Perceptions of Answer Quality in an Online Technical Question and Answer Forum

Kerry Hart and Anita Sarma
Computer Science and Engineering Department
University of Nebraska-Lincoln, Lincoln, NE, USA
{khart, asarma}@cse.unl.edu

ABSTRACT

Software developers are used to seeking information from authoritative texts, such as a technical manuals, or from experts with whom they are familiar. Increasingly, developers seek information in online question and answer forums, where the quality of the information is variable. To a novice, it may be challenging to filter good information from bad. Stack Overflow is a Q&A forum that introduces a social reputation element: users rate the quality of posted answers, and answerers can accrue points and rewards for writing answers that are rated highly by their peers. A user that consistently authors good answers will develop a good 'reputation' as recorded by these points. While this system was designed with the intent to incentivize high-quality answers, it has been suggested that information seekers—and particularly technical novices—may rely on the social reputation of the answerer as a proxy for answer quality. In this paper, we investigate the role that this social factor—as well as other answer characteristics—plays in the information filtering process of technical novices in the context of Stack Overflow. The results of our survey conducted on Amazon.com's Mechanical Turk indicate that technical novices assess information quality based on the intrinsic qualities of the answer, such as presentation and content, suggesting that novices are wary to rely on social cues in the Q&A context.

Categories and Subject Descriptors

D.2.0 [**Software Engineering**]: General

General Terms

Human Factors

Keywords

Q&A forum, StackOverflow, reputation

1. INTRODUCTION

Software development is a social activity. In a traditional development environment, social interactions occur within the well-defined context of a team, where interactions are repeated and trust can be earned. However, developers are increasingly likely to leverage external social resources to solve problems. For example, the Question and Answer forum StackOverflow.com allows a developer to query strangers for solutions to a problem or, on the other hand, to seek out pre-answered queries that match the problem at hand.

The popularity of Stack Overflow attests to its efficacy: as of 2013, the site boasts 2.3 million registered users who have provided over 10 million answers to 7.8 million questions [1]. It is an essential source for not only the experienced programmer, but also the technical novice—the person who, while perhaps having a strong technical background, is seeking information on a topic with which he or she is not familiar. These novices often arrive at a particular StackOverflow.com question/answer pair through an external website, such as Google, and so find themselves in a technical forum with which they are not necessarily familiar.

We are interested in how a technical novice finds quality information in such a forum. After all, he or she faces a social problem: is a given answer trust-worthy? The asker could try out each given solution one at a time, testing each for correctness. However, such a process may be time-consuming. Instead, a question asker might use available social cues about the "expert" who provided the answer to prioritize or filter answers, focusing his or her attention on a subset of answers.

The social reputation scheme in Stack Overflow can be appropriated to support this filtering process. To incentivize contributions, questions perceived to be 'good' are up-voted by the community, awarding points to the question asker. The same is true for answers, awarding answerers who author 'good' responses. Furthermore, the earned points are associated with concrete benefits. Points unlock privileges such as the authority to edit posts and moderate responses. For certain achievements (e.g., authoring an answer that receives more than 25 up votes), special 'badges' (bronze, silver, gold) can be earned. These are reported next to the user's point score.

This element not only serves to incentivize contributions, but also acts as a measure of one's usefulness to the community (the more questions you have answered, edited, moderated, etc.—and the higher that these contributions are rated—the more likely you are to be helpful to your peers). Because the answerer's user reputation is publicly displayed alongside their posts, it may serve as a signal to someone trying to evaluate the quality of the answer. It is intuitive to assume that an answer posted by someone with a track record of quality contributions should, presumably, be more reliable than one posted by someone with a limited track record.

In this paper, we aim to understand the role of social cues in how novice users filter and select answers on Stack Overflow. Since a key element in Stack Overflow is the reputation points accrued to its users, it is reasonable to hypothesize that social reputation may play an important role in how information provided by users is perceived in the forum. We therefore ask:

RQ1: To what degree does social reputation affect perception of answer 'quality'?

Second, when investigating how novice users judge quality we also consider question length. It is reasonable to hypothesize that longer answers may be more comprehensive, and so would be considered to be of higher quality; however, it is equally reasonable to suspect that longer answers fatigue the reader and, therefore, might be considered to be of lower quality. To better understand the relationship

CHASE'14, June 2 – June 3, 2014, Hyderabad, India
Copyright 2014 ACM 978-1-4503-2860-9/14/06...$15.00
http://dx.doi.org/10.1145/2593702.2593703

between answer length and perceived quality by the novice user, we ask a second research question:

RQ2: To what degree does answer length affect the perception of answer 'quality'?

The answers to these questions can be leveraged to improve the quality of discourse on Stack Overflow—and, by extension, the efficacy of the information-sharing process on technical forums in general. Furthermore, we aim to better understand the role of social signals in quality perception. Past studies have observed that users in online communities are acutely aware of their social following and curate their activities in these forums accordingly [2]. Similarly, studies have shown that users often like to follow the activities of experts to learn from them [3].

We performed a survey on a sample set of Java related questions and their answers through the Amazon Mechanical Turk. The goal was to understand the importance of social factors to a technical forum. Second, we attempt to better understand what additional factors contribute to how readers perceive answer quality. Our results indicate that social reputation did not play a significant role in how novice users judged the quality of an answer. Instead it was the presentation styles (completeness and conciseness) that appeared to be the deciding factor in their choice.

2. BACKGROUND

A considerable body of research devoted to the study of StackOverflow.com has developed recently to investigate such diverse topics as how the 'gaming' elements incentivize contributions [4], exploring the 'activeness' of contributors [5], and the automated identification of topical 'experts' [6] [7]. At the present, however, the subject of how users (and novices in particular) perceive answer quality has received little attention. At least one study has sought to better understand why certain questions elicit highly-rated answers and others do not, but it did not investigate the determinants of user-perceived quality [8].

Studies have shown that users with high social status tended to be more active members of the forum. This indicates that users who have achieved high status and are at the core of the community have done so on the merits of their quality of contributions. Research has been conducted as to how reputation can be earned and how high-reputation members provide largest contributions (Bosu, Corley, Heaton, Chatterji, Carver, & Kraft), but the precept that posts of reputed contributors are trusted more by the community remains untested. The role that social reputation can play as an information filtering mechanism has remained underexplored in the context of technical Q&A forums. We attempt to further our understanding of this aspect by investigating how users perceive answer quality.

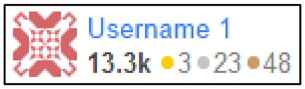
3. STUDY DESIGN

We administered a survey through Amazon's Mechanical Turk engine. The survey format was well suited for the task—rating the quality of Stack Overflow answers—and allowed us to reach a large number of respondents. Participants were screened to verify technical competency with a simple Java proficiency test that we created. We chose to use MechTurk to leverage its large pool of technical novices willing to answer the survey, which may reflect the sort of technical novices seeking out information on Stack Overflow or through web searches.

This study uses a mixed-methods methodology. First, we quantify the effect of the factors (social rating and answer length, in particular) on perceived answer quality. Second, we consider participant comments to better understand the quality perceptions of the survey takers.

The dependent variable is a survey-derived measure of perceived answer quality. Survey respondents scored sample answers based on a Likert scale, which ranged from 1-5 (where 1 is 'very low' quality and 5 is 'very high' quality). We made no effort to define 'quality' to the survey respondent. We did not wish to bias the results towards particular traits that we, the researchers, were predisposed to consider important.

We considered two independent variables: social reputation of the question answerer and the length of the answer provided. First, the answerer may have a *high* or a *low* social reputation. These reputations (*high* and *low*) are given in the standard Stack Overflow social reputation format: a number for earned points (earned from peers voting up the user's questions, answers, and edits) and three 'medal' counts reflecting specific achievements (these achievements may or may not be reflected in the score, depending upon the achievement). An example is given in Figure 1.



Username 1

13.3k ●3 ●23 ●48

Figure 1. Example user reputation, as displayed below a posted answer.

We simulated the user reputation values. To create the "high" reputation values in our data, we started with the reputation of the top StackOverflow contributor for the quarter (140k) and added noise. For the "low" reputations, we added noise to a representative low value (20k). This procedure was repeated for all four numbers (rating and gold, silver, and bronze medals). Each answer, therefore, is associated with a fictitious answerer of a unique "high" or "low" reputation. This procedure makes the ratings appear more 'natural' and help prevent the survey respondent from discovering the precise variable of interest (which may skew results). If users consider reputation to be an important factor in answer 'quality', then we expect the survey respondents to choose answers with associated *high* user reputations.

The second independent variable measures answer verbosity. Short answers (1–2 sentences and a code snippet) are classified as *terse* and lengthy answers (1-3 paragraphs and one or more code blocks) as *verbose*. The resulting experimental design is a 2x2 factorial (Table 1).

Table 1. Experimental design

		Verbosity	
		Terse	Verbose
User reputation	High	T, H	V, H
	Low	T, L	V, L

Question and answer groups were sampled from Stack Overflow. All questions concern general Java functionality. We use Java because of its popularity in both industry and academia as well as because of the number and quality of Java-related questions and answers on Stack Overflow (a total of 556,276 on January 20, 2014). Six questions were chosen. For each question, two answers were selected. There are a total of six *terse* answers and six *verbose* answers; two questions have only *terse* answers, two have only *verbose* answers, and two have one *verbose* and one *terse*. This design allows us to test for interaction effects. Note that all questions met a minimum threshold of quality: they answered the question correctly and intelligibly.

The two answers associated with each question were differentiated by a *high* user reputation and a *low* user reputation. We counterbalanced to control for individual differences in answer by administering two variations of the survey. The first variation presented the *high* and *low* user reputations in one order (e.g., answerer *a* has *high* reputation and answerer *b* has a *low* one) and the second variation reversed the order (answerer *a* is now *low* and answerer *b* is *high*).

Stack Overflow layout and format of questions and answers was preserved. However, the answer rating was erased. The respondents saw the Stack Overflow question and two related Stack Overflow answers. The result is a total of twelve data points per survey: one survey respondent 'quality' score per each of the twelve answers. Furthermore, respondents were asked to briefly justify their ratings after each answer pair. These responses are used to better understand the reasoning behind the ratings.

At the end of the survey, a series of questions were asked to better understand the role that specific answer characteristics (presence of prose, code snippet, code blocks) played in helping users determine 'quality' (the various attributes will be rating on a 1-5 Likert scale). A space for additional prose explanation was given to allow respondents to explain any additional characteristics that mattered to them.

The survey was administered online. Survey respondents were recruited through Amazon's Mechanical Turk engine and were paid $3.00 for their time..

4. ANALYSIS AND RESULTS

A total of 48 MechTurk users participated in our survey. Of these, 34 successfully passed the Java proficiency test, which comprised a series of three Java questions. The following analysis is conducted on these 34. Two of which were about Java constructs and the other one required identifying a Java program snippet from a set of snippets in other languages. As specified in the previous section, two versions of the survey were conducted to better control for variations between answers within answer pairs. One survey version received 16 responses; the other survey version received 18 responses.

Table 2 reports the results of an Analysis of Variance (ANOVA) test on the factors: Answer Length and Answerer Reputation. Length is statistically significant (*p*-value <0.001); Social Reputation and the interaction effect are not (p-values of 0.58 and 0.52, respectively). Now let us review our research questions:

Effect of social information on answer quality perception: analysis of the survey results indicate that social information as presented by stack overflow made no statistically discernable impact on the perception of answer quality. There are two possible explanations: (1) most respondents did not consider social information when making their assessments, or (2) respondents considered social information in opposing ways, cancelling out the effects.

A qualitative analysis of survey respondents' comments shows that the former hypothesis appears to be correct: of the 34 respondents, only *one* respondent mentioned social rating as a factor. This particular respondent considered social standing to be the deciding factor on two of the five questions; for example, *I prefer Answer #2.1 It is because it is answered by an Java expert with more reputation score and more gold batches.*

The statistical insignificance combined with the almost-total lack of any survey responses on this issue suggests that measures of social standing are of little, if any, importance to the typical technical user. Evidence, to be presented shortly, suggests that users primarily consider answer *content* and, secondarily, *presentation*. Source (the answerer) is not important. Therefore, we conclude:

Conclusion 1: Social factors have little impact on perceptions of answer quality in the StackOverflow.com Q&A forum.

Let us now consider the second research question:

Effect of answer length on perception of answer quality: Our results show that answer length has a statistically discernable impact on the perception of answer quality. To better understand this relationship, we perform a simple Student's t-test to compare the mean answer rating of Verbose answers to the mean answer rating of the Terse answers. With a p-value of 0.0002, it was found that there exists a statistically distinguishable difference between the Terse and Verbose answers (matching the results of the ANOVA table). The mean answer rating of the Terse answers was 3.50, lower than the 3.93 of the Verbose answers. Taken at face value, these results suggest that users, on average, prefer longer answers. Yet, the story is not so simple.

Let us consider the results of the question "If answer length was important to you, what length did you prefer?" For a full third of our respondents (13), size did not matter whatsoever. Even more interestingly, those that considered length to be important were evenly split between those that preferred shorter answers (10) and those that preferred longer (11) answers. To reconcile this disparity we further investigate the context of the survey answers.

The statistical results do not capture whether a short answer is short because it successfully simplified a difficult concept or because it elided essential details (and thereby confused the reader); by the same token, a long answer might be a deliberate guided walkthrough or a dense wall of superfluous information.

Table 2. ANOVA results

Factor	DF	F	Pr(>F)
social reputation	1	0.31	0.58
length	1	14.41	0.0001 ***
social reputation * length	1	0.41	0.52
Residuals	404		

Significance: ***$\alpha < 0.001$ **$\alpha < 0.01$ *$\alpha < 0.05$

A qualitative analysis of the comments provided by the survey respondents reflect this fact. A common theme among the respondents' comments was *completeness*, best encapsulated by one respondent who said: "*All else being equal, I preferred answer that were more explanatory, which tends to be longer. But I didn't like them just BECAUSE they were longer"*. (There were forty-six such similar comments.) Although, because of the nature of the exit questions, participants did not explicitly define the term, associated comments placed emphasis on two distinct types of completeness. The first might be defined as *intrinsic* completeness: how well did the answer itself use code examples, example output, explanatory prose, references to documentation, and references to common errors to convey a sense of thoroughness? The second might be defined as *extrinsic* completeness: How well did the answer review all technical aspects of the problem at hand? Therefore, we conclude:

Conclusion 2: Answer length is important insofar as longer answers tend to be more thorough, which is preferred.

Other respondents, however, emphasized *conciseness*, choosing answers that they felt "got to the point." (There were eleven such comments). It must be carefully noted, however, that conciseness was secondary to thoroughness in the minds of many respondents. For example, one respondent stated: "*Answer 1 had a better example with output, while still being concise",* and another said: "*The* *first was better because the second tended to provide too much unneeded information"*. In other words, presentation, while of significance, was subordinated to content.

The significance of presentation was further underscored by a common emphasis on *clarity* (mentioned in nineteen comments). For example, one respondent said: "*This one is tough, both are well written and have great clarity. I'd give a tiny slight edge to 5.1 just because I prefer the writing style, but honestly I love that 5.2 includes the null string exception note"*. However, as in the earlier case clarity was also rarely defined by the respondents.

Table 3. Summary of Follow-up responses

Factor	Mean	Confidence Interval	Group	Standard Deviation
Code Snippets	3.88	(3.56, 4.20)	a	0.91
Code Blocks	4.00	(3.74, 4.26)	a	0.74
Prose	3.65	(3.25, 4.04)	a b	1.12

Confidence Intervals computed using Student's-T test at $\alpha = 0.05$.
Grouping determined by TukeyHSD test.

To better understand it, we consider the responses to the follow-up questions (which appear at the survey's end) that directly asked the users about three specific factors: importance of code snippets, importance of code blocks, and the importance of prose explanation. According to Table 3, all three elements are of moderately-high importance, with no one element of singular significance (according to the results of the Tukey Honest Significant Difference test, the ratings are statistically indistinguishable at the $\alpha = 0.05$ threshold). The comments reinforce this interpretation. Thirty-four comments mention code examples as important; thirty-nine comments emphasize the prose, of which 60% emphasize its explanatory purpose and the remainder focus on the quality of the prose itself. Furthermore, half of the respondents mentioned the importance of both code *and* prose in their comments. Therefore, we conclude the following:

Conclusion 3: Presentation is important. Both code (whether in snippet or block format) and prose are essential elements of a 'high quality' answer.

5. THREATS TO VALIDITY

Our respondents had a strong self-selection bias: all are active members on the Amazon MechTurk website.

Our study suffers from generalizability problems. Second, users of the StackOverflow website are—presumably—familiar with its associated social rating system. That rating system may be obtuse to the uninformed, using a series of 'medal' counts and an aggregated count. Survey respondents may have simply ignored the social information because they did not understand it. Furthermore, the presentation of the user reputation—the aggregated score, medal counts, size of text and icons used to represent this information, and the placement of the information—is unique to StackOverflow. Social reputation may be more strongly considered by users of site that more prominently displays social reputation information.

Finally, our respondents were not seeking answers to their own questions. They may have been more willing to overlook deficiencies in the answers, and they had less incentive to carefully scrutinize all available information (such as social reputation) when rating an answer.

6. CONCLUSIONS

We can reasonably conclude that the social rating system used by StackOverflow has no discernable effect on an average novice user. The fact that *only one* respondent mentioned social factors in their comments is telling: the impact of social factors is so muted that a subtle approach, like ours, is unlikely to reveal their impact on answer perception.

It is interesting that novices appear to ignore indicators of social reputation to instead focus on the contents (e.g., elements of presentation) of the answer. In theory, novices suffer from information deficiency—by definition the least qualified to rate a post based on its technical content—and so should have the most to gain by leveraging quantified presentations of the social reputation of an information provider (e.g., answerer). This may be a result of the relatively complicated user reputation system used by StackOverflow (a system that relies not only a simple raw number but on a system of medals that may not be immediately understood by a novice) or because technical novices do not consider social reputations to be a reliable indicator of answer quality. Further research is needed to better understand which of these two underlying factors best explains why users do not consider social rating to be important.

We can also reasonably conclude that length has no simplistic effect on the perceived quality of technical answers among novices. Users are more interested in thoroughness and conciseness, which while related to length, are also intimately connected with writing style, the type and difficulty of the technical content, and the perceptions and preferences of the reader. Better understanding the relationships between these factors may be fruitfully explored in future research via interviews with StackOverflow users.

We can also conclude that both code and prose explanations are considered to be essential elements to a good answer. How these elements are used in conjunction with one another to affect a good presentation is important, and a topic for further research.

7. ACKNOWLEDGEMENT

This work is partially supported by NSF awards: IIS-1110916, CCF-1253786, and DGE-0903469.

8. REFERENCES

[1] Stack Exchange: http://stackexchange.com/sites.

[2] J. Marlow, L. Dabbish and J. D. Herbsleb, "Impression formation in online peer production: activity traces and personal profiles in github," CSCW 2013, pp. 117-128.

[3] L. Dabbish, H. C. Stuart, J. Tsay and J. D. Herbsleb, "Leveraging Transparency," IEEE Software, vol. 30, no. 1, pp. 37-43, 2013.

[4] S. Grant and B. Betts, "Encouraging user behaviour with achievements: an empirical study," MSR 2013, pp. 65-68.

[5] V. S. Sinha, S. Mani and M. Gupta, "Exploring activeness of users in QA forums," MSR 2013, pp. 77-80.

[6] A. Pal, S. Chang and J. A. Konstan, "Evolution of Experts in Question Answering Communities," AAAI Conference on Weblogs and Social Media 2012, pp. 274-281.

[7] B. V. Hanrahan, G. Convertino and L. Nelson, "Modeling problem difficulty and expertise in stackoverflow," CSCW-Companion 2012, pp. 91-94.

[8] C. Treude, O. Barzilay and M. Storey, "How do programmers ask and answer questions on the web?" ICSE 2011, pp. 804-807.

[9] A. Bosu, C. S. Corley, D. Heaton, D. Chatterji, J. C. Carver and N. A. Kraft, "Building reputation in StackOverflow: an empirical investigation," MSR 2013, pp. 89-92.