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*—Question-answering (Q&A) websites offer a plethora of meaningful knowledge that remains untapped. Prior studies mainly focused on providing the best answer to the questioner. However, there is a shift towards extrapolating value for a broader audience by analyzing community dynamics. The voting and reputation feature of Q&A sites ensure the accuracy and quality of the content.**

In our paper, we focused on questions with their corresponding set of answer, rather than individual question-answer pairs. In addition, we used reputation to find information about levels of community involvement and incentives for successful contributions and positive behavior. We considered the dynamics of the community activity from a question and full site level that shaped the set of answers, how answers and votes arrive over time and how the dynamics influence the outcome. We formulated two prediction tasks to help us predict the long-lasting value of a question and predict whether a question has been sufficiently answered. Based on the high performance of our prediction tasks, we concluded that there are certain features in community dynamics on Stack Overflow that can provide valuable information.

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.

I. Introduction

Over the last decade, Q&A communities have evolved into a large repository of community-driven knowledge. Notably, communities such as Quora and Stack Overflow have evolved into active and mature communities. In our study, we focus on the Stack Overflow community as it contains one of the most active Q&A communities for developers. The questioner is the user who posed the question. On Stack Overflow, there is a significant fraction of domain experts who can provide answers to questioners with long-lasting value. Since Stack Overflow is a web-based Q&A community, the interactions are stored, so that they can be viewed at any time, which means that the content increasingly has lasting value for users.

Seeing that there is opportunity for long-lasting value for consumers and producers from Q&A communities, techniques can be used to analyze and extrapolate useful information about the community dynamics. Consumers of information are users who utilize the Q&A community to consume knowledge. Producers of information are the domain experts that provide answers to difficult questions on the site. We can guide consumers of information to questions with the potential of having long-lasting value. In addition, we can help producers of information potentially identify questions that have not been successfully answered yet.

Prior works have focused on using question-answer pairs for their analysis. In addition, prior work have proposed approaches to retrieve high quality question-answer pairs with the goal of helping people who have similar questions [7].

A holistic view of question-answering sites. Rather than the question-answer pair approach, we alternatively extract information from the community activity by considering questions together with their corresponding set of answers. We view community activity from two levels:

1) **Question level:** We focus on community activity from a question level by using questions with their corresponding set of answer because individual questions have the potential to generate multiple high quality answers. For example, a question such as,"How do you add a remote repository using Git?", can produce multiple high quality replies. We conjecture that questions combined with all their corresponding set of answers can create long-lasting value for questions on sites such as Stack Overflow.
2) **Full site level:** We focus on community activity from a full site level by using the reputation feature from Stack Overflow because reputation provides holistic information about: 1) Levels of community involvement. 2) Incentives for successful contributions and positive behavior. Community involvement and reputation can show us the dynamics of how users provides answers to new questions and how the community approves or disapproves the answer.

Overview of Results. To investigate the potential applications of studying community dynamics from a question and full site

Group assignment for CISC 880.

level for users on Q&A sites we develop two tasks. The first task is to *predict the long-lasting value* in order to help guide consumers of information to questions with the potential of having long-lasting value. We predict the long-lasting value by computing the question activity within a small time frame after the question is posed. The second task is to *predict whether a question has been sufficiently answered* in order to help producers of information potentially identify questions that may need their contribution.

We use approaches that are constructed by the data from Stack Overflow to address the two tasks. We first identify latent information from the Stack Overflow community. Stack Overflow questions and answers can receive positive and negative votes from community members, which determines the quality of the answer. In addition, the questioner can accept one of the given answers. These factors contribute to a user's *reputation score*, which we use for our analysis.

We identify two principles that provide an organizing framework and features for our two prediction tasks:

1) **Expertise level:** There is a wide range of expertise level that influences the sequence of contributions to a question, with experts generally responding first. The sequence is comparable to a *reputation pyramid*, where experts or elites are at the top of the pyramid and the question trickles down in a top-down manner.
2) **Higher activity level:** Questions with higher activity level signifies the potential interest in the question and the potential of benefitting all answerers based on the evaluation of their answer from the community and their reputation increase. Higher activity questions associate with multiple answerers and can hint at the type of lasting value.

For predicting whether a question will have long-lasting value, we use features based on the answer arrival dynamics within an hour after the question is posed. In doing so, we can classify whether the questions pageviews will be high or low, one year later. We find that number of answers, sum of answer scores, number of questioner's questions, length of highest-scoring answer to arrive, number of comments on highest-scoring answer, and number of comments on highest-reputation answerer's answer are the most powerful features, which shows that attracting a diverse set of answers obtain greater value on Stack Overflow.

For identifying questions that have not sufficiently been answered yet, we predict the questions that offer bounty for a better answer because when a question is not sufficiently answered they will resort to offering bounty. In result, we find that powerful features can lead to an effective prediction.

The main goal of our paper is to use Stack Overflow to provide insights about question-answering sites by leveraging the performance of the features from the two prediction tasks to suggest that community dynamics can provide more information than simple question-answer pairs.

II. Background

A. The Stack Overflow Community

Stack Overflow is one of the most active and successful Q&A sites, where over 90% of the questions receive a response that is accepted by the questioner. More than 80 Q&A sites were influenced and have adopted the same Q&A paradigm as Stack Overflow. In addition, Stack Overflow exhibits qualities that exist in the other Q&A sites: 1) Complex questions on a certain domain. 2) An active community. 3) Significant number of experts.

Experimentation Results With different values of k, we show the results in table. It is also interesting to observe that as the value of k goes high, users rich in reputation provide bounties others questions.

III. 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 [1], [2], [3]. The output of these studies has helped to develop the network-based ranking algorithms which identify users who have higher expertise [4], [3], [5], [6].

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 [7], [8], [9], [10], [11]. 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 [12] 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 [13] 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 [14], [15], [16], 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 [17], [18], [19]. 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.

TABLE I
STACK OVERFLOWS REPUTATION SYSTEM.

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

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 [20] 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 [21] 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.

IV. DATASET DESCRIPTION

In this section, we describe how we collect the dataset that we used to answer our research questions.

A. Data Preprocessing

There are 8 types of posts by users in Stack Overflow [22]. In this study, we are interested in the Question and Answer posts that are the primary motivation of the platform which drives the community activity. Any user with a registered account can post a Question in the site. Following the posting of the Question other users in the platform can post their Answer. There are further activates and actions that can be performed on a Question and Answer. There are 38 different types of actions in Stack Overflow [22]. For our study we are interested in the Comment and Vote. Comments are posted under a post and has a score given by users. There are 16 types of votes in Stack Overflow [22]. We are interested in 4 types of votes for our study, AcceptedByOriginator, UpMod AKA upvote, DownMod AKA downvote and BountyStart. AcceptedByOriginator means that an answer to a question has been accepted as the correct answer to that question by the questioner. Upvote indicates how many distinct users found a question or answer to be useful. Downvote indicates how many distinct users found a question or answer to be not useful. The difference between the upvote and downvote on a post is considered the score of that post. BountyStart is a special vote put on a question only. Providing an answer to a question after a bounty has been placed on it gives the answerer an opportunity to receive more Reputation from it if his/her answer is accepted by the questioner. User activity in the Stack Overflow is indicated by their Reputation. Users receive positive reputation points for upvotes in their posts and negative reputation points for downvotes. A detailed chart of reputation change is shown in Table I. The reputation system in Stack Overflow is modelled to motivate users to post rich content and be active in the site. Different actions in the site is privileged to higher reputation users. Users require a minimum of 15 reputation points to vote on posts. A minimum of 75 points to put a bounty on a question. The amount of bounty offered is deducted from the user who put the bounty. The lowest bounty that can be offered is 50 and maximum is 500. Questioner decides who gets the bounty by accepting an answer at any point in time after the bounty was set. We collected data about all question and answers between August 31, 2008 and December 31, 2010. The dataset consisted details of 299,398 distinct users, 1,096,144 questions, 2,632,009 answers, 4,657,336 comments and 10,143,364 votes. The dataset was archived in a xml format, from there we have extracted the required information and populated a SQL database. More details about the data is shown in Table II.

V. DESCRIPTION OF TASKS

In this section, we aim to cater to the two fundamental questions prevalent in Q&A sites, i.e. predicting the long-term value of the question and predicting whether a question has been adequately answered. This is done by comparing performance gains for different features used to predict the individual tasks.

A. Predicting long-term value of a question

Q&A sites serve as an information warehouse providing long term information to multiple users and viewers. Questions with a long-lasting value attract a lot a community attention. The early signs displayed by a question are adequate to demonstrate its long-term effects. A question as a fast phase when it will gather community attention and a slow phase where it will present its long-term value to the community. With more page views, the value if a question is high, as more people refer to the question for information gain.

B. Predicting whether a question has been sufficiently answered

The second task focuses on those questions which have not been sufficiently answered and turn them into valuable resources. A questioner decides to offer bounties on a question, he feels is inadequately answered. On the other hand, questioner accepts an answer when he is convinced with the response. This phenomenon is further explored with predicting whether a question will be answered in task 2.

VI. COMMUNITY DYNAMICS OF QUESTION ANSWERING

Except answering to a proposed question, voting is another significant mechanism in Stack Overflow. Community users can evaluate the quality or usefulness of an answer by giving them a vote up or vote down. In this section, we focus on both

TABLE II
STATISTICS OF THE STACK OVERFLOW DATASET.

Data	Status (2008 to 2010)	Status (2015 to 2017)
Users	440K (198K questioners, 71K answerers)	8.2M (1.9M questioners, 1.2M answerers)
Questions	1M (69% with accepted answer)	15M (55% with accepted answer)
Answers	2.8M (26% marked as accepted)	24M (35% marked as accepted)
Votes	7.6M (93% positive)	100M (89% positive)
Favorites	775K actions on 318K questions	9.3M actions on 963K questions

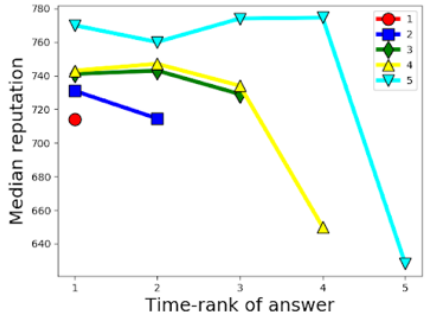

Fig. 1. Median reputation versus answer time-rank. Questions with a total of 1 to 5 answers plotted (one curve each). High reputation users tend to answer early. (From 2008/07/31 to 2010/12/31)

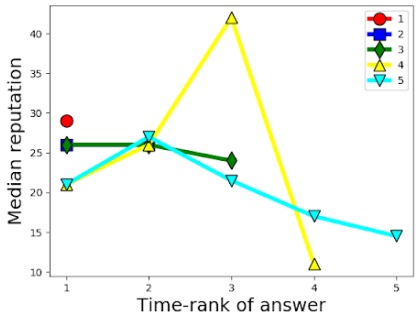

Fig. 2. Median reputation versus answer time-rank. Questions with a total of 1 to 5 answers plotted (one curve each). High reputation users tend to answer early. (From 2015/07/31 to 2017/12/31)

investigate answering and voting mechanisms. Specifically, we want to find out how answerers' reputation correlates to the answers' arrival time. In addition, different scenarios from various questions' activeness are also our concern. To investigate these concerns, we reveal underlying scenarios behind Stack Overflow community, the result can be useful for researchers to possess a deeper understanding and insights on Stack Overflow.

A. A reputation pyramid

A reputation mechanism adopted in Stack Overflow, it can be accumulate by answering questions, collecting more vote scores or winning a bonus. Usually, reputation is a significant standard to evaluate an answerer's expertness. In Stack Overflow, answerers are encouraged to answer a question quickly. In this scenario, we would expect an answerer with high reputation to be the provide timely correct answers.

From Fig 1 we calculate the median reputation of different answerers in a time-rank sequence, which means the answers of a question arrive successively. We find in most cases, the early answerers in the time-rank sequence usually have a relatively high reputation, and the later answerers are inclined to have lower reputations. This decrease is especially obvious when analyzing low speed answerers' reputation. The fact indicates that answerers' reputation can be utilized in evaluating whether if a question has been sufficiently answered, this will serve as the foundation of following prediction work. In 2017 data, from Fig 2 we also find in most cases, the high reputation users arrive earlier when question is posted. This trend is not that obvious as previous dataset.

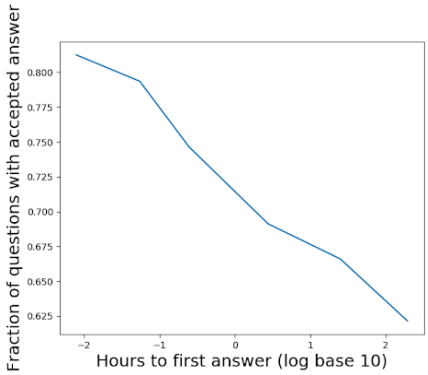

Fig. 3. Fraction of questions with the accepted answer as a function of the time for the first answer to arrive. The longer the wait to get the first answer, the less likely it is for any answer to be eventually accepted. (From 2008/07/31 to 2010/12/31)

A similar scenario is also discovered by Anderson et al. [23], they find the higher reputation a answerer owns, the quicker he/she will respond to a question. They consider wall-clock time, and they find users' respond time is approximately 5 minutes. Among all answerers hit the target around 5 minutes, high reputation users take a larger proportion of the group.

From the finding listed above, we can envision a pyramid structured relationship among Stack Overflow users. Users with high reputation occupy the top of the pyramid, when new questions are proposed, they usually be noticed, considered or answered by top level users. Then lower level user will involve in such Q&A section is the question remains to be unsolved.

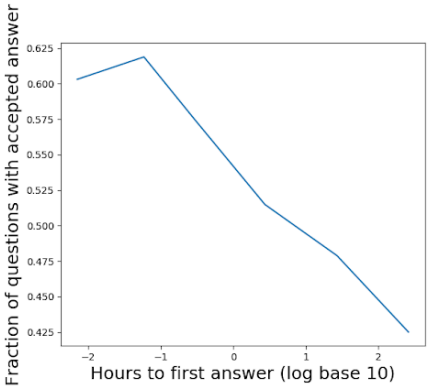

Fig. 4. Fraction of questions with the accepted answer as a function of the time for the first answer to arrive. The longer the wait to get the first answer, the less likely it is for any answer to be eventually accepted. (From 2015/07/31 to 2017/12/31)

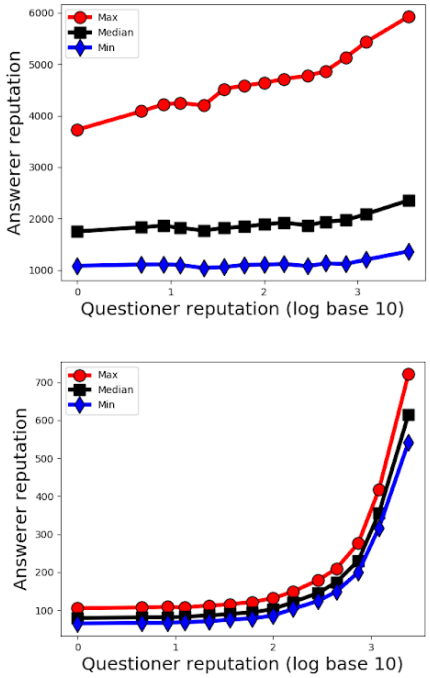

Fig. 5. Max, median, min answerer reputation as a function of the questioner reputation.

In Fig 3 we find the longer time it takes for an answer to arrive, the less likely that such question will accept an answer at the end. Similar trends can also be found in Fig 4.

Activity in Stack Overflow is driven by the incentive of reputation. Users with variant reputation points creates a pattern of behaviour among themselves. Overtime this pattern could be seen with respect to the number of total users. In Figure 5 the relation between the questioner and answerer reputation is plotted. Questions from high reputation users are answered more by high reputation answerers. The questioners reputation has been plotted with log base 10. And the answerer reputation is plotted separate. Except the above rare case, the median reputation of users is the same.

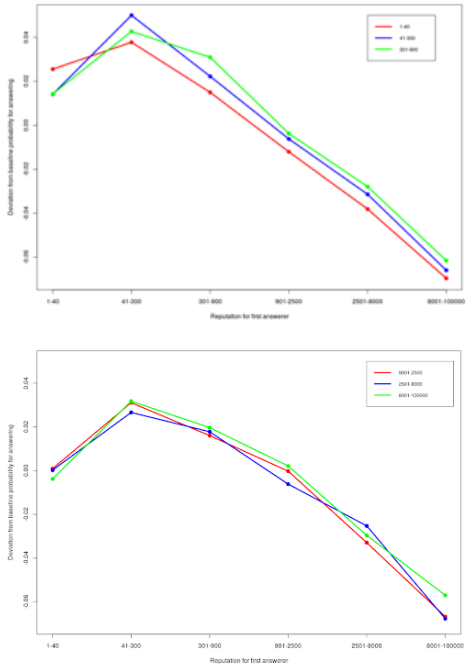

Fig. 6. Each curve shows how, given the reputation of the second answerer on a two-answer question, the likelihood of answering second deviates from a uniform baseline as a function of the reputation of the first answerer. The curves on the left (showing the bottom three reputation levels) slope downward, 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

For all the previous analysis, time orders of answers are ignored. Now our goal is to investigate the relationship between ordered reputation and time ordered answers. In this regard, we have taken the questions having two answers and the first answer arriving after 6 minutes. We also divided users into 6 different reputation group ordering according to their reputations. Then we put 3 groups from the bottom as low reputation groups and 3 groups from the top as high reputation groups. After that, we calculated the probability of answering a question when the first answer is answered by someone lower or higher reputation. From Figure 6 we can understand that irrespective of the second answerer reputation, it is less probable that a question is answered by the second answerer when the first answerer reputation is in the higher range.

Users get reputation point for each answer they post. If an answer has 3 positive votes and 1 negative votes, then the reputation change can be measured using Table 2. The calculation follows as 3*10 + 1 * -2 = 28. If the questioner accepts that same answer, then the answerer gets an additional 15 reputation points. The reputation won by an answer depends on the timing of that answer in that question. From Figure 7(a), we see that the reputation won by the first answer is always higher respective to the answers arriving later. And the last answer hast least reputation change, which indicates as time

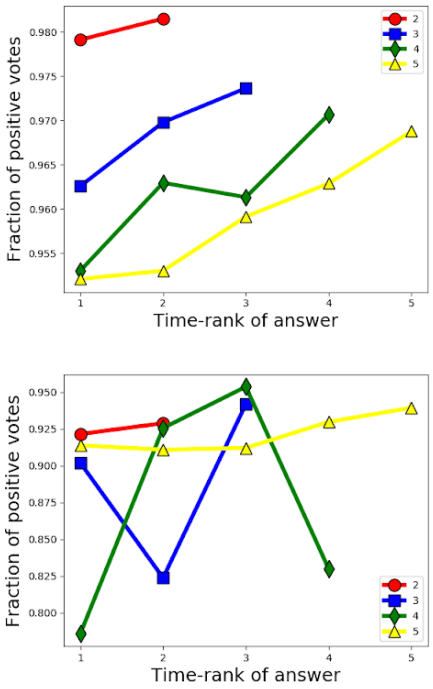

Fig. 7. (left) Average reputation points won, (right) Fraction of positive votes on a answer as a function of answer-rank i when the total of k were given to the question.

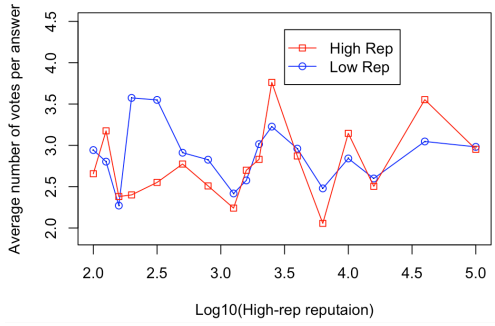

Fig. 8. Average number of votes per answer for both answerers on a 2-answer question as a function of the higher answerer reputation. Lower reputation fixed between 75-125. High reputation plotted on a logarithmic (base 10) scale. (From 2008/07/31 to 2010/12/31)

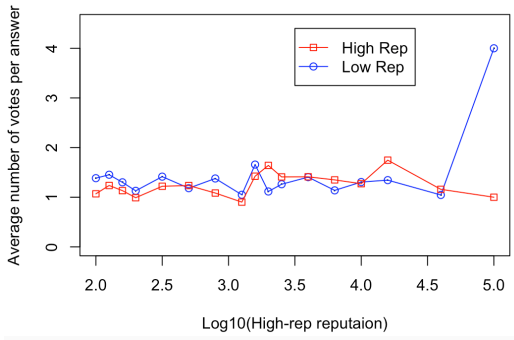

Fig. 9. Average number of votes per answer for both answerers on a 2-answer question as a function of the higher answerer reputation. Lower reputation fixed between 75-125. High reputation plotted on a logarithmic (base 10) scale. (From 2015/07/31 to 2017/12/31)

passes, activity on a questions page decline. But opposite to that, the fraction of votes in the later arriving answers are higher. This increase in positive votes along with the number of answers might be an indication that questioners might want to gather more answers on it before accepting one.

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 Q&A 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 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 Q&A 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 Q&A 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

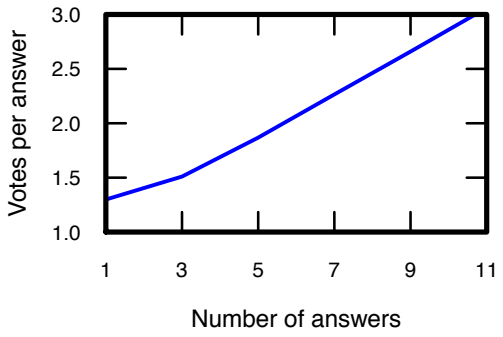

Fig. 10. Number of votes per answer as a function of the number of answers on the question. (From 2008/07/31 to 2010/12/31)

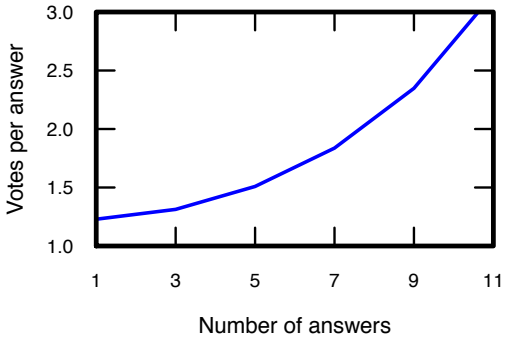

Fig. 11. Number of votes per answer as a function of the number of answers on the question. (From 2015/07/31 to 2017/12/31)

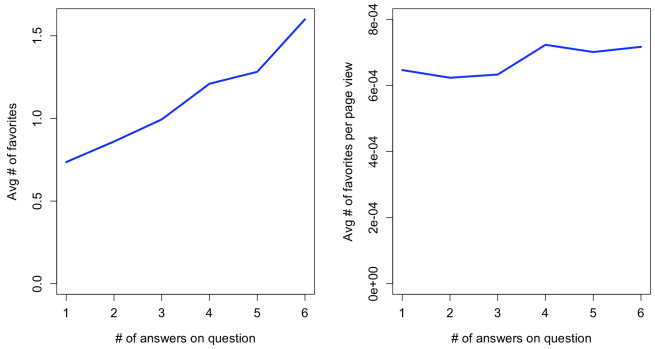

Fig. 12. (Left) Number of question favorites plotted against the number of answers over all questions where the maxi- mum answer vote score is 5. (Right) Same plot normalized by pageviews. (From 2008/07/31 to 2010/12/31)

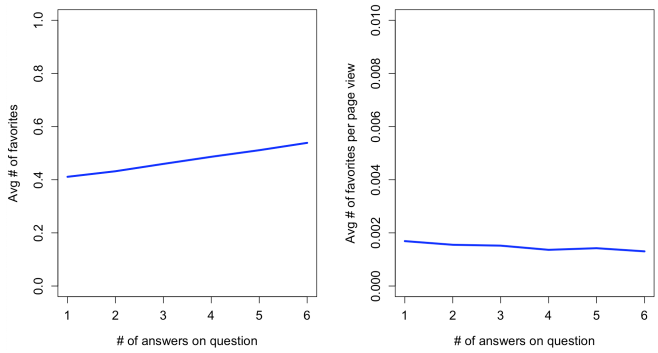

Fig. 13. (Left) Number of question favorites plotted against the number of answers over all questions where the maxi- mum answer vote score is 5. (Right) Same plot normalized by pageviews. (From 2015/07/31 to 2017/12/31)

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 Fig8 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 relationship 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 10 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 11, 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 [23], 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 12 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 favoriting 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 Q&A 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 12 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 13 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.

VII. Prediction Tasks

In this section, we grab the phenomenon surrounding around community process of Stack Overflow question-answering. We demonstrate prediction of value of a question to the community and to the questioner in two subtasks: Predicting the long-term value of a question and predicting whether the has been sufficiently answered or not.

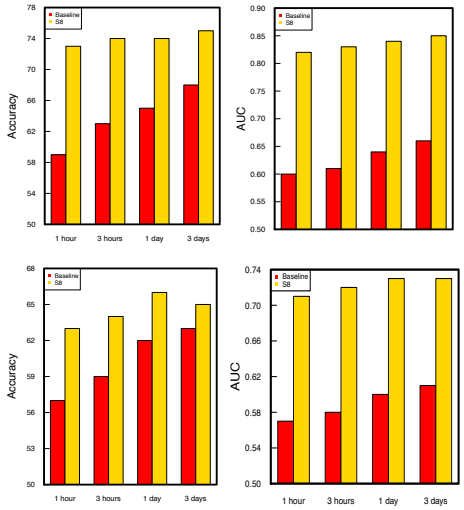

Fig. 14. Results of pageview prediction. Notice strong absolute and also relative performance of our method. (left) Accuracy, (right) Area under ROC curve. Top row is for quartile division of page view, bottom row is for half division of page view.

A. Predicting long-lasting value

The community value of a question can be reflected in the long-term impact of a question. While, multiple features like favourite count , Page Views, Up Votes and Down Votes reflect the impact of the question, page view has chosen to be a balanced feature as it is the reflector of a questions value. Unlike up votes, down votes and favourite count; page view has a significantly large value. However, the number of page views can pose to be noisy and need pre-processing before prediction tasks. The features that have been found out from the previous sections include: questioners factors, activity and Q/A factors, community factors, and temporal factors as shown in Table III.

Experiment Setup To perform this prediction, we split the page view value into high and low. In the first case, this division is done by the mean, and data is split into half across the mean. In the second case, page view is split into 4 quadrants, taking the top most and the bottom most quartiles as high and low respectively. Quartile split provides reinforcement to the learnings and important features found in the case where page views are split by half. We reduced the questions with pageviews more than 5 million, to remove outliers and for generalizability of results.

Providing adequate time for a questions pageview value to mature itself is vital for the experiment. All the questions posted within the same month, 1 year before the last update of the stack-overflow data dump (31st December 2011) have been considered for the study, i.e. we considered questions with creation date between 1st December 2010 and 31st December 2010. Most features required to predict the long-term value of

TABLE III
FEATURES USED FOR LEARNING

Questioner features (SA)	questioner reputation, # of questioners questions and answers, questioners, percentage of accepted answers on their previous questions.
Activity and Q/A quality measures (SB)	# 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 (SC)	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 (SD)	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.

TABLE IV
FEATURE COEFFICIENTS FOR PREDICTION TASK 1

Feature	Coefficient
Number of answers +0.61	+ 0.09
Sum of answer scores +0.47	+ 4.72
# of questioners questions (log scale) -0.46	- 0.21
Length of highest-scoring answer +0.38	+ 0.08
Questioners reputation (log scale) +0.31	+ 0.05
Time for highest-scoring answer to arrive +0.22	+ 0.34
# comments on highest-scoring answer +0.19	+ 0.23
# comments on highest-reputation answerers answer +0.17	+ 0.07

a question arrive within a specified time of when question was posted. To further analyse this, we have predicted the model taking all features in 1, 3, 24 and 72 hour time-frames.

From the initial set of 27 features, we have performed correlation analysis, redundancy analysis and backward feature selection to obtain a core set of 8 features that sufficiently predict the long-term value of a question. In Fig 14, we found that maximum reputation answer, sum of reputation, mean reputation were highly correlated to median reputation. Median reputation has been considered to reduce the total number of branches in the dendrogram. Redundancy analysis results concluded that the minimum time-gap between answers was a redundant feature. From the remaining set of 23 features, we performed downward feature selection to find a core set of 8 features as given in Table IV. The coefficients in the figure reflect the effect of the features in predicting the high and low values of quartile page views for questions with features computed up to 3 days.

The results conclude that number of answers, sum of answer scores, number of questioners questions, length of highest scoring answer, time for highest scoring answer to arrive, number of comments on highest scoring answer, and comments on the highest reputation answerers answer impact the long-lasting value of a question. A comparison has been made with respect to baseline features. The model built with baseline features considers two features, sum of upvotes sum of downvotes of a question, and the number of favourites of the question. In Fig **??** the AUC value increases less than 2

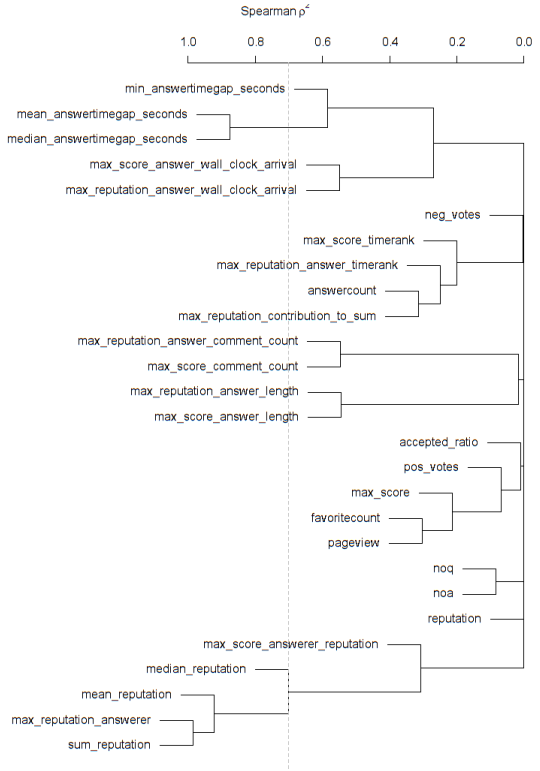

Fig. 15. Hierarchical clustering of variables according to Spearmans correlation

AUC units from one hour to three-hour time-frame. Similar is the case for accuracy value. The results demonstrate that most of the information required to predict the value of the question is obtained in 1 hour of the creation time of the question.

B. Predicting whether the question has been sufficiently answered

Bounties are questioners way of expressing whether their questions have been sufficiently answered or not. Users with reputation lower than the minimum value required by Stack

TABLE V
NUMBER OF QUESTIONS AFTER WHICH BOUNTY IS OFFERED

k	#records
1	7810
2	3208
3	1406
4	640
5	276
6	138
7	110
8	44
9	34
10	22
11	12
12	10

TABLE VI
IMPORTANT FEATURES FOR PREDICTING TASK 2

Sa‘	questioner reputation, #of questioners questions, and # of questioners answers
Sb‘	# favorites on question, maximum answer score, maximum answerer reputation, and positive and negative question votes
Sc‘	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
Sd‘	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.

Exchange for posting bounties, do not experience the freedom to stress the value of their question. For each question with bounty, we observe k number of answers after which bounty was offered; and take a random sample of questions for which an answer was accepted after k number of answers. To make the experiment fair, only the non-bounty questions, with questioners reputations lesser than the minimum required amount (75 reputation points) have been considered. The set of features Sa, Sb, Sc, Sd have been considered, taking Sa as the baseline. Since, the value of k is constant for the experiment, answerers features remain constant per value of k. Addition of answerers features have been done in the original set of features Sa, Sb, Sc, Sd to update to Sa‘, Sb‘, Sc‘, Sd‘ as shown in the table.

- Sa‘ = Sa
- Sb‘ = Sb
- Sc‘ = Sa + (Sum of Upvotes + Sum of Down-Votes + Sum Answerers Reputation for each answer)
- Sd‘ = Sd + (Time gap between answers)

Experimentation Results With different values of k, we show the results in table. It is also interesting to observe that as the value of k goes high, users rich in reputation provide bounties others questions. The result is shown in Table V.

Due to the absence of a baseline (as present in prediction task 1), we take agglomerative addition of features Sb‘, Sc‘ and Sd‘ in a baseline Sa‘ feature set to compute AUC and Accuracy metrics as shown in Fig 15.

We found that a core set of 18 features that impact this

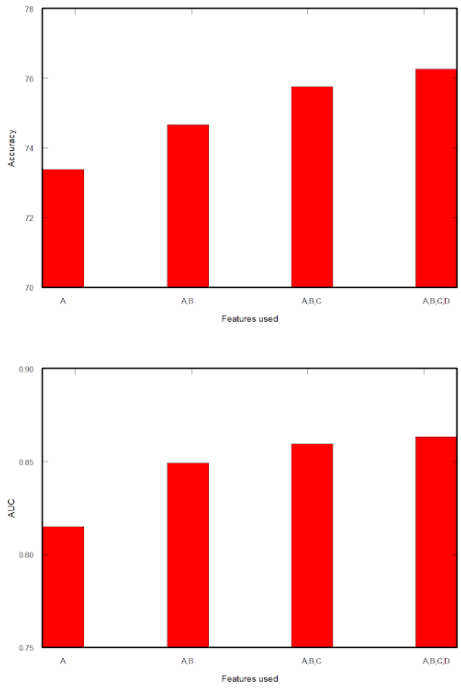

Fig. 16. Each curve shows how, given the reputation of the second answerer on a two-answer question, the likelihood of answering second deviates from a uniform baseline as a function of the reputation of the first answerer. The curves on the left (showing the bottom three reputation levels) slope downward, 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

prediction in Table VI. It is noted that addition of set Sd‘ in Sa‘b‘c‘ increases both Accuracy and AUC by less than 2 units, indicating low impact of time-based features in predicting whether a question has been sufficiently answered.

VIII. CONCLUSION

The paper provides insight on Q&A websites where questions not only provide immediate ailment to the questioners concerns but provide long term value information store for the community. The paper discusses community processes and their implications in question answering. We also find the features important for predicting the long-lasting value of a question, and the features that implicate whether a question has been adequately answered. As information content is increasingly growing in question-answering sites, the paper helps in providing insights in terms of answer content and the processes that produce them.

REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, “Knowledge sharing and yahoo answers: everyone knows something,” in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 665–674.

[2] J. Preece, B. Nonnecke, and D. Andrews, “The top five reasons for lurking: improving community experiences for everyone,” *Computers in human behavior*, vol. 20, no. 2, pp. 201–223, 2004.

[3] K. K. Nam, M. S. Ackerman, and L. A. Adamic, "Questions in, knowledge in?: a study of naver's question answering community," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2009, pp. 779–788.

[4] P. Jurczyk and E. Agichtein, "Discovering authorities in question answer communities by using link analysis," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007, pp. 919–922.

[5] J. Yang, L. A. Adamic, and M. S. Ackerman, "Crowdsourcing and knowledge sharing: strategic user behavior on taskcn," in *Proceedings of the 9th ACM conference on Electronic commerce*. ACM, 2008, pp. 246–255.

[6] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 221–230.

[7] Q. Liu, E. Agichtein, G. Dror, E. Gabrilovich, Y. Maarek, D. Pelleg, and I. Szpektor, "Predicting web searcher satisfaction with existing community-based answers," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011, pp. 415–424.

[8] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan, "Predictors of answer quality in online q&a sites," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 865–874.

[9] Y. Liu, J. Bian, and E. Agichtein, "Predicting information seeker satisfaction in community question answering," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 483–490.

[10] E. Agichtein, Y. Liu, and J. Bian, "Modeling information-seeker satisfaction in community question answering," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 2, p. 10, 2009.

[11] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, "A framework to predict the quality of answers with non-textual features," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006, pp. 228–235.

[12] C. Shah and J. Pomerantz, "Evaluating and predicting answer quality in community qa," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010, pp. 411–418.

[13] R. Kumar, Y. Lifshits, and A. Tomkins, "Evolution of two-sided markets," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 311–320.

[14] F. Wu and B. A. Huberman, "Novelty and collective attention," *Proceedings of the National Academy of Sciences*, vol. 104, no. 45, pp. 17 599–17 601, 2007.

[15] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, "Characterizing and modeling the dynamics of online popularity," *Physical review letters*, vol. 105, no. 15, p. 158701, 2010.

[16] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[17] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee, "How opinions are received by online communities: a case study on amazon. com helpfulness votes," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 141–150.

[18] J. Leskovec, D. P. Huttenlocher, and J. M. Kleinberg, "Governance in social media: A case study of the wikipedia promotion process." in *ICWSM*, 2010, pp. 98–105.

[19] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 641–650.

[20] Y. R. Tausczik and J. W. Pennebaker, "Predicting the perceived quality of online mathematics contributions from users' reputations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 1885–1888.

[21] H. Oktay, B. J. Taylor, and D. D. Jensen, "Causal discovery in social media using quasi-experimental designs," in *Proceedings of the First Workshop on Social Media Analytics*. ACM, 2010, pp. 1–9.

[22] "Database schema documentation for the public data dump and sede." [Online]. Available: https://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede

[23] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Discovering value from community activity on focused question answering sites: a case study of stack overflow," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 850–858.