A Crowd-Powered Socially Embedded Search Engine

Jin-Woo Jeong[1,2], Meredith Ringel Morris[1], Jaime Teevan[1], and Dan Liebling[1]

[1]Microsoft Research, Redmond, WA, USA, [2]Hanyang University, Ansan, Republic of Korea

selphyr@hanyang.ac.kr, {merrie, teevan, danl}@microsoft.com

Abstract

People have always asked questions of their friends, but now, with social media, they can broadcast their questions to their entire social network. In this paper we study the replies received via Twitter question asking, and use what we learn to create a system that augments naturally occurring "friendsourced" answers with crowdsourced answers. By analyzing of thousands of public Twitter questions and answers, we build a picture of which questions receive answers and the content of their answers. Because many questions seek subjective responses but go unanswered, we use crowdsourcing to augment the Twitter question asking experience. We deploy a system that uses the crowd to identify question tweets, create candidate replies, and vote on the best reply from among different crowd- and friend-generated answers. We find that crowdsourced answers are similar in nature and quality to friendsourced answers, and that almost a third of all question askers provided unsolicited positive feedback upon receiving answers from this novel information agent.

Introduction

Social networking site (SNS) question asking is a common phenomenon in which people use status updates to ask questions of their contacts on social networking sites like Facebook (Lampe et al., 2012; Morris et al., 2010a; Panovich et al., 2012) or Twitter (Efron and Winget, 2010; Li et al., 2011; Paul et al., 2011b). As an example, Figure 1 shows a Twitter user asking his followers, *What type of iPhone case to get????* A friend responds, *the mophie!! It charges your phone* [1]

SNS question asking offers complementary benefits to traditional keyword-based search (Morris et al., 2010b). Some search engines attempt to capture the benefits of both approaches by bringing question asking into the search experience. For example, Bing (2012) offers users the ability to ask their network about recently issued queries. Alternatively, algorithmically-generated content can be brought into the question asking experience. Hecht et al. (2012) call search tools that monitor social networking sites and provide automated answers *socially embedded search engines*. Socially embedded search engines allow people to express their information needs in natural language, receive system-generated and friend-generated answers in a unified context, and get responses in spite of unfavorable conditions (e.g., to a small network or during off-peak hours (Morris et al., 2010b; White et al., 2011)).

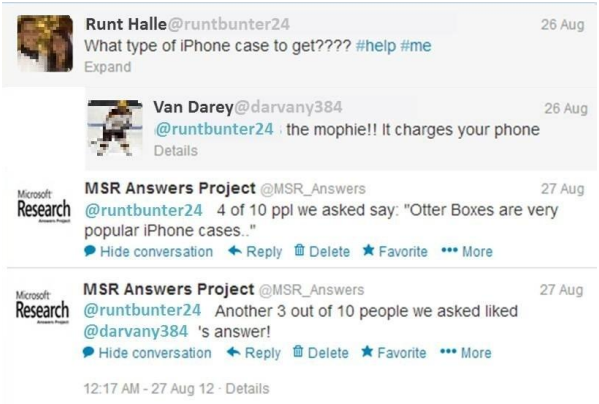

Figure 1. A public Twitter conversation in which MSR Answers responds to a question, offering a new answer and feedback on an answer given by the asker's friend.

However, Hecht et al.'s (2012) exemplar of a socially embedded search engine, SearchBuddies, encountered several challenges upon deployment. It was more suited to answering factual questions than the more common (Morris et al., 2010a) subjective ones, and it sometimes offered answers that did not match social norms and expectations. In this paper we address these obstacles by creating a socially embedded search engine, called *MSR Answers*, that incorporates human judgment by using crowd workers to create high-quality answers to public Twitter questions. For example, in Figure 1 MSR Answers suggests the Otter Box iPhone case. MSR Answers' crowd-powered nature enables it to offer answers like this that benefit from human insight. The answers are also likely to conform to social norms, since crowd workers can identify questions that may be inappropriate to answer (such as those meant for

[1] Tweets are quoted verbatim, including spelling errors. Usernames and photos have been altered to protect privacy).

humorous or rhetorical effect) or require special considerations (such as questions about sensitive topics).

To inform the design of MSR Answers, we begin by analyzing Twitter Firehose data to understand the Q&A ecosystem, characterizing the questions that receive answers and the answers they receive. We then describe MSR Answers, which uses the crowd to identify question tweets, create candidate replies, and vote on the best reply from among crowd- and friend-generated answers. We show that even without a pre-existing social connection, crowdsourced answers are similar in nature and quality to the naturally occurring answers. Almost a third of all question askers provided unsolicited positive feedback after receiving replies from MSR Answers. The two primary contributions of this work are (1) a large-scale analysis of Twitter Q&A exchanges that significantly extends existing knowledge by focusing on the answers people receive; and (2) a process for generating high-quality crowdsourced answers. Our results show that a crowd-powered socially embedded search engine can improve the SNS question asking experience.

Related Work

Although this paper focuses on SNS question asking, there are many ways people ask questions online. For example, a person can post a question on a community Q&A site such as Yahoo! Answers (Adamic et al., 2008; Liu et al., 2011) and receive answers from members of the public. SNS question asking differs from community Q&A in that users post their questions as part of a broader dialog with a specific subset of Internet users, such as their Facebook friends or Twitter followers. The conventions and restrictions of various social networking sites impose limits (e.g., on the length of a post) that change the nature of Q&A dialogues in these media.

SNS question asking has been studied in many contexts, including on Facebook (Lampe et al., 2012; Morris et al., 2010a&b; Panovich et al., 2012; Teevan et al., 2011), on Twitter (Efron and Winget, 2010; Java et al., 2007; Li et al., 2011; Paul et al., 2011b), in the enterprise (Thom et al., 2011), and across cultures (Yang et al., 2011; Liu & Jansen, 2013). Morris et al. (2010a) found most questions posted to social network sites are subjective, seeking recommendations and opinions. Paul et al. (2011b) showed many are also rhetorical, with the asker expecting social answers rather than informative ones. The prevalence of subjective and rhetorical questions on social network sites has been a challenge for socially embedded search engines like SearchBuddies (Hecht et al., 2012), a Facebook agent that algorithmically suggests URLs in response to questions. Our crowd-powered system handles these nuanced scenarios because people are kept in the loop.

Identifying questions that merit replies is important for socially embedded search engines because extraneous answers can detract from the user experience (Hecht et al., 2012). Researchers have taken several approaches to question identification, employing heuristics (Efron and Winget, 2010; Hecht et al., 2012), machine learning (Li et al., 2011), and crowdsourcing (Paul et al., 2011a&b), all of which tend to incorrectly identify many rhetorical questions as questions. The failure of Paul et al.'s (2011a&b) crowdsourcing approach is particularly surprising, given crowd workers were asked whether or not tweets ending in a "?" were questions. Even ignoring the 68% of tweets that were judged to not be questions, 42% of the remaining tweets labeled as true questions by crowd workers were still subsequently coded by Paul et al. as rhetorical. One of the contributions of this paper is the introduction of improved heuristics and crowdsourcing pipelines for identifying answerable questions on Twitter.

Harnessing the labor of crowd workers on services such as Amazon's Mechanical Turk is an emerging mechanism for supporting information seeking. For example, pay-for-answer Q&A sites provide monetary incentives to their answerers (Li et al., 2013). The Tail Answers (Bernstein et al., 2012) system uses crowd sourcing to supplement search result lists with succinct natural language answers to queries. Twiage (Van Kleek et al., 2012) is a game-with-a-purpose for identifying high quality, naturally-occurring tweets that might be shown in response to subsequent users' information needs. Franklin et al. (2011) proposed handling complex database queries that request non-instantiated data by generating data from crowd workers. VizWiz (Bigham et al., 2010) is a Q&A tool for blind users that enables them to submit photographs to crowd workers and receive audio answers; some questions are factual, while others are subjective (Burton et al., 2012). Unlike these systems, where users explicitly send their questions to a crowd-powered search system, we monitor public question tweets to serendipitously provide crowd-generated answers.

Some systems make use of *friendsourcing*, meaning they use members of the user's own online social network as workers. For example, CrowdSearcher (Bozzon et al., 2012) enables users to explicitly federate search engine queries to subsets of their online social network, and SearchBuddies (Hecht et al., 2012) and Cognos (Ghosh et al., 2012) suggest contacts who might have the expertise to answer a question. Q&A tools like Aardvark (Horowitz and Kamvar, 2010) and IM-an-Expert (White et al., 2011) use expertise information to route questions to relevant members of a user's extended online social network. We analyze the properties of friendsourced answers on Twitter, studying the answers a question asker's network provides naturally and comparing them with crowdsourced content. We also collect crowd feedback on friendsourced answers to let the

question asker know how popular their friends' responses are, since prior work shows users feel more confident in an answer's quality when they receive confirmation of it from multiple sources (Morris et al., 2010b).

Nichols and Kang (2012) ask Twitter users to answer questions on topics like the wait times at airport security checkpoints. Our project does the inverse, sending answers rather than questions to selected Twitter users. Nichols and Kang's project showed that Twitter users can react favorably to unsolicited tweets if these tweets are carefully crafted and targeted. Similarly, Aiello et al. (2012) found that in a book-recommendation network, a bot delivering quality, non-random content became popular. These findings suggest that a system offering unsolicited answers to users on Twitter may be feasible.

Answers to Twitter Questions

To guide the construction of crowdsourced answers, we begin by looking at the answers people already receive on Twitter. Our goals are to successfully identify information-seeking tweets, understand the properties of the tweets and users that receive answers, and characterize those answers. The analysis is conducted over one month of data (starting May 9, 2012) collected from the Twitter Firehose, which is a complete stream of all public tweets. We restrict our analysis to English-language tweets.

Question and Answer Identification

Question identification on Twitter is hard (Paul et al., 2011b). To identify a candidate set of questions, we applied a number of filters that have been used by other researchers. We begin with tweets that end with a question mark (Efron and Winget, 2010; Paul et al., 2011b), and exclude retweets, mentions (@username) (Paul et al., 2011b), and tweets containing URLs (Li et al., 2011). Such filtering has high recall, in that it identifies a large fraction of information seeking tweets, but also has low precision, meaning a high proportion of the tweets identified are not actually questions (Paul et al., 2011b).

Because unnecessary answers to status updates that are not questions can be perceived very negatively (Hecht et al., 2012), question identification should target precision over recall. To reduce false positives we adopted an additional heuristic filter based on hashtags (i.e., phrases beginning with the "#" symbol, which are used by authors to label tweets). We limited our question tweets to those with a hashtag that indicated an information need, including: *#advise*, *#help*, *#ideas*, *#recommendation*, *#suggest*, *#lazyweb*, *#needanswer*, *#needopinions*, *#pleaserespond*, and any using these as prefixes (e.g., *#suggestplease*, *#helpme*). The use of these hashtags may indicate the asker is likely to be receptive to discovery of the tweet by non-network members.

	With Response	Without Response	Total
Askers	17,506 (25.3%)	51,671 (74.7%)	69,177
Tweets	29,995 (35.0%)	55,744 (65.0%)	85,739

Table 1. The proportion of question-asking users and question tweets in our dataset that received responses.

This approach identified 85,739 question tweets posted by 69,177 distinct users in one month. Examples include, *Ladies, do you like guys with beards or them to be clean shaved? #help* and *I want to dye my hair again. I just don't know what color. #suggestions?*

While the tweets identified with this approach are not necessarily representative of all Twitter questions, they allow us to compare questions that receive responses with those that do not. Responses to tweets made using Twitter's "reply" feature can be identified by finding tweets starting with @askername whose in_reply_status_id metadata field matches the unique identifier of the original tweet (replies not composed in this manner are not captured in our analysis). Table 1 shows the proportion of askers and tweets in our question dataset that received responses. About a third (35.0%) of the question tweets we identified received a response. However, only 25.3% of the users who asked questions received a response, meaning some users were more likely than others to get answers. The following sections explore the properties of tweets and of askers that were associated with receiving responses, and characterizes the content of those replies.

Response Patterns by Tweet

We examined the characteristics of the question tweets that did and did not receive responses. Because this comparison required qualitative labeling of the data, we randomly sampled 500 question tweets with responses and 500 question tweets without responses to label and analyze. We used crowd workers from the CrowdFlower service to label each tweet according to whether or not it was rhetorical, the topic of the tweet, the type of answer expected, and what (if any) additional contextual information about the asker would be helpful in crafting an answer.

Each labeling operation was redundantly completed by five crowd workers (three for the simpler rhetorical nature labeling), and a final label was chosen using CrowdFlower's voting formula which weights each worker's choice using a trust score based on past performance. Workers completed labeling tasks in batches so that they could gain skill at applying a particular label (e.g., classifying batches of 50 tweets by topic, or by answer type, or by rhetorical nature). Examples of tweets in each label category were shown in the instructions, and "gold standard" items (for which our team had verified the correct label) were em-

	Rhetorical	Not Rhetorical
With Response	33 (6.6%)	467 (93.4%)
Without Response	146 (29%)	354 (71%)

Table 2. Proportion of rhetorical and not rhetorical question tweets with and without responses.

Topic	Response	None	Example
Technology	20.6%	18.3%	*How do you log out of twitter on iphone? #help*
Personal	16.7%	18.3%	*To keep or not to keep my monroe piercing in??? #help*
Professional	14.1%	14.0%	*Trying to decide on colleges A&t, wssu, or nccu? #help*
Entertainment	13.3%	12.3%	*making a summer playlist, #suggestions ?*
Shopping	7.3%	3.7%	*I need a jean jacket SO BAD. Where do I find one? #help*
Health	6.2%	4.0%	*Does anyone know good sore throat remedies? #please #help*
Place	6.0%	5.2%	*Where can I park for free in Newport for work?! #help*
Restaurants & Food	4.9%	5.7%	*Is there a good Japanese restaurant around here? #helpaguyout*
Home & Family	1.5%	1.2%	*Can i put heels in the washing machine ? Lol :/ #help*
Other	9.4%	17.2%	*Does Ahmed #Shafik have a website? Would you please #help me get his presidential program?*

Table 3. Distribution of topics for the 467 question tweets with responses and the 354 question tweets without.

bedded among those to be labeled. Labels from workers who did not match our gold standard items were rejected. The combination of weighted voting based on past performance, detailed examples and instructions, batch-labeling to develop labeling skill, and gold standard tasks were used to address the inaccuracies in crowd-labeling of tweets encountered in prior work, e.g. (Paul et al., 2011b).

Rhetorical Questions: Table 2 shows the proportion of rhetorical (*Why am I up this early? #help*) and non-rhetorical (*Anyone know of a good remedy to get rid of cramps?! #Helpppppppp*) questions in each set. Rhetorical questions, which by definition do not invite an answer, are less likely to receive a response than non-rhetorical questions ($\chi^2(1, N = 500) = 123.5$, $p < .000$). Because rhetorical questions fundamentally differ from information-seeking questions, we removed these questions from subsequent labeling and analysis tasks.

Question Topic: We provided CrowdFlower workers with a list of question topics adapted from Morris et al.'s (2010a) survey of SNS Q&A practices, and had them select which topic best describes a given question tweet. Table 3 shows the distribution of topics and provides an example from each topic.

The topic distribution was significantly different between tweets with and without responses ($\chi^2(9, N = 467) = 42.2$, $p < .000$). Technology and personal (fashion, cosmetics, etc.) question topics were the most prevalent among questions both receiving and lacking responses. Questions falling into the category "other" (i.e., difficult-to-classify questions about unusual topics) were more commonly unanswered. Shopping-related questions, on the other hand, were more common among those receiving answers.

With Response		Without Response	
Type	**%**	**Type**	**%**
Personal opinion	32.1	Personal opinion	28.1
Provide fact	23.1	Provide fact	22.6
Explain how to	13.1	Social support	16.9
Recommend item	11.6	Explain how to	11.5
Business name or URL	11.4	Recommend item	8.9
Social support	4.9	Business name or URL	8.0
Definition	1.9	Personal anecdote	2.0
Personal anecdote	1.7	Definition	2.0

Table 4. Distribution of expected answer types for the 467 question tweets with responses and the 354 without.

Type	Example Question
Personal opinion	*About to get my nails done. What colour shall I get? #SuggestionsPlease*
Provide fact	*What name is given to patterns of DNA? #helpme*
Explain how to	*How do you see tweets that you've favourited?? #helpme*
Recommend item	*Who knows an app. that blocks incoming texts? #PleaseRespond*
Business name or URL	*Any suggestions on decent coffee shop in Birmingham with wi-fi for a meeting? #lazyweb*
Social support	*who wants to drink tonight? this day is driving me nuts and I only work one shift for the day #help*
Definition	*What does smh mean? #help #clueless*
Personal anecdote	*Has anyone done any sort of road trip from LA though Utah/colarado/down to Texas before? I'm planning a trip but need some advice! #help*

Table 5. Example question tweets having each expected answer type, as categorized by crowd workers.

Expected Answer Type: We also examined how the type of answer expected by the question related to whether or not the question received a response. We adapted the question type schemes from prior studies of SNS Q&A (Morris et al., 2010a; Paul et al,. 2011b) and asked CrowdFlower workers to choose which answer types the asker seemed to want in response to their tweet. For example, the question in Figure 1 solicits an opinion, while the one in Figure 2 requests "how to" instructions. Table 4 shows the classification results and Table 5 an example for each type.

The type of answer expected was significantly different for tweets with and without responses ($\chi^2(7, N = 467) = 54.4$, $p < .000$). Questions seeking opinions were the most common in each set, followed by those requesting facts. However, questions that seemed to expect social support (rather than information) commonly went unanswered.

	Context Required	Context Not Required
With Response	88 (18.84%)	379 (81.16%)
Without Response	88 (25.14%)	262 (74.86%)

Table 6. Proportion of question tweets requiring additional context beyond the tweet itself to formulate a response.

Context: To better understand the requirements for generating answers to question tweets, we also asked CrowdFlower workers to indicate whether or not additional context beyond the text of the tweet itself was required to create an answer. We provided workers with a checklist of the types of context that we could conceivably extract from a user's Twitter profile (username, short bio, location, photo, and recent tweets), and asked them to mark which of those items might be relevant to answering the question.

As can be seen in Table 6, questions requiring context beyond the tweet itself were less likely to receive responses (($\chi^2(1, N = 467) = 9.6$, $p = .002$)). The most commonly required types of context were location (58.5% of those needing context, e.g., *Good bars to go to on a Tuesday night? #veg #help #birthday*) and photo (29.5% of those needing context, e.g., *Ahhh do I keep on growing my hair or get it cut?!?! #help*).

Response Patterns by Asker

In addition to looking at which questions received responses, we also looked at which users received responses. To do this, we analyzed four user properties, including:

- Follower count (how many people follow the user),
- Friend count (the number of people the user follows),
- Tweet count (the number of tweets posted by the user),
- Favorite count (the number of tweets the user has starred as being among their favorite).

We hypothesized that a higher follower count would lead to more responses, since larger networks have been shown to increase answer likelihood in other networks (Morris et al., 2010b; White et al., 2011; Liu & Jansen, 2013). Additionally, high friend, tweet, or favorite counts may indicate an active Twitter user who has built up social capital and thus may be more likely to receive responses per norms of reciprocity (Uehara, 1995; Yang et al., 2011). Because these properties were not normally distributed, we used non-parametric statistical tests in our analysis.

First, we measured the median values of each property for users who did and did not receive responses to their questions. Mann-Whitney *U* tests were conducted to see whether there were significant differences between users who received answers versus those who did not. Users were significantly more likely to get answers if they favorited tweets more often ($z = -15.6$, $p < .000$), posted more tweets ($z = -30.6$, $p < .000$), had more followers ($z = -45.7$, $p < .000$), or had more friends ($z = -27.4$, $p < .000$).

For the 17,506 users who received a response to their question, we further examined whether the number of responses was affected by the user properties. The number of answers was weakly correlated with the properties, with the number of followers representing the most substantial correlation. Users received more answers if they favorited tweets more ($r = .05$), posted tweets more ($r = .13$), had more followers ($r = .22$), or had more friends ($r = .13$). All correlations (Spearman's rho) are significant ($p < .000$). For those users receiving a response, the median response time was just under 8 minutes (471.5 seconds).

Type	Example Answer
Personal opinion	*TANGERINE of course ma love x* (F)
Provide fact	*no existing iPhones have 4G capability, only the newest iPads.* (F)
Business name or URL	*I've had good luck w/ craigslist* (F)
Recommend item	*Logitech Bedside Dock for iPad* (C)
Explain how to	*go to the settings, there's a twitter icon and you can delete your account from there* (F)
Clarifying questions	*What sites are blocked? Have you tried torrenting?* (F)
Humor	*Some aliens try to contact you, lol.* (C)
Do not know	*Sorry, I can't help you* (C)
Suggesting others	*Did you ask @ANONYMIZED ? He is the single, most effective food consultant bangalore tweep*. (F)
Search by yourself	*You can google it.* (C)
Personal anecdote	*YES! They are great, but saying that i had them when i was young, so might be more of hinderance when your older* (F)
Definition	*H is for Hawas Mode. HSDPA - High Speed Data Packet Access. :-)* (F)
Also requesting answers	*Let me know if you get one.* (F)
Social support	*You deserve a little luck. Go for it and have fun!* (C)
Talking to oneself	*wish I had time to bake* (F)
Advertisement	*we specialise in move in/move out cleans! You can contact our Perth office on (08) 93645488* (F)
Other	"*This is an inappropriate question to answer.*" (C)

Table 7. Examples tweets from friends (F) or the crowd (C) of each answer type, as classified by crowd workers.

Characterization of Answers

We also looked at the content of the 1,298 answers that the 467 answered questions received. We observe additional types of answers beyond those anticipated based on our initial analysis of the questions and previous work (Morris et al. 2010a). For this reason, we expanded our taxonomy of answer types from those derived from analysis of questions (Table 5) to include those shown, with examples, in Table 7. These include answers that suggest approaches for finding an answer (such as to search on one's own or ask someone in particular), request additional information

Type	Friendsourced	Crowdsourced
Personal opinion	518 (39.9%)	593 (42.1%)
Provide fact	168 (12.9%)	165 (11.7%)
Business name or URL	154 (11.9%)	165 (11.7%)
Recommend item	105 (8.1%)	88 (6.3%)
Explain how to	81 (6.2%)	59 (4.2%)
Clarifying questions	70 (5.4%)	62 (4.4%)
Humor	62 (4.8%)	44 (3.1%)
Do not know	0 (0.0%)	93 (6.6%)
Suggesting others	12 (0.9%)	38 (2.7%)
Search by yourself	3 (0.2%)	44 (3.1%)
Personal anecdote	40 (3.1%)	6 (0.4%)
Definition	21 (1.6%)	17 (1.2%)
Also requesting answers	25 (1.9%)	8 (0.6%)
Social support	19 (1.5%)	13 (0.9%)
Talking to oneself	6 (0.5%)	0 (0.0%)
Advertisement	3 (0.2%)	0 (0.0%)
Other	11 (0.8%)	12 (0.9%)

Table 8. Distribution of answer types for answers generated naturally and by our crowdsourcing pipeline.

(clarification questions), or benefit the respondent (advertisements or requests for the answer as well). As before, CrowdFlower workers were shown the answer tweet and asked to select the answer type that best described it, with gold-standard tasks to eliminate poor labelers. Five workers labeled the same item, and answers were combined using CrowdFlower's trust-weighted voting formula. The results of this labeling can be found in Table 8. People were much more likely to provide personal opinions (39.9% v. 32.1%) than expected from analysis of the questions, and much less likely to provide explanations (1.6% v. 13.1%) or facts (12.9% v. 23.1%)

Three crowd workers also labeled the "information source" of each response, meaning whether the answer could be found in the body of the tweet itself, or if it required visiting an external URL. We also included an option for responses without answers (e.g., that were purely social). We anticipated that due to the short format of tweets, answers would often refer to external URLs. However, over 70% of the tweets contained the answer completely within the text of the tweet.

Crowdsourced MSR Answers

Through this analysis we found commonalities and differences among the questions and askers that received responses and that did not. We saw that most questions were not answered, indicating a service that can automatically provide answers could be useful. This might particularly benefit users with small networks or who are less active on Twitter, as these users were particularly unlikely to receive responses. We also saw that subjective answers were common, and might benefit from human-based answers rather than those generated algorithmically as in prior work (Hecht et al., 2012). The fact that most questions did not require additional context beyond the tweet itself suggests that crowd workers might be able to generate answers.

Our MSR Answers system uses these findings to generate crowdsourced answers to Twitter questions in three stages. In Stage 1, non-rhetorical question tweets are identified from the Twitter feed; In Stage 2, crowd workers compute a response to the question; And in Stage 3, the crowd- and friendsourced answers are voted on by crowd workers to select the best answer.

Stage 1: Question Identification

The MSR Answers system identifies question tweets by applying the question filters described earlier to the Twitter stream. Specifically, tweets must end in *?*, contain no retweets, mentions, or URLs, and including certain hashtags. When a question is identified, it is sent to CrowdFlower to be labeled as rhetorical or not. Only those that are labeled as not rhetorical are answered. Although some questions may be easier to answer than others, we do not filter on any other attributes (e.g., question difficulty).

The crowd-based labeling scheme is modified somewhat so as to be suitable for operating on a single tweet at a time, rather than on a large batch. In the modified scheme, each task included five items to label, four of which are gold-standard tweets whose rhetorical nature we already know. A worker needs to label at least three of the four gold-standard items correctly for their rating of the target tweet to be accepted. This heavy use of gold standard items allowed us to use only a single worker's labeling judgments, at a cost of 5 cents per tweet.

Stage 2: Answer Generation

To generate answers, crowd workers are given the question tweet and the publically available components of the asker's Twitter profile (username, photo, location, and bio), and asked to compose a response to the question. The instructions indicate that answers should not exceed 125 characters. Although Twitter has a longer (140) character limit, this allows room to prepend @username and other introductory text to the answer.

For quality control, we include a set of simple predefined question and answer pairs as a gold standard. An example of a gold standard question is, "How often is the FIFA world cup held? #helpsuggest" The gold standard answer is defined by a regular expression. For the example question above, any answer containing "4" or "four" is considered correct. Therefore, a crowd worker can pass the quality control stage if they enter an answer containing "4," "4 years," or "four years." If a crowd worker correctly answers the gold standard tweets, and their answer to the

target tweet are within the character limit, we accept their response as a candidate answer. Once we receive three candidate answers for a tweet, we stop requesting answers. We refer to these as the *crowdsourced answers*. The cost of generating three crowd answers for a tweet is 30 cents.

Stage 3: Answer Selection

In parallel, we monitor the public Twitter feed to see if the asker receives any natural, friendsourced responses. To provide question askers with the best possible information and feedback on their friends' responses, crowd- and friendsourced answers are voted on by crowd workers to select the best answer. Ten crowd workers vote on which of the potential answers best answers the original question, at a cost of 6 cents per vote. For quality control, we include the gold standard tasks of re-categorizing the tweets from our earlier, offline analysis.

Based on the voting outcome, our system then posts a public tweet as a reply to the asking user. If no friendsourced answers are received, the agent posts the highest rated crowdsourced answer; otherwise, the agent posts two tweets, one about the most popular crowdsourced answer and one about the most popular friendsourced answer, according to the scheme depicted in Table 9. In total, the three-stage process cost is $0.95 to generate the final crowdsourced answer to a question tweet.

Case	Answer Template
Only crowd-answers available	[BEST CROWD ANSWER TEXT]
Crowd- and friend-answers available, Crowd-answer voted best	Tweet 1: *X out of Y ppl we asked say:* [BEST CROWD ANSWER TEXT] Tweet 2: *X out of Y people we asked liked @FRIEND's answer!*
Crowd-and friend-answers available, Friend-answer voted best	Tweet 1: *X out of Y people we asked liked @FRIEND's answer best!* Tweet 2: *X out of Y ppl we asked say:* [BEST CROWD ANSWER TEXT]

Table 9. How the three possible outcomes of crowd workers' voting for answers translate to the final answer tweet(s).

Comparing Crowd and Friend Answers

We begin by evaluating the quality of the answers the crowd generates in response to people's public Twitter questions using this process. To do this, we created three crowdsourced answers (the output of Stage 2) for the 467 non-rhetorical questions that received responses discussed earlier in the "Response Patterns by Tweet" section. We compared these 1,407 crowdsourced answers with the 1,298 naturally occurring answers. Table 7 shows examples of both answer types. The answers were sent to workers on CrowdFlower for labeling regarding answer quality, answer type, and information source, as described below.

Answer Quality

We initially tried asking the users who originally tweeted the questions to assess quality by sending them a tweet inviting them to take a short survey containing both crowd-generated and naturally occurring answers. However, 20 such invitations yielded no responses and two users blocked us as spammers, so we turned to the crowd instead. The use of paid judges to evaluate result quality is a common practice in information retrieval research, with the crowd a viable alternative to in-house judges (Alonso & Mizzaro, 2012). Five crowd workers rated each answer on the 6-point Likert scale depicted in Table 10. Raters were not told the origin of the answer, but could view the question tweet. The inter-rater reliability among judges was a Fleiss' Kappa of .41, indicating moderate agreement.

Score	Example
0: Spam or advertisement	Q: *Will anyone please help me make a video on my computer and come like now? #help!* A: *Need more folks to like your video?! http://t.co/6FpPIGgv*
1: Not related	Q: *Getting an iPhone this weekend :D #finally switch to Fido or stay with Telus?!! #decisions #help* A: *the future is friendly :) lol*
2: Related but not helpful	Q: *How long is it likely to take me to get from the Tate Modern (ish) to Finsbury Square (ish) in a cab at lunchtime today? #lazyweb* A: *More than a minute and less than a year. Depends on traffic.*
3: Related, but cannot trust the answer source	Q: *can anyone recommend a good website hosting service? #HELP* A: *here you are a list: http://www.consume-rankings.com/hosting/*
4: Related, good answer	Q: *DOES ANYONE KNOW HOW TO GET RID OF SEA LICE?! #help #dyingoverhere* A: *Clean the affected areas with vineger.*
5: Related, high quality answer	Q: *#GenuineKoschan - For mobile internet, there's a G - GPRS. E - Edge. 3G - 3G. But, I also get an H. What is that? #help* A: *H stands for HSDPA. HSPDA is the good one. This one supports up to 14.4MB/s.*

Table 10. Likert scale used to rate answer quality, and example answers receiving each rating. A "high quality" answer is distinguished from a merely "good" one by the presence of additional detail beyond the minimum sought by the asker.

On average, both friendsourced and crowdsourced answers received a score of 3.6. A Mann-Whitney *U* test found that there was no significant difference between the quality of these answer types ($z = -1.6$, $p = .11$).

Answer Preference

In addition to having workers individually rate answer quality, we also collected information about people's preferred answer to each question. Judges were shown the original question tweet and the set of crowdsourced and friendsourced answers (blind to the answers' origins), and were asked to select the most useful answer.

Crowd workers might find the answers that other crowd workers write more appealing than a different rating de-

mographic might. For this reason, rather than use crowd workers to compare answers, we recruited 29 university students (aged 18 to 31, 41% female). Due to the time-consuming nature of doing these ratings, we collected them for 247 (about half) of the question tweets, sampled at random. Each participant completed the task for 30 to 60 questions, and each of the 247 questions was rated by 4 or 5 different participants, for 1165 ratings total.

In 66.2% of the cases, a crowdsourced answer was selected as the preferred answer. If our null hypotheses is that raters would show no preference for any answer type (which would lead to choosing a crowd-generated answer as the preferred one 51.8% of the time, their prevalence in the data set), we can see that this hypothesis is disproven, with crowdsourced answers being preferred significantly more than chance ($\chi^2(1, N = 1165) = 97.0$, $p < .000$).

Answer Content

As was done with the friendsourced answers, CrowdFlower workers were also asked to determine the type of content contained in the crowdsourced answer, with gold-standard tasks to eliminate poor labelers. Five workers labeled the same item, and answers were combined using CrowdFlower's trust-weighted voting formula. Table 7 shows examples from this labeling of the different answer types.

Table 8 shows the distribution of answer types for both crowdsourced and friendsourced answers. Although the two distributions were significantly different ($\chi^2(14, N = 1407) = 9087.2$, $p < .000$), the most common types (opinion, fact, business name, recommendation) were similar in prevalence for both crowd- and friend-answers. Answer types indicative of low effort (such as suggestions for the question asker to search for the answer) were more common in the crowd-generated answers, but comprised a relatively small percent of answers overall. Additionally, 93 (6.6%) responses stated the crowd worker did not know the answer. This could be addressed by allowing workers to opt out of answering questions or filtering for expertise (e.g., White et al. 2011).

We also observed that crowd workers were twice as likely to only provide a pointer to an external link (7% of answers) as friends were (3%). This difference was significant ($\chi^2(4, N = 1407) = 71.2$, $p < .000$).

Deployment of MSR Answers

While there are some differences between friendsourced and crowdsourced answers, in general the naturally occurring responses to question tweets were largely similar in nature and quality to those produced through the crowdsourcing pipeline. However, although crowd-answers are of reasonable quality it is possible that friendsourced answers have value not readily identifiable by third-party raters (e.g., increasing social ties between asker and answerer). For this reason, we deployed MSR Answers and studied its use in real world situations.

Deployment Details

MSR Answers' user interface consisted of a public Twitter account, whose user image was a logo with our research projects' name. The account's profile said, "Answering public Twitter questions since 2012! This account is part of a <anonymized institution> project." The profile included a URL to a web page explaining the privacy and legal policies associated with the project. Below we describe how the system worked, and how users responded to it.

The system was set up to monitor the live public Twitter feed using Twitter's API over a ten day period in August 2012 to gather initial data about end-user reactions. Figures 1 and 2 show example interactions between the system and Twitter users. The system answered 100 questions during this time period. Questions were identified using the heuristics described earlier, with no additional filtering other than the rhetorical question filtering that occurs in stage one of our crowdsourcing pipeline. Among these questions, questions expecting opinions or recommendations accounted for 66%, while those seeking factual information (e.g., how-to, definition, etc.) accounted for 34%. From the perspective of topic, "personal" questions (29%) were the most prevalent, followed by technology (19%), shopping (16%), and entertainment (14%).

The median time to produce an answer was 162.5 minutes (mean 188.8) from the time a tweet was posted. The median time of each phase was 3.02 minutes for Stage 1, 77.4 minutes for Stage 2, and 82.1 minutes for Stage 3. The latency could be reduced with tactics such as those employed by quickturkit (Bigham et al., 2010), or by altering the payment structure.

User Response

Although our system's tweets only answered questions and did not request a response, 33% of the 100 users who were answered responded to MSR Answers, almost all positively. Eleven users replied to the tweet with a message thanking the system for the answer, seven users retweeted the response to their followers, seven users marked the response as "favorite" tweet, four users replied indicating that they found the response humorous, three users replied with follow-up questions (see Figure 2), and one user began following the MSR Answers account. The total response count is more than 33 because some users responded in multiple ways (e.g., both thanking and retweeting).

Answers for questions asking personal opinions regarding fashion were the ones that received the most positive acknowledgments. This provides further evidence that subjective responses can be appreciated even when coming

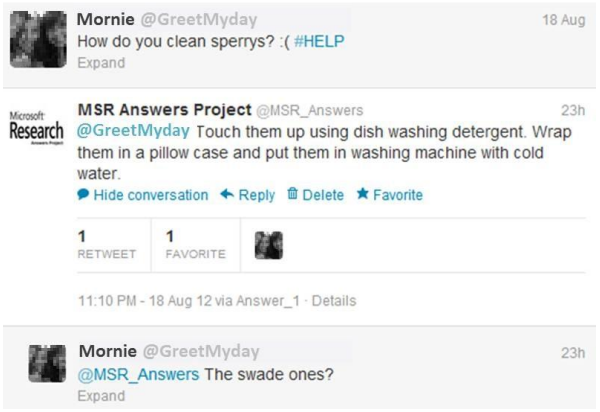

Figure 2. This user provides unsolicited feedback after receiving our system's response, including marking the tweet as a favorite, retweeting it, and replying.

from strangers. The answers offering business recommendations to shopping questions and "how-to" instructions to technology questions were the second and third most appreciated ones. In general, answers containing more concrete information tended to receive more thanks.

We also received two replies indicating that the tweet was embarrassing and two expressing disagreement with the system's answer. These negative responses appeared to result from a mismatch between crowd workers and question askers. For example, one question expecting a nice dinner date venue for adults received a response suggesting a restaurant more suitable for teens.

The unsolicited, largely positive feedback suggests users valued the socially embedded search service provided by MSR Answers, and that the crowd-answers were of good quality. The dearth of negative responses is also noteworthy. In particular, we were surprised that none of the 100 people we answered blocked the MSR Answers account or reported it as a spam account, despite the fact that it was not a member of their network and sent unsolicited messages. Blockage has been a problem affecting other Twitter projects that contact strangers (Nichols and Kang, 2012).

Using a separate Twitter account we sent a link to a follow-up survey to the 100 users whose questions the system answered. We sent the survey links one week after the initial interaction to avoid interfering with spontaneous, unsolicited reactions. Only four users completed the follow-up survey. Those four indicated that they found the system's answer to be helpful, trustworthy, and understandable. They were all surprised to receive our answer, but indicated that it was not "creepy" and did not look like spam.

Conclusions and Future Work

This paper described the motivation, design, and deployment of MSR Answers, a novel user experience in which question askers receive unsolicited, crowd-generated replies to public Twitter questions. MSR Answers is motivated by the result of a study of naturally occurring Twitter answers. We analyzed the differences between the questions that did and did not receive answers, and found that most questions went unanswered, with less active users and those with smaller networks receiving the fewest replies. These results suggest there is an opportunity to improve the question asking experience through automated question answering. We presented a crowdsourcing pipeline to identify and answer Twitter questions, taking advantage of human computation to create answers that are, like naturally occurring answers, often subjective in nature. A multi-pronged evaluation of the generated answers revealed that they were largely similar in style and quality to the existing answers from friends. Although crowdsourced answers may lack some of the difficult-to-measure social benefits of friendsourced answers, when we deployed MSR Answers nearly one third of the 100 users who received responses provided unsolicited positive feedback.

The results of our study of a working crowd-powered socially embedded search engine are promising, but there is much that can be done to continue to understand and support SNS question asking. One area for future investigation is which users or questions would most benefit from automated answers. Crowd-powered socially embedded search engines seeking to balance budgetary constraints with impact might choose to prioritize answering questions whose topics or types receive fewer answers, or answering users whose profiles suggest a low probability of receiving a reply. There are also many nuanced reasons why a crowd-powered approach may or may not be appropriate for a particular question or user. A recent study by Brady et al. (2013) found that blind adults prefer to have their information needs addressed by paid crowd workers rather than their online social networks; it is possible the opposite is true for other specific demographic sub-groups or for particular question genres.

The scarcity of negative feedback suggests the interaction design is generally socially conformant (Hecht et al., 2012). Even better social conformance could be achieved by identifying sensitive topics (such as health, religion, finance, sex (Morris et al., 2010a)) in the initial labeling stage, and not responding to such questions. MSR Answers currently only creates answers to publicly posted questions. Extending the system to address network-only posts (e.g., most posts on Facebook) would require further study into privacy-preserving use of crowd workers.

The answer-generation process used by MSR Answers could be improved to meet additional constraints. For example, cheaper answers could be produced by using fewer workers, and better answers could be produced by providing workers with additional context. Although MSR Answers already provides the crowd workers with public profile information about the question asker, workers could be

given additional topical or domain information. Such information could be generated algorithmically or gathered during the initial question labeling stage. Topic labels could be used to recruit domain experts, or expected answer type labels could be used to create custom answer-generation instructions.

Question askers may also want additional context to better understand the crowdsourced replies they receive. We observed that the few negative reactions to MSR Answers appeared to arise from a mismatch between the person asking the question and the people generating the answer. Demographic information of the crowd workers that contributed to the answers could help people interpret the value of their advice. For example, an asker may like to know whether the "8 of 10 people" voting for coloring one's hair red are men or women. Matching characteristics of the asker and potential answerers may also lead to more appropriately tailored responses.

Question askers' naturally occurring feedback could be used to further optimize the system. For example, users' attempts to ask follow-up questions by tweeting directly at the system suggests that it might be beneficial to respond to direct questions in addition to serendipitously answering public inquiries. It is likely that, should socially embedded search engines become commonplace, user interaction styles and reactions will evolve over time.

References

Adamic, L.Z., Zhang, J., Bakshy, E. & Ackerman, M.S. Everyone knows something: Examining knowledge sharing on Yahoo Answers. *WWW 2008*.

Aiello, L.M., Deplano, M., Schifanella, R. & Ruffo, G. People are strange when you're a stranger: Impact an influence of bots on social networks. *ICWSM 2012*.

Alonso, O. and Mizzaro, S. Using crowd sourcing for TREC relevance assessments. *IP&M*, 18, 2012, 1053-1066.

Bernstein, M.S., Teevan, J., Dumais, S., Liebling, D. & Horvitz, E. Direct answers for search queries in the long tail. *CHI 2012*.

Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S. & Yeh, T. VizWiz: Nearly real-time answers to visual questions. *UIST 2010*.

Bing Team. Introducing the new Bing: Spend less time searching, more time doing. *Bing Search Blog*, May 10, 2012.

Bozzon, A., Brambilla, M. & Ceri, S. Answering search queries with CrowdSearcher. *WWW 2012*.

Brady, E., Zhong, Y., Morris, M.R. & Bigham, J. Investigating the appropriateness of social network question asking as a resource for blind users. *CSCW 2013*.

Burton, M.A., Brady, E., Brewer, R., Neylan, C., Bigham, J.P. & Hurst, A. Crowdsourcing subjective fashion advice using VizWiz: Challenges and opportunities. *ASSETS 2012*.

Efron, M. & Winget, M. Questions are content: A Taxonomy of Questions in a Microblogging Environment. *ASIS&T 2010*.

Franklin, M.J., Kossman, D., Kraska, T., Ramesh, S. & Xin, R. CrowdDB: Answering queries with crowdsourcing. *SIGMOD 2011*.

Ghosh, S., Ganguly, N., Sharma, N., Benevuneto, F. & Gummadi, K. Cognos: Crowdsourcing search for topic experts in microblogs. *SIGIR 2012*.

Hecht, B., Teevan, J., Morris, M.R. & Liebling, D. SearchBuddies: Bringing search engines into the conversation. *ICWSM 2012*.

Horowitz, D. and Kamvar, S.D. The anatomy of a large-scale social search engine. *WWW 2010*.

Java, A., Song, X., Finin, T. & Tseng, B. Why we Twitter: Understanding microblogging usage and communities. *KDD 2007*, workshop on web mining and social network analysis.

Lampe, C., Vitak, J., Gray, R. & Ellison, N. Perceptions of Facebook's value as an information source. *CHI 2012*.

Li, U., Kim, J., Yi, E., Sung, J. & Gerla, M. Analyzing Answerers in Mobile Pay-for-Answer Q&A. *CHI 2013*.

Li, B., Si, X., Lyu, M.R., King, I. & Chang, E.Y. Question identification on Twitter. *CIKM 2011*.

Liu, Q., Agichtein,E., Dror, G., Gabrilovich, E., Maarek, Y., Pelleg, D. & Spektor, I. Predicting web searcher satisfaction with existing community-based answers. *SIGIR 2011*.

Liu, Z. and Jansen, B.J. Analysis of Factors Influencing the Response Rate in Social Q&A Behavior. *CSCW 2013*.

Morris, M.R., Teevan, J. & Panovich, K. 2010a. What do people ask their social networks, and why? A survey of status message Q&A behavior. *CHI 2010*.

Morris, M.R., Teevan, J. & Panovich, K. 2010b. A comparison of information seeking using search engines and social networks. *ICWSM 2010*.

Nichols, J. & Kang, J-H. Asking questions of targeted strangers on social networks. *CSCW 2012*.

Panovich, K., Miller, R. & Karger, D. Tie strength in question & answer on social network sites. *CSCW 2012*.

Paul, S.A., Hong, L. & Chi, E.H. 2011a. What is a question? Crowdsourcing tweet categorization. *CHI 2011*, workshop on crowdsourcing and human computation.

Paul, S.A., Hong, L. & Chi, E.H. 2011b. Is Twitter a good place for asking questions? A characterization study. *ICWSM 2011*.

Teevan, J., Morris, M.R. & Panovich, K. Factors affecting response quantity, quality, and speed for questions asked via social network status messages. *ICWSM 2011*.

Thom, J., Helsley, S.Y., Matthews, T.L., Daly, E.M. & Millen, D.R. What are you working on? Status message Q&A within an enterprise SNS. *ECSCW 2011*.

Uehara, E. Reciprocity reconsidered: Gouldner's "moral norm of reciprocity" and social support. *Journal of Social and Personal Relationships*, 12(4), 1995, 483-502.

Van Kleek, M., Smith, D.A., Stranders, R. & Schraefel, m.c. Twiage: A game for finding good advice on Twitter. *CHI 2012*.

White, R.W., Richardson, M. & Liu, Y. Effects of community size and contact rate in synchronous social Q&A. *CHI 2011*.

Yang, J., Morris, M.R., Teevan, J., Adamic, L. & Ackerman, M. Culture matters: A survey study of social Q&A behavior. *ICWSM 2011*.