Effects of Community Size and Contact Rate in Synchronous Social Q&A

Ryen W. White
Microsoft Research
Redmond, WA 98052 USA
ryenw@microsoft.com

Matthew Richardson
Microsoft Research
Redmond, WA 98052 USA
mattri@microsoft.com

Yandong Liu
Carnegie Mellon University
Pittsburgh, PA 15213 USA
yandongl@cs.cmu.edu

ABSTRACT

Social question-and-answer (Q&A) involves the location of answers to questions through communication with people. Social Q&A systems, such as mailing lists and Web forums are popular, but their asynchronous nature can lead to high answer latency. Synchronous Q&A systems facilitate real-time dialog, usually via instant messaging, but face challenges with interruption costs and the availability of knowledgeable answerers at question time. We ran a longitudinal study of a synchronous social Q&A system to investigate the effects of the rate with which potential answerers were contacted (trading off time-to-answer against interruption cost) and community size (varying total number of members). We found important differences in subjective and objective measures of system performance with these variations. Our findings help us understand the costs and benefits of varying contact rate and community size in synchronous social Q&A, and inform system design for social Q&A.

Author Keywords
Synchronous social Q&A, community size, contact rate

ACM Classification Keywords
H.5.3 [Information Interfaces and Presentation] – Group and Organization Interfaces

General Terms
Experimentation, Human Factors, Measurement.

INTRODUCTION

Many systems to support collaborative question answering have been developed and evaluated, including large databases of questions with associated answers (e.g., [2]), forums, list servers, and chat systems that distribute an instant message to a group of people with registered interest in the question topic (e.g., [26]). Web question-and-answer (Q&A) sites, where askers pose questions and others answer them, are popular and can provide high quality answers [12] but can have high latency even when the number of users (the *community size*) is large [13]. Mailing lists broadcast questions to all members, potentially interrupting and overloading everyone and forcing them to adopt strategies to manage discussion threads [19]. Groupware systems have been developed that facilitate synchronous communication with others, normally via instant messaging (IM), c.f. [17]. IM, typically used through a client-side application, allows users to exchange short textual messages with other users with extremely low latency. IM operates within organizations and on the Web, but typically depends on users to identify knowledgeable answerers [27] or directs questions to members of the asker's social network [14]. Automated expertise location (e.g., [3]) can be more effective than letting users choose question recipients based on shared expertise information or social relationships, leading to benefit from diversity [32]. Combining IM and expertise location in synchronous social Q&A systems has potential to facilitate rapid task completion. However, since askers expect answers immediately, the designers of these systems face important challenges associated with interruption costs, answerer availability at question time, and question load balancing. Researchers have theorized about the impact of community size and other factors on member contributions [19,24,33], but a key difference in targeted question-answering is that users are not aware of other community activity, potentially reducing reputation effects and facilitating contributions from more members.

In this paper we describe a large-scale study on the effects of varying the community size (with target sizes of 25, 50, and 100 members) and the rate at which experts in those communities are contacted with a question (two or five at a time) on asking, answering, and overall perceptions of system utility. We use a synchronous Q&A system called *IM-an-Expert*, developed by the authors, that receives questions via IM, automatically identifies candidate answerers by ranking all users by representations of their interests and expertise, routes questions to those available and most able to answer, and mediates the conversation between the asker and answerer. IM-an-Expert operates under the principle that all users are "experts" and ask or answer questions; to ask, users must be willing to answer. We use objective measures such as the fraction of questions answered and time to answer and subjective measures such answer quality and overall utility. The findings improve our understanding of cost-benefit trade-offs in synchronous social Q&A.

The remainder of the paper is structured as follows. We describe related work on social Q&A, the IM-an-Expert system, and our study, including research questions, participants, and methodology. We then present our findings, discuss them and their implications, and conclude.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
Copyright 2011 ACM 978-1-4503-0267-8/11/05...$5.00.

RELATED WORK

We divide our discussion of related work into two parts: first we discuss research on identifying experts, a key element of targeted question-answering, then we focus on methods to contact others and the effect of community size.

Identifying experts: Social matching systems bring people together for interaction in physical and online spaces [30]. Systems that afford implicit social matching let users navigate information spaces to find the desired facts. However, the spaces are constructed so that when users need information beyond that already recorded, pointers are provided to people who can help. Tools such as *Phoaks* [29] use "social navigation" [10] involving the implicit or explicit recommendations of other users to help find relevant content. Phoaks harvests recommended Web pages from newsgroup messages, provides additional information, and lets users explore to locate and contact the recommender. *Designer Assistant* [31] arranges knowledge of software design as a hierarchical series of questions, each tagged with the person in the organization who is most familiar with that particular aspect of the system so that the asker can then target their question directly to the expert.

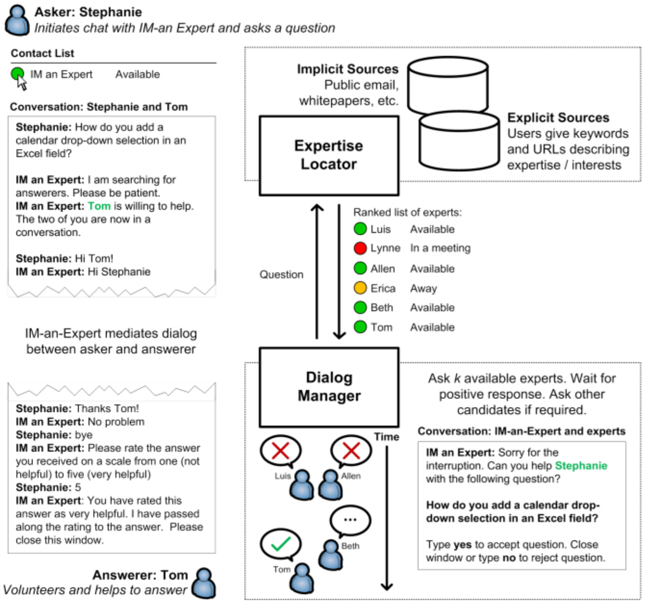
In newsgroups and in collections of Frequently Asked Questions (FAQ), people find answers to specific questions presented in natural language. Ackerman and McDonald address this issue with *Answer Garden* [1] and later *Answer Garden 2* [2]. By allowing users to forward questions to an expert, both systems bridge FAQs and newsgroups. Answer Garden organizes knowledge around a hierarchy of questions and answers; users traverse the hierarchy to locate their question and the corresponding answer. In the system, questions and answers also are labeled with the subject matter expert. If a question is unanswered, a user can contact the responsible expert. Phoaks, Answer Garden, and the Designer Assistant attempt to satisfy users' information needs and facilitate social interaction when existing information spaces are inadequate. Unlike IM-an-Expert, no explicit user profiles are created; instead, users are matched through shared interest/expertise in a topic.

Expertise location systems help find people with specific, desired knowledge [3]. These systems harness technology to locate people and engage with them in order to benefit from their knowledge. Such systems are becoming very popular in both research [18,22,27] and enterprises. *ReferralWeb* [18] and *Expertise Recommender* [22] leverage two types of profiles, one concerning expertise and the other, social relations. Both systems obtain expertise information by data mining. ReferralWeb mines public Web documents for knowledge about potential experts through content analysis to identify salient topics and social relationships. Expertise Recommender mines work products and byproducts such as software source control systems and technical support databases. ReferralWeb does not provide any explicit support for the interaction process—seekers are expected to use whatever means they find appropriate to communicate with the experts. Expertise Recommender provides a simple IM system for users logged into the system [22], one of many ways to communicate in an organizational setting. *BlueReach* [27] is a real-time expertise sharing and capture application that connects an asker to another who can provide the answer. BlueReach provides a browseable directory of expertise categories and subcategories to allow users to locate the right expert and initiate an IM dialog, while safeguarding experts' time. Expertise location systems still assume that answers should come from experts—often experts whose role is to help others—and use previously-created content to reduce their workload. This ignores the powerful concept of peer-support [11], and the advantages that sharing needs and conducting public discussions have for enhancing community cohesion. Of course, peer support has challenges around incentivizing participation and load balancing.

Contacting others: A range of strategies have been employed, ranging from broadcasting messages to all users to targeting individual users, c.f. [26]. We focus specifically on how IM has been used for that purpose. IM is popular in virtual communities, and despite its drawbacks for archiving and navigation, IM is informal and provides instant support for negotiation of meaning [23,26], characteristics needed for free flow and sharing of tacit knowledge. Among other things, researchers have studied the role of responsiveness in IM communication [4], the support for multi-tasking during communication that IM provides [16] and the effect of task type on IM interruptions [9]. *Mimir* [13] a market-based Q&A service, employs a strategy (i.e., broadcasting all questions to all users) that is similar to an email distribution list, since they do not filter question recipients based on personalization. *Aardvark* [14] is an IM-based synchronous social Q&A system that removes the need for users to select the target of the question prior to asking. Instead, Aardvark automatically routes incoming questions to the asker's social network. *ReachOut* [26] combines publish/subscribe technology from listserv with narrowly-focused topics to help reduce information overload. It uses IM for awareness, but has persistence and supports question/user targeting. Systems such as *Babble* [7] and *Well* [25] also use chat to facilitate collaboration.

The effect of community size has been studied previously in real-time collaborative settings [24], although group sizes were small and the purpose was cooperative work not question-answering. Social psychologists have also investigated the effect of group size in collaborative activities (e.g., research on "social loafing" where reduced individual effort is observed in groups [33]).This is based on the assumption that group interactions are visible to all members, something that may not be true in synchronous social Q&A.

The research described in this section illustrates the wealth of related work in this area. However, there are still important unanswered questions, especially for synchronous social Q&A, which continues to emerge as a useful question-answering method (e.g., [14]). To better understand how asker and answerer needs can be met, we perform a large-scale study of a synchronous social Q&A system to


Asker: Stephanie
Initiates chat with IM-an Expert and asks a question
Contact List
IM an Expert Available
Conversation: Stephanie and Tom
Implicit Sources
Public email, whitepapers, etc.
Expertise Locator
Explicit Sources
Users give keywords and URLs describing expertise / interests
Question
Ranked list of experts:
Luis Available
Lynne In a meeting
Allen Available
Erica Away
Beth Available
Tom Available
IM-an-Expert mediates dialog between asker and answerer
Dialog Manager
Ask *k* available experts. Wait for positive response. Ask other candidates if required.
Conversation: IM-an-Expert and experts
Time
Luis
Allen
Beth
Tom
Answerer: Tom
Volunteers and helps to answer

Stephanie: How do you add a calendar drop-down selection in an Excel field?

IM an Expert: I am searching for answerers. Please be patient.
IM an Expert: Tom is willing to help. The two of you are now in a conversation.

Stephanie: Hi Tom!
IM an Expert: Hi Stephanie

IM an Expert: Sorry for the interruption. Can you help Stephanie with the following question?

How do you add a calendar drop-down selection in an Excel field?

Type **yes** to accept question. Close window or type **no** to reject question.

Stephanie: Thanks Tom!
IM an Expert: No problem
Stephanie: bye
IM an Expert: Please rate the answer you received on a scale from one (not helpful) to five (very helpful)
Stephanie: 5
IM an Expert: You have rated this answer as very helpful. I have passed along the rating to the answer. Please close this window.

Figure 1. Interaction flow in IM-an-Expert.

establish the effect of varying community size and the rate with which we send questions to experts in that community. Before describing the study, we describe the system used.

IM-AN-EXPERT

IM-an-Expert is an automated service that receives questions via IM, locates and contacts potential answerers with expertise or interest in the question topic, and mediates the dialog between asker and answerer. Figure 1 illustrates the flow for a typical question submitted to the IM-an-Expert service. The asker initiates an IM conversation with IM-an-Expert, typically via their contact list, and poses the question. The question is used to retrieve a candidate set of experts based on profile information (described later). A small group of those experts who are currently available (not busy, away, in a meeting, in a call, etc. available via presence information from the IM client) are contacted via IM, in descending order of their expertise, to determine whether they are willing to help answer the question. If and when an answerer accepts, other requests are canceled. If a candidate answerer does not respond in time the service asks others. IM-an-Expert then mediates the conversation between the asker and answerer. Once the conversation ends, the asker is asked to optionally rate the quality of the answer they received on a scale from one (not helpful) to five (very helpful). These ratings are communicated to the answerer by IM-an-Expert as an IM. To support the functionality described in this paragraph, the system has two main components: (i) *expertise locator*: selects users who are most likely able to answer a question, and (ii) *dialog manager*: handles question processing and the management of communication throughout the Q&A process.

Expertise Location

To locate subject matter experts, IM-an-Expert searches user profiles created from both explicit and implicit sources. Profiles were work-related (e.g., SharePoint) or non-work related (e.g., hobbies). We now describe the profile sources in more detail.

Self-reported knowledge (explicit): We provided a Web interface where users could create and update an explicit profile. The interface elicited keywords about which users were knowledgeable. Emphasis could be placed on keywords by purposely repeating them. For example, a math expert could enter “math math math” to emphasize that area. Although this was not the ideal design (e.g., we could have provided a rating scale to weight each keyword), the repetition concept was easily understood and keywords could be entered rapidly, which was important given the profile creation overhead. The interface also allowed users to provide URLs of pages about them (e.g., their homepage). Periodically, and prior to profile indexing, the service downloaded text from these URLs and added it to profiles.

Email sent to mailing lists (implicit): For privacy reasons we could not access all user email. Instead, we used email sent to mailing lists within our organization, available on a range of topics. We crawled and indexed these archives, collecting over 300,000 emails for around 30,000 people. Email was preprocessed to exclude headers and quoted text so that each profile contained only the user’s authored text. Non-work related topics were obtained explicitly through users’ provision of keywords and also from crawling distribution lists on non-work topics such as home ownership concerns.

Text (excluding common stopwords) from both sources was combined to form a textual representation of each user’s interests and expertise. We now describe the ranking procedure that ranks users with respect to the incoming question.

Expert Ranking

Selecting those who may *best* provide needed information is the *expert finding* problem, well investigated by the information retrieval (IR) community e.g., [5]. Since the purpose in this investigation was not to develop the optimal user ranking function, but rather to study social Q&A, we used TF.IDF [28], an established ranking function used extensively in IR research and practice, normalized based on profile length (to avoid overweighting large profiles). TF.IDF ranks profiles by combining the frequency of question words within each profile and the inverse of the frequency of those same words across all profiles. TF.IDF’s efficiency also met our real-time ranking needs.

In their profiles, users could configure the minimum time between questions, β, (default is 20 minutes) and limit the maximum number of questions received per day (default is 15). Additionally, the TF.IDF score for each employee in relation to a question is multiplied by a decay function which has low value if this user was recently contacted. This helps balance question load across all users, an important factor in online communities [6]. The function is:

$$Decay = \begin{cases} 0, \Delta t \le \beta \\ 1 - e^{-\Delta t/\alpha}, \Delta t > \beta \end{cases}$$

where Δt is the number of minutes passed since the last question, and α is set to 120 / maximum number of questions received per day. Setting α in this way means that a user can receive their maximum number of questions in two hours with a relatively severe decay of $1 - 1/e$, or over an eight-hour workday with a less severe decay of $1 - 1/e^4$.

To prevent users knowledgeable about popular topics from being overloaded, each question is routed to at most 20 users. This set comprises the top-five users whose score exceeds a threshold, with the remainder coming from backfilling with randomly-selected users contacted least recently.

Dialog Management

A dialog management component coordinates the flow of messages between the asker and answerers. A session is initiated by a user asking a question via IM. The system determines the other users most likely to be able to answer the question and sends an IM to the top-k of them (in our study k = 2 or 5), asking if they are willing to help answer the question. Targeting only a small number of participants per question (rather than broadcasting IMs, as has been described previously [13]) helps ameliorate the effects of IM interruptions. The system only asks users who are currently available to answer. The IM status used by IM-an-Expert is set automatically by Microsoft Office Communicator based on users' calendars (for the "in a meeting" status) and computer activity (for "available" vs. "away from computer"). Users could also set their status manually.

If a potential answerer does not respond within 60 seconds or indicates that she cannot help at the moment (by either closing the IM window or sending "no" in response to the invitation), the system contacts the next user on the list, until up to 20 users have been asked. The system does not inform the user whether they have been selected based on expertise or randomly (though this feature has since been added). Expired candidate answerers who kept their dialog window open still have the option to accept the question until another user agrees to help answer.

Once a user indicates that she is willing to help, other potential answerers are thanked for their time and informed that they are no longer needed. The system then acts as a bridge between the asker and answerer, relaying messages from one to the other. In this way, the system has full control over the conversation, for instance, it can mask the identity of the asker or answerer, record the conversation, and watch for termination. This also lets us record IM dialog for later analysis and sharing (with permission).

We use the IM-an-Expert service to answer important questions on the effects of contact rate and community size. We now describe a user study to examine these effects.

USER STUDY

We conducted a large-scale user study within Microsoft Corporation. In this section we present the research questions and information about participants and their recruitment, experimental groups, and the methodology employed.

Research Questions

We were interested in the effect on subjective and objective measures of social Q&A performance of the rate with which we contact experts and the community size. The study asks the following two research questions:

$\boldsymbol{RQ1}$**:** How does the contact rate per question (k) affect system effectiveness and participant preferences? A smaller contact rate gives higher ranked experts more time to respond to a question before contacting lower-ranked experts.

$\boldsymbol{RQ2}$**:** How does community size (n) impact system effectiveness and participant preferences?

Answers to these questions help us to better understand the impact of these factors in synchronous Q&A, with a view to designing better social Q&A systems.

Participants and Recruitment

Participants were members of our organization. Email invitations were sent to randomly-chosen members whose user profiles from mailing lists alone exceeded one kilobyte in size. This seemed sufficient for expertise location even if participants elected not to manually update their profile. All participants worked on the same campus. We enforced this since we did not want communication to be affected by time zone differences and geographic proximity that has been shown to be important in computer-mediated collaboration [8]. Email invitations requested that volunteers be available for the two-week duration of the study and provided an overview of our expectations (e.g., that they would complete pre- and post-experiment surveys). We offered monetary rewards of 100 USD to five randomly-selected volunteers as a participation incentive, unrelated to how much they used their assigned system. The participation rate was 8.9%, which was reasonable given the substantial time commitment that was expected from our volunteers.

Participants were asked to complete a pre-experiment questionnaire that captured basic demographic information and described their experiences with IM. 64% of the volunteers (195 male and 58 female) completed this survey. The most frequent age bracket was 25 years to 34 years. Previous work has shown that users' frequency of use of instant messaging affects aspects of their interaction behavior such as message length (e.g., [16]). To establish the experience level of our participants, we asked about how frequently participants used IM generally. Responses were provided on a five-point scale: 1=*never*, 2=*rarely*, 3=*occasionally*, 4=*fairly often*, and 5=*very often*. The mean rating was 4.5 (median 5), suggesting that participants were highly familiar with IM. Instant messaging is used extensively in our organization: all machines have Microsoft Office Communicator enabled by default, and participants sent and received instant messages to and from the IM-an-Expert system, and each other, through that IM client.

Experimental Groups

We used a between-groups experimental design. For each variant of k, we constructed groups of 25, 50, and 100 participants, and randomly assigned volunteers to them. This design required 350 participants spanning six experimental groups; we had 402 volunteers in total. Since participation was remote, we expected some attrition due to changing work priorities or other factors. To help counteract this, we assigned a small number of additional participants to each group (e.g., we assigned 30 participants to each group meant to have only 25). This totaled 400 of the 402 volunteers. We started the study with all groups above 65% of their target capacity; attrition affected all experimental groups equally. Table 1 shows the target number of participants in each group, the number invited, the actual number that started the experiment in each group, and the percentage of actual participants with respect to the target size.

Table 1. Target group sizes and actual number participants. Note that groups with target size 100 have somewhat lower participation rates due to inviting only 10% more than the target size, vs. 20% for the smaller groups. This was done to ensure the small groups were of adequate size.

k	**Group**	**Community size (*n*)**			
		Target	**Invited**	**Actual**	**%Target**
2	1	25	30	20	80%
	2	50	60	38	76%
	3	100	110	69	69%
5	4	25	30	23	92%
	5	50	60	43	86%
	6	100	110	66	66%

We deployed a separate version of IM-an-Expert to each of the six groups and varied the *k* parameter per our design.

Methodology

The study lasted two weeks and participation was remote. Following recruitment we:

(i) Asked participants to take a pre-experiment survey.

(ii) Randomly-assigned participants to experimental group. Participants only knew of the existence of their own group, helping to ensure the realism of the study.

(iii) Asked participants to visit their profile page and provide keywords and URLs that describe their interests and their expertise. The provided keywords and URL contents were indexed prior to the start of the study and re-indexed daily to capture any profile updates.

(iv) Participants were instructed to "consider using IM-an-Expert as a resource for answering questions" for the duration of the study. Examples of questions posed by participants included: "When is the next public holiday in New York?" and "What is a good elliptical trainer under $300?" Dialogs and system events were logged.

(v) Two weeks from the start date, participants were informed that the study was complete and were asked to complete a post-experiment survey. 182 participants (70% of those that completed the pre-experiment survey) did so. Attrition was spread evenly across groups.

In the next section we present experimental findings.

FINDINGS

We report on findings from the post-experiment survey and analysis of the logs gathered during the study on each of the six systems. To reduce redundancy in the presentation of results, we present findings in terms of question asking, question answering, and participants' overall perceptions of the systems, rather than by research question. We use non-parametric statistical testing at $p < .05$. In addition to monitoring the effects of our two independent variables, we also watched for interaction effects between the variables. Since no significant interactions were noted, we do not report any further results on interaction effects. Ratings on five-point scales are converted to numeric form such that a rating of one is a negative response (e.g., *highly ineffective*) and a rating of five is a positive response (e.g., *highly effective*).

General Usage of IM-an-Expert Systems

Around 50% of the participants asked and answered questions in the two-week duration of the study (with around 35% of participants performing both activities). The average number of questions asked per participant was just over one, and a quarter of participants asked and answered half of the questions.[1] On average, at question time 20-25% of participants were available to answer (not in meeting, busy, etc.), and dialogs between askers and answers lasted six minutes and comprised 10 dialog turns, evenly distributed between askers and answerers.

Asking Questions

We analyzed system performance in asking questions via survey responses, in situ relevance ratings, and log analysis.

Participant Perceptions

We asked participants about their experiences with asking questions using their assigned system in a post-experiment survey. Although the survey was distributed at the end of the two-week study (and system usage may not have been fresh in participants' minds), the survey was necessary to measure aspects of overall satisfaction, effectiveness, timeliness, etc. which were not directly captured in the interaction logs or in answer ratings (which were optional and not always provided). The survey may also be unaffected by politeness obligations that may bias the in-situ answer ratings, especially in a small community, where the asker and the answerer may be socially connected. The survey asked:

- How effective was the system for asking questions? (1=*highly ineffective* to 5=*highly effective*)
- How successful were you at receiving answers? (1=*highly unsuccessful* to 5=*highly successful*)
- What was the quality of the answers you received? (1=*poor* to 5=*excellent*).
- What fraction of questions had satisfactory answer? Participants reported the number of questions asked and how many of those that were answered as *satisfactory*, *somewhat satisfactory*, or *not satisfactory*. We report the fraction of questions with *satisfactory* answers in Table 3.
- Rate the timeliness of the satisfactory answer(s) that you received. (1=*extremely slow* to 5=*extremely fast*).

Table 2 summarizes participant responses.

[1] This participation rate is significantly higher than reported in other studies of online communities [6,19]. This may relate to the directed nature of IM, social obligations to respond to IMs, or participant engagement with the study.

Table 2. Participant perceptions of asking, per group.

Measures	Comm. size (n)	Contact rate (k)		
		2	5	All
Asking effectiveness	25	3.29	3.00	*3.13*
	50	3.48	3.12	*3.36*
	100	3.85	3.38	*3.69*
	All	*3.51*	*3.23*	
Success	25	1.86	2.25	*2.07*
	50	2.86	2.46	*2.67*
	100	3.39	2.69	*3.16*
	All	*2.96*	*2.66*	
Answer quality	25	2.00	2.50	*2.27*
	50	2.68	2.58	*2.62*
	100	3.30	2.88	*3.16*
	All	*2.84*	*2.68*	
% questions answered satisfactorily	25	58.7%	63.5%	*61.0%*
	50	62.8%	65.0%	*63.9%*
	100	63.4%	67.3%	*65.4%*
	All	*62.5%*	*65.7%*	
Timeliness of satisfactory answers	25	3.77	3.95	*3.86*
	50	3.95	4.11	*4.03*
	100	4.08	4.52	*4.30*
	All	*4.03*	*4.33*	

Our analysis revealed that participants' ratings of *asking effectiveness* increased with community size and increased as contact rate shrank (community size: $\chi^2(2)=11.75$, $p < 0.01$; contact rate: $\chi^2(2)=5.05$, $p=0.08$). Similar trends were observed for *success* and *answer quality*, both of which favored larger groups and fewer people contacted per question. Despite the low ratings, which perhaps reflected the relatively small community sizes and contact rates, $k=2$ led to more *success* (2.96 with $k=2$ vs. 2.66 with $k=5$), and improved *answer quality* (2.84 with $k=2$ vs. 2.68 with $k=5$). An explanation for lower answer quality when $k=5$ is that there was insufficient expertise for some questions or that it was difficult to find the relevant experts. To verify that our system was finding experts, we measured the correlation between an answer's rating and the corresponding answerer's expertise score, and found this to be positive. We therefore believe that at least one of the reasons that a high contact rate leads to lower answer ratings is that less-knowledgeable experts are contacted sooner. The benefit from increasing community size surpassed the benefit from targeting more expert users. As one participant pointed out: "*The system is only as good as the experts listening to the question. I felt my answer wasn't provided because there wasn't a base of experts enough*" (sic). This confirms our intuition that when a range of questions is expected, it may be more important to increase community size than improve expertise location, especially in small communities.

We also asked participants about the fraction of questions that were answered satisfactorily and the timeliness of the satisfactory answers received. Our findings showed some important differences with measures of success and quality discussed in the previous paragraph. Although the fraction of questions answered satisfactorily and in a timely manner increased with the community size, increasing k appeared to lead to better performance across all groups and for both measures ($\chi^2(2) \geq 9.3$, $p \leq 0.01$). In explaining these findings, it is worth paying special attention to the question wording, especially the inclusion of the word "satisfactory." It may be that contacting more—and perhaps less expert—users is a way to get a reasonable answer to a question in less time, but it may not be a way to get a high-quality answer. Asking more users in parallel may lead to a reduction in answer time, but it also seems likely that in doing so the question will get picked up by less capable answerers, volunteering to answer the question before experts. This highlights a critical challenge in the design of synchronous social Q&A systems: satisfying askers' desire for high quality answers and their desire for low answer latency.

It is also worth noting that even small differences in the expertise of the users contacted (top two versus top five) were perceived by participants, even retrospectively when completing the post-experiment survey. In addition to asking participants for general perceptions of system effectiveness, the system also allowed askers to provide assessments on the answer received *in situ* at the conclusion of the IM dialog. We now analyze these ratings in more detail, to see if they are impacted by community size or contact rate.

Answer Ratings

At the conclusion of the answer dialog, the asker could rate the answer on a five-point scale ranging from one (*not helpful*) to five (*very helpful*). These ratings were provided to the system via IM dialog. Around 75% of all IM conversations had a rating from the asker. Table 3 presents the average ratings from each of the groups, for each value of k.

Table 3. Answer ratings, per group.

Comm. size (n)	Contact rate (k)		
	2	5	All
25	3.33	3.25	*3.29*
50	3.50	3.32	*3.43*
100	3.61	3.52	*3.58*
All	*3.56*	*3.42*	

Two things should be noted from our findings: (i) answer ratings seem higher than the *answer quality* ratings reported in Table 2, and (ii) answer rating increases as the community size increases and there is a slight (although still statistically-significant) answer rating increase with a more targeted contacting strategy (both $\chi^2(2) \geq 6.45$, $p \leq 0.04$). For (i), it might be that participants' overall impressions were more negative than the average of their individual experiences, or that they regarded answer *helpfulness* as different from *answer quality*. Finding (ii) may be attributable to the system contacting less knowledgeable experts (lower in the ranked list) as k increases. While it may appear astounding that small differences in contact rate translates into noticeable differences in answer rating, there may be an explanation. Although the difference in k is small, community sizes in our study were also small, and there may only be a small number of experts on any topic (if any) available at question time. This creates a situation where if anyone but the top one or two experts answer, then they will have a markedly lower level of expertise. Indeed, when the expert who responded was in the first rank position, the average assigned answer rating was 3.55. Whereas, when the expert is fifth in the list, the rating was 3.40, and when the expert is randomly selected, the average rating was 3.23. Although these differences are not significant (Kruskal-Wallis analysis: $\chi^2(2)=5.49$, $p=0.07$), our expertise locator ranks users sensibly and in a way that was noticed by participants.

Log Analysis

From analyzing usage logs gathered during the study, we could determine the number of questions asked, the fraction of questions asked that received answers, and the median time to receive an answer. These statistics are summarized in Table 4. We define answering as another participant indicating that they are willing to answer, rather than the statement in the dialog where an answer is received; automatically identifying answers in dialog may be challenging.

Table 4. Objective measures of asking (m: mins, s: secs).

Measures	Comm. size (n)	Contact rate (k)		
		2	5	*All*
Total number of questions	25	23	20	*43*
	50	54	59	*113*
	100	84	77	*161*
	All	*161*	*156*	
Average answer percentage	25	64.2%	66.4%	*65.2%*
	50	67.0%	68.8%	*68.2%*
	100	68.3%	71.2%	*70.6%*
	All	*67.2%*	*69.9%*	
Average time to answer	25	4m 31s	4m 2s	*4m 8s*
	50	3m 58s	3m 34s	*3m 40s*
	100	3m 12s	2m 57s	*3m 5s*
	All	*3m 37s*	*3m 22s*	

The findings in Table 4 show that the number of questions posed by our participants grows linearly with actual group size (shown in Table 1), and that there are no differences in the number of questions between groups with differing k. This was confirmed by a two-factor Kruskal-Wallis analysis (with Bonferroni correction), with *community size* and *contact rate* as factors ($\chi^2(2)=2.04$, $p=0.36$). Interestingly, a larger fraction of questions are answered as group size increases, with only a small difference in answer ratio with different k values. This is true even though the ratio of queries asked to actual group size remains fairly constant. Although these differences are not significant ($\chi^2(2)=4.82$, $p=0.09$), they do suggest a greater-than-linear benefit from adding more users to the system, which we may be due to a broader range of expertise from which to draw upon.

The time to receive an answer decreases as community size increases; each doubling in size leads to a reduction in time-to-answer of around 30 seconds. This is promising, but we may witness diminishing marginal improvements as community size grows, since these and other measures become subject to ceiling effects. We also note a reduction in time to receive an answer as we increase k from 2 to 5 and a decrease in time-to-answer as community size grows (both significant, $\chi^2(2) \geq 7.83$, $p \leq 0.02$). However, as we will show, increasing the contact rate means venturing further down the ranked list and may not lead to improvements in *answer quality*, since users added may have less expertise.

In this section we highlighted interesting effects on the answer rate as community size increases, an apparent trade-off between answer quality and timeliness associated with contact rate, and noticeable improvements in relevance with larger groups and/or more targeted routing.

Answering Questions

The experience for answerers is also important in the success of social Q&A systems. Two factors that may affect answering are the interruption cost that answerers may incur by being part of the community and the relevance of the questions they receive. We analyzed logs and asked questions in the post-experiment survey question answering.

Interruptions and Effectiveness

To estimate interruption cost we used log data and responses to the post-experiment survey. Using the log data, we computed the number of participants interrupted per question. This was the number of participants who received a question via IM, regardless of whether they eventually answered. Since participants' perceptions of the interruptions the received are also important, we asked participants "To what extent did you feel bothered by incoming questions?" with response options ranging on a five-point scale from *extremely bothered* (rating=1) to *not bothered at all* (rating=5). To measure participants overall perceptions of how effective the system was, we also asked participants "How effective was the system for answering questions?" with a response range from *highly ineffective* (rating=1) to *highly effective* (rating=5). Table 5 summarizes our findings.

Table 5. Measures of answering, per group.

Measures	Comm. size (n)	Contact rate (k)		
		2	5	*All*
% of group interrupted	25	14.5%	25.2%	*19.9%*
	50	9.6%	17.4%	*13.6%*
	100	7.7%	14.2%	*10.8%*
	All	*10.4%*	*17.7%*	
Felt unbothered by questions	25	3.59	3.34	*3.44*
	50	3.92	3.78	*3.87*
	100	4.23	4.09	*4.15*
	All	*4.20*	*4.00*	
Answering effectiveness	25	3.44	3.35	*3.40*
	50	3.47	3.36	*3.42*
	100	3.52	3.40	*3.44*
	All	*3.50*	*3.37*	

The findings of the analysis show that *answering effectiveness* is near-constant as community size increases, but is higher for cases where fewer users are contacted per question ($\chi^2(2)=9.2$, $p=0.01$). As group size increased, the fraction of the group that was contacted per question decreased, as the number contacted remained at around six. Indeed, an advantage of targeted Q&A methods over broadcast Q&A media such as mailing lists is that the number of users interrupted per question can remain constant irrespective of community size. This was reflected in participants' perceptions, with ratings of how *un*bothered they were questions increasing with group size ($\chi^2(2)$=4.24, *p*=0.12). Participants felt less bothered with lower *k* ($\chi^2(2)$=7.91, *p*=0.02), perhaps because lower contact rate interrupted fewer users.

Question Relevance

In addition to interruption cost, the experience for the answerer may relate to question relevance. Question routing is critical to the effectiveness of any expertise location system. In targeted systems such as IM-an-Expert, irrelevant questions are likely to go unanswered and negatively affect the perceptions regarding system effectiveness of those asked. As such, we investigated question relevance *from the answerer's perspective*. The post-experiment survey asked participants: "Approximately what percentage of questions asked were relevant to you?" and provided the following list of response options: 0%, 1-10%, 11-20%, 21-30%, 31-50%, 71-90%, 91-100%. More response granularity was provided early in the scale (i.e., between 0% and 30%) as we did not expect a large fraction of questions to be relevant given the small number of participants relative to the breadth of potential question topics. We summarize participant responses for variations in *k* in Figure 2. Differences for variations in group size were observed but small, so for simplicity we elect not to report those findings.

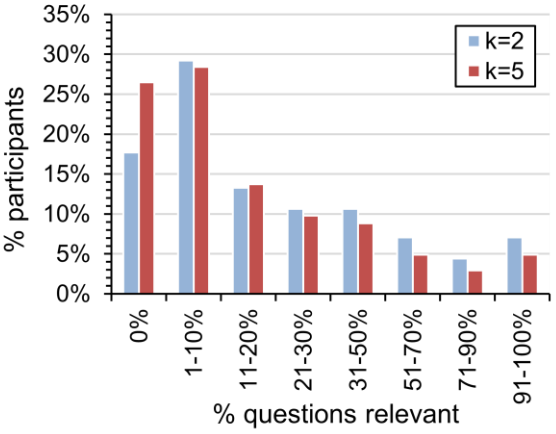


Figure 2. Percentage of questions relevant for different *k*.

We see that *k*=2 outperforms *k*=5, with fewer participants receiving no relevant questions, perhaps since the lower contact rate afforded a more targeted selection of users.

Question routing did well for some participants and not well for others. We investigated possible reasons for this, and found one likely cause. Although requested to do so, 39.8% of all IM participants did not provide any keywords or URLs describing their interests for use in building their profile. This meant that for those participants we relied solely on their public email, which may only partially represent their interests. Recall from our description of participant selection that all participants had an email-based profile that was at least one kilobyte in size. Interestingly, 43 of the 54 participants (80%) who rated the relevance of the questions posed via IM as 0% also did not provide explicit profile information. Since the profiles of those participants may not have been complete or accurate, this may have resulted in less relevant questions. It therefore seems important to explore low-overhead ways of gathering profile information from other sources or explicitly from users.

To find out more about the effects of irrelevant questions, we probed participants to understand reasons for not answering certain questions posed to them. Popular responses were "I didn't know the answer" (55% for *k*=2 and 49% for *k*=5) and "Question was not relevant to me" (22% for *k*=2 and 27% for *k*=5). This points to an interesting difference between topic knowledge and expertise that affects targeted Q&A solutions. A participant may have knowledge of the question topic (and have provided this explicitly to the system), but may lack the *expertise* to answer.

In this section we have shown that the interruption rate is lower for larger groups, that these and relevance differences are perceptible to users, and that question irrelevance in our system may be associated with a lack of profile information or matching based on topic knowledge not expertise level.

Overall Perceptions

In addition to studying participants' experiences with asking and answering questions, we were also interested in exploring participants' general perceptions of the synchronous social Q&A system, and how that is affected by group size and contact rate variables. In the post-experiment survey, we asked: "How useful do you think this system is in general for getting questions answered?" and provided response options from *highly useless* (rating=1) to *highly useful* (rating=5). Figure 3 summarizes participant responses.

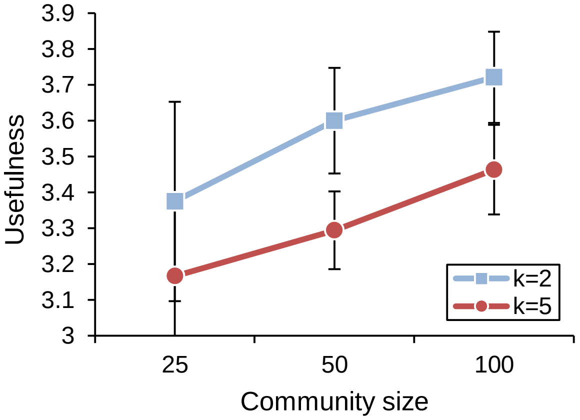


Figure 3. Usefulness for different community sizes and *k*-values. Error bars depict standard error of the mean (±SEM).

Figure 3 illustrates that participants found the *k*=2 systems more useful overall, and that this trend held for groups of all sizes. Differences were significant for the larger groups (50 and 100) using post-hoc testing (both $p \leq 0.04$). It might be that limiting the number of users asked per question allowed the system to balance the apparent needs of the askers (for high quality answers) with the needs of the answers (for fewer interruptions and more relevant questions), leading to a better overall experience. Although the k=5 systems led to a greater fraction of questions being answered, a greater fraction answered satisfactorily, and more timely answers, the k=2 systems were more useful. Users may be willing to wait longer for a high-quality answers but may be adverse to more IM interruptions.

Figure 3 illustrates that participants found the k=2 systems more useful overall, and that this trend held for groups of all sizes. Differences were significant for the larger groups (50 and 100) using post-hoc testing (both $p \leq 0.04$). It might be that limiting the number of users asked per question allowed the system to balance the apparent needs of the askers (for high quality answers) with the needs of the answers (for fewer interruptions and more relevant questions), leading to a better overall experience. Although the k=5 systems led to a greater fraction of questions being answered, a greater fraction answered satisfactorily, and more timely answers, the k=2 systems were more useful. Users may be willing to wait longer for a high-quality answer but may be adverse to more IM interruptions.

Although the ratings appear to increase with each jump in group size, it is worth noting that the maximum usefulness rating is five and there will be a ceiling effect that is not visible in Figure 3. Further study with even larger communities is required to understand these and similar effects.

DISCUSSION AND IMPLICATIONS

We have studied the effects of community size (25, 50, or 100 members) and contact rate (two at a time or five at a time) for synchronous social Q&A. The study analyzed the effects of these variables in terms of objective and subjective measures, and from the standpoint of askers, answers, and all members' general perceptions of utility. Every metric showed improvement with increased community size, including increased fraction of questions answered, asking effectiveness, answer quality, and answer ratings, along with a corresponding decrease in the time to receive an answer, number of users who were bothered by incoming questions, and fraction of the community that was interrupted. However, there appears to be an interesting tradeoff between answer quality and timeliness associated with contact rate; a higher k leads to quicker answers but a somewhat lower answer quality. We also showed that users receiving irrelevant questions may be associated with a lack of profile information or matching based on topic knowledge but not expertise level. Finally, the overall perception of usefulness was maximized by larger community sizes and a reduced contact rate.

Askers prefer community sizes and contact rates that facilitated answers that were timely (k=5) and of high quality (k=2). Answerers prefer community sizes and groups that provided relevant questions and fewer interruptions (k=2 only). More of these preferences—specifically high quality answers, relevant questions, and fewer interruptions—would be satisfied with a lower k, perhaps explaining why in our analysis of overall perceptions, groups with k=2 found IM-an-Expert more useful than those with k=5. An interesting research direction is on balancing askers' desire for timely responses with low interruption costs. Some progress has been made in this area via market schemes [13].

Although IM may be regarded as an intrusive method our findings showed that our participants rarely felt bothered by the IM conversation invites (indeed, 40% of all participants felt "not bothered at all"). By varying the number of people contacted per question and the size of the groups, user perceptions of interruption changed and could be potentially controlled as an objective of the design of social Q&A systems. There has been significant research on the costs of interruption demonstrating that decreased interruptions leads to more efficient task completion [15]. More user studies are needed to understand how users react to unexpected notifications, determine the best modes of interruption, and determine how to select the contact rate that balances the efficiency gains for the asker vs. answer quality and interruption cost of the answerer. Further work is also needed on the utility of using answerers' IM status in determining whether to contact them with a question invite.

Some questions posed during the study were general in nature, and could be answered by many community members, regardless of its size. For such questions, success may depends more on answerer availability than expertise location. In practice, approximately 20-25% of the participants were available to answer at any given point in time (as indicated by their IM client status). Larger community sizes create more opportunity to route questions to available experts. Improvements in the performance of the system as group size grows may be due to a number of factors, including an increase in the number of available users from which experts could be chosen, and a simply a large community with a broader range of question expertise.

It is interesting to note the distinction that participants made between the topical relevance of the question and their ability to answer it. It was often the case the questions were relevant but the candidate answerer was not sufficiently expert in the subject matter to be able to answer. In future iterations of the profile page we will include a way for users to express both areas of interest and expertise, and their level of competence within each one, perhaps on a rating scale with options ranging from *novice* to *expert*.

CONCLUSIONS

We have presented an investigation of the impact of community size and contact rate on the effectiveness of synchronous social Q&A. Over 400 volunteers participated in our experiment that studied the impact of our independent variables on the performance of an operational Q&A system, in terms of asking, answering, and overall user perceptions. Our findings suggest that as community size increases, system performance increases in a number of objective and subjective measures. We also found that answerers prefer a low contact rate, askers generally prefer a low contact rate, apart from when focused on timeliness or the fraction of answers that were satisfactory. In addition, in terms of overall utility, participants found systems with a low contact rate to be more useful. To satisfy most users, it seems that synchronous social Q&A systems should use low contact rates and large communities. More research is needed on the asker quality-timeliness tradeoff. Since no ceiling effects were observed in our measures as community size grew, further study with larger communities is needed to understand performance limits as Q&A systems scale.

REFERENCES

1. Ackerman, M.S. 1994. Augmenting the organizational memory: a field study of answer garden. *Proc. CSCW*, 243–252.
2. Ackerman, M.S. and McDonald, D.W. 1996. Answer Garden 2: Merging organizational memory with collaborative help. *Proc. CSCW*, 97–105.
3. Ackerman, M.A. and McDonald, D.W. 1998. Just talk to me: a field study of expertise location. *Proc. CSCW*, 315–324.
4. Avrahami, A., Fussel, S., and Hudson, S. 2008. IM waiting: timing and responsiveness in semi-synchronous communication. *Proc. CSCW*, 285–294.
5. Balog, K., Azzopardi, L., and De Rijke, M. 2006. Formal models for expert finding in enterprise corpora. *Proc. SIGIR*, 43–50.
6. Beenen, G., Ling, K., Chang K., Wang, X., Resnick, P., and Kraut, R. 2004. Using social psychology to motivate contributions to online communities. *Proc. CSCW*, 212–221.
7. Bradner, E., Kellogg, W. and Erickson, T. 1999. The adoption and use of babble: a field study of chat in the workplace. *Proc. CSCW*, 139–158.
8. Bradner, E. and Mark, G. 2002. Why distance matters: effects on cooperation, persuasion and deception. *Proc. CSCW*, 226–235.
9. Czerwinski, M., Cutrell, E., and Horvitz, E. 2000. Instant messaging and interruptions: influence of task type on performance. *Proc. OZCHI*, 356–361.
10. Dourish, P. and Chalmers, M. 1994. Running out of space: models of information navigation. *Proc. HCI.*
11. Greer, J.E., McCalla, G.I., Cooke, J.E., Collins, J., Kumar, V.S., Bishop, A., and Vassileva, J.I. 1998. The intelligent helpdesk: supporting peer-help in a university course. *Proc. Intelligent Tutoring Systems Conference*, 494–505.
12. Harper, F.M., Raban, D., Rafaeli, S., and Konstan, J.A. 2008. Predictors of answer quality in online Q&A sites. *Proc. CHI*, 865–874.
13. Hsieh, G. and Counts, S. 2009. Mimir: a market-based real-time question and answer service. *Proc. CHI*, 769–778.
14. Horowitz, D. and Kamvar, S.D. 2010. The anatomy of a large-scale social search engine. *Proc. WWW*, 431–440.
15. Iqbal, S.T. and Horvitz, E. 2007. Disruption and recovery of computing tasks: field study, analysis and directions. *Proc. CHI*, 677–686.
16. Isaacs, E., Walendowski, A., Whittaker, S., Schiano, D.J., and Kamm, C. 2002. The character, functions, and styles of instant messaging in the workplace. *Proc. CSCW*, 11–20.
17. Jensen, C., Farnham, S., Drucker, P., and Kollack, P. 2000. The effect of communication modality on cooperation in online environments. *Proc. CHI*, 470–477.
18. Kautz, H., Selman, B., and Shah, M. 1997. Referral-Web: combining social networks and collaborative filtering. *Communications of the ACM*, 40(3): 63–65.
19. Ludford, P.J., Cosley, D., Frankowski, D., and Terveen, L. 2004. Think different: increasing online community participation using uniqueness and group dissimilarity. *Proc. CHI*, 631–638.
20. Malone, T.W., Grant, K.R., Turbak, F.A., Brobst, S.A., and Cohen, M.D. 1987. Intelligent information sharing systems. *Communications of the ACM*, 30(5): 390–402.
21. Masmoodian, M. and Apperley, M. 1996. The effect of group size and communication modes in CSCW environments. *Proc. OZCHI*, 42.
22. McDonald, D.W. and Ackerman, M.S. 2000. Expertise recommender: a flexible recommendation architecture. *Proc. CSCW*, 231–240.
23. Nardi, B., Whittaker, S. and Bradner, E. 2000. Interaction and outeraction: instant messaging in action. *Proc. CSCW*, 79–88.
24. Oliver, P.E. and Marwell, G. 1998. The paradox of group size in collective action: a theory of the critical mass. II. *American Sociological Review*, 53(1):108.
25. Rheingold, H. 2000. *The virtual community: Homesteading on the electronic frontier*. MIT Press.
26. Ribak, A., Jacovi, M., and Soroka, V. 2002. “Ask before you search” peer support and community building with ReachOut. *Proc. CSCW*, 126–135.
27. Singley, K., Lai, J., Kuang, L., and Tang, J.-M. 2008. BlueReach: harnessing synchronous chat to support expertise sharing in a large organization. *Proc. CHI*, 2001–2008.
28. Spärck-Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1): 11–21.
29. Terveen, L., Hill, W., Amento, B., McDonald, D., and Creter, J. 1997. Phoaks: a system for sharing recommendations. *Communications of the ACM*, 40(3): 59–62.
30. Terveen, L. and McDonald, D.W. 2005. Social matching: a framework and research agenda. *ACM Transactions on Computer-Human Interaction*, 12(3): 401–434.
31. Terveen, L.G., Selfridge, P.G., and Long, M.D. 1995. Living design memory: framework, implementation, lessons learned. *Journal of Human-Computer Interaction*, 10(1): 1–37.
32. Whittaker, S. 1996. Talking to strangers: an evaluation of factors affecting electronic collaboration. *Proc. CSCW*, 409–418.
33. Williams, K.D. and Karau, S.J. 1991. Social loafing and social compensation: the effects of expectation of co-worker performance. *Journal of Personality and Social Psychology*, 61(4): 570-581.