifiers agreed on the classification, and the number of instances where the final prediction was conversational.

5.7 Summary Discussion And Design Implications

In this chapter, we investigated the phenomenon of conversational question asking in social Q&A sites, finding that few questions of this type appear to have potential archival value. We explored the use of machine learning techniques to automatically distinguish conversational and informational questions, finding that an ensemble of techniques is remarkably effective, and in the process learning about categorical, linguistic, and social differences between the question types.

Not included in this chapter are several classification methods that did not work well. For instance, we built a classifier from quantitative features extracted from the text (e.g., question length, and Flesch-Kincaid grade level) that barely outperformed our baseline classifier: apparently conversational and informational questions "look" similar! Also, we tried a suite of additional features to supplement the social network model, none of which improved performance. However, there is potentially more signal to be mined from the text of a question than we have realized. Natural language-based methods could extract additional features to

supplement our naïve "bag of words" feature set, though NLP programmers beware: the language used in Q&A sites often barely resembles English! (LOL)

We acknowledge that looking at questions in isolation is not necessarily the best way to classify a Q&A thread. For example, the question *"Why is the sky blue?"* might appear to be informational in intent, until you realize that this question has been asked over 2,000 times in Yahoo Answers, and often receives no replies with serious answers to the question. Although we addressed the classification problem at the time a question is asked, it is certainly possible to classify a question hours or days after it is asked, utilizing information about the answerers and the text of the answers.

However, classifying questions at the time they are asked can allow for effective automated tagging, and for the development of systems that direct questions to the appropriate place. Such classification could lead to the development of interfaces that support both social users (who drive traffic numbers and advertising revenue) and informational users (who generate searchable content). Q&A sites could adjust user reward mechanisms based on question type, offer different "best answer" interfaces to enhance the fun of participating in conversational Q&A, or create search tools that allow users to perform informational search (emphasizing keywords) or conversational search (emphasizing timeliness).

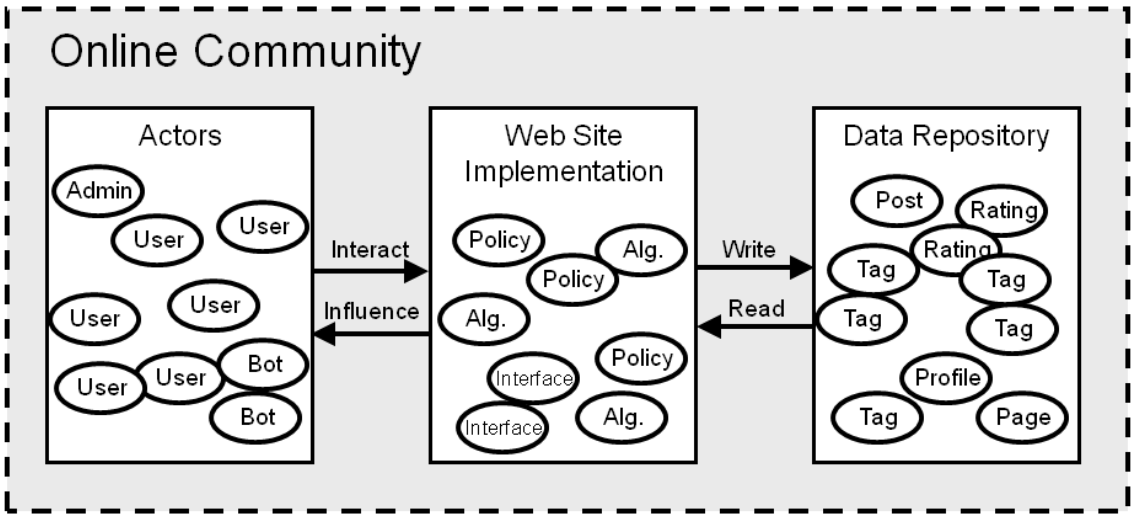
Social Q&A is certainly not the only online domain struggling to better understand how to simultaneously attract committed users and encourage high quality contributions. Wikis, discussion forums, blogs, and other forms of social media are an increasing fraction of the searchable Web; across all of these types of sites, it is important to understand the implications of users' roles, motivations, and intentions. In this work, we use a combination of human evaluation and machine learning techniques to explore Q&A questions, discovering that conversational question asking is an important indicator of information quality. We hope that these methods will prove useful to others in the research community as they pursue related questions across different domains to better understand how to identify and harvest archival-quality user-contributed content.

Chapter 6

Context, Contributions, and Future Work

In the past four chapters, we've looked at research studies that examine the interplay between social design and user contributions in online communities. In each chapter, we included a focused discussion of the most relevant related work and contributions. In this chapter, we broaden our perspective to situate the research, discuss the contributions in a larger context, and present opportunities for future work.

We began this thesis with a conceptual diagram:



6-1. A conceptual diagram that summarizes our perspective on online communities. We identify three primary components: the actors who participate in the community, the Web site implementation that mediates that participation, and the repository that stores the results of that participation.

This diagram represents an online community as a socio-technical entity – it is at once a place where people interact, and an engineered technology. The relationship between the primary components in this diagram represents an interesting situation: the designer who controls the implementation details might strongly influence user behavior, even if it is the users themselves

who generate most of the viewable content. We are interested in both the social and the technical aspects that come together in the study of social design practices.

We might summarize this thesis by situating our contributions and methodologies within this conceptual diagram. In Chapters 2 and 3, we extracted users' contributions from a data repository as the basis for creating new personalized interfaces; we subsequently measured the extent to which these interfaces influenced users' behavior. In Chapter 4, we looked at the influence of different policies on user behavior by posing as users ourselves, and measuring the differences in responses to our questions; this method allowed us to examine the influence of different policies on user interactions. Finally, in Chapter 5, we built a system for automatically classifying users' questions as conversational or informational; this work exists primarily at the level of the data repository.

There are several other ways to summarize and frame the contributions in this thesis. In the coming sections, we focus on four perspectives:

- *Theoretical Perspective*. Theories of social influence offer a lens through which we may interpret our findings. We introduce this theory and its relationship to this thesis, then examine the potential implications of our empirical findings.

- *Methodological Perspective*. Methods for studying online communities continue to evolve. We discuss known techniques from the literature that inform our work, and novel techniques that we contribute in this thesis.

- *Computational Perspective*. Algorithms are emerging to support increasingly specific goals in support of online communities. We discuss known techniques for personalizing interfaces and recommending content, and summarize new techniques that we contribute.

- *Empirical Perspective*. The online communities research field is amassing a corpus of results that inform the decisions made by site designers. We discuss some of the most important results from the field, and summarize our highest-level contributions.

6.1 Theoretical Context

One theme unifying this thesis is the idea that the representation of the social aspect of an online community plays an important role in influencing the actions of community members. This theme is highly related to the idea of "social influence" – the study of how people's behavior affects the actions of others. Examples of social influence include people's propensity to reciprocate others' actions, and their propensity to use others' reactions to a phenomenon as a proxy for forming their own reactions (Cialdini 1998).

Broadly, we may divide the social influence literature into two categories: compliance and conformity (Cialdini 2004). Compliance research is concerned with the use of explicit appeals or tactics to change individuals' behavior, while conformity is concerned with the more subtle effects of unintentional influence.

Conformity

Users may contribute to an online community for a variety of reasons. They might choose actions that maximize benefits (e.g., the benefits of having fun or learning new information) while minimizing costs (e.g., the costs of spending time and effort), as predicted by a rational agent model (e.g, Smith 1776). However, while this model is capable of making bold predictions about human actions, its predictions turn out to be relatively inaccurate. There are several reasons for this shortcoming: we humans are not rational in all situations, we do not necessarily have stable or consistent preferences, and we do not have perfect knowledge of costs and benefits in making decisions (Rabin 1993, Camerer 2004).

One prominent theory of "conformity" (Bernheim 1994) has emerged to help explain how social pressures can account for peoples' lack of rationality. This theory incorporates the notion that people care about their status, and that in cases when the status implications of an action are more important than the cost/benefit implications, people will "conform" to some homogeneous standard. This theory has been empirically tested in a number of domains, such as littering (Kallgren 2000), charitable contributions (Croson 2005), and recycling (Schultz 1999). In each case, researchers found that people were influenced by information about others' actions.

This is closely related to the idea of "social comparisons" (Suls 2002) which theorizes that people reference others' performances in order to set a baseline for their own behavior. Festinger, in his classic work on social comparison (Festinger 1954) theorized that we compare ourselves to others who are better off for guidance, while we compare ourselves to others who are worse off to increase our self-esteem. Lockwood and Kunda (Lockwood 1997) find additional evidence that comparing ourselves to others who are better off can inspire us, as long as success feels attainable.

One notion that is central to how we evaluate ourselves in a community context is "reputation" (Tyler 1999). While in real world communities, we might use signs of respect as a proxy for reputation (Cremer 2005), online communities have the potential to make reputation explicit (e.g., Resnick 2000). However, determining the effect of reputation on behavior is complex, and depends on knowledge of a person's propensity to weight reputation in decision making (Cremer 2005) and the distribution of reputation across a community (Whitmeyer 2000).

Compliance

Social influence may be applied directly. For example, people might simply make requests of one another. While this may seem trivial, Langer (1978) provided empirical evidence that people will often adhere to requests, even if the reason for the request is unrelated or inconsequential. The way in which requests are delivered is important. Garner, for example, found that personalized notes were more effective than non-personalized notes in soliciting time consuming work (Garner 2005).

Another mechanism for soliciting effort from individuals is the use of positive or negative incentives. Several types of organizational incentives were discussed in (Clark 1961), such as material, social, goal-oriented, and status-oriented rewards. Incentives also exhibit nuanced effects, though. Several studies have found that the use of monetary rewards can displace voluntary effort (Fehr 2002; Deci 1999). Additionally, studies have found that while positive rewards tend to encourage an elite few, negative incentives (punishments) encourage broad-participation at the cost of potential discord (Oliver 1980).

6.2 Theoretical Contributions

Several empirical results from this thesis inform the body of work on social influence. In particular, little work has been done validating the predictions from this theoretical area in the context of online communities, where the idea of a "social interaction" is perhaps very different from the context of offline communication.

In chapter 2, we studied a conformity technique: the use of social comparisons. We found that upward comparisons – where we told members that they had rated fewer movies than other members – had a motivating and focusing effect on those members. Members receiving a message showing they had rated fewer movies than other members like them rated more movies than members receiving a similar message with no social comparison. Lockwood (1997) found that upwards comparisons can be motivating, as long as success is attainable. This study further validates this finding. Wheeler and Miyake (1992) found that upwards comparisons can decrease peoples' sense of well being. However, we found no statistically significant differences among members receiving different directions of comparison, when asked whether they agreed "I didn't care".

In chapter 3, we investigated a compliance technique: asking members to write or read posts in a discussion forum. We found evidence supporting Langer (1978), who found that the simple act of asking for something often leads to compliance. We asked users to write posts, and they did so at a rate higher than members who were not asked. Further, we found evidence that the language of the request has a strong effect on the results. Using the simple phrase "check out the following new post" led to much lower click-through rates than other phrases emphasizing that the Web site's algorithm was making a recommendation. This language, using phrases like "our system predicts" and "we think" also emphasized the credibility of the Web site in understanding the preferences and history of the user..

However, in Chapter 4, we found a surprising lack of results with regard to the use of language in making requests of other users. The presence of statements of gratitude, or statements indicating that the person making the request has already looked for help elsewhere, had either no effect or effects that changed greatly depending on the particular community. Just as Garner's use of a sticky note (Garner 2005) increased compliance, we imagine that appending a

short message of gratitude would do so. However, perhaps this is not the case in an online community, where the potential for reciprocity is dampened. And, perhaps the statement that the person making the request had already looked for information made the request itself appear more difficult, offsetting any other benefits from increased liking or curiosity.

One more surprising result came from Google Answers, an online community where paid researchers and unpaid community members work together to answer questions. Several research studies have shown that extrinsic motivations crowd out intrinsic motivations (Deci 1999). Thus, we would hypothesize that the presence of the paid researchers would dampen the motivations of the unpaid community to help with the work ("why should I help, if I'm not getting paid like them?"). However, we qualitatively observed that this was simply not the case. In one of our questions, an unpaid community member went so far as to challenge the veracity of a well-written and complete answer made by a member of the research staff. The researcher, despite already having earned his or her money, then spent additional time correcting the answer, with ongoing discussion from several other members from the community. This interaction points to the potential for building strong communities that are based on a mixed paid/volunteer system.

6.3 Methodological Context

This thesis relies on a variety of research methods. Most of these methods build on prior work – there's a long history of research in the area of online communities! In this section, we describe several of the research methods most related to our work, to better develop the larger context in which this work was conducted.

Many general-purpose, data-driven techniques for describing online communities have been used in prior work. Several researchers (e.g., Whittaker 1998; Schoberth 2003) have developed and used metrics for analyzing and comparing discussion forums, email lists, or other conversation-oriented online communities. These metrics allow researchers to understand aspects of a community such as its size (e.g., number of members, number of posts) and activity level (e.g., number of active members/day, number of posts/day).

Other researchers (e.g., Mislove 2007) have developed methods for overlaying social networks onto online communities. From these networks, it is possible to compute a variety of social networking metrics – metrics such as the "indegree" and "outdegree" of users. Social networking techniques have the advantage that they can be applied to any online community where users interact, and that the metrics in use in these types of analysis are relatively context-independent.

However, specialized methods for summarizing particular types online communities have emerged in response to the need for understanding unique forms of behavior that are difficult or impossible to measure with general-purpose methods. For example, researchers studying community use of tagging features have developed several metrics, including ones that describe the number of unique tags used (Golder 2006) and the distribution of tags across specialized "tag classes" (Sen 2006). Researchers interested in Q&A sites have looked at specialized metrics such as the number of questions (Agichtein 2008) and the percentage of answers rated as "best answer" (Adamic 2008).

Often, pure quantitative representations of online communities cannot adequately represent the type or nature of user interactions that take place. One way of characterizing the type of interactions is through the use of human coding, where people – often subject-matter experts – assign labels to posts, users, or other entities. Historically, coded data has served as both a research outcome or as a means to further quantitative analysis. For example, Rafaeli et al. (1994; 1997) used coding to better understand how confrontational users are online, while Burke et al. (2007) coded messages for the presence of introductory messages to investigate downstream effects of that action.

One class of data-driven summarization that has helped to shape our understanding of social interactions in online communities is data visualizations. These tools are used to help researchers more quickly identify patterns of communication. Viegas et al. (2004) developed "history flow visualizations" to demonstrate how wiki pages are edited over time, and Turner et al. (2005) adopted "Tree Map" visualizations to locate differences between Usenet communities.

Equally important to the field of online community research is the practice of qualitative methods such as interviews, document (e.g., forum post) analysis, or observations to study

online communities (Mann 2000). Qualitative methods can provide deep insights into the behaviors and motivations of people, and generate a richness of description that is difficult to attain through quantitative methods (Hine 2000; Dourish 2006). For instance, Maloney-Krichmar and Preece (2005) conduct a multi-year study of an online health-related community, using observations and interviews to derive insights into such things as the community's role structure and how membership in the community affected users' real-world existence.

Beyond techniques for observing natural interaction in online systems, there is a history of work based on controlled interventions. One common form of intervention is between-subjects – commonly known as split testing, or A/B testing – where groups of users are treated differently in some controlled fashion, and effects are measured. For instance, to learn about the utility of goal-setting in an online context, Beenen et al. (2004) used email to deliver control messages to one group of subjects, and messages with specific performance goals to another group. Another form of intervention is within-subjects, where a change is introduced to an online community, and differences in user behavior are observed. For instance, Cheng and Vassileva (2005) observe the differences in user activity before and after the introduction of an incentive mechanism in a file sharing system.

6.4 Methodological Contributions

Many of the methodologies just described had a direct influence on the techniques applied in this thesis. For the most part, we employed quantitative methodologies, focusing on controlled experimentation and data analysis using standard summarization techniques. In some cases, we relied on user surveys, coding, and qualitative observations. We did develop some new novel techniques, which in this section we discuss as contributions.

In Chapter 4, we extended a methodology developed by Janes et al. (2001) for developing and injecting questions into a variety of Q&A sites. The original methodology was not entirely suitable for our study, as we were interested in public questions and answers that are continually indexed by search engines. Because we wished to avoid "cross-contamination", where a Q&A exchange at one site influenced the outcome on another site through search results, we did not ask the same question in more that one site. Instead, we developed a set of templates from

which we could write sets of questions that were approximately equal from the perspective of the experimenter (in terms of length, difficulty, and the experimental criteria, for example). These questions could then be randomly distributed among the experimental sites. We ensured the consistency and quality of our questions by evaluating them with a panel of six coders.

In Chapter 5, we leveraged the technique of ensemble machine learning methods to meet our dual goals of learning about the structural properties of questions and building an accurate classifier. Though machine learning methods have been used in the study of online communities, they are not commonly used as a means to gaining an understanding of the thing being modeled. Rather, they are typically used as a means for determining which features are predictive of outcomes (e.g., Arnt 2003), or as a means of automating the process of hand-classification (e.g., Terveen 1997; Burke 2007). We demonstrate the utility of studying the features themselves: in the process of building our three specialized question classifiers, we learned about the categorical, textual, and social-network properties that distinguish conversational and informational questions. This methodology could be extended to a wide variety of domains in the area of online communities.

6.5 Computational Context

In this section, we briefly describe the computational techniques that are most central to system designers hoping to shape user participation: user modeling, personalization, and machine learning algorithms.

This thesis repeatedly draws on the idea that online content may be personalized and recommended to users. Computational techniques for personalizing a user's experience in a Web site are highly varied. One influential conceptual model for organizing these techniques identifies two primary system tasks: first, building a "user model" of the user's preferences and history of actions; second, interpreting this model to personalize the content that appears on the screen (Brusilovsky 1996). Web personalization has many applications – including the personalization of advertisements and product offerings – that are summarized in Kobsa's (2001) article on the topic.

Perhaps the most active part of this research area concerns the recommendation of content (Resnick 1997). Though recommendation algorithms such as collaborative filtering were originally developed in response to the problem of "information overload" (e.g., Maes 1994, Resnick 1994), more recently, these algorithms have been adapted to a wider variety of situations. For instance, Geyer et al. (2008) experiment with a variety of content- and social network-based recommendation algorithms that are designed to increase user contributions to "about you" forms on user profile pages. Chen et al. (2009) experiment with related algorithms for suggesting potential "friends" on social networking sites. Cosley et al. (2006) investigate methods for "intelligent task routing" – algorithms for matching users with potential contributions in online communities.

Supervised machine learning (Mitchell 1997) is another technique that can be used to personalize the user experience. In particular, however, its ability to automatically categorize user-generated data has begun to attract interest in the research community. For instance, some systems do not have enough active users to provide fair moderation, so Arnt and Zilberstein (2003) developed machine learning techniques for automatically inferring the quality of posts. Similarly, Adamic et al. (2008) use machine learning techniques to predict "best answers" in an online Q&A site.

6.6 Computational Contributions

This thesis contributes several computational techniques that address real problems for systems that rely on user-generated content.

In Chapter 3, we presented a computational framework for the display of personalized messages that encourages users to visit or post in a discussion forum. Our notion was that an invitation for user participation might be made more compelling by referring to another system entity. For instance, an invitation to read a discussion post might include the title of a movie that is being discussed. To pick the entity for inclusion in the recommendation – post text, movies, and user names – we designed and evaluated several algorithms that chose entities for recommendation based on the user's history of activity in the site. Our most successful algorithm recommended

posts based on the degree of disagreement between the viewer of the invitation and the author of a discussion post.

In Chapter 5, we developed an ensemble machine learning framework for classifying questions as either conversational or informational. This framework is based on a collection of three feature sets that are common in online communities: the category in which a contribution is made, the text of a contribution, and the social network of the user making the contribution. We discussed methods for collecting these features, and described our implementation of an ensemble learner. Though ensemble methods have been heavily studied in the machine learning community (e.g., Polikar 2006), we contribute the idea of using meta-classification as a means to learn about the entity being modeled while also producing high-accuracy classifications.

6.7 Empirical Context

The fourth perspective we will take in contextualizing this thesis is empirical – the body of results specific to the study of user contributions and social design in online communities. Though we have done our best to cite related work and describe our contributions in each individual chapter, this section gives us the chance to understand the context and contributions from a broader, integrative perspective.

While online communities have been studied since their emergence in the 1980s (e.g., Rheingold 2000), this thesis is mostly concerned with modern interpretations of online community (i.e., after 1993, when the Mosaic Web browser was released). While people haven't changed much in this time, technologies and expectations have, leading to new sets of problems and styles of interactions. For a discussion of the research literature on early online communities, see Jones (1997).

One early technology – Usenet – closely resembles much discourse online today. Kollock and Smith (1996) present one of the early studies on how cooperation can emerge in an online community that, theoretically, should suffer from the free rider problem. They frame contributions to online communities as *public goods* (Olson 1971), which are free to consume but costly to produce. Importantly, they identify several components of cooperation in online

communities that stand in contrast with cooperation in real communities – e.g., the costs of bad behavior are amplified as the scale of participation grows, and the costs of coordination are low.

Several other studies of Usenet are important to this thesis. Whittaker et al. (1998) conducted one of the first large-scale studies of online interaction, finding evidence of what has come to be known as the *power law of participation* (e.g., Mayfield 2006). They analyze a large data set of Usenet posts to model interaction based on a variety of (intrinsic) user characteristics such as demographics and (extrinsic) design factors such as the presence of news group moderation. They present several results concerning the influence that users exert on one another, such as how cross-posting and short messages have a positive effect on the number of replies received. More recently, Turner et al. (2005) performed an extremely large-scale investigation of Usenet, presenting visualizations of factors such as the number of new threads and the number of replies per message to classify different sub-communities.

In one recent study, Mislove et al. (2007) scraped data from four major online communities (Flickr, LiveJournal, Orkut, and YouTube) to better understand the nature of participation, finding evidence across these communities that (1) there exists a power-law distribution of contributions per user, (2) connections between members resembles a "small-world" graph structure (Watts 1998), and (3) at the center of the social network is a densely connected core of high-degree nodes. Other studies have looked at social networking sites to understand the bridge between online and offline social networks. For example, Ellison et al. (2006) examined the college students' use of social networking site Facebook, finding that use correlates with the presence of social capital, benefiting users who are well-connected.

Understanding user motivations has been a common theme of recent work on online communities. For instance, researchers have investigated users' motivations for becoming power users in Wikipedia (Bryant 2005), contributing book reviews to Amazon (Peddibhotla 2007), writing open source code (Lakhani 2003), and tagging photos on Flickr (Ames 2007). As it turns out, users act in response to lots of different motivators, such as altruism, fun, and empathy. Moore and Serva (2007) observe that different designs of online communities create different sets of motivators.

Related to the notion of user motivations is the idea that users tend to specialize, or take on particular *roles* within a community. User roles may be explicitly defined through the use of

policy in some communities such as Wikipedia; in other cases, roles simply emerge from behavioral patterns (Gleave 2009). The study of roles is useful to designers who seek to design for higher quality or better contributions, because roles provide rough categorizations that may be used to understand *who* designs are targeting, and what goals these users hope to accomplish. Research to date has examined the role structure in a variety of systems, including online discussion forums (Fisher 2006; Welser 2007), wikis (Guzdial 2000), and tagging systems (Thom-Santelli 2008). Gleave et al. (2009) proposed research techniques for the identification of roles in online communities.

In this thesis, we have focused on the quality and quantity of user contributions. The notion of "quality" can be measured at different levels. For instance, Priedhorsky et al. (2007) developed a metric for understanding the "value" contributed by a single edit to a Wikipedia article, in order to better understand where the quality in the online encyclopedia comes from. Giles (2005) ran a direct comparison between Wikipedia and Britannica, and used the rate of errors in scientific articles as a metric through which we can infer quality. One increasingly common mechanism for inferring the quality of online contributions is to ask users their opinions; researchers have subsequently begun to explore this design space. For instance, Lampe and Resnick (2004) critically assess the strengths and weaknesses inherent in one online discussion forum's distributed moderation system.

Other work has gone beyond the detection or prediction of quality to the development of design principles that can encourage quantity or quality. For instance, Ludford et al. (2004) constructed a MovieLens discussion forum where they experimented with ways of injecting messages of uniqueness and similarity to promote higher levels of participation. Cosley et al. (2005) researched policy-level design, by testing interfaces for editing the MovieLens database that used different levels of peer- and expert-based oversight. Other researchers studied interfaces for encouraging movie ratings by providing users with specific goals (Beenen 2004) or by showing the "value" of each rating to the system (Rashid 2006). Other studies have looked at the effect of awarding non-monetary incentives such as points, stars, or badges (Cheng 2005; Cheng 2006; Hummel 2005). These studies have demonstrated that incentives are in fact a powerful way of motivating additional contributions from users, but that some users will actively provide lower quality contributions to game the system.

6.8 Empirical Contributions

Clearly, this thesis exists in the context of a deep and diverse set of empirical findings. How does our research fit in? Generally, we have favored exploration over validation, leading us to explore new designs and ask novel research questions. In this section, we summarize the highest-level domain-specific findings in this thesis.

In Chapter 2, we presented survey data that further informs our understanding of *why* users contribute ratings to online recommendation systems: because it's fun. Users told us that their second most important reason why they rated movies was for the fun of it, and their second most important reason to use the system was so that they *could* rate movies!

Another significant finding from this study is that email newsletters with *upwards comparisons* are motivating to members. These newsletters, which told members that they had rated fewer movies than other members like them, caused members to rate more movies in the following week, and also caused them to self-report that they *wanted* to do something to change their below-average standing. On the other hand, messages telling members that they are average or above average had little measurable effect. This result informs our understanding of common design patterns in practice today, such as leaderboards, which tell members how they compare with one another.

In Chapter 3, we tested a variety of personalization algorithms to better understand the most effective ways of encouraging members to visit or post in a discussion forum. Overall, we found that the algorithms designed to emphasize the social aspects of the discussion forum were the most successful. The algorithm that produced the strongest effect was called *disagree* – it showed members a recent post where they happened to disagree with the author about their rating of a movie mentioned in the post.

In Chapter 4, we investigated high-level community design as a predictive factor in the quality of responses to users' requests. We found evidence that unstructured workflows in open (i.e., Web 2.0) communities are able to perform as well as – or better than – more traditional workflows that support differentiated roles and system privileges. In addition, we found qualitatively that the open community features in one site, Google Answers, improved upon the answers of the paid research staff. In general, we found that sites that depend on single "expert"

respondents are more prone to leaving questions unanswered, are less timely in their responses, and provide a lower diversity of answers.

On the other hand, we found qualitative evidence that open community sites like Yahoo Answers tend to provide highly variable answer quality. Some of our experimental requests were responded to with a seemingly random mixture of good and bad responses. As the research community moves towards a better understanding of how social contributions in online communities should be archived for future information retrieval, it is essential to better understand how to deal with this inherent variability in quality.

In Chapter 5, we developed a new classification for questions asked online that distinguishes between conversational and informational content. We showed that this classification is intuitive: two random human judges were able to agree on the type of a random question drawn from any of three social Q&A sites over 87% of the time. We also showed that this classification is useful: our judges rated informational questions as much more likely to lead to the creation of archival quality information. It is quite possible that this classification can be used to classify other types of contributions to online communities, such as forum posts or blog post comments.

We also discovered several differences between conversational and informational questions that are both simple and strongly predictive. First, the language of informational questions tended to be inwardly focused, using words like “I”, while the language of conversational questions tended to be focused on the potential audience, using words like “you”. Also, users who posted conversational requests tended to have a much more densely interconnected social network than users who posted informational questions. It is possible that these differences are common across a variety of social media, and can be leveraged by researchers and practitioners to better understand the sociality of user generated content.

6.9 Future Work

Clearly, there are many directions for future work – and we have already discussed several of them in Chapters 2-5. In this section, we broaden our thinking, and discuss three bigger areas that this thesis points to as both realizable and important.

Structural Incentives for Quality

We first point to the obvious need for a better understanding of incentive mechanisms for promoting quality. We found in Chapter 4 that the paid Google Answers community provided higher quality answers, on average, than the communities from several free sites. However, while paid incentives have been heavily studied, structural incentives – e.g., points, badges, or other non-monetary awards – are not well understood. Several studies have shown that structural incentives can increase the quantity of contributions. For instance, we found in Chapter 2 that social comparisons can be used to boost the contributions of under-performers, and Cheng and Vassileva (2005) found that awarding reputation markers increased sharing behavior. However, there has not been work investigating how to use these mechanisms to motivate users to make higher *quality* contributions.

There are several ways to approach this area of future work. First, it would be useful to run controlled studies of existing mechanisms, measuring the impact of commonly used features such as reputation markers on outcomes such as the quantity and quality of contributions. Though many sites actively use these features (often, in fairly standard ways), it is unclear whether they have a positive effect on site outcomes, especially on the average quality of a user contribution. Second, it is important to understand the range of behaviors that *can* be rewarded, and how appropriate incentive-based interventions can lead users to consider the quality of their contributions when they act. Finally, it will be useful to explore new designs. For instance, we might develop adaptive mechanisms (i.e., "market controls") that can understand how much different types of contributions are needed, and self-adjust to most strongly reward the contributions that are most valuable to the community.

Mixed Paid and Free Communities

In certain cases, online communities appear capable of defying the predictions of established theories from the social sciences. For instance, online communities often don't suffer from the

free rider problem, but instead encourage widespread cooperation (Kollock 1996). In Chapter 4, we identified another example: the unpaid community in Google Answers actively supported the paid research staff. However, the theoretical literature, backed by a variety of empirical results, shows that extrinsic motivations crowd out intrinsic motivations (Deci 1999). How was Google successful in getting the unpaid community members to contribute in this context?

Based on our observations, we think that it is the case that there are designs that successfully encourage paid and unpaid community members to successfully work side by side. However, it is completely unclear what makes these designs work, and whether these designs only work for certain types of communities, for certain types of users, or for certain types of interactions. Future work should investigate these questions by more deeply investigating existing communities where paid and unpaid labor coexist, and by conducting lab or field experiments to better understand ideal conditions where the users' intrinsic motivations are not dampened by the presence of money.

Mining Archival Information from Social Processes

Finally, we believe there is potential to meaningfully extend our work in Chapter 5, where we began to understand how to mine archival quality content from naturally occurring social processes. There is an opportunity to take Ackerman and Malone's (1990) ideas about Answer Garden – a tool for the aggregation of knowledge through expert-based Q&A – and to use them in the context of social Q&A sites, selecting interactions for inclusion in the garden based on their informational properties.

We have made just a little progress towards meeting this goal, by dividing the world of social contributions into two categories: conversational and informational. It is certain that this simple division should be expanded to more nuanced and generalizable categorizations that better understand users' motivations and abilities to contribute. Also, because we have only looked at user intent, there is substantial room to better understand and measure the actual informational content of the interactions themselves. Has this question been answered? Is this post correct?

Current research about Wikipedia tells us part of the story – the part about a community that has come together to achieve an informational resource of unprecedented breadth and high quality. The part that has not been told is of the untold terabytes of informational data that have been generated through discussion forums, email lists, and other online communities. Some of the

discussions in these communities have led to the creation of information; others never received any response. Yet from the perspective of the search engine, they all appear the same. Future researchers should work towards an understanding of which contributions contain useful information, as well as mechanisms that help users find useful content and improve the precision of search results.

6.10 Conclusions

Online communities are deeply complex, the result of large numbers of people interacting through intricately designed algorithms and interfaces. This complexity makes the creation of "theory" that explains or predicts user behavior difficult. Indeed, Jenny Preece observes that "No particular theory or set of theories currently dominates research on online communities." (Preece 2005) Even if we are to put aside theory in hopes of discovering "best practices", broadly applicable results are difficult to come by. As Joel Spolsky writes, "small software implementation details result in big differences in the way the community develops, behaves, and feels." (Spolsky 2003)

Despite this complexity, online communities offer unique opportunities for researchers. What other environment allows access to such detailed historical information about people's activities? Log files of user behavior and databases of contributions allow researchers to look back in time with a scope previously unavailable in physical environments. Also, online communities are flexible environments that may be modified in precise ways. By modifying software, researchers can conduct controlled experiments involving many subjects over a short period of time.

In this thesis, we have presented a set of studies that investigate the interaction between social design and user contributions to online communities. We asked questions from the level of the individual user contribution to the level of community-wide workflow, to understand how social design affects communities, users, and the quality and quantity of user-generated content. We hope that some of our findings will inform or inspire system designers and researchers, to enable them to create richer experiences for users and better outcomes for their community.

References

Ackerman, M., Malone, T. Answer Garden: A Tool for Growing Organizational Memory. In *Proc. Office Information Systems* (1990).

Adamic, L., Zhang, J., Bakshy, E., Ackerman, M. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In *Proc. WWW* (2008).

Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G. Finding High-Quality Content in Social Media. In *Proc. WSDM* (2008).

Algesheimer, R., Dholakia, P. Do Customer Communities Pay Off? *Harvard Business Review* (2006).

Ames, M., Naaman, M. Why We Tag: Motivations for Annotation in Mobile and Online Media. In *Proc. CHI* (2007).

Arnt, A., Zilberstein, S. Learning to Perform Moderation in Online Forums. In *Proc. Web Intelligence* (2003).

Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., Kraut, R. Using Social Psychology to Motivate Contributions to Online Communities. In *Proc. CSCW* (2004).

Bernheim, D. A Theory of Conformity. *The Journal of Political Economy*, 102 (1994).

Box, G., Hunter, W, Hunter, S., Hunter, W. *Statistics for Experimenters*. John Wiley & Sons, New York (1978).

Brusilovsky, P. Methods and Techniques of Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 6 (1996).

Bryant, S., Forte, A., Bruckman, A. Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. In *Proc. GROUP* (2005).

Bryk, A. S., and Raudenbush. *Hierarchical Linear Models: Applications and Data Analysis Methods* (1992).

Burke, M., Joyce, E., Kim, T., Anand, V., Kraut, R. Introductions and Requests: Rhetorical Strategies that Elicit Response in Online Communities. In *Proc. Communities and Technologies* (2007).

Butler, B. When is a group not a group: An empirical examination of metaphors for online social structure. *Social and Decision Sciences* (1999).

Buunk, B., Collins, R., Taylor, S., VanYperen, N., Dakof, G. The Affective Consequences Of Social Comparison: Either Direction Has Its Ups And Downs. *Journal of Personality and Social Psychology*, 59(6), 1990.

Chae, M., Lee, B. Transforming an Online Portal Site into a Playground for Netizen. *Journal of Internet Commerce*, 4 (2005).

Chaiken, S., Liberman, A., and Eagly, A. H. Heuristic and Systematic Information Processing Within and Beyond the Persuasion Context. In Uleman, J. S. and Bargh, J. A. (Eds.), *Unintended Thought*. Guilford Press, (1989).

Chen, J., Geyer, W., Dugan, C., Muller, M., Guy, I. Make New Friends, But Keep the Old: Recommending People on Social Networking Sites. In *Proc. CHI* (2009).

Chen, Y., Harper, F., Konstan, J., Li, X. Social Comparisons and Contributions to Online Communities: A Field Experiement on MovieLens, Unpublished Manuscript (2009).

Cheng, R., Vassileva, J. User Motivation and Persuasion Strategy for Peer-to-Peer Communities. In *Proc. HICSS* (2005).

Cheng, R., Vassileva, J. Design and Evaluation of an Adaptive Incentive Mechanism for Sustained Educational Online Communities. *User Modeling and User-Adapted Interaction*, 16 (2006).

Cialdini, R. *Influence: The Psychology of Persuasion*, Collins (1998).

Cialdini, R., Goldstein, N. Social Influence: Compliance and Conformity. *Annual Review of Psychology*, 55 (2004).

Clark, P., Wilson, J. Incentive Systems: A Theory of Organizations. *Administrative Science Quarterly*, 6 (1961).

Cosley, D., Frankowski, D., Kiesler, S., Terveen, L., Riedl, J. How Oversight Improves Member-Maintained Communities. In *Proc. CHI* (2005).

Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *Proc. CHI* (2006).

Cremer, D., Tyler, T. Am I Respected or Not?: Inclusion and Reputation as Issues in Group Membership. *Social Justice Research*, 18 (2005).

Croson, R., Shang, J. Field Experiments in Charitable Contribution: The Impact of Social Influence on the Voluntary Provision of Public Goods. *Knowledge@Wharton* (2005).

Deci, E., Koestner, R., Ryan, R. A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin*, 125 (1999).

Dourish, P. Implications for Design. In *Proc. CHI* (2006).

Drenner, S., Harper, M., Frankowski, D., Riedl, J., and Terveen, L.. Insert Movie Reference Here: A System to Bridge Conversation and Item-Oriented Web Sites. CHI Notes, in *Proc. CHI* (2006).

Edelman, B. Earnings and Ratings at Google Answers. Unpublished Manuscript (2004).

Ellison, N., Lampe, C., Steinfield, C. Spatially Bounded Online Social Networks and Social Capital: The Role of Facebook. *International Communications Association* (2006).

Fawcett, T. ROC Graphs: Notes and Practical Considerations for Researchers. *HP Labs Tech Report HPL-2003-4* (2004).

Fehr, E., Falk, A. Psychological Foundations of Incentives. *European Economic Review*, 46 (2002).

Festinger, L. A Theory of Social Comparison. *Human Relations*, 7 (1954).

Fisher, D., Smith, M., Welser, H. You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. In *Proc. HICSS* (2006).

Frey, B., Meier, S. Social Comparisons and Pro-social Behavior - Testing 'Conditional Cooperation' in a Field Experiment. *American Economic Review*, 94(5) (2004).

Garner, R. Post-It Note Persuasion: A Sticky Influence. *Journal of Consumer Psychology*, 15 (2005).

Gazan, R. Specialists and Synthesists in a Question Answering Community. In *Proc. American Society for Information Science and Technology*, 43 (2006).

Geyer, W., Dugan, C., Millen, D., Muller, M., Freyne, J. Recommending Topics for Self-Descriptions in Online User Profiles. In *Proc. RecSys* (2008).

Giles, J. Internet Encyclopaedias Go Head to Head. *Nature* 438 (2005).

Gleave, E., Welser, H., Lento, T., Smith, M. A Conceptual and Operational Definition of 'Social Role' in Online Community. In *Proc. HICSS* (2009).

Golder, S., Huberman, B. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32 (2006).

Greenspan, R. Surfers Prefer Personalization. http://www.webcitation.org/5IsMR19Gt. 2004.

Grossman, L. Time's Person of the Year: You. *Time Magazine* (2006). http://www.webcitation.org/5Lh4GdH4l

Guzdial, M., Rick, J., Kerimbaev, B. Recognizing and Supporting Roles in CSCW. In *Proc. CSCW* (2000).

Gyöngyi, Z., Koutrika, G., Pedersen, J., Garcia-Molina, H. Questioning Yahoo! Answers. In *First Workshop on Question Answering on the Web* (2008).

Hanneman, R., Riddle, C. *Introduction to Social Network Methods*. Riverside, CA: University of California, Riverside (2005).

Harper, F., Li, X., Chen, Y., Konstan, J. An Economic Model of User Rating in an Online Recommender System, In *Proc. User Modeling* (2005).

Harper, F., Li, X., Chen, Y. & Konstan, J. Social Comparisons to Motivate Contributions to an Online Community. In Persuasive Technology, 148-159 (2007a).

Harper, F., Frankowski, D., Drenner, S., Ren, Y., Kiesler, S., Terveen, L., Kraut, R., Riedl, J. Talk Amongst Yourselves: Inviting Users to Participate in Online Conversations. In *Proc. IUI* (2007b).

Harper, F., Raban, D., Rafaeli, S., Konstan, J. Predictors of Answer Quality in Online Q&A Sites, In *Proc. CHI* (2008).

Harper, F., Moy, D., Konstan, J. Facts or Friends? Distinguishing Informational and Conversational Questions in Social Q&A Sites. In *Proc. CHI* (2009).

Hill, W., Stead, L., Rosenstein, M., and Furnas, G. Recommending and Evaluating Choices in a Virtual Community of Use. In *Proc. CHI* (1995).

Hine, C. *Virtual Ethnography*, SAGE Publications (2000).

Hitwise. U.S. Visits to Question and Answer Websites Increased 118 Percent Year-over-Year (2008). http://www.webcitation.org/5alK5xpWh

Hsieh, G., Counts, S. mimir: A Market-Based Real-Time Question and Answer Service. In *Proc. CHI* (2009).

Hummel, H., Burgos, D., Tattersall, C., Brouns, F., Kurvers, H., Koper, R. Encouraging Contributions in Learning Networks Using Incentive Mechanisms. *Journal of Computer Assisted Learning*, 21 (2005).

Janes, J., Hill, C., and Rolfe, A. Ask-an-Expert Services Analysis. *Journal of the American Society for Information Science Technology,* 52(13) (2001).

Jones, Q (1997). Virtual-Communities, Virtual Settlements & Cyber-Archaeology: A Theoretical Outline. *Journal of Computer Mediated Communication*, 3 (1997).

Jurczyk, P., Agichtein, E. HITS on Question Answer Portals: Exploration of Link Analysis for Author Ranking. In *Proc. SIGIR* (2007).

Kallgren, C., Reno, R., Cialdini, R. A Focus Theory of Normative Conduct: When Norms Do and Do Not Affect Behavior. *Personality and Social Psychology Bulletin*, 28 (2000).

Kluger, A., Denisi, A., The Effects Of Feedback Interventions On Performance: A Historical Review, A Meta-Analysis, And A Preliminary Feedback Intervention Theory. *Psychological Bulletin*, 119(2) (1996).

Kobsa, A. Generic User Modeling Systems. *User Modeling and User-Adapted Interaction*, 11 (2001).

Kollock, P., Smith, M. Managing the Virtual Commons: Cooperation and Conflict in Computer Communities. In *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives* (1996).

Kramer, C. Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications. *Biometrics*, 12 (1956).

Kullback, S., Leibler, R. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22 (1951).

Kuncheva, L., Whitaker, C. Measures of Diversity in Classifier Ensembles. *Machine Learning*, 51(2) (2000).

Lakhani, K., Wolf, R. Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects. *MIT Sloan Working Paper* No. 4425-03 (2003).

Lampe, C., Resnick, P. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proc. CHI* (2004).

Lampe, C., Johnston, E. Follow the (Slash) Dot: Effects of Feedback on New Members in an Online Community. In *Proc. GROUP* (2005).

Langer, E., Blank, A., Chanowitz, B. The Mindlessness of Ostensibly Thoughtful Action. *Journal of Personality and Social Psychology*, 36 (1978).

Ledyard, J. Public Goods: A Survey of Experimental Research. In *The Handbook of Experimental Economics*, Princeton University Press, 1994.

Lee, J. H., Downie, J. S., Cunningham, S. J. Challenges in Cross-Cultural/Multilingual Music Information Seeking. In *Proc ISMIR* (2005).

Leibenluft, J. A Librarian's Worst Nightmare: Yahoo! Answers, where 120 million users can be wrong. *Slate Magazine*, Dec. 7, 2007 (2007).

Ling, K., et al. Using social psychology to motivate contributions to online communities. *Journal of Computer Mediated Communication* 10(4) (2005).

Liu, Y., Bian, J., Agichtein, E. Predicting Information Seeker Satisfaction in Community Question Answering. In *Proc SIGIR* (2008).

Lockwood, P., Kunda, Z. Superstars And Me: Predicting The Impact Of Role Models On The Self. *Journal of Personality and Social Psychology*, 73(1) (1997).

Loewenstein, G. The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116 (1994).

Ludford, P., Cosley, D., Frankowski, D., Terveen, L. Think Different: Increasing Online Community Participation Using Uniqueness and Group Dissimilarity. In *Proc. CHI* (2004).

Maes, P. Agents That Reduce Work and Information Overload. *Communications of the ACM*, 37 (1994).

Maloney-Krichmar, D., Preece, J. A Multilevel Analysis of Sociability, Usability, and Community Dynamics in an Online Health Community. *ACM Transactions on Computer-Human Interaction*, 12 (2005).

Mann, C., Stewart, F. *Internet Communication and Qualitative Research : A Handbook for Researching Online* (New Technologies for Social Research series), SAGE Publications (2000).

Mayfield, R. Power Law of Participation. Blog entry in Ross Mayfield's Weblog (2006). http://ross.typepad.com/blog/2006/04/power_law_of_pa.html

McClennen, M., Memmott, P. Roles in Digital Reference. *Information Technology and Libraries* 20(3) (2001).

McDonald, D. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In *Corpus Processing For Lexical Acquisition*, 1996.

Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, B. Measurement and Analysis of Online Social Networks. In *Proc. IMC* (2007).

Mitchell, T. *Machine Learning* (1997).

Moore, T., Serva, M. Understanding Member Motivation for Contributing to Different Types of Virtual Communities: A Proposed Framework. In *Proc. Computer Personnel* (2007).

Nam, K., Ackerman, M., Adamic, L. Questions in, Knowledge iN?: A Study of Naver's Question Answering Community. In *Proc. CHI* (2009).

Nonnecke, B. and Preece, J. Lurker Demographics: Counting the Silent. In *Proc. CHI* (2000).

Oliver, P. Rewards and Punishments as Selective Incentives for Collective Action: Theoretical Investigations. *The American Journal of Sociology*, 85 (1980).

Olson, M. *The Logic of Collective Action: Public Goods and the Theory of Groups*, Harvard University Press (1971).

Peddibhotla, N., Subramani, M. Contributing to Public Document Repositories: A Critical Mass Theory Perspective. *Organization Studies*, 28 (2007).

Polikar, R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6 (2006).

Pomerantz, J., Nicholson, S., Belanger, Y., Lankes, R. D. The Current State of Digital Reference. *Information Processing and Management*, 40(2) (2004).

Preece, J., Nonnecke, B., and Andrews, D. The Top Five Reasons for Lurking: Improving Community Experiences for Everyone. *Computers in Human Behavior*, 20 (2004).

Preece, J., Maloney-Krichmar, D. Online Communities: Design, Theory and Practice. *Journal of Computer-Mediated Communication*, 10 (2005).

Priedhorsky, R., Chen, J., Lam, S., Panciera, K., Terveen, L., Riedl, J. Creating, Destroying, and Restoring Value in Wikipedia. In *Proc. GROUP* (2007).

Raban, D., Harper, F. Motivations for Answering Questions Online. In *New Media and Innovative Technologies* (2008).

Rabin, M. Incorporating Fairness into Game Theory and Economics. *The American Economic Review*, 83(5) (1993).

Rafaeli, S., Sudweeks, F., Konstan, J., Mabry, E. ProjectH Technical Report (1994). http://www.it.murdoch.edu.au/~sudweeks/papers/techrep.html

Rafaeli, S., Sudweeks, F. Networked Interactivity. *Journal of Computer-Mediated Communication*, 2 (1997).

Rafaeli, S., Raban, D., Ravid, G. Social and Economic Incentives in Google Answers. ACM Group 2005 Workshop: Sustaining Community: The role and design of incentive mechanisms in online systems (2005).

Rafaeli, S., Raban, D., Ravid, G. How Social Motivation Enhances Economic Activity and Incentives in the Google Answers Knowledge Sharing Market. *International Journal of Knowledge and Learning*, 3(1) (2007).

Rashid, A., Ling, K., Tassone, R., Resnick, P., Kraut, R., Riedl, J. Motivating Participation by Displaying the Value of Contribution. In *Proc. CHI* (2006).

Ren, Y., Kraut, R., Kiesler, S. Applying Common Identity and Bond Theory to Design of Online Communities. *Organization Studies*, 28 (2007).

Resnick, P., Iacovou, N., Sushak, M., Bergstrom, P., Riedl, J. Grouplens: An Open Architecture For Collaborative Filtering Of Netnews. In *Proc. CSCW* (1994).

Resnick, P., Varian, H. Recommender Systems. *Communications of the ACM*, 40 (1997).

Resnick, P., Kuwabara, K., Zeckhauser, R., Friedman, E. Reputation Systems. *Communications of the ACM*, 43 (2000).

Rheingold, H. *The Virtual Community: Homesteading on the Electronic Frontier*, revised edition, The MIT Press (2000).

Ridings, C. M., and Gefen, D. Virtual community attraction: Why people hang out online. *Journal of Computer Mediated Communication*, 10(1) (2004).

Rousch, W. What's the Best Q&A Site? *Technology Review*, December 3, 2006 (2006).

Sahami, M., Dumais, S., Heckerman, D., Horvitz, E. A Bayesian Approach to Filtering Junk E-mail. In *Learning for Text Categorization* (1998).

Sang-hun, C. South Koreans Connect Through Search Engine. *New York Times*, July 5, 2007 (2007).

Schafer, J., Konstan, J., Riedl, J. ECommerce Recommendation Applications. *Data Mining and Knowledge Discovery* (2001).

Schoberth, T., Preece, J., Heinzl, A. Online Communities: A Longitudinal Analysis of Communication Activities. In *Proc. HICSS* (2003).

Schultz, P. Changing Behavior With Normative Feedback Interventions: A Field Experiment on Curbside Recycling. *Basic and Applied Social Psychology*, 21(1) (1999).

Sen, S., Lam, S., Rashid, A., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F., Riedl, J. tagging, communities, vocabulary, evolution. In *Proc. CSCW* (2006).

Sen, S., Harper, F., Lapitz, A., Riedl, J. The Quest for Quality Tags. In *Proc. GROUP* (2007).

Shah, C., Oh, J., Oh, S. Exploring Characteristics and Effects of User Participation in Online Social Q&A Sites. *First Monday*, 13 (2008).

Simon, H. Designing Organizations for an Information Rich World. In Greenberger, M. (Ed.) *Computers, Communications and the Public Interest*. The Johns Hopkins Press, Baltimore, MD, 1971.

Smith, A. *The Wealth of Nations* (1776).

Spolsky, J. Building Communities With Software (2003). http://www.webcitation.org/5fKF74QxU

Suls, J., Martin, R., Wheeler, L. Social Comparison: Why, With Whom, and With What Effect? *Current Directions in Psychological Science*, 11(5) (2002).

Terveen, L., Hill, W., Amento, B., Mcdonald, D., Creter, J. Phoaks: A System for Sharing Recommendations. *Communications of the ACM*, 40 (1997).

Thom-Santelli, J., Muller, M., Millen, D. Social Tagging Roles: Publishers, Evangelists, Leaders. In *Proc. CHI* (2008).

Turner, T., Smith, M., Fisher, D., Welser, H. Picturing Usenet: Mapping Computer-Mediated Collective Action. *Journal of Computer-Mediated Communication*, 10 (2005).

Tyler, T., Smith, H. Justice, Social Identity, and Group Processes. In *The Psychology of the Social Self*, Lawrence Erlbaum Associates, Inc. (1999).

Viegas, F., Wattenberg, M., Dave, K. Studying Cooperation and Conflict Between Authors with History Flow Visualizations. In *Proc. CHI* (2004).

Watts, D., Strogatz, S. Collective Dynamics of 'Small-World' Networks. *Nature*, 393 (1998).

Webster A.S., Vassileva J. Visualizing Personal Relations in Online Communities. Workshop on Social Navigation and Community Based Adaptation Technologies, Adaptive Hypermedia (2006).

Welser, H., Gleave, E., Fisher, D., Smith, M. Visualizing the Signatures of Social Roles in Online Discussion Groups. *The Journal of Social Structure*, 8 (2007).

Wheeler, L., Miyake, K. Social Comparison In Everyday Life. *Journal of Personality and Social Psychology*, 62(5) (1992).

White, M. Diffusion of an Innovation: Digital Reference Service in Carnegie Foundation Master's (Comprehensive) Academic Institution Libraries. *Journal of Academic Librarianship*, 27(3) (2001).

Whitmeyer, J. Effects of Positive Reputation Systems. *Social Science Research*, 29 (2000).

Whittaker, S., Terveen, L., Hill, W., Cherny, L. The Dynamics of Mass Interaction. In *Proc. CSCW* (1998).

Witten, I., Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques* (2005).

Yule, G. On the Association of Attributes in Statistics, In *Proc. Royal Society of London*, 66 (1900).

Zajonc, R. B. Attitudinal Effect of Mere Exposure. *Journal of Personality and Social Psychology*, 9, Monograph Supplement (1968).

Zipf, G.K. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA (1949).

Appendix A: Sample Experimental Questions

In chapter 4, we developed a set of 126 questions to ask at seven Q&A sites. These questions were derived from 18 templates. Below, we show one example question from each of these 18 templates. Note that while the length and tone of these questions varies substantially between templates, questions were as internally consistent as possible within a given template.

For the first template, we show all seven versions; for subsequent templates, we show one example.

Template 1

Topic	Type	Prior Effort	Gratitude
Technology	Factual	Yes	Short

I'm developing a web app in PHP, and I've found that when people click the "back" button in the browser, sometimes things break. Basically, when they click a link, I update the database, then they go back, and it looks like the change didn't happen. I found a bunch of information about how I can disable the back button in the browser, but I don't like this approach, as it seems that users should be able to click the back button and the application should just do the right thing. Is there any way to force the browser to refresh when the user clicks "back"? It's be great if it worked on both IE and Firefox. Thanks.

Recently I have seen a lot of sites which do a trick - they show and hide text in the web page without a page reload. I looked at the URL of the link, and it looks like javascript is being used somehow. I did a google search to figure out how to do this on my web site - I'd like to show and hide help text when people click on a link - but I didn't turn up any answers. Could you tell me how to do this trick? Appreciate it.

I run an Apache web server, and I recently added some perl CGI scripts to run a small, interactive part of my web site. My problem is that my web server is an old computer that gets easily overworked. When more than a few people access the CGI pages at the same time, the site starts to noticably slow down. I spent some time trying to make the perl scripts more efficient, but after that effort, I became convinced that the script itself can't be made much faster. I then turned to the Apache documentation to see if I could tweak it to run scripts faster, but no luck. Any ideas for how to make perl scripts run faster on my server? Thanks.

My web site has exceeded my server's capabilities. I have loaded the server up with 4 gigs of RAM, and it has a fast dual-core processor, but the site has started to feel slow. What I'd like to do is spread the load of the web site across several machines. The problem is that my other (currently unused) machines are much less capable in their performance, so I don't want to just direct an equal number of users to each machine. But I searched the web and the approaches that I found for configuring Apache seemed best suited for balanced load sharing. What I'd like is free software that manages the site by sending page requests to the server with the most excess capacity. Is there such software? Thank you.

I'm trying to code up a servlet for my company's intranet that keeps track of which people are online using the site. The problem is that when people forget to click "log out" and just close the browser (which they often do) the web site continues to think that they're logged in. Now, we're using a Tomcat web server, and I know it has a notion of user sessions - and these sessions expire after a certain amount of time (I think 30 minutes). I poked around the documentation, and didn't turn up any information on how to get code to run automatically when sessions expire. Is this even possible? If so, how do you do it? Thanks!

Not exactly sure why this is happening, but a couple of folks have started to use bots which crawl my site. No, it's not search engine bots- they show up in the server logs as user agents NetSpider, WebReaper, and a few others. I looked at ways of blocking crawlers, and I found a way to modify my server's .htaccess file to disallow these agents from visiting. But this seems very easy to work around...what I'd really like is a way to gradually throttle traffic. Thus, as a client requests unreasonable numbers of pages, we decide to limit them to one page per 5 seconds or something like that. Any ideas on ways to do that using an Apache server? I've got full access to the server, so I can install modules, etc, if that helps. Many thanks.

I am building a MySQL database-driven web application - my goal is scalability. I want to make sure that the app can take a slashdotting and live through it. Now, the site is pretty fast, but some database queries take time to complete. In researching ways to speed these up, I discovered that MySQL has a caching mechanism. I enabled it, and things appear faster. However, it seems like it's on auto-pilot -- I don't seem to have control over which queries are cached and which ones are not. Instead, it tries to cache all queries, including very infrequent ones that take up most of the cache. Is there a way to control the MySQL caching so that it only works for the queries I want it to work for? Thanks a bunch!

Template 2

Topic	Type	Prior Effort	Gratitude
Technology	Opinion	Yes	None

Ok, what's going to happen with file sharing over the next 4-5 years? I've been following the technology from the Napster boom and bust, through BitTorrent movie sharing, and have been trying to get a sense of whether the technology is going to morph into something with serious business or paying consumer applications. Does this technology have a future outside of sharing copyrighted material, and if so, what is it?

Template 3

Topic	Type	Prior Effort	Gratitude
Technology	Advice	Yes	Long

I'm heavily involved in a local church with about 1000 members, and I've been asked to explore getting us a good e-mail system to support a variety of mailing lists. Our goal is to have at least one membership-wide list (with opt-out), and many special interest lists (with opt-in). I've started by talking with other organizations, and have been advised to use an outside service (to avoid problems with being tagged as a spammer, with security, and with bounce-back messages); at the same time, I'd also like as clean a mechanism as possible for being able to do "one-time" changes of e-mail address or additions/deletions to multiple lists. Thanks so much; I appreciate any recommendations/advice you can offer.

Template 4

Topic	Type	Prior Effort	Gratitude
Business	Factual	Yes	Long

Where can I find good case studies on outsourcing to Russia? I've looked at a number of the usual places (magazine articles, "biased" websites such as russoft.com, etc.), but would really like a more in-depth set of objective case studies. Thanks for any help you can give.

Template 5

Topic	Type	Prior Effort	Gratitude
Business	Opinion	Yes	Short

Is it a good idea to offer plants to employees if they're willing to take care of them? Folks in my local business group are split between those that say it is great for morale (and clears the air!) and those that say you'll end up with water stains on papers and other problems. What do you think; is this a good thing to do? Thanks.

Template 6

Topic	Type	Prior Effort	Gratitude
Business	Advice	Yes	None

Hi, I'm hoping you can help me think of some good ways to invest my money. I'm in my early 30s, and all of my money is currently in my bank account. It's mostly laziness that has prevented me from investing it so far, but also that I don't know where I should invest it. I would be willing to invest if I knew how to find businesses that are both environmentally friendly (e.g. windpower companies) and that would be a sound investment. I did some google searching on "environmental investments", but I didn't know how much to trust the sources of information. I would be happier if I heard firsthand from somebody. I don't have too much money to invest, but I have enough to buy mutual funds, and enough that I don't want to invest it all in the same company.

Template 7

Topic	Type	Prior Effort	Gratitude
Entertainment	Opinion	Yes	None

Which actress has the first female line in a talking movie? I found on Wikipedia that Al Jolson had the male line, but I can't find any record of which female was first?

Template 8

Topic	Type	Prior Effort	Gratitude
Entertainment	Opinion	Yes	Long

Yes - I am a Food TV junkie. So I ask this. Who is the most skilled chef on TV today? I looked on wikipedia, and saw that Iron Chef Sakai won the "king of Iron Chefs" competition - does this mean that he's the most skilled chef on TV? If not, who is? Thanks for your opinion!

Template 9

Topic	Type	Prior Effort	Gratitude
Entertainment	Advice	Yes	Short

Hi! My 14-year-old son just started at a new school and has become very interested in computer games. I've looked around a bit, but I don't have a good sense as to which ones I should discourage or even forbid. So I guess my question has two parts: First, given that he's interested in games that he's going to play with a bunch of school friends (over the network, and at "game parties"), are there some I should be particularly concerned about? Second, he's mentioned two popular games in particular -- World of Warcraft and Prey -- are these games I should be concerned about? I guess I should mention that he's generally a good kid, and I'm not worried about him misbehaving in general. At the same time, I want to make sure I'm being responsible. Thanks.

Template 10

Topic	Type	Prior Effort	Gratitude
Technology	Factual	No	Short

Anyone know how to send SMS from a Samsung i500 (PalmOS) Phone? I'm willing to install software if I need to, but really miss not being able to send text messages. Thanks.

Template 11

Topic	Type	Prior Effort	Gratitude
Technology	Opinion	No	None

I'm interested in comparing hard drives. Which SATA hard drive on the market today is the best value, in your opinion? There are so many models available today that it's hard to know which one to purchase.

Template 12

Topic	Type	Prior Effort	Gratitude
Technology	Advice	No	Long

Hi, I'm a movie buff, but not really very tech-savvy. I own about 100 movies on DVD (and a bunch still on VHS), and I rent about 2 a week on top of that; I display them on an 8-foot screen using a digital projector that displays high def but is not widescreen (Sharp XR-10X). So, my question is should I get an HD-DVD player, a Blu-Ray player, or wait it out? Thanks for your advice, I appreciate it.

Template 13

Topic	Type	Prior Effort	Gratitude
Business	Factual	No	Long

I am interested in learning more about putting together a benefits package for my small business. Can you point me towards some resources - online or offline - where I can learn more about best practices? Thanks very much, I appreciate the help.

Template 14

Topic	Type	Prior Effort	Gratitude
Business	Opinion	No	Short

Can you recommend good books on volunteer management? Thanks!

Template 15

Topic	Type	Prior Effort	Gratitude
Business	Advice	No	None

I'm starting a new graphic design firm (after working at another firm for 15 years) and seek advice on what to look for when it comes to renting office space. I know we need open spaces, the ability to have good light, etc. Where I really need help is in figuring out contract details I should care about -- in other words, how do I make sure that things work well whether my firm grows quickly or only slowly, and what else should I watch out for? If it matters, I'm located in the Washington DC suburbs.

Template 16

Topic	Type	Prior Effort	Gratitude
Entertainment	Factual	No	None

Are there videos available of the Tonight show from the Steve Allen and Jack Paar years? For sale, online or even just in a museum, I'd like to find them.

Template 17

Topic	Type	Prior Effort	Gratitude
Entertainment	Opinion	No	Long

What are the "don't miss" restaurants to eat at in Las Vegas? Don't worry about price or style of food, I just want to know the best places to go for a great meal. Thanks for your help; I appreciate it.

Template 18

Topic	Type	Prior Effort	Gratitude
Entertainment	Advice	No	Short

I'd like to introduce new games into poker night with my buddies. We've been playing quarter-ante texas hold 'em, seven-card stud, and a few versions of pass-the-trash for a few years now, and it's good fun. But I'd like to shake things up here and there. What other card games do you think I should try to work into a night of poker? Appreciate it.

Appendix B: Sample Questions From Q&A Sites

In chapter 5, a group of volunteers coded questions that have been asked in online question and answer sites. In this appendix, we provide several examples of questions, and how they were coded.

Answerbag

Conversational

Should I spy on my teenage daughters (4 of them) internet activity, I have the tools to do so, but resist it?

Informational

My ps2 series 7 only seems to play games and dvds in black and white please help

Coder Disagreement

Is it me or do ipods vibrate?

Ask Metafilter

Conversational

How immoral is "the other woman?"

Ethics 101: What are the moral implications of being "the other woman?" If someone leaves their lover/partner/wife/husband for you, have you done anything wrong?

Note: Don't worry, I'm not Amber Frey. I'm just genuinely curious.

Informational

Keeping a harmonious relationship with your super

I've been living in New York for only about six months, and was a bit surprised when one of my friends asked if I tipped my superintendent. He told me after the fact that upon your first meeting, you should tip anywhere from $50-100 (and possibly more) to ensure any of your needs are met further along the line, in addition to providing a small tip whenever any small bit of work is done in the apartment. Am I alone in never hearing of this practice? Do you tip your super, and if so, how much?

Coder Disagreement

Festive Pants?

Idiom filter: Party Pants. I was watching Cien Mexicanos Dijeron (the Unavision version of Family Fued) with my girlfriend, when they got to the final stage, where two people try to answer quick questions with the most popular answers. We didn't manage to catch the first contestant's answer, nor the question, and the second contestant was clearly just spitballing with her answer of what we believe was "cepillo de dientes" (toothbrush). According to the the host, the most popular answer was "pantalones festivos." Festive pants? What the hell?

First off, we assume that this is an idiom, and the woman groaned like she couldn't imagine not giving festive pants as an answer. So what the hell are pantalones festivos? Second, what catagory could both a toothbrush and festive pants fall into? Her answer of a toothbrush got, like, three points, which means that it wasn't totally alien, just not a good guess at all.

This has led to rampant and oft-hilarious speculation in my circle of friends, but even those who took Spanish at university levels are still totally clueless as to what pantalones festivos means. (And we do have that right – it was spelled out as the most popular option on the board, though sometimes they do abbreviate common phrases so it might be missing a word or two at the end or something). And there's only one result on Google, which doesn't provide much context.

Yahoo Answers

Conversational

Which performance was your favorite on last nights American Idol? (Men's Top 12)?

And on a sidenote, DANGIT! I wish Josiah Leming was on the show, especially over that 17 year old kid with Van Halen hair.

Informational

Can you recommend me an internet site for a Dad to be!?

I am looking for a good internet forum for new dads which gives advice and guidance as im bricking it...lol. my girlfreind is 15 weeks 4 days pregnant

Coder Disagreement

Can anyone help me with revenge/love?

ok, there are four guys in my class who constantly call and harass me, I like two of them and they like me back but, anyway they called me pretending to be my secrect admirer and other stupid pointless stuff, I need something to do to them, if you need to say more contact me at: [email address]

Appendix C: One Highly Rated and One Low Rated Question of Each Type

In chapter 6, we found evidence that informational questions generally have higher potential archival value than conversational questions. This appendix lists an example of one highly rated and one low rated question of each type.

Archival value 5.0, informational:

> *When playing tracks in iTunes there's always an annoying gap between tracks while the computer catches up with the next song. Is there a plugin or hack to pre-buffer the next song so I get a seamless audio experience? I remember Winamp being able to do this.*

Archival value 1.0, informational:

> *Where is Washington D.C located?*

Archival value 4.0, conversational (the highest-rated conversational question):

> *DO you ever wonder why pictures and videos can't be included in our questions?*

Archival value 1.0, conversational:

> *Why won't my dogfish bark and my catfish meow?*