2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Analysis of the Reputation System and User Contributions on a Question Answering Website: StackOverflow

Dana Movshovitz-Attias*
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA
Email: dma@cs.cmu.edu

Yair Movshovitz-Attias*
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA
Email: yair@cs.cmu.edu

Peter Steenkiste
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA
Email: prs@cs.cmu.edu

Christos Faloutsos
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA
Email: christos@cs.cmu.edu

***Abstract*—Question answering (Q&A) communities have been gaining popularity in the past few years. The success of such sites depends mainly on the contribution of a small number of expert users who provide a significant portion of the helpful answers, and so identifying users that have the potential of becoming strong contributers is an important task for owners of such communities.**

We present a study of the popular Q&A website StackOverflow (SO), in which users ask and answer questions about software development, algorithms, math and other technical topics. The dataset includes information on 3.5 million questions and 6.9 million answers created by 1.3 million users in the years 2008-2012. Participation in activities on the site (such as asking and answering questions) earns users reputation, which is an indicator of the value of that user to the site.

We describe an analysis of the SO reputation system, and the participation patterns of high and low reputation users. The contributions of very high reputation users to the site indicate that they are the primary source of answers, and especially of high quality answers. Interestingly, we find that while the majority of questions on the site are asked by low reputation users, on average a high reputation user asks more questions than a user with low reputation. We consider a number of graph analysis methods for detecting influential and anomalous users in the underlying user interaction network, and find they are effective in detecting extreme behaviors such as those of *spam* users. Lastly, we show an application of our analysis: by considering user contributions over first months of activity on the site, we predict who will become influential long-term contributors.

I. Introduction

Question answering (Q&A) sites provide a platform for online users to share and exchange knowledge on a variety of topics. They are part of many knowledge sharing domains, such as blogs, wikis and video sharing networks. Some sites, e.g., Yahoo! answers, encourage users to ask questions on any topic while others, such as StackOverflow (SO) and Quora, are specialized communities focused on more specific domains. All knowledge sharing sites build on the power of human expertise and the motivation of individuals to provide answers and exchange information, however, participation in technically oriented sites such as StackOverflow requires a high level of understanding of their domain

Q&A sites provide long lasting value not only to active users who ask and answer questions. They are also an archive of knowledge organized around specific queries that can later be accessed, for example through web search, by many. Studies find varying answer quality especially when comparing answers on paid versus free Q&A sites [1], motivating the need to identify and incentivize the participation of *expert* users who can provide high quality answers. It has been claimed these experts are the main source of answers, as well as of helpful answers in many knowledge sharing communities [2], [3].

On StackOverflow expertise and user participation is recognized and rewarded through a detailed reputation system. Users gain reputation for asking good questions, answering helpful answers, voting on the answer/question quality of others, and through several other site activities (see Section IV-A for details). This reputation scheme facilitates an easy recognition of experts as users with high reputation. Moreover, the site's reputation system has evolved since it was launched in order to reward users who answer questions more than the users who ask. In an effort to incentivize experts, the SO community also offers users greater privileges in site management as they accumulate more reputation, as well as other bonuses (e.g., honorary qualification badges) to experienced users. It is clear then, that there is a need in identifying and encouraging the activity of experts, and differentiating between levels of users.

Knowledge sharing websites, including Q&A sites, are often studied similarly to social networks where traditional *friendship* relationships are replaced with interactions leading to information exchange. User interactions on StackOverflow are initiated by a user asking a question, they typically continue when another user answers the question, and may extend further through the exchange of insightful comments. Additionally, a user who asks a question can indicate which was the most helpful answer, and other users can vote on whether they find any answer useful. Analysis of the graph emerging from the different types of user interactions provides insight into the activity patterns of users, and in particular, of experts.

In this paper, we provide an analysis of user interaction and participation on StackOverflow. We analyze the reputation scheme used on SO and the distribution of user reputation, and we find the effects that changes in this scheme have had over

* Equal contribution

ASONAM'13, August 25-29, 2013, Niagara, Ontario, CAN

the past four years since the site was launched. We examine the contribution of SO users to the system over time from the moment of creating an account, describing the different activity patterns of experts versus non-experts. We explore the use of PageRank and Singular Value Decomposition as indicators of user expertise and highlight their importance in detecting anomalous users. Finally, building on the evidence from our analysis we approach the task of identifying potential expert users based on their activity in the first few months of activity on the site. Our results indicate that experts can be reliably identified based on their site participation in the first month.

II. Related Work

Knowledge sharing networks play an important role in the daily activities of many Internet users and so have been analyzed by many studies. Cheng et al. [4] measured the statistics of YouTube videos, such as growth trend and active life span. In [5] the authors studied the Blogspace, showed the formation of micro-communities, and detected bursty communities of blogs that are topically and temporally focused. In contrast Leskovec et al. [6] found that blogs do not exhibit a bursty behavior but a weekly periodicity. Richardson and Domingos [7] mined a consumer review website to choose viral marketing plans.

Recently, more attention has been given to analysis of Q&A based communities. Adamic et al. [8] analyzed the Yahoo! Answers Q&A forum to understand the knowledge sharing activity in its different forum categories. They clustered categories according to their content and the patterns of interactions of their users, and predicted within a given category if an answer will be chosen as the best answer by the question asker. Q&A based knowledge sharing communities can also be studied in the context of information retrieval, in which a question is a query, and the answers are its results. When viewed in this perspective several works have tried to find the answers that are most relevant for a given question [1], [9], [10].

StackOverflow, and other members of the Stack Exchange network, form one of the most popular Q&A based knowledge sharing communities on the web, and as such have been the topic of a number of studies. Tausczik and Pennebaker [11] showed that user reputation is consistently related to the perceived quality of their answer. In [12], Anderson et al. did a through analysis of StackOverflow's knowledge creation process, and predicted the long-term value of a question and answers session and whether a question has not yet been able to receive a good answer. Most related to our work are the studies by Pal et al. [3], [13] in which the authors identified expert users on StackOverflow and on the TurboTax Live community, a Q&A service focused on tax related discussions. Their method is based on a probabilistic model that captures the selection bias users have in answering questions. Their data consisted of 2 types of *experts*: (1) a hand-labeled set of 100 users, and (2) the top 10% of users who answered more than 10 questions. In Section V we compare their results to our predictions despite the difference in our definition of expert users.

Expert identification approaches have been studied for other knowledge sharing networks, e.g., newsgroups [14] and email networks [15]. Matrix and tensor operations have been instrumental in identifying patterns of influential users in graph-like networks [16], [17]. Social networks, question answering networks and the Web can be formalized in the form of adjacency matrices representing interactions between users, or links between pages. One can then apply link analysis methods such as PageRank [18], HITS [19] and variants of those [20], [21], [22] to find hubs and other types of influential nodes. Multivariate analysis, SVD [23] and other spectral analysis methods [24], [25] are useful in finding patterns and outliers, for example, the network value of a user is closely related with the first eigenvectors of the network adjacency matrix [26]. Bouguessa et al. [27] used a feature-based approach, where they modeled user authority scores as a mixture of gamma distributions, based mainly on the number of answers provided by a user as a measure of their expertise. In this work, we explore both the use of spectral methods as an indicator of expertise as well as a feature-based model of user interactions on StackOverflow in the first months of their activity since joining the site.

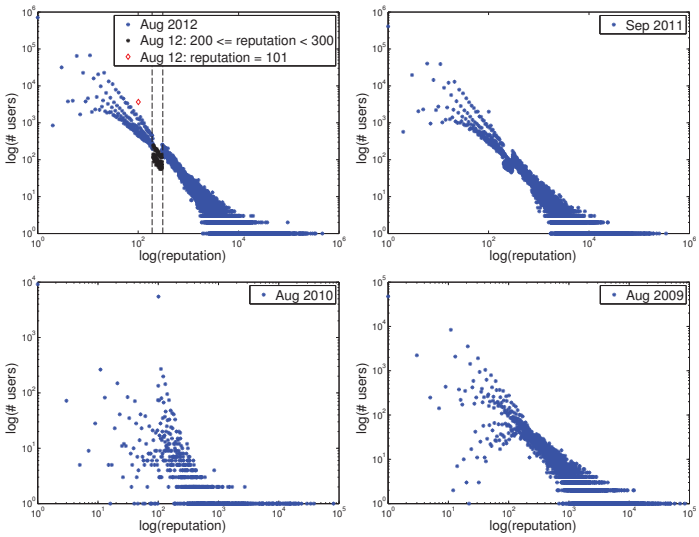


Fig. 1: Number of users versus user reputation (log scale) over the years 2009-2012. A change in the reputation rules starting Mar 2010 lead to different reputation distributions in later years. Several deviations from the log-logistic pattern of the reputation function are highlighted on the data from August 2012. Best viewed in color.

III. Dataset Description

We used data from the popular Q&A website StackOverflow which allows users to ask and answer questions related to software development, tools and other technical areas such as math and algorithms. We downloaded a complete dataset of actions performed on the site since it was launched in August 2008 until August 2012. The data includes 3,453,742 questions and 6,858,133 answers posted by 1,295,620 users. In order to analyze the change in the system characteristics over the years, we also downloaded three additional similar data dumps from the past 4 years including snapshots of StackOverflow up to the dates September 2011, August 2010 and August 2009.

Action	Reputation change
Answer is voted up	+10
Question is voted up	+5
Answer is accepted	+15 (+2 to acceptor)
Question is voted down	-2
Answer is voted down	-2 (-1 to voter)
Experienced Stack Exchange user	onetime +100
Accepted answer to bounty	+bounty
Offer bounty on question	-bounty

TABLE I: StackOverflow reputation scheme. Users are rewarded more reputation for giving good answers than for asking good questions.

To indicate the quality of answers, Stackoverflow allows a user that asked a question to select one of the posted answers as an accepted answer, suggesting that this is the most helpful response. Similarly, other users can upvote any answer as an indication of its helpfulness. The dataset contains 2,148,455 accepted answers (meaning 62% of questions have an accepted answer) and a total of 5,542,193 votes for questions and 13,058,295 votes for answers.

Performing actions on the site earns users reputation, which is an indicator of the value of that user to the site and is officially described as a "measurement of how much the community trusts you"[1]. Our data includes the reputation of each user (ranging between a reputation of 1 for a starting user and 465,166 for the user with the highest reputation in the system in the data dump from August 2012). The reputation scheme used by StackOverflow is described and analyzed in detail in Section IV-A.

IV. Characterization of user interaction and Activity

In this section we analyze patterns of user activities and interactions on StackOverflow. We later show how this analysis can be used for identifying expert and helpful users in the site. We describe the role of the system's reputation scheme in incentivizing users to post helpful answers. We consider the activity of users on the site from the moment of creating a user account and throughout the years of using the site, showing that early activity is indicative of long-term contributing users. Finally, we provide a PageRank and SVD analysis over the SO interaction network and identify anomalous users in the system.

A. Reputation Scheme

StackOverflow users can perform a variety of actions on the site, including, asking and answering questions, commenting on answers and question, acknowledging helpful answers by selecting an accepted answer or up/down voting answers, and selecting favorite questions which allows to easily access them in the future. By participating in these activities users gain reputation according to the scheme detailed in Table I, which is described on the site[1].

As can be seen in Table I, reputation is gained mainly when the user's answers are selected as accepted, upvoted, or are the answers to a question with a bounty (reputation that transfers directly from one user to another). Using this reputation scheme means that users with high reputation are normally users that provided many helpful answers, and so we consider a user reputation to be a measure of their expertise. The top 1% of users (13087 users) have reputation greater or equal to 2400 and we consider them to be the current expert users of StackOverflow.

Figure 1 shows the distribution of users over user reputation in intervals of one year over the past four years. A significant change in the distribution can be seen between 2009 and 2010 that can be explained by a change in the reputation rules which was aimed at rewarding users who provide the best possible answers, rather than users that ask good questions. The main change, which was implemented as a retroactive recalculation of reputation score for all users, included lowering the reputation bonus given for an upvoted question from +10 to +5, while the bonus for upvoted answers remained at +10. This change was meant to discourage users who ask many questions simply to gain reputation through upvotes, while still rewarding those users that provide helpful answers, supporting our assumption that under the current reputation system reputation can be considered as a measure of expertise. In Section IV-C we show that despite this change in the reputation scheme, some users are still able to achieve a relatively high reputation score by mainly asking a lot of questions.

All four distributions in Figure 1 show a log-logistic pattern with some notable deviations. Mainly, the lower-end of each of the distributions (reputation ≤ 200) is discretized showing a mixture of a number of log-logistic functions, each formed by the different possible ways of earning reputation. This suggests that new users to the system, that have not yet accumulated significant reputation, follow a number of different participation habits which lead them to earning reputation in different ways. For example, some users may start out asking many questions, while others may be initially occupied only with providing answers. We propose that early participation patterns can indicate who will become a significant contributor of helpful answers over time, and in Section V we use this information to predict expert users.

StackOverflow is one of the more popular of several Stack Exchange question answering websites which follow a similar interaction format, where the other websites include questions on more specific technical areas such as math, latex, electronics and more. Experienced users, that have gained at least 200 reputation on one of their Stack Exchange accounts, are rewarded with a one-time 100 reputation increase to each account. This rule is meant to encourage experienced users to participate in more than one Stack Exchange community, and it leads to two anomalies in the reputation distribution of users as highlighted in Figure 1(Aug 2012). First, users that have gained 200 reputation on a site other that StackOverflow and later open a StackOverflow account will have a starting reputation of 101, making this reputation more frequent than expected according to the log-logistic function (marked by a red $\Diamond$). Similarly, users that have first gained 200 reputation on StackOverflow and only later opened a second account in another Stack Exchange site will receive a 100 reputation increase to their StackOverflow reputation which leads to a lower-than-expected number of users with reputation between 200-299, as highlighted by the dashed lines in the figure. The Stack Exchange cross-site bonus was launched at July 2009,

[1] http://stackoverflow.com/faq#reputation

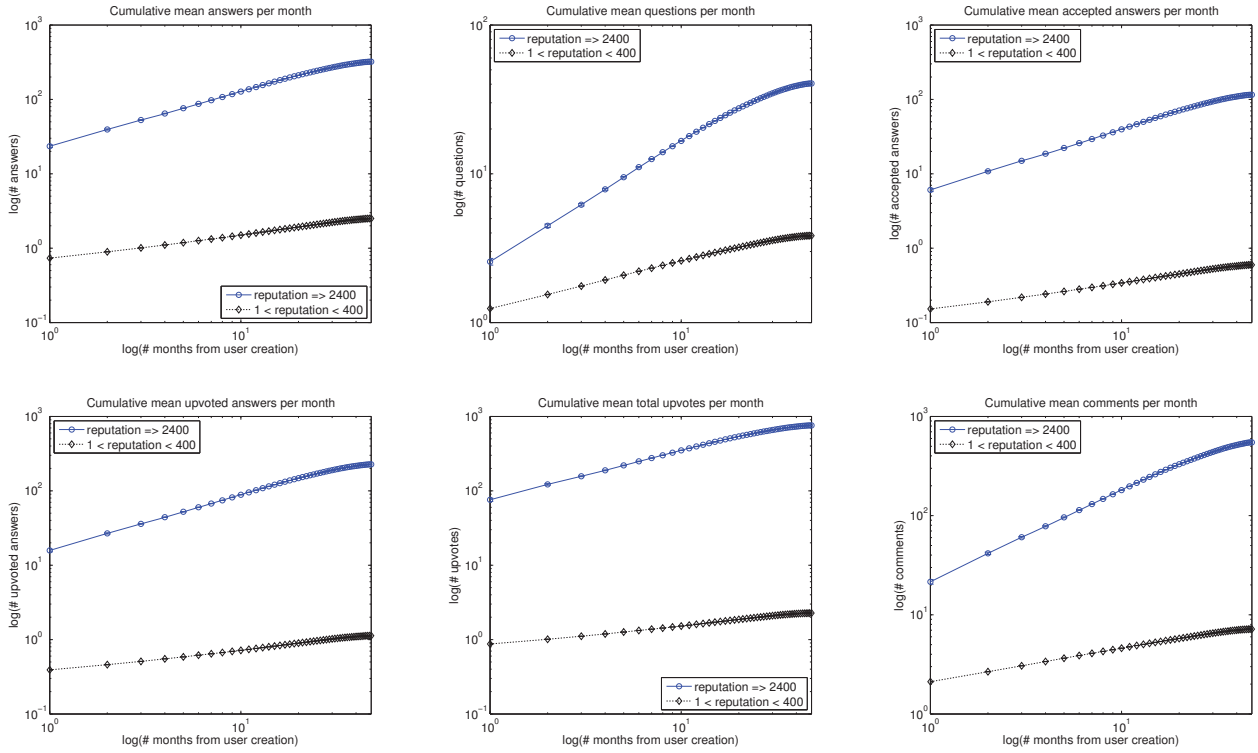


Fig. 2: Mean cumulative user contribution (answers, questions, accepted answers, upvoted answers, total upvotes, and posted comments) for high (*) and low ($\diamond$) reputation users in their activity months since the creation of the user account, with a confidence interval for $\alpha = 0.01$ ($CI < 1$ for most data points). There is a clear difference in the level of contribution of high and low reputation users throughout their time on the site. The initial months of activity are indicative of the long-term level of contribution of the user.

and so this pattern is more prominent only in later years, especially in the distributions from years 2011 and 2012.

B. User Contribution Over Time

In the previous section we examined the StackOverflow reputation system, and concluded that new users to the system follow different activity patterns which lead to varying reputation gain patterns. We now propose that different initial participation can indicate who will become a major contributor to the site over time.

Next, we analyze the contribution of users to the site from the moment they create an account and throughout their use of the system. Contribution is measured by analyzing the participation of users through posting questions, answers and comments, as well as the value of their activity as measured by the number of their answers selected as accepted answers or upvoted and the total number of upvotes they received. The contributions are evaluated over each month of activity of the users starting from the moment they created a StackOverflow account.

Figure 2 shows the mean cumulative user contribution in each of these measures, distinguishing high reputation users (with reputation ≥ 2400) from users with low reputation (< 400). We exclude from this analysis users with reputation 1 that have not performed a single reputation-gaining activity since creating an initial account. In the figure, we can see the different contribution patterns of high and low reputation users with respect to each of the activities measured, where high reputation users contribute more activity and value. This observation is consistent over time and notably also over the first months of activity, supporting our proposal that new user activity is indicative of their long-term contribution.

Interestingly, each of the activity curves exhibits log-linear growth, indicating that both high and low reputation users follow a predictable pattern of behavior when using the site, where different patterns are consistent within the high/low reputation communities. Note also that while the gap in the number of answers and helpful answers contributed in the first months by high reputation users is higher than the gap for number of questions posted, high reputation users ask on average 2-3 times more questions per month than low reputation users.

C. User Interaction and Graph Connectivity

Next we provide an analysis of user interactions on StackOverflow by considering the underlying graph structure, where nodes represent users and edges represent an interaction between two users. We construct three adjacency matrices based on three types of interactions, between a user that asked a question and: 1) any user who answered ($A_{answer} \in \mathbb{R}^{m_1 \times n_1}$), 2) the user who answered the accepted answer, if any ($A_{accept} \in \mathbb{R}^{m_2 \times n_2}$), and 3) any user who answered an answer that was upvoted ($A_{upvote} \in \mathbb{R}^{m_3 \times n_3}$). Each adjacency

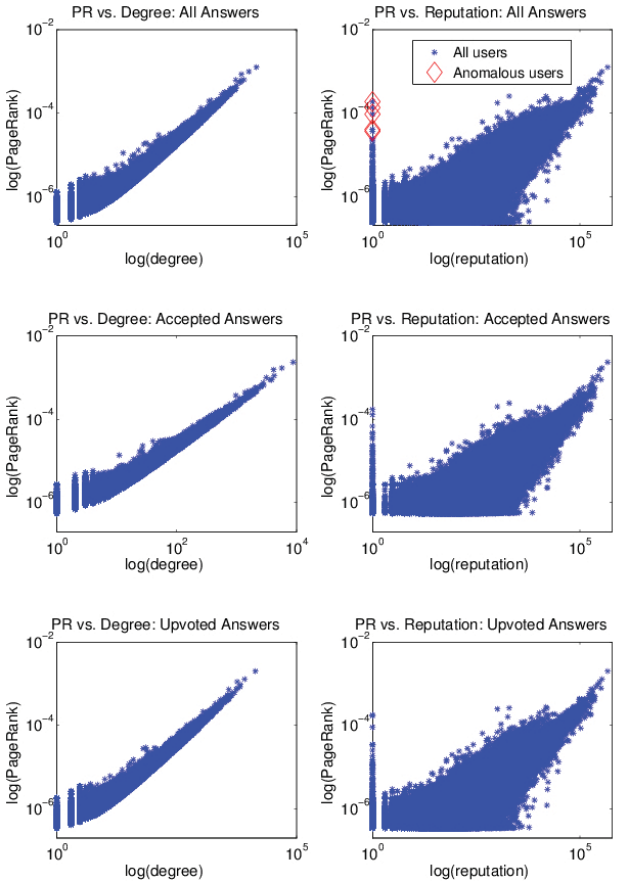


Fig. 3: PageRank versus degree (left) and reputation (right) of StackOverflow users, based on the interaction of a user who asked a question with: a user who answered (top row), the user who answered the accepted answer (middle row), and a user who answered an answer that was upvoted (bottom row). Note that the PageRank is correlated with the user degree, as expected. This is in contrast to the user reputation which is not as well correlated with the PageRank value. However, the PageRank distribution is helpful in detecting anomalous users. Some examples are highlighted of users that have high PageRank while their reputation is 1. These have been found to be *spam* users (see full discussion in Section IV-C).

matrix $A \in \mathbb{R}^{m \times n}$ represents the interaction of m questioners with n answerers.

We apply PageRank [18] to all three graphs in order to identify important nodes (users). The latter two graphs represent a more meaningful interaction, since the answerer is acknowledged of providing useful information, and therefore we might expect the PageRank values to be more indicative of "helpful" users. Figure 3 shows the PageRank of SO users versus their degree and reputation in the three interaction graphs. Note the similarity of PageRank distribution over either degree or reputation in all three graphs, suggesting that this measure is more directly affected by the volume of a user's answers than by their usefulness. This can also be seen by the fact that the user degree is highly correlated with their PageRank, as is expected since PageRank is calculated based on graph connectivity, however, the PageRank value is less correlated with the user reputation. While users with very high reputation (> 10000) have high PageRank, many experienced users with reputation in the thousands have considerably lower ranks.

We note the effectiveness of PageRank in detecting anomalous users in extreme cases. For example, in Figure 3 we highlight several users with high PageRank that have a reputation of 1. We examined the five highest PageRanked users with a reputation of 1, by locating those users in the online version of StackOverflow. We found that four out of the five currently have a considerably high reputation (ranging from 3K-47K) in the online site. The fifth user has a reputation of 1 and his account has been temporarily suspended due to *problematic behavior*. As detailed on the Stack Exchange blog[2], user accounts may be suspended if they are suspected of serial upvoting or downvoting, or if they pose some other disruption to the site. Suspended accounts have their reputation adjusted to 1 for the duration of the suspension. For the four users that currently have high reputation on the online version of SO, we found that they had high reputations scores in an earlier snapshot of the SO data (taken at 2011), confirming our suspicions that they were suspended during the most recent snapshot that we have, but later had their reputation restored.

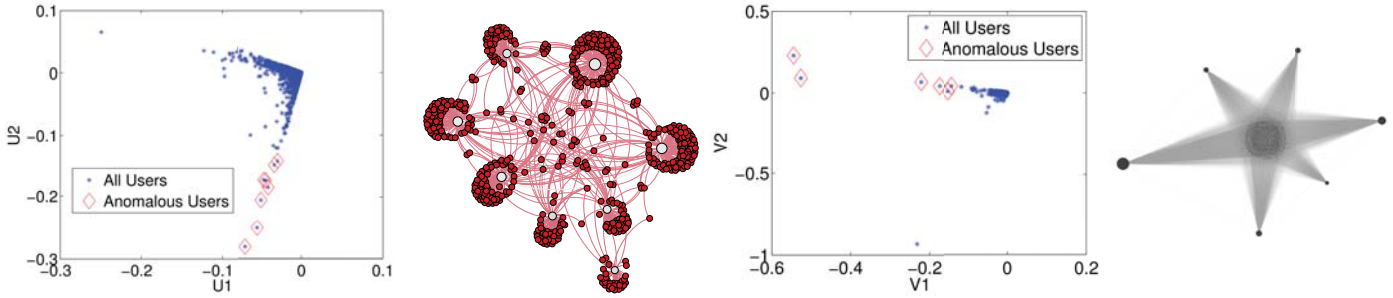
Population	Mean Z-Score
All users	-0.04
High reputation ($r \geq 2400$)	11.97
Mid reputation ($400 \leq r < 2400$)	2.35
Low reputation ($1 < r < 400$)	-0.5
Anomalous answerers	108.63
Anomalous questioners	-9.84

TABLE II: Mean Z-score ($\frac{a-q}{\sqrt{a+q}}$) for sub-populations of StackOverflow, including anomalous answerers and questioners found by an SVD analysis over the adjacency matrix A_{accept}, representing the interaction of a user asking a question with the user who answered the accepted answer. Note that the mean Z-scores of the anomalous users differ greatly than either of the background populations aggregated by reputation level.

Next we compute the Singular Value Decomposition (SVD) [23] of the StackOverflow interaction graphs. The SVD of an adjacency matrix A decomposes it into three matrices such that $A = U \times \Sigma \times V^T$. We can identify anomalous questioner users by examining pairs from the first columns of U, the eigenvectors of AA^T, and anomalous answerer users by pairs from the first columns of V, the eigenvectors of A^TA. We report our findings over the adjacency matrix of accepted answer interactions A_{accept}.

Figure 4 shows in (a) $U1$ versus $U2$, the first and second eigenvectors of $A_{accept} \times (A_{accept})^T$, highlighting a subset of eight anomalous questioner users, and in (b) the ego-net of the anomalous users. We find that the anomalous users have a relatively high reputation ranging 1161-3333, which has been earned mainly by asking many questions. As can be seen in Figure 4b the anomalous users are the centers of hubs, connected to many answering users. Following Zhang et al. [28] we examine a measure of the ratio of a user's answers (a) to

[2]http://blog.stackoverflow.com/2009/04/a-day-in-the-penalty-box/



(a) $U1$ **versus** $U2$**, first and second eigenvectors of** $A_{accept} \times (A_{accept})^T$

(b) Ego-net of anomalous questioner users.

(c) $V1$ **versus** $V2$**, first versus second eigenvalues of** $(A_{accept})^T \times A_{accept}$

(d) Ego-net of anomalous answerer users.

Fig. 4: (a) The projection of the first and second eigenvectors of $A_{accept} \times (A_{accept})^T$, highlighting eight anomalous questioner users ($\diamond$), and (b) the ego-net of these users where the anomalous questioner users (white nodes) can be found in the hub centers of connected answerer users (red nodes). All anomalous users have relatively high reputation and a higher rate of questions to answers than the rest of the SO community. (c) The projection of the first and second eigenvectors of $(A_{accept})^T \times A_{accept}$, highlighting six anomalous answerer users ($\diamond$), and (d) the ego-net of these users where the anomalous answerer users are the points of the star-shaped network containing a central hub of around 29K nodes of questioner users. Node size is proportional to its degree. All anomalous users have high reputation and a higher rate of answers to questions than the rest of the SO community.

questions (q), called Z-score, formulated as $\frac{a-q}{\sqrt{a+q}}$. In Table II we report the mean Z-scores of all StackOverflow users, of a number of sub-populations aggregated by reputation level, and of the anomalous questioners found in this analysis. We see that the anomalous population has a lower answer to question ratio than any other sub-population we have examined.

Figure 4 also shows a similar analysis using the first and second eigenvectors $(A_{accept})^T \times A_{accept}$ in (c), highlighting a subset of six anomalous answerer users, and in (d) their ego-net. The six anomalous answerers are among the highest reputation users on SO, with reputations ranging 194,943-465,166. The magnitude of the importance of these nodes is evident by their ego-net (Figure 4d), where they are the six points of the star-shaped network containing a central hub of around 29K nodes of questioner users. Table II shows the mean Z-score of the anomalous answerers group. They have an especially high rate of answers versus questions, suggesting they have gained their reputation mainly by providing many helpful answers.

We conclude that an SVD analysis is useful in detecting extreme cases of users who have been influential in the network. In the next section, we develop a classifier for early detection of potentially expert users according to their initial site activities.

V. Identifying Expert Users

We now show how the analysis we have done of user behavior and interaction on StackOverflow can be used to predict users that will go on to become experts based on their participation profile in the first few months of activity. In Section IV-B we have shown that expert users contribute more to the site than non-experts throughout their time on SO. This indicates that one can predict expert users based on their early interactions with the site.

A. Experimental Setup

We formulate this task as a classification problem. Given information of a user's activity on SO in the first N months, we classify this user into one of two classes: *expert*, or *non-expert*. As motivated above, we consider *experts* to be users who are predicted to accumulate a reputation of at least 2400 and thus make significant contributions to SO. *Non-expert* are those users who will make moderate contributions and will not ultimately accumulate high reputation. Following this definition, we filter the 1.3 million site users in our training and testing data, and leave only those that have been members for at least one year. Our ground-truth labeling is based on the reputation these users have accumulated after a year of using the site as we consider this period of time long enough to determine who has become an expert user.

From the filtered set we sample users into a training and testing set, such that in each one a third of the users have a reputation score r such that $1 < r < 400$ (low), a third have reputation $400 \leq r < 2400$ (mid), and a third have reputation $r \geq 2400$ (high). Note that we exclude from this analysis users that have reputation score of 1. As this is the starting reputation this indicates that the user performed no site activities for at least one year and so can be trivially classified as non-expert.

Based on our analysis from Section IV we designed a simple *User Activity Model* (UAM) using the following features:

- **Answers:** Number of answers the user has authored in the first $1, \ldots, N$ months of activity.
- **Questions:** Number of questions the user has authored in the $1, \ldots, N$ months of activity.
- **Accepted:** Number of answers the user has authored in the first $1, \ldots, N$ months of activity that have been accepted by the question asker.
- **Upvoted:** Number of answers the user has authored

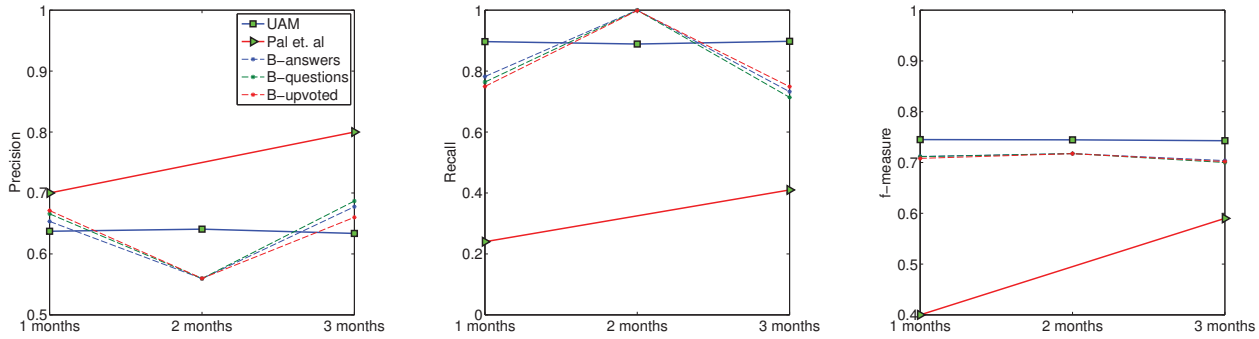


Fig. 5: Expert user classification performance. From left to right: precision, recall, and f-measure of our approach (UAM) compared with Pal et al. [13] and three baselines that use a single feature for classification (these are the most important features in the UAM classifier). Results are shown for classifiers based on user activities in their first one, two, and three months of using StackOverflow, and use a threshold of 0.5 on the predicted probability of belonging to the *expert* class (for UAM and baselines). Figure 6 shows a more detailed precision-recall analysis.

in the first $1, \ldots, N$ months of activity that have been upvoted.

- **Upvotes:** Number upvotes that the user has received in the first $1, \ldots, N$ months of activity.
- **Comments:** Number of comments the user has authored in the first $1, \ldots, N$ months of activity.
- **QA Ratio:** Ratio of questions to answers.
- **AA Ratio:** Ratio of accepted answers to answers.
- **UA Ratio:** Ratio of upvoted answers to answers.

B. Classification Performance

Figure 5 shows the performance of Random Forest Classifiers using features describing the activity of users in their first one, two, and three months of activity on StackOverflow. The results include a comparison with three baseline classifiers which use only a single feature – the selected features are the most important features in the UAM classifier based on 3 months of user activity. All results are based on a threshold of 0.5 on the predicted probability of belonging to the *expert* class. We include a comparison with the results of Pal et al. [3] who predict expert users on SO using a Question Selection Model [13], following the hypothesis that expert users are more selective when choosing questions to answer in order to maximize the amount of help they provide. We report precision (p), recall (r) and f-measure ($\frac{2pr}{p+r}$) for each of the models.

Using a prediction threshold of 0.5, our classifier consistently achieves higher recall and higher f-measure, but lower precision than Pal et al., over data from each of the first three months. Figure 6 includes a more detailed precision-recall analysis. It shows the precision and recall values given by varying the threshold on the predicted probability of belonging to the *expert* class from 0 to 1, with an Area Under the Curve of 81%. We include the precision and recall values reported by Pal et al. using 1 and 3 months of user activity. As can be seen in the figure, for a given precision/recall our classifier achieves higher recall/precision.

It is interesting to note that our classifier achieves similar results in precision, recall and f-measure when using the

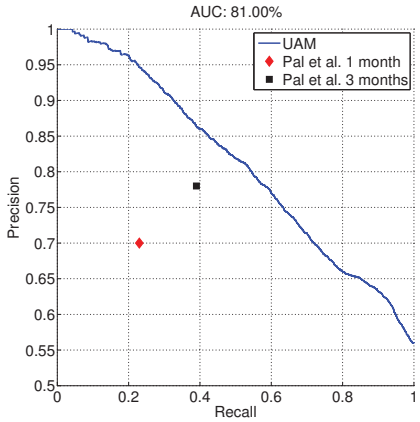


Fig. 6: Precision-Recall curve for the UAM classifier using the first 3 month of user activity. Our results are compared with the precision and recall reported by Pal et al. [13] using 1 (◇) or 3 (□) months of user activity.

activity of users from the first one, two or three months of activity. This observation is in line with the results described in Section IV-B which show a log-linear growth in user activity for all the presented measures, for both high and low reputation users. This suggests that users follow a consistent activity pattern, at least in the first months of activity, meaning that their activity in the first month is as indicative of their expertise as their activity in the first two months. Considering this data for a practical setting of identifying expert users indicates that experts can be reliably identified within a month of site usage.

VI. Conclusions

Q&A based knowledge sharing communities are becoming a principal source of information for Internet users. Understanding the way users interact on these sites is important in order to facilitate the flow of information between users. Several studies have shown that the majority of answers on knowledge sharing sites are authored by a small group of experts. Detecting expert users soon after they start interacting with the site can help site owners improve the information pro-

duced by their community by promoting the experts' answers, or help them retain these users by offering special privileges.

In this paper, we presented an analysis of the StackOverflow community with data from the initial launch of the site in 2008 until 2012. Our study focuses on the behavior and contribution of expert users to the site versus non-experts. We examined the reputation scheme used by StackOverflow which is the current method for rewarding expert users. We present an analysis of user participation patterns, showing that expert users differ in their interaction profile from non expert users from their very first actions on the site. They ask more questions, answer many more questions, and their answers are more likely to be accepted or upvoted. We find that both experts and non-experts exhibit log-linear growth in their engagement on the site, suggesting that their initial activity when joining the site is indicative of their long-term contribution. Using SVD analysis of the interaction graph underlying the StackOverflow network, we detect users with extreme ratios of answers versus questions, and we demonstrate that a PageRank analysis of this network is not well correlated with user expertise, but is effective in detecting anomalous users.

Leveraging the different behavior patterns of experts versus non-experts we have designed a classifier that detects expert users based on their early activity on the site with recall of 0.7 for precision of 0.7. We expect that more analysis of StackOverflow user activity will indicate more ways in which expert users differ from non-experts, and allow for earlier detection of experts.

ACKNOWLEDGMENT

The authors would like to thank StackOverflow for providing the data used in this paper.

REFERENCES

[1] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan, "Predictors of answer quality in online q&a sites," in *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. ACM, 2008.

[2] H. T. Welser, E. Gleave, D. Fisher, and M. Smith, "Visualizing the signatures of social roles in online discussion groups," *Journal of Social Structure*, vol. 8, 2007.

[3] A. Pal, F. M. Harper, and J. A. Konstan, "Exploring question selection bias to identify experts and potential experts in community question answering," *ACM Transactions on Information Systems (TOIS)*, vol. 30, 2012.

[4] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of youtube videos," in *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*. IEEE, 2008.

[5] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," *World Wide Web*, vol. 8, 2005.

[6] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading behavior in large blog graphs," *SIAM International Conference on Data Mining (SDM)*, 2007.

[7] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002.

[8] 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.

[9] 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 on Information Retrieval*, 2011.

[10] 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.

[11] Y. R. Tausczik and J. W. Pennebaker, "Predicting the perceived quality of online mathematics contributions from users' reputations," in *Proceedings of the 2011 annual conference on Human factors in computing systems*. ACM, 2011.

[12] 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.

[13] A. Pal and J. A. Konstan, "Expert identification in community question answering: exploring question selection bias," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010.

[14] D. Fisher, M. Smith, and H. T. Welser, "You are who you talk to: Detecting roles in usenet newsgroups," in *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, vol. 3. IEEE, 2006.

[15] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom, "Expertise identification using email communications," in *Proceedings of the twelfth international conference on Information and knowledge management*. ACM, 2003.

[16] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos, "Less is more: Compact matrix decomposition for large sparse graphs," *Computer Science Department*, 2007.

[17] J. Sun, D. Tao, and C. Faloutsos, "Beyond streams and graphs: dynamic tensor analysis," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," in *Stanford InfoLab*, 1999.

[19] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, 1999.

[20] G. Zhou, K. Liu, and J. Zhao, "Topical authority identification in community question answering," *Pattern Recognition*, 2012.

[21] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010.

[22] T. Cheng, X. Yan, and K. C.-C. Chang, "Entityrank: searching entities directly and holistically," in *Proceedings of the 33rd international conference on Very large data bases*. VLDB Endowment, 2007.

[23] G. H. Golub and C. F. Van Loan, *Matrix computations*. Johns Hopkins University Press, 1996, vol. 3.

[24] R. Kannan, S. Vempala, and A. Vetta, "On clusterings: Good, bad and spectral," *Journal of the ACM (JACM)*, vol. 51, 2004.

[25] P. Indyk, N. Koudas, and S. Muthukrishnan, "Identifying representative trends in massive time series data sets using sketches," in *Proceedings of the 26th International Conference on Very Large Data Bases*, 2000.

[26] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001.

[27] M. Bouguessa, B. Dumoulin, and S. Wang, "Identifying authoritative actors in question-answering forums: the case of yahoo! answers," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008.

[28] 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.