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

Aaditya Bhatia, Abdullah Ahmad Zarir, Daniel Lee, Kundi Yao, Md Hasan Ibrahim
Software Analysis and Intelligence Lab (SAIL)
Queen's University
Kingston, Canada
aaditya.bhatia, a.zarir, 18dil, 18ky10, ibrahim.mdhasan@queensu.ca

Abstract—

I. Introduction

II. Related Work

Researchers have been studying community Q&A (question and answer) forums for a quite a long time from a different point of views. Few of the research groups focused on studying the users, their characteristics in community Q&A forums as well as their motivations behind contribution [?], [?], [?]. The output of these studies has helped to develop the network-based ranking algorithms which identify users who have higher expertise [?], [?], [?], [?].

Another group considered these Q&A forums simply as the data source of questions and answers. They retrieve this information and treated the questions as query and the answers as the results [?], [?], [?], [?], [?]. The aim of these kinds of studies is to find out a question according to the search query and propose the best answer relevant to that query. This approach can be considered as a trial to remove unnecessary information from the searched question pages and focusing on the best answers. Sometimes this kind of problems is treated as a classification problem which tries to predict whether the quality of the provided answer is high [?] in a given question. However, in our case, we found out that users are benefited from good answers of users with different level of expertise (according to their reputation). For each question, the answers build a thread of discussion which provides competing approaches. If any of these answers are read in isolation, it will lose its value. Models of inquiry noting networks as zero-sum two-sided markets of inquiry askers and answers have also risen [?] with the objective of clarifying the elements and steadiness of Q&A communities.

Our work on predicting long-term value of the questions and the difficulty level of the questions are on the side of novelty and popularity of contents online [?], [?], [?], which can also be considered as a part of the role of search engines in discovering online contents in a broader sense. Another line of research focused on discussion, voting and the feedback of the

users in community Q&A forums [?], [?], [?]. Although these researches mainly aimed at finding out the behavior of users on voting, our focus is more on the identifying the quality of questions and answers with early community-based indicators.

Finally, researchers have studied the Stack Overflow and similar kinds of Q&A community forums like Stack Exchange to observe the relationship between user reputation and the quality of received answer [?] in the past. Oktay et al. have tried to reveal the use of different quasi-experimental designs to build causal relationships in social networking sites by studying the dynamic arrival [?] character of Stack Overflow answers. The observation from this research suggests that even if the best answer is accepted by the question owner, answers keep arriving which is somewhat related to our work. From our study, we found that these efforts by the user community can provide information which may not be necessary for the question asker to meet her current need.

III. Dataset Description

IV. Description of tasks

A. Predicting long-term value of a question

B. Predicting whether a question has been sufficiently answered

V. Community dynamics of question answering

Except from proposing a question and

A. A reputation pyramid

B. The activity level of a question

In the previous section, we illustrate the structure of users' reputation pyramid, which explains how answers arrival process relates to the user's reputation, as well as other phenomena from our observation. In this section, we would investigate more on voting, another significant mechanism of Stack Overflow. Except for merely answerers and questioners, other users can also involve in those QA sections by commenting, voting, etc.. The voting mechanism is not only applicable to answers but also to questions, and it is also considered as an important factor to reflect community involvement and evaluation. From our observation of Stack Overflow, we notice
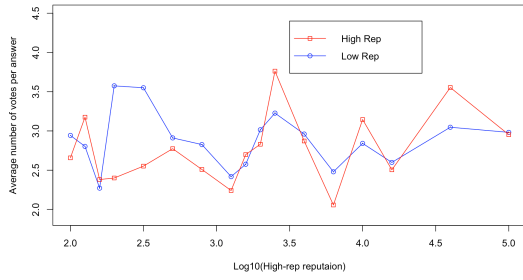
Fig. 1. Average number of votes per answer for both answerers on a 2-answer question as a function of the higher answerer reputation. Lower reputation fixed between 75-125. High reputation plotted on a logarithmic (base 10) scale. (From 2008/07/31 to 2010/12/31)
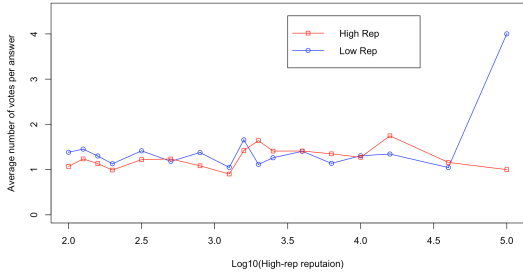

Fig. 2. Average number of votes per answer for both answerers on a 2-answer question as a function of the higher answerer reputation. Lower reputation fixed between 75-125. High reputation plotted on a logarithmic (base 10) scale. (From 2015/07/31 to 2017/12/31)
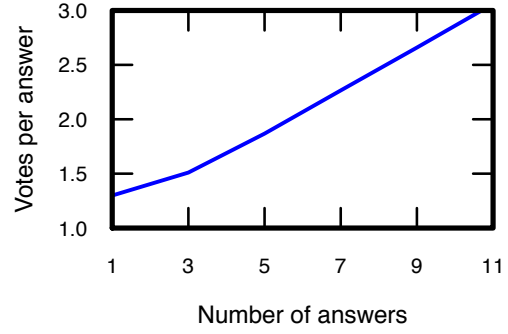

Fig. 3. Number of votes per answer as a function of the number of answers on the question. (From 2008/07/31 to 2010/12/31)
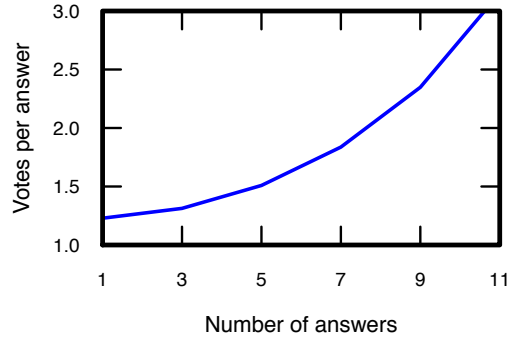

Fig. 4. Number of votes per answer as a function of the number of answers on the question. (From 2015/07/31 to 2017/12/31)

that questions with more answers are more likely to benefit from community involvement: answers will get higher votes and questions will receive more favorites. Highly active QA processes are more inclined to benefit from the community, instead of a competence among answerers themselves. Based on our observation, our goal is to validate whether if the activity level is able to explain the feedback and evaluation from the community activities.

Higher activity produces benefits. Like we discussed, unlike the features of general QA sites, Stack Overflow is programming-oriented, the questions are generally hard to be answered by majority community users. Furthermore, based on our observation from the previous section, answerers' arrival time is related to acceptance rate and answerers' reputation, it motivates answerers to answer a question quickly once it is posted. However, is there a competitive relationship existing among answerers?

In order to answer the question above, we choose all the questions with exactly two answers as our target dataset. Suppose r_i is the reputation of the i-th answerer, and v_i is the number of vote score of the i-th answer. If there exists a trend where v_1 goes up while v_2 decreases, we consider there's a competitive relationship between two answerers. Now we set the value of r_i unchanged as the x-axis, and compare the average vote score for both answerers. To begin with, we collected data from all questions with two answers, and separate answers according to answerers' reputation. Since the dataset is highly biased, majority users have very limited

reputation, which means they are either non-active users or possibly non-questioner or answerer. Adopting data from this part of users can deviate our result, we set a threshold where lower reputation users of the two answerers should have reputation between 75 and 125. In order to make the graph easier to interpret, first we scale x-axis, which represents higher reputation users' reputation, using a logarithmic(base 10). For each reputation score, we use an average value to present higher and lower answerers' reputation respectively. Instead of representing vote score for each reputation point, we smooth the curve by splitting the vote scores into groups, according to different reputations in the x-axis, and calculate the average value of the group. From Fig1 we notice that in most cases, answer vote scores from high and low reputation users have a similar trend, they either decrease or increase together, but high reputation answerers' answer votes have higher variance. The corresponding pattern reveals that in most cases, the two answerers do not have a competitive relationship. Similarly, this trend can also be noticed during 2015 to 2017. However, we notice the later dataset have relatively less average vote score, this could probably indicate that although there are a great amount of newly introduced answers, the quality of such answers are not as good as before, or the question has been sufficiently answered, newly added answers does not contribute much to the question.

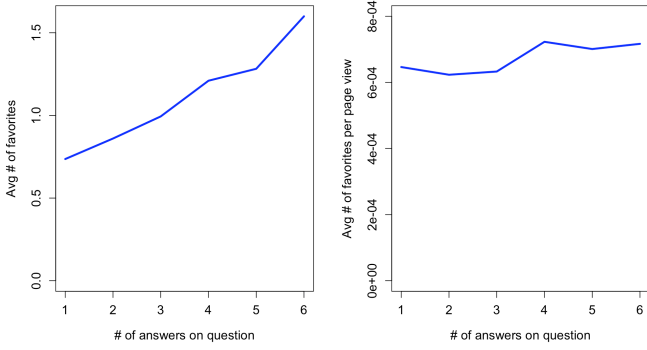In the following step, we aim to investigate the relation-

Fig. 5. (Left) Number of question favorites plotted against the number of answers over all questions where the maxi- mum answer vote score is 5. (Right) Same plot normalized by pageviews. (From 2008/07/31 to 2010/12/31)
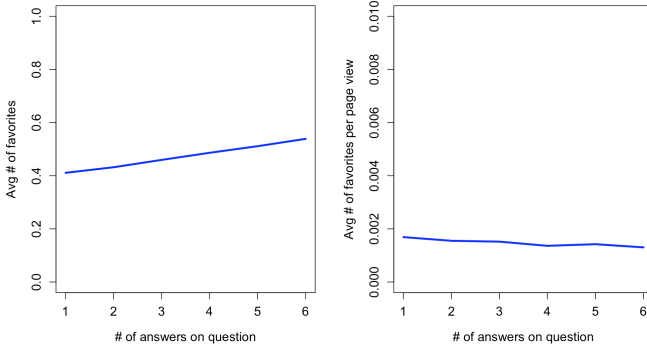

Fig. 6. (Left) Number of question favorites plotted against the number of answers over all questions where the maxi- mum answer vote score is 5. (Right) Same plot normalized by pageviews. (From 2015/07/31 to 2017/12/31)

ship between the number of answers in each question, to the number of votes per answer. This work is to testify the correctness of the competition theory. If there exists a competitive relationship among answers, we would expect a decreasing number of votes per answer, as the number of answers increases. From Fig 3 we find that from 2008 to 2010, the average number of votes display an ascending trend as the number of answers increases. This fact shows that the hotspot questions (questions with more answers) are more likely to benefit from community attention. Correspondingly, we also notice in Fig 4, where data is collected from 2015 to 2017, we find a similar upward trend for votes per answer as a function of the number of answers. From both of the two datasets we notice that, instead of competing for votes, all answers get more profit from communities involvement, especially for hotspot questions.

Apart from voting volcanism as an important source for evaluating community attention, favoring is also another channel to assess the quality of a question. In this part, our concern is whether questions with more answers receive more favorites from the community. In other words, we expect a hotspot question inclines to show its value by gaining more favorites. From previous study [**?**], Anderson et al. found that among the highest voted answers of all questions, the highest voting

score usually receives five votes. In this case, we first narrow our scope to the data from 2008 to 2010, to find all questions with the maximum answer vote score exactly equals to 5. From Fig 5 we can find that the average number of favorites to a question is positively correlated to the number of answers in that question. As a question gaining more answers from community activities, it is probably because the same question is also shared by a wide range of groups. If there is an existing question proposed a commonly possessed confusion, answerers will get more votes for sharing their own experience and solution, and the question will be favored if users are satisfied to acquire instant solutions.

Although we illustrate hotspot questions are more likely to gain a higher number of favorites, the favoring process usually takes a relatively long time, after a question is posted and even sufficiently answered. During this time period, the potential answerers or users have same concerns will visit those QA pages, this will lead to a significant number accumulation in each question's favorite number. In order to avoid the interference from the number of page views, we normalize the data by dividing the number of favorites from the previous step by the number of page views. From Fig 5 we find the average number of favorites per page view number is quite stable, which means users do not have a significant favoring preference on hotspot questions.

Then we choose questions proposed between 2015 and 2017, where all the data are collected right after the end of 2017, this should reduce the error caused by analyzing questions with more than two years' favorites accumulation. From Fig 6 we notice, even though the trend is similar as we illustrate before. However, we notice a significant scale change between those two figures. From 2015 to 2017, compared with the earlier dataset, we notice a significant decrease in the average number of favorites for all the questions. What's more, we also find the average number of favorites per page view has increased in these years. From our perspective, since we only study the questions posted in recent two years, those questions have relative low favorites number and page views. We think it is because Stack Overflow is already a mature knowledge sharing community after around ten years of development, most of the commonly possessed questions have been sufficiently answered. The newly introduced questions are either duplicating with previous questions or orienting a specific context and requiring expertise in that distinct domain. The duplicated questions will be marked and merged with previously proposed questions, and expertise questions are not accessible to the majority of community users cause they are targeting a special usage scenario.

In this section, we can conclude that answers are not competing with others, all the answers will profit from continuous community involvement. Also, for hotspot questions with many answers, the more answer it gets, the more possible that both questions and their corresponding answers will receive more benefits.

VI. Prediction tasks

A. *Predicting long-lasting value*

B. *Predicting whether the question has been sufficiently answered*

VII. Conclusion