Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow

Ashton Anderson
Stanford University
ashton@cs.stanford.edu

Daniel Huttenlocher
Cornell University

Jon Kleinberg
Cornell University

{dph, kleinber}@cs.cornell.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

Question answering (Q&A) websites are now large repositories of valuable knowledge. While most Q&A sites were initially aimed at providing useful answers to the question asker, there has been a marked shift towards question answering as a community-driven knowledge creation process whose end product can be of enduring value to a broad audience. As part of this shift, specific expertise and deep knowledge of the subject at hand have become increasingly important, and many Q&A sites employ voting and reputation mechanisms as centerpieces of their design to help users identify the trustworthiness and accuracy of the content.

To better understand this shift in focus from one-off answers to a group knowledge-creation process, we consider a question together with its entire set of corresponding answers as our fundamental unit of analysis, in contrast with the focus on individual question-answer pairs that characterized previous work. Our investigation considers the dynamics of the community activity that shapes the set of answers, both how answers and voters arrive over time and how this influences the eventual outcome. For example, we observe significant assortativity in the reputations of co-answerers, relationships between reputation and answer speed, and that the probability of an answer being chosen as the best one strongly depends on temporal characteristics of answer arrivals. We then show that our understanding of such properties is naturally applicable to predicting several important quantities, including the long-term value of the question and its answers, as well as whether a question requires a better answer. Finally, we discuss the implications of these results for the design of Q&A sites.

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: Systems and Software.
General Terms: Experimentation, Human Factors.
Keywords: Question-answering, reputation, value prediction.

1. INTRODUCTION

Question-answering sites — in which people pose questions to a community of Internet users — have evolved steadily over the past half-decade. One direction this evolution has taken is the development and maturation of sites such as Stack Overflow and Quora built around focused communities in which a significant fraction of the participants have deep expertise in the domain area. One consequence of this trend is that the content on these question-answering sites increasingly has lasting value: since questions and answers are saved on the site and often prominently ranked via search engines, people in the future who may not even be a priori aware of the site can be directed to the information there. Thus, rather than viewing each answer principally in terms of the immediate information need of the question-asker, the focus in recent years has broadened to further include the potential long-lasting value to people in the future who might have a similar question.

Given these developments, there is a clear opportunity to add value for both the producers and consumers of information on these sites by developing techniques that can analyze and extract valuable information from the community dynamics taking place. For consumers of information, there is the potential to identify and highlight questions of lasting value as soon as possible after they have appeared on the site, so that users can be directed to them. For producers of information — experts who are able to answer difficult questions on the site — there is the potential to identify questions that have not yet been successfully answered, so as to highlight them for increased attention.

A number of interesting lines of recent work have pursued related issues through a focus on the question-answer pair as a basic unit of analysis. Recent work in information retrieval, for example, has proposed methods by which high-quality question-answer pairs can be extracted and hence used for people who have the same (or similar) questions [15].

A systemic view of question-answering sites. Here we develop an alternate approach for extracting information from the activity on question-answering sites. Rather than considering free-standing question-answer pairs, we consider questions together with their set of corresponding answers. There are two aspects to this view — one at the question level, and another at the full site level.

First, as questions on these sites become more complex, single questions often generate multiple good answers produced by different experts who explore distinct aspects of the problem. As one of many prototypical examples, a question like "How do you format a JSON date in jQuery?" on Stack Overflow generates multiple useful responses; the answerers and subsequent commenters then differentiate among the several approaches and debate their relative merits. In this respect, the full set of answers constitutes an investigation of issues relevant to the original question that would be lost if any one of the answers — even a very good one — were viewed in isolation. Thus, when one talks about the creation of long-lasting value on a site like Stack Overflow, we claim that it is the question as well as all the corresponding answers that *together* bring long-lasting value to the site.

KDD'12, August 12–16, 2012, Beijing, China.

Second, in order to understand whether the question has been sufficiently answered, there is useful information residing in the community dynamics that govern the site as a whole — the reputation mechanism on a site like Stack Overflow both provides information about levels of community involvement, as well as provides incentives for effective contributions and good behavior. As we will see, properties such as reputation and community involvement also serve as correlates for further forms of behavior, including the dynamics of how users arrive to answer new questions that are posed, and how their answers receive approval or disapproval from the community.

Overview of Results. To make progress on the issues discussed above, we formulate two concrete tasks that capture the potential applications of this type of analysis for users of question-answering sites. The first task, motivated by the goals of information consumers on these sites, is the *prediction of long-lasting value*: given the activity on a question within a short interval after it was posed, can we tell whether it will continue to draw attention long into the future? The second task, motivated by the potential to elicit further contributions from experts, is the *prediction of whether a question has been sufficiently answered*: given the answers to a question so far, and the activity around the question, can we tell whether the needs of the question-asker have been met yet?

We develop approaches to these tasks using data from Stack Overflow, which is an ideal domain for considering these issues for several reasons. The first is due to its scale and adoption; it is one of the most successful focused question-answering sites on the web, and has a very active user community in which more than 90% of the questions posed receive an answer that is formally accepted by the question-asker. But beyond the activity on the site itself, Stack Overflow has played a major role in shaping the current paradigm for on-line question-answering, as more than 80 other Q&A sites have adopted the same basic platform. For our purposes, what is important is that Stack Overflow exhibits a set of basic properties that are now present in a wide range of focused Q&A sites: complex questions on a focused domain, active engagement by its users, and a substantial number of experts.

In order to address our two basic tasks on Stack Overflow, we begin by identifying sources of latent information in the community activity on the site that can be used for analysis. There is a rich pattern of behavior on Stack Overflow that generates such information: for each question on Stack Overflow, the answers (as well as the question itself) can receive positive and negative votes from members of the community, signaling evaluations of quality; independently of this, the user who posed the question may at any point decide to *accept* one of the answers. All of these contribute to a user's numerical *reputation score* on the site. Meanwhile, the question itself acquires attention from users other than its answerers, as people vote on the question and the answers, and arrive to view it from outside the site.

From our analysis of these processes, we identify two important but subtle principles that help drive the process of question-answering in this domain. These principles provide an organizing framework as well as specific features for our approach to the two tasks we have defined. The first principle is that the wide range of expertise levels lead to a kind of aggregate sequencing of the contributed answers to a question, with the most expert users generally moving first. Thus, although there is no explicit structure on the site that formalizes this dynamic, we can think of users as conceptually organized into a kind of latent "pyramid," with expert users at the top; a question enters at the top of the pyramid, where it is first considered by the elites, after which it progressively filters down through the reputation levels if it remains unanswered. This mental image is a simplification, but it is a useful guide for thinking about how expertise, answer speed, and content quality inter-relate.

The second principle we identify is that a higher activity level around a question not only signals the potential interest in the question, but in aggregate it also tends to benefit *all* the answerers of the question, in terms of the evaluation and reputation increases they receive. Thus, although a question-asker can only formally accept one of the answers, it is too simple a view to consider the multiple answers as existing in a state of pure competition. Rather, high activity tends to correspond to the presence of multiple answers that receive endorsement from the community more broadly, and hence hints at the type of lasting value we are seeking.

Following our discussion of the evidence for these two principles, we show that features based on this view lead to performance on our two tasks that improves significantly on natural baselines. More precisely, for predicting whether a question will have long-lasting value, we find that features of the answer arrival dynamics within as little as an hour after the question is posed can be effective at predicting whether the number of pageviews to the question will be high or low *a year later*. Our formulation of these features is motivated by the latent expertise pyramid discussed above. Moreover, we find that the number of answers to the question, and related measures, are particularly powerful features for this task, reinforcing our premise that questions on a site such as Stack Overflow acquire greater value when they attract a diverse set of answers.

For our second task, identifying questions that have not been resolved to the satisfaction of the question-asker, we establish a way of evaluating our predictions by making use of instances in which the questioner returns to offer a "bounty" for a better answer to the question. Here too we find that features based on the underlying community processes can lead to effective prediction; on the other hand, it is interesting that the actual speed of answer arrival is much less informative for this task.

Overall, our goal is to contribute to a broader investigation of this perspective on question-answering sites, and our performance on these tasks suggests that features arising from the community dynamics on a site such as Stack Overflow can provide important information beyond simply considering individual question-answer pairs.

2. RELATED WORK

Community question answering websites have been studied from several different perspectives. The first is the study of user communities, where research has investigated users, their interests and motivation for contribution [1, 19, 4]. Insights from such studies informed the design of network-based ranking algorithms for identifying users with high expertise [10, 17, 25, 26].

The second is the perspective of information retrieval where a question is viewed as a "query" and answers could be thought of as "results" [15, 8, 16, 2, 9]. One goal of this line of work is to take a question with multiple answers and extract the answer of best quality or the answer that is most related to a particular search query. This can be viewed as an attempt to "declutter" the question-answering pages by focusing on one "best" answer for each question. The exact problem is often formalized as a classification task of trying to predict whether a single given answer is of high quality (under various notions of quality [21]) with respect to a particular question. In our work, however, we recognize that users get significant benefit from good answers produced by diverse experts. In this respect the full set of answers constitutes a discussion of competing approaches that would be lost if any one of the answers were viewed in isolation. Models of question answering communities as zero-sum two-sided markets of question askers and answers have

Users	440K (198K questioners, 71K answerers)
Questions	1M (69% with accepted answer)
Answers	2.8M (26% marked as accepted)
Votes	7.6M (93% positive)
Favorites	775K actions on 318K questions

Table 1: Statistics of the Stack Overflow dataset.

also emerged [11] with the goal of explaining the dynamics and stability of Q&A communities.

Broadly related to our prediction tasks of long-term question value and question hardness is the work on novelty and popularity of online content [24, 20, 22], which is wrapped up with the broader theme of the role of search engines in online content discovery [6]. Another more distantly related line of work is on deliberation, voting and explicit user feedback in online communities [5, 12, 14]. While this line of work mainly seeks to predict user voting behaviors [3, 7, 13], our work attempts to identify early community-based indicators of question and answer quality.

Lastly, the Stack Overflow and the related Math Overflow question answering communities have been studied in the past for correlating user reputation and the perceived answer quality [23]. Recently, Oktay et al. studied the dynamics of Stack Overflow answerer arrivals [18] with a focus on demonstrating the use of several quasi-experimental designs to establish causal relationships in social media. Their observation relevant for our work here is that even after the best answer has been identified by the question asker, answers to the question keep arriving. In the light of our findings here this can be interpreted as an effort by the Stack Overflow community to provide answers that go beyond the current information need of the question asker.

3. DATASET DESCRIPTION

General question answering sites such as Yahoo! Answers, Quora, and others support many different types of interaction: expertise sharing, discussion, everyday advice, and moral support [1]. On the other hand, focused Q&A sites, like Stack Overflow, the programming-related Q&A site we study, differ from these broad interest sites in that all questions are meant to be objective and factually answerable – most subjective questions are frowned upon by the Stack Overflow community. Stack Overflow questions are generally hard, in the sense that relatively few people can provide a sufficient answer. Deep expertise and domain knowledge is thus often essential to providing a good answer. As mentioned in the introduction, this type of focused Q&A model has been extremely successful.

Stack Overflow's success is largely due to the engaged and active user community that collaboratively manages the site. Content is heavily curated by the community; for example, duplicate questions are quickly flagged as such and merged with existing questions, and posts considered to be unhelpful (unrelated answers, commentary on other answers, etc.) are removed. As a result of this self-regulation, content on Stack Overflow tends to be of very high quality. We obtained a complete trace of all the actions on the Stack Overflow website between its inception on July 31, 2008 and December 31, 2010. The data is publicly available off the Stack Overflow site and the basic statistics are shown in Table 1.

There is a rich set of actions a user can perform on Stack Overflow, which grows as a user builds up reputation on the site. The most basic actions are asking and answering questions. Both questions and answers can be upvoted or downvoted by other users. The basic mode of viewing content is from the *question page*, which lists a given question along with all the answers to the question and their respective votes. The vote score on an answer, the difference between the number of updates and downvotes it receives, determines the relative ordering in which it is displayed on the question page. The questioner can select an answer as the *accepted answer* at any point in time, indicating that it was the "best" answer to his/her question. Users may comment on other questions and answers and also vote on the comments. Any user may mark a question as a *favorite*, bookmarking it for future reference.

Action	Reputation change
Answer is upvoted	+10
Answer is downvoted	-2 (-1 to voter)
Answer is accepted	+15 (+2 to acceptor)
Question is upvoted	+5
Question is downvoted	-2 (-1 to voter)
Answer wins bounty	+bounty amount
Offer bounty	-bounty amount
Answer marked as spam	-100

Table 2: Stack Overflow's reputation system.

The reputation system on Stack Overflow is designed incentivize users to produce high-quality content and to be generally engaged with the site. Table 2 shows how reputation is gained and lost. Some actions have effects on two users' reputations, *e.g.* if user A downvotes an answer by user B, then B loses 2 reputation points and A loses 1. The ability to vote on answers is not granted to new users, but is earned relatively quickly, requiring 15 points for the right to upvote and 125 for the right to downvote. A user also has the ability to offer a *bounty* on their question if they want to provide an additional incentives for good answers. The questioner funds the award with their own reputation (it must be between 50 and 500 reputation points). A bounty can be offered only after two days have elapsed since the question was asked, and a bounty period of 1 week begins. At any time the questioner may decide to award the bounty to one of the answers.

4. DESCRIPTION OF TASKS

We first introduce the two prediction tasks that motivate our analyses. Both are drawn from practical problems that occur naturally on Q&A sites: the first is predicting the long-term interest and value of a question page; the second is predicting whether a question has been sufficiently answered or not. In both cases, we describe quantitative proxies for these properties that we use in prediction.

Our primary goal in formulating these tasks is to use them as an analysis framework, assessing how the information about community processes can be used to determine value on Q&A sites. As such, they are structured to explore relative performance gains from different types of information, rather than for optimizing raw performance per se.

4.1 Predicting long-term value of a question

As we discussed in the introduction, Q&A sites have increasingly shifted from revolving around satisfying the questioner's information need to building up repositories of useful knowledge about a given question. Thus, predicting which question pages have lasting value and garner a lot of attention — as well as understanding which properties are associated with lasting value — is of central interest to maintainers of a question answering community. Question pages that show early signs of long-term value could be displayed more prominently on the site or could be recommended to experts to contribute answers. The insights we derive in the next section can provide effective approaches for this task.

First, we note that surprisingly good performance on this task is possible due to the fact that the time scales on which social processes for each question occur are in fact a bit complex: the typical question has a "fast" phase when it acquires answers and votes, and a "slow" phase in which members of the community indicate its longer-term value — both by visiting the question page and through the mechanism of *favoriting*. The majority of answers and votes

on both questions and answers occur within the first day after the question is asked (and the median *response time*, how long it takes for a question to be first answered, is just 12 minutes across Stack Overflow's entire history).

However, we find strong evidence that, although most of the votes and answers arrive within a day of the question creation, question pages are of lasting value: for example, only 37% of favorites on a question arrive within the same time frame. After this initial period, favorites accumulate extremely gradually over time. This is consistent with a two-phase view of the lifecycle of a question page — first there is a "construction" phase, when most of the answering and voting (signaling of quality) take place, after which follows a long period of existence in mostly static form as a potentially valuable public resource to future would-be questioners.

4.2 Predicting whether a question has been sufficiently answered

Our second task forms a natural complement to the first: whereas before we aim to predict the long-lasting value of the question, now we try to tell if a question has been satisfactorily answered or not. This would be obviously useful on Q&A sites: attention could be directed towards currently unsatisfactory question pages to help turn them into useful resources.

On Stack Overflow, a questioner can decide to offer a reputation award (a *bounty*) on her question. If this happens, it is safe to assume that the question has not been answered to the questioner's satisfaction yet — otherwise she would not spend her reputation points on the bounty. On the other hand, if the questioner accepts an existing answer, we can say that the questioner is satisfied. In this task, we consider predicting if a bounty will be offered on a question or if the questioner will accept an existing answer.

Now that we've introduced our motivating tasks, predicting two complementary properties of questions, we explore the various community processes that lead to the creation of question pages. After this exploration, we will show that the information we derive from these processes helps us accurately predict both properties.

5. COMMUNITY DYNAMICS OF QUESTION ANSWERING

The Stack Overflow community responds to questions in two main ways: by answering them, and by voting on the answers and the question. We observe these two processes, answering and voting, as occurring simultaneously. In this section, we investigate some of the basic principles that govern these community processes at work. We group this analysis into two parts, corresponding to the answering and voting processes, respectively: (1) the ways in which reputation interacts with the arrival of users to answer a given question; and (2) the consequences of a question's overall level of activity. In these two parts, we identify some basic and recurring phenomena that will be useful when we develop techniques for our prediction tasks in the following section.

5.1 A Reputation Pyramid

There is an incentive to answer questions quickly on Stack Overflow, since many question-askers will accept the first answer that they deem satisfactory, thereby conferring reputation on the answerer. Hence, we expect to see that the higher a user's reputation, the faster he or she answers questions.

In Figure 1 we examine how median answerer reputation varies with the time-rank of an answer for questions with a fixed number of answers. We find that the highest-reputation answerers do usually occur earlier in the time-ordering of answers on a question.

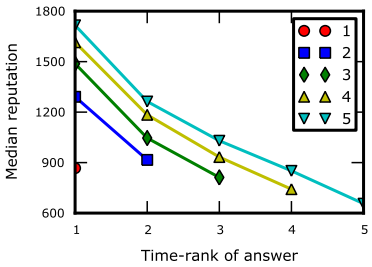


Figure 1: Median reputation versus answer time-rank. Questions with a total of 1 to 5 answers plotted (one curve each). High reputation users tend to answer early.

Reputation clearly decreases with increasing rank within a question, which is evidence of a direct relationship between reputation and answer speed.

Instead of the time-order of answers, we can also consider wall-clock time — how fast users of various reputation levels respond to questions. Here we find the same relationship: the higher the reputation, the quicker the user is to reply to a question. The typical (i.e. modal) response time is approximately the same — around 5 minutes — for all levels of reputation; the main difference is that high-reputation users hit this target of 5 minutes on a much larger fraction of the questions they answer.

These results suggest the conceptual picture mentioned in the introduction, in which users are organized in a *reputation pyramid*, with the highest-reputation users at the top and the lowest-reputation users at the bottom. A question enters the system at the top of the pyramid, where it is first considered by the highest-reputation users, then progressively percolates down through the reputation levels if it remains unanswered. This is a simplified picture of answering dynamics, but it is a useful conceptual picture for thinking about how answer speed, reputation, and content quality inter-relate. We stress that we are not claiming such an explicit vertical organizational structure exists on Stack Overflow; rather we are pointing out that many of the patterns we observe in this section are consistent with this picture of implicit behavior. For example, it helps explain the finding shown in Figure 2: the longer a question goes unanswered, the more likely it is that no satisfactory answer will be given (i.e. no answer will be accepted). Our picture of a question descending downward through reputation layers suggests this effect may at least partially be due to lower-reputation users becoming disproportionately likely to give a first answer the longer the question goes unanswered. The fact that lower reputation users give lower-quality answers on average (as measured in votes from other users) could then contribute to the observed relationship. We note that there could well be other factors contributing to the effect seen in Figure 2, including the fact that questions on which the first answer is slow to arrive may be more difficult or more idiosyncratic. These results suggest that high-value questions tend to be answered quickly and by high-reputation users, trends that we'll exploit in our prediction tasks.

These connections between reputation and answer speed show that the incentives arising from Stack Overflow's reputation system are producing behavior beneficial to the site. High-reputation users achieve their reputations largely by answering questions quickly and correctly, and presumably gain utility by doing so. From the questioner's perspective, the order in which answerers usually answer questions (high to low reputation) is ideal, since the questioner's expected time to receive a good answer is minimized.

Homophily by reputation. We observe that all reputation levels gain the majority of their reputation from receiving upvotes on

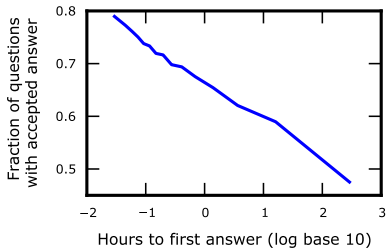


Figure 2: Fraction of questions with the accepted answer as a function of the time for the first answer to arrive. The longer the wait to get the first answer, the less likely it is for any answer to be eventually accepted.

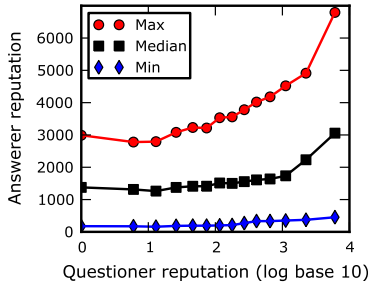


Figure 3: Max/median/min answerer reputation as a function of the questioner reputation.

their answers. This highlights an interesting fact about Stack Overflow: users who have attained upper-tier reputations are generally "answer-dominant": they gain the bulk of their reputation from answering others' questions well, and do not ask many questions of their own. Also, we see that the "elites", those users who have achieved over 100K reputation, gain significantly more of their reputation from their answers being accepted than other reputations levels, and correspondingly less from receiving upvotes (plot not shown due to space constraints). This fact about elites, however, appears to be largely due to an idiosyncrasy of the reputation system on Stack Overflow: once users gain 200 points in a day, they can only gain more from either having their answers selected or winning bounties. Since only the highest-reputation users hit this daily cap on a regular basis, we see a shift in the source of their reputation from upvotes to having answers accepted.

Having established that high-reputation users tend to gain most of their reputation from answering questions, one can ask whether there is any further stratification in *which* questions the high-reputation users answer. For example, one might have conjectured that there is a hierarchy of questioners and answerers, with high-status answerers reserving their efforts for questions from high-status questioners. But we do not see strong evidence for such a hierarchy; for example, Figure 3 shows that except at the highest levels of reputation, the median reputation of a question's answerer depends very little on the reputation of the user asking the question, and the maximum reputation among the answerers increases only weakly in the questioner's reputation. Thus, the real picture seems to be that high-reputation users are fairly omnivorous in the questions they answer.[1]

Although there isn't a strong connection between the questioner and answerer reputations, there could still be correlations between the reputations of answerers who answer the same questions. Indeed, our mental picture of a question floating down through different reputation levels suggests this might be the case — and we now show that it is.

A first approach to doing this is to compute the correlation coefficient between the reputations of co-answerers — pairs of users who answer the same question. To determine whether the correlation coefficient is indicative of homophily by reputation (i.e. the tendency of users with similar reputations to answer the same question) we compare it to the correlation coefficient for reputations of co-answerers in a randomized baseline. For this baseline, we consider the bipartite graph formed by questions on one side and answerers on the other, and with a link between a question Q_i and an answerer A_j if A_j answered Q_i. We then randomly rewire the bipartite graph while preserving the degrees on the left and right; this gives us our randomized baseline pattern of co-answering. The correlation coefficient between the reputations of the real co-answerers is 0.11 and the correlation coefficient between the reputations of the co-answerers in the randomized baseline is 0.031 (we use reputations on a log scale). This calculation shows that answerers with similar reputations are much more likely to answer the same question than would be expected by random chance given the distributions of answers by reputation. Thus, it seems that answerers in a given reputation level are attracted to the same sorts of questions, and that the source of this attraction is not the reputation of the questioner (due to Figure 3).

This previous calculation ignores the time-ordering in which the answers arrive. We now carry out a computation to answer the following question: What are the characteristics of *ordered* reputations on questions? Let r_i denote the reputation of the answerer who authors the i-th answer to arrive. Our question is: when a user with reputation r_1 first answers a question, what is then the conditional distribution over reputations of the second answerer (provided there is a second answerer)? In Figure 4 we show this conditional distribution, subtracted from the overall distribution of r_2 for the full population restricted to the set of response times in the figure (we restricted to questions where first answer comes in 6 minutes after the question). As the figures show, when the first answerer has high reputation, then high reputations are overrepresented in the population of second answerers; and correspondingly, when the first answerer has low reputation, the second answer has an elevated chance of having low reputation as well. Thus, this provides another indication of homophily by reputation among the answerers of a question. This is another phenomenon that provides useful information for the tasks we introduced in Section 4.

Interleaved processes of question answering and voting. Recall our observation from the beginning of this section about answers and the votes they receive, that one should think of the arrival of answers and votes as simultaneous. We find that they are in fact interleaving — both accumulate during the initial "fast" phase after a question is posed. The effect of this can be seen in Figure 5(a), which shows the reputation gained by an answer when it is the i-th answer to arrive out of k total answers. The linear decrease in i and the fact that the line is shifted upward for larger k can both be explained by the fact that answers and votes are arriving in an interleaved fashion: this means that earlier answers have more time to receive votes (hence the linear decrease in i), and as k grows, it means that the arrival process goes on longer, resulting in more votes for all answers.

There are some other aspects of this arrival process that stand largely as open questions, however. For example, we see in Figure 5(b) that the fraction of positive votes for the i-th answer out of k *increases* with i. (Note that all the fractions are very close to 1, so this is a distinction involving small differences.) There are

[1]Indeed, this may be almost by necessity: if relatively few high-reputation users ask questions, then one cannot acquire a very high reputation by restricting one's activities to questions from this subset.

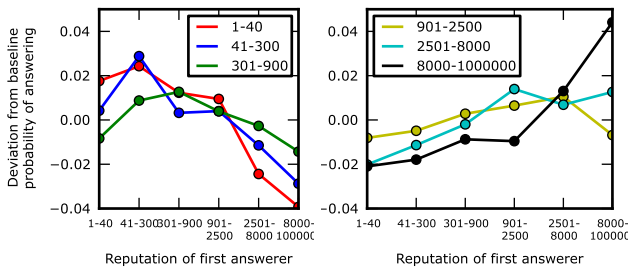


Figure 4: Each curve shows how, given the reputation of the second answerer on a two-answer question, the likelihood of answering second deviates from a uniform baseline as a function of the reputation of the first answerer. The curves on the left (showing the bottom three reputation levels) slope downward, indicating lower reputation levels are more likely to answer questions second if the first answerer also has low reputation; and the curves on the right (showing the top three reputation levels) slope upward, illustrating an analogous homophily by reputation effect.

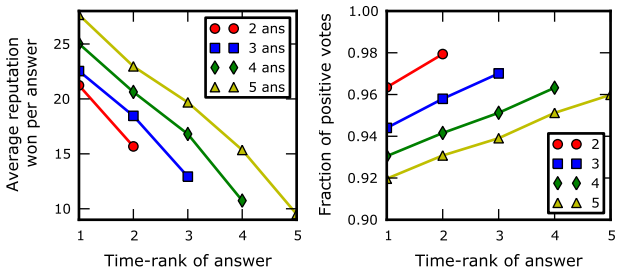


Figure 5: (left) Average reputation points won, (right) Fraction of positive votes on a answer as a function of answer-rank i when the total of k were given to the question.

several possible conjectures for this increasing behavior; for example, the fact that early answers receive negative votes may be the reason that the user posing the question allows answers to continue accumulating before accepting one of them.

5.2 The Activity Level of a Question

In the previous section we analyzed various aspects of the answer arrival process, and how our model of the reputation pyramid explains many of the phenomena we observe. Now we turn our attention to the other interleaved community process on questions: voting. Specifically, we consider the level of activity on a question, and its consequences for how both the answers and the question itself are evaluated by the community. Interestingly, when a question receives many answers, all the answerers benefit in terms of reputation gained, and the question receives more favorites over time. Thus, instead of viewing the answers as competing with one another for the community's limited attention, it appears that an essentially opposite view is more apt, and one in keeping with a central premise of the paper — that heightened activity around a question leads to greater value.

Higher activity produces benefits. As we previously discussed, questions on Stack Overflow are supposed to be answerable factually and objectively (if they're not, then they're marked as a "Community Wiki" question and actions on them do not count towards one's reputation score). This creates an incentive to answer quickly, since it is likely that the first correct answer may well be accepted. (Of course, as highlighted in the introduction, there are better and worse ways to give a factual answer — being grounded in fact is

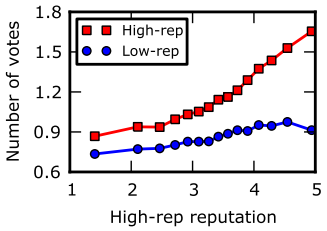


Figure 6: Average number of votes per answer for both answerers on a 2-answer question as a function of the higher answerer reputation. Lower reputation fixed between 75-125. High reputation plotted on a logarithmic (base 10) scale.

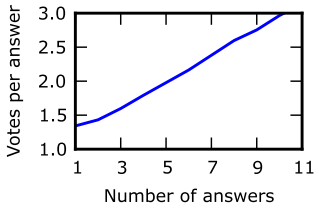


Figure 7: Number of votes per answer as a function of the number of answers on the question.

not the only prerequisite for being a good answer). However, this raises an interesting question: Does this mean that answers compete with each other for votes — that answers act as players in a zero-sum game? Or is it that answers are judged more by their inherent quality than in relation with the other answers to the question? Or perhaps there is some other phenomenon at work.

We answer this question with the following experiment. Let r_i denote the reputation of the i-th answerer, as before, and let v_i denote the vote score of the i-th answer. We now fix r_2 and observe how v_2 varies as a function of r_1. If the "competition" theory is correct, then we would expect v_2 to drop as r_1 is increased, since higher-reputation users are more likely to produce the correct answer quickly. If answers are judged by their inherent quality alone, we would expect v_2 to stay roughly the same. We show the relationship in Figure 6. Surprisingly, v_2 goes *up* as r_1 is increased. Clearly the answers are not playing a zero-sum game. We observe this effect because questions with high-reputation users participating receive more attention than average. As r_1 increases, both $v_1/(v_1+v_2)$ and v_2 increase. This means the high-reputation user increasingly "wins" the question on a relative scale (as measured in the share of the votes), but the lower-reputation user still gains on an absolute scale (by gaining more votes).

Next, we examine how the number of votes per answer varies with the number of answers on the question. Again, if the first theory (that answers compete for votes) is correct, then we would expect that the number of votes per answer decreases with the number of answers since more answers are competing to be the best. As can be seen in Figure 7, the opposite is again true: the more answers there are, the higher the votes-to-answers ratio. This is primarily due to the fact that questions with many answers receive a large amount of attention. Still, the fact that answers on question with many answers get more votes on average than votes with fewer "competitors" reinforces the point that answering and voting on Stack Overflow is not a zero-sum game. Even if answers compete for "individual" votes, we've seen that the increased attention on your answer both from having higher-reputation competitors and from having more competitors more than makes up for any competition between answers.

Finally, we show that questions with more answers are also viewed as more valuable by the community over the long time scale dur-

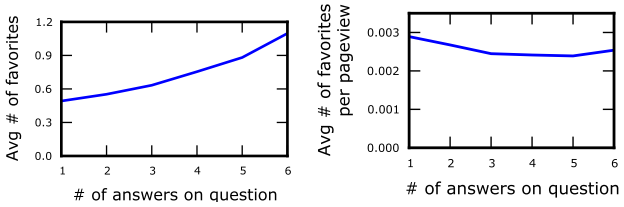


Figure 8: (Left) Number of question favorites plotted against the number of answers over all questions where the maximum answer vote score is 5. (Right) Same plot normalized by pageviews.

ing which the question is favorited. In Figure 8 we consider questions in which we fix the evaluation of the "best" answer — that is, we look at questions in which the highest-voted answer always received exactly five votes. Given a best answer of fixed quality, does it help to have additional answers? Figure 8 shows that the number of users favoriting the question indeed increases monotonically in the number of answers.

Again, this increase seems largely due to the increased activity around a question — though it is worth noting that this is increased activity in a highly generalized sense, since the favoriting is occurring over a significant time period, long after the burst of answering and voting on the question has subsided. In Figure 8 we see that the probability a user chooses to favorite a question they are viewing remains roughly constant with the number of answers. Thus, the correct way to view this may be as follows: having more answers increases the number of viewers of the question in the long term, and it does so with no downside in the *rate* of favoriting — each given user is still equally likely to favorite the question.

This again fits closely with a central theme of the introduction. Rather than viewing multiple answers to a question as a form of "clutter" that needs to be cleaned up, we see here that questions with multiple answers produce benefits in the form of increased attention, and they do this without suffering any loss in the chance that a reader of the question will choose it as a favorite.

6. PREDICTION TASKS

The previous section presented a set of principles governing the community process of question-answering on Stack Overflow. Now we show that this new understanding is directly applicable to the two prediction tasks introduced in Section 4. First we introduce the features that our analyses suggest would be helpful for the tasks.

Features used for learning. Overall, we explore four different classes of features (27 features in all) describing static and dynamic properties of the answering process to a given question. Note that the actual models we present in this section will not necessarily include all the features. Our aim here is to illustrate the space of features that arise from our findings in the previous sections. We then focus on building explanatory models using only the most essential features. The full set of features we consider is as follows:

- **Questioner features (S_A)**, 4 features total: questioner reputation, # of questioner's questions and answers, questioner's percentage of accepted answers on their previous questions.
- **Activity and Q/A quality measures (S_B)**, 8 features total: # of favorites, # of page views, # positive and negative votes on question, # of answers, maximum answerer reputation, highest answer score, reputation of answerer who wrote highest-scoring answer,
- **Community process features (S_C)**, 8 features total: average answerer reputation, median answerer reputation, fraction of sum of answerer reputations contributed by max answerer reputation, sum of answerer reputations, length of answer by highest-reputation answerer, # of comments on answer by highest-reputation answerer, length of highest-scoring answer, # of comments on highest-scoring answer.
- **Temporal process features (S_D)**, 7 features total: average time between answers, median time between answers, minimum time between answers, time-rank of highest-scoring answer, wall-clock time elapsed between question creation and highest-scoring answer, time-rank of answer by highest-reputation answerer, wall-clock time elapsed between question creation and answer by highest-reputation answerer.

6.1 Predicting long-lasting value

Recall from Section 4.1 that our first task is to predict long-lasting value of a question together with its answers.

Experimental setup. We use the number of pageviews of a question with its answers in a given time-frame as our measure of the amount of attention the question receives and thus as a proxy of its long-term value. The number of times the question was favorited would be an alternate proxy for value; however, we consider the number of pageviews (after controlling for question age) to be a better choice for several reasons: the number of pageviews is a large and robust number, whereas the number of favorites is quite sparse for most questions, and hence a noisy indicator of the community's reaction to the question. Moreover, only registered users can favorite a question, whereas the full Internet population contributes to its pageviews — hence the latter measure is consistent with our goal of viewing the question with its answers as having value that transcends the question-answering community itself.

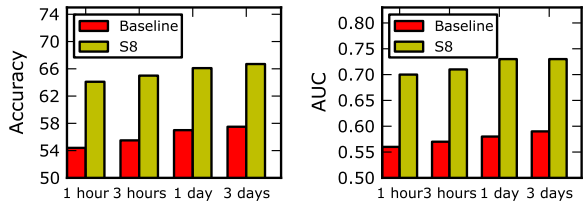
To control for the age of the question and Stack Overflow's ever-increasing popularity, we restrict our attention to questions created in the same month and predict the number of pageviews one year later. We only consider questions in which the first answer arrives within an hour after the question was asked (otherwise the question almost certainly doesn't receive a lot of attention). We formulate the task of predicting pageviews as a binary classification task[2], and report our performance on two setups: in the first case, the response variable is whether the question's number of pageviews is in the bottom or top quartile of the questions in our controlled sample (thus excluding the middle half of the dataset), and in the second case the response variable is whether the question's number of pageviews is in the bottom or top half of the questions (thus using the entire dataset). The full dataset consists of 28,722 examples, and our dataset is balanced (the response variable is split 50-50) in both cases by construction.

Since the practical application of this task would be to predict question quality early on in a question's lifetime, we perform this task using only the information available up to a given amount time after the question is asked. The time-frames we consider are 1, 3, 24, and 72 hours after the question is posted[3].

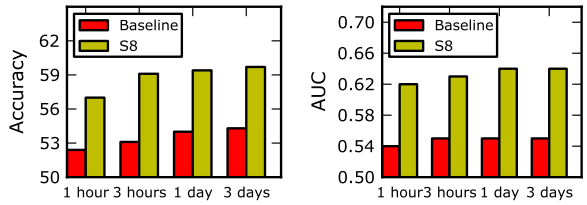
We performed feature selection and found a core set of 8 features that we use in this task: questioner reputation (log scale), # of questions the questioner has asked (log scale), # of answers on the question, sum of scores on answers to question, # of comments on highest-reputation answerer, and 3 features of the highest-scoring answer: length, # of comments on it, and how long after question creation it was written. We call this set S_8.

[2]Formulating this task as a regression problem puts too much emphasis on predicting the exact number of pageviews, and detracts from the main object of interest to a Q&A site operator: whether a page will be of high or low value in the long run.

[3]Another interesting feature is the number of pageviews in this short time-frame; however, we do not have access to such data.



(a) Predicting top quartile vs. bottom quartile



(b) Predicting top half vs. bottom half

Figure 9: Results of pageview prediction. Notice strong absolute and also relative performance of our method. (left) Accuracy, (right) Area under ROC curve.

We standardize all the features. For interpretability and ease of comparison, we use a logistic regression classifier, perform 10-fold cross validation and report classification accuracy and the area under the ROC curve (AUC).

To establish a baseline set of features we notice that on Stack Overflow there are two main mechanisms by which users can directly evaluate a question's quality: they can upvote or downvote it, and they can mark it as a favorite. Since these are ways Stack Overflow directly asks users about question quality, they should be strong predictors on this task. Therefore, we compare our features against these two "crowd-sourced" features: # of favorites on the question, # of positive minus negative votes on the question.

Experimental results. Figure 9a gives the results of the prediction task with the top-vs-bottom quartile setup. Surprisingly, the baseline only gives less than 5% improvement in accuracy over random guessing 1 hour after the question was asked.[4] This means that the number of votes and the favorites, which are two direct signals of question interestingness and value from the Stack Overflow user community, only slightly improves over random guessing. Extending the time window to 3 days — after which most answers, votes, and a large fraction of favorites have arrived (see Section 5) — only improves this to a 7.5% accuracy gain over random guessing. However, the community process and dynamics features (S_8) inspired by our previous analyses double the improvement to 9-10% in accuracy and 14-15 AUC points in each time frame. We achieve 0.70 AUC using the S_8 features available in the hour after the question was asked; the baseline only scores 0.56 AUC in the same time frame. Figure 9b shows that the results are similar when we don't exclude the middle half of examples, except that the absolute performance levels drop since the output variable is inherently noisier (i.e. differentiating between the 52nd and 48th percentiles is harder than differentiating between the 77th and 23rd percentiles).

To fully appreciate the result we emphasize that we are extracting features available only 1 hour after the question was asked and are predicting question pageviews 1 year in the future. We find it remarkable that only after a single hour there is already enough signal to predict the long-term question-page value, and that much of this signal comes from community features beyond simply looking at direct evaluations of the question.

[4]One way to reconcile this weak improvement with the clear relationship between pageviews and favoriting is to recall the point made earlier that favorites are quite sparse, and so fail to provide information about many questions.

Feature	Coefficient
Number of answers	+0.61
Sum of answer scores	+0.47
# of questioner's questions (log scale)	-0.46
Length of highest-scoring answer	+0.38
Questioner's reputation (log scale)	+0.31
Time for highest-scoring answer to arrive	+0.22
# comments on highest-scoring answer	+0.19
# comments on highest-reputation answerer's answer	+0.17

Table 3: Relative importance of features for predicting long-lasting value of the question.

Incorporating the rich contextual information found on the question page significantly helps predict eventual question attention and quality – even over directly asking Stack Overflow users. The relative importance of the S_8 features (aside from those included in the baseline) is listed in Table 3.

The single most important feature is the number of answers. This is already a strong indication that considering all of the answers instead of just one is helpful. Since the baseline knows the number of votes and favorites, this isn't purely an attention effect. The second-most informative feature provides another variation on our central theme of multiple good answers providing value over and above the best answer: the sum of the answer scores is more informative than any single feature of the highest-scoring answer. Interestingly, the effect of community interaction in the form of comments on answers also has significant predictive power. The fact that the time for the highest-scoring answer to arrive has a significant positive coefficient (meaning the *longer* it takes to arrive, the better the eventual question quality) is intriguing, and perhaps related to Figure 5, which shows that later answers get positive votes at a slightly higher rate. These results show that our findings in Section 5 significantly help predict the long-term value of question pages.

6.2 Predicting whether the question has been sufficiently answered

We now show that these same features also help on our second task: predicting if a question has been satisfactorily answered.

Experimental setup. Every question that eventually has a bounty offered on it (a "bounty question") has some number k of answers on it at the moment when the bounty is offered. Let B_k be the set of bounty questions with k answers prior to when the bounty is offered, and let A_k denote the set of non-bounty questions for which exactly k answers arrive before the questioner decides to accept one of them (and in which the questioner had less reputation points than required offer a bounty). Our classification task is then to use the information on a question page with k answers to predict whether it is a member of A_k or B_k. Since bounty questions are quite rare (there are 13K of them in total), we take a random sample of questions from A_k so that our dataset is balanced. We performed this prediction task for different values of k and the results were all qualitatively similar. We report our results for $k = 3$.

In contrast to the previous task, there is no natural baseline to compare to since Stack Overflow lacks a direct mechanism by which the user community could explicitly express that a question has not been satisfactorily answered thus far. Thus, instead of arbitrarily choosing certain features to act as a baseline, we study the performance of the four feature classes described at the outset.

Features used for learning. We again started with our full feature set $S_A \cup S_B \cup S_C \cup S_D$, and since we are given the exact number of answers so far (k), we also considered few additional features. Specifically, we computed the individual # of positive and negative vote counts for each answer (added to S_C), answerer reputations of

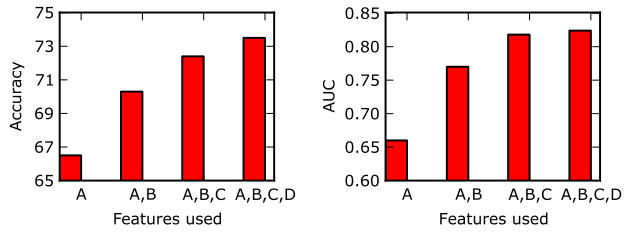


Figure 10: Results of bounty prediction. (left) Accuracy, (right) Area under the ROC curve.

each answer (added to S_C), and time between the answers (added to S_D) as additional features.

After conducting feature selection, we reduced our features to a smaller set of 18 features: 3 from S_A (questioner reputation, # of questioner's questions, and # of questioner's answers), 5 from S_B (# favorites on question, maximum answer score, maximum answerer reputation, and positive and negative question votes, 6 from S_C (average answerer reputation, # positive votes on last answer, # negative votes on 2nd answer, length of highest-scoring answer, length of answer given by highest-reputation answerer, and # comments on highest-scoring answer), and 4 from S_D: average time difference between answers, time difference between last 2 answers, time-rank of highest-scoring answer, and time-rank of answer by highest-reputation answerer. We add a prime to the names of the feature sets (*e.g.*, S'_A) to denote the subsets we chose.

Experimental results. The results for $k = 3$ are reported in Figure 10. Notice that the properties of the questioner on their own (S'_A) have good predictive power because high-reputation questioners can more easily afford to offer a reputation bounty. The page activity and question and answer quality measures (S'_B) are also useful predictors, as expected. But adding features incorporated from our study of the community processes governing Stack Overflow (S'_C and S'_D) gives a gain of nearly 5 AUC points over S'_A and S'_B. Again, we see that taking into account the rich interaction on the question page improves performance over even strong categories of features.

7. CONCLUSION

As question-answering sites grow in complexity, with more involved questions that are increasingly addressed by multiple experts, it becomes useful to think not just in terms of a "best" answer to a question but in terms of a set of answers and the community processes that produce them. We have seen how Stack Overflow, a site that exemplifies these aspects of question-answering, has a rich temporal structure that we have been able to use to identify important properties of a question together with its corresponding set of answers—specifically which questions and their answers are likely to be of lasting value, and which ones are in need of additional help from the community.

Our goal in this paper has been to start exploring the foundations for reasoning about community processes in question-answering. We anticipate that further analysis could potentially suggest richer ways of assessing expertise among users, identify a more intricate spectrum of genres among the questions that appear, and quantify more fully the role that incentives and competition play within a community as it answers questions.

Acknowledgements. We thank Stack Overflow for providing their data. This research has been supported in part by NSF CNS-1010921, IIS-1016909, IIS-1149837, IIS-1159679, IIS-0910664, CCF-0910940, and IIS-1016099, a Google Research Grant, a Yahoo Research Alliance Grant, the Albert Yu & Mary Bechmann Foundation, Boeing, Allyes, Samsung, Yahoo, an Alfred P. Sloan Fellowship, and a Microsoft Faculty Fellowship.

8. REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and Yahoo Answers: everyone knows something. *WWW*, 2008.
[2] E. Agichtein, Y. Liu, and J. Bian. Modeling information-seeker satisfaction in community question answering. *ACM Trans. Knowl. Discov. Data*, 3(2009).
[3] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Effects of user similarity in social media. *WSDM*, 2012.
[4] C. Aperjis, B. A. Huberman, and F. Wu. Human speed-accuracy tradeoffs in search. *HICSS*, 2011.
[5] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, L. Lee. How opinions are received by online communities: a case study on Amazon.com helpfulness votes. *WWW*, 2009.
[6] S. Fortunato, A. Flammini, F. Menczer, A. Vespignani. Topical interests and the mitigation of search engine bias. *Proc. Natl. Acad. Sci. USA*, 103(34):12684–12689, 2006.
[7] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. *WWW*, 2004.
[8] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan. Predictors of answer quality in online Q&A sites. *CHI*, 2008.
[9] J. Jeon, W. Croft, J. Lee, S. Park. A framework to predict the quality of answers with non-textual features. *SIGIR*, 2006.
[10] P. Jurczyk E. Agichtein. Discovering authorities in question answer communities by using link analysis. *CIKM*, 2007.
[11] R. Kumar, Y. Lifshits, and A. Tomkins. Evolution of two-sided markets. *WSDM*, 2010.
[12] J. Leskovec, D. Huttenlocher, J. Kleinberg. Governance in social media: A case study of the Wikipedia promotion process. *ICWSM*, 2010.
[13] J. Leskovec, D. Huttenlocher, J. Kleinberg. Predicting positive and negative links in online social networks. *WWW*, 2010.
[14] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. *CHI*, 2010.
[15] Q. Liu, E. Agichtein, G. Dror, E. Gabrilovich, Y. Maarek, D. Pelleg, I. Szpektor. Predicting web searcher satisfaction with existing community-based answers. *SIGIR*, 2011.
[16] Y. Liu, J. Bian, E. Agichtein. Predicting information seeker satisfaction in community question answering. *SIGIR*, 2008.
[17] K. K. Nam, M. S. Ackerman, and L. A. Adamic. Questions in, knowledge in?: A study of naver's question answering community. *CHI*, 2009.
[18] H. Oktay, B. J. Taylor, and D. Jensen. Causal discovery in social media using Quasi-Experimental designs. *SIGKDD Wkshp Soc. Media Analytics*, 2010.
[19] J. Preece, B. Nonnecke, D. Andrews. The top five reasons for lurking: Improving community experiences for everyone. *Computers in Human Behavior*, 20(2004).
[20] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, A. Vespignani. Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett.*, 105(2010).
[21] C. Shah, J. Pomerantz. Evaluating and predicting answer quality in community QA. *SIGIR*, 2010.
[22] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *CACM*, 53(2010).
[23] Y. R. Tausczik and J. W. Pennebaker. Predicting the perceived quality of online mathematics contributions from users' reputations. *CHI*, 2011.
[24] F. Wu and B. A. Huberman. Novelty and collective attention. *Proc. Natl. Acad. Sci.*, 104(45):17599–17601, Nov. 2007.
[25] J. Yang, L. Adamic, M. Ackerman. Crowdsourcing and knowledge sharing: Strategic user behavior on taskcn. *EC*, 2008.
[26] J. Zhang, M. Ackerman, L. Adamic. Expertise networks in online communities: Structure and algorithms. *WWW*, 2007.