8-2013

An empirical analysis of a network of expertise

LE TRUC VIET
Singapore Management University, trucviet.le.2012@phdis.smu.edu.sg

Minh Thap NGUYEN

An Empirical Analysis of a Network of Expertise

Truc Viet Le
School of Information Systems
Singapore Management University
80 Stamford Rd., Singapore 178902
E-mail: trucviet.le.2012@phdis.smu.edu.sg

Minh Thap Nguyen
School of Information Systems
Singapore Management University
80 Stamford Rd., Singapore 178902
E-mail: mtnguyen.2012@phdis.smu.edu.sg

***Abstract*—In this paper, we analyze the network of expertise constructed from the interactions of users on the online question-answering (QA) community of Stack Overflow. This community was built with the intention of helping users with their programming tasks and, thus, questions are expected to be highly factual. This also indicates that the answers one provides may be highly indicative of one's level of expertise on the subject matter. Therefore, our main concern is how to model and characterize the user's expertise based on the constructed network and its centrality measures. We used the user's reputation established on Stack Overflow as a direct proxy to their expertise. We further made use of linear models and principal component analysis for the purpose. We found out that the current reputation system does a decent job at representing the user's expertise and that focus matters when answering factual questions. However, our model was not able to capture the other larger half of reputation which is specifically designed to reflect a user's trustworthiness besides their expertise. Along the way, we also discovered facts that have been known in earlier studies of the other/same QA communities such as the power-law degree distribution of the network and the generalized reciprocity pattern among its users.**

I. Introduction

The web has undoubtedly given rise to new forms of knowledge production on an unprecedented scale that involves the mass collaboration among its users. One of the most interesting forms is the online question-answering (QA) community. QA communities serve an essential role in the production of informal knowledge, one that focuses on users helping one another. Stack Overflow is such an online QA community that provides a popular platform where programmers from a wide range of expertise post and answer questions related to various programming tasks. Understanding how an online QA community in general, and Stack Overflow in particular, is used could help better improve the user experience on these sites such as recommending questions to expert users in order to reduce the response time gap. For example, recently, Treude *et al.* [1] manually labeled 385 questions in Stack Overflow and group them into 10 categories based on their contents. They also analyzed how tags are used on Stack Overflow.

In this empirical study, we wish to better understand how users interact with one another on the website through rigorous analysis of the network constructed from the interactions of users on Stack Overflow. However, our primary goal is to characterize a user's expertise in the community using centrality measures of the constructed network. Through this, we wish to investigate how much the network accounts for its user's expertise and what this has to say about the current reputation system in use on Stack Overflow.

II. Related Work

Zhang *et al.* [2] were perhaps the first to study online QA communities from the perspective of network science. The authors modeled the Sun Java Forum as a directed network containing $13,739$ nodes and $333,314$ edges, where each node is a user and an edge describes the questioning/answering relationship between them. They found out that the network exhibits an uneven bow tie structure with many more askers than answerers. Furthermore, the indegree distribution follows the power law but the outdegree counterpart does not. The authors also used centrality measures to classify the users into five levels of expertise. They found out that the simple centrality measures such as degree centralities correlate significantly with the rankings produced by two human experts who were hired to perform the task.

Adamic *et al.* [3] followed up the study by analyzing the online QA community of Yahoo! Answers. They constructed a similar network as the Sun Java Forum's. The authors used k-means clustering to classify the dataset into three broad and non-overlapping categories: programming, marriage, and wrestling. The indegree distribution of each network (corresponding to each category) follows the power law while the outdegree does not as in [2]. The authors then used entropy to measure how a user tends to answer questions in diverse topics in order to predict how good a given user is at answering a question belonging to a certain topic. They found out that focus matters when it comes to answering factual questions (e.g., science and technology) but does not when it comes to answering discussion-typed questions (e.g., family and relationships).

Nam *et al.* [4] followed the thread by studying the Korean language QA community of Naver Knowledge-iN. In this study, apart from the characterization of user expertise, the authors were additionally interested in the behavioral aspects of the users such as motivations, roles, and usage. They discovered that altruism, learning, and competency are the frequent motivations for top answerers. In addition, since the system uses points to reward users for each correct answer and establish reputation, users do behave strategically in selecting which questions to answer to maximize their gains.

Recently, Anderson *et al.* [5] studied the Stack Overflow community with the purpose of characterizing and discovering long-lasting valued questions and answers in the community in order to promote their prominence and reduce the search effort. Their most relevant result is the proposed "reputation pyramid" model of answering behavior, i.e., when a question is posted, it is first attempted by the highly reputed users and

then less advanced users will gradually take time to answer – "high-reputation users tend to answer questions early." [5].

The most recent and related work is the one by Wang *et al.* [6], in which the authors studied the same dataset on Stack Overflow [7] but from the behavioral aspects. They discovered that most users only ask one questions (77.3%). Only about 23.1% of them ask two or more questions, and only 1.6% ask more than five questions. Moreover, about 2.3% of the users do not answer any questions and about 35.2% answer two or more questions. Only 7.8% of the users answer more than five questions. Thus, the majority of the users only ask questions but do not answer any (83.2%). In terms of reciprocity, the authors found out that users tend to help one another regardless if they have been helped by them before, i.e., that Stack Overflow tends to benefit the community as a whole. Coincidentally, Hua *et al.* [9] also recently analyzed an online professional social network of medical doctors and have similar findings in which the network does not form tightly knit communities and users tend to help one another in a generalized reciprocal manner. Hence, generalized reciprocity seems to be a common feature of many online expertise-sharing communities.

III. The Stack Overflow Dataset

We obtained a large and rich dataset from the MSR 2013 Challenge [7] that captures the interactions between all users on Stack Overflow from August 2008 to August 2012. The dataset contains more than 10.3 million posts where each post is either a question posed or an answer to some question. The total number of users involved in the dataset is approximately 1.3 million users. Because it is such a large dataset, we had to sample from it to get a smaller and more manageable subset.

A. Snowball Sampling

We used snowball sampling to sample from the original dataset. Snowball sampling was used here because it is the most appropriate sampling method to identify a hidden population out of a large sample, e.g., identifying experts in a certain field, in an exploratory research. Refer to [8] for more details on snowball sampling and its theoretical justification for hidden population identification. Let S be the final sample obtained for our study, the procedure of the sampling method is as follows:

1) Sample randomly 200 initial "seed" questions. Let this be set S_0,
2) Sample all the answers to the seed questions. Let this be set S_1,
3) For each author of those answers, sample all their questions. Let this be set S_2,
4) Finally, sample all the answers to the questions in the previous step and let this be set S_3. Then $S = S_0 \cup S_1 \cup S_2 \cup S_3$.

This procedure resulted in set S that contains $39,610$ posts (both questions and answers). The sample contains $7,248$ questions and $32,362$ answers to those questions. There are $8,978$ unique users involved in those posts and the interactions between them are represented by a network or graph G which is described in the following section.

B. Network of Interactions

We constructed the network $G = (V, E)$ that represents the interactions between users in the sampled dataset in which a directed edge $(i \to j) \in E$ implies that user j has answered at least one of user i's questions. The set of nodes V represents the set of users of degree at least 1 – i.e., isolated nodes were filtered out – and the set of edges E represents the interactions between them. In total, G has $n = 8,908$ nodes and $|E| = 27,842$ directed edges. This is thus a relatively sparse graph. A directed edge (i, j) is further weighted by the number of interactions between i and j, e.g., how many questions of i that j has answered. In other words, G is a weighed directed graph. Table I summarizes some of the basic network statistics of G. Of noteworthy here is that the average clustering coefficient is rather low at 0.056, which may be indicative of the lack of tightly knit communities in this network. This vaguely suggests that users generally do not tend to help each other to return favor (or reciprocating help).

n	$8,908$
$\|E\|$	$27,842$
Avg. Degree	3.126
Avg. Weighted Degree	3.48
Network Diameter	10
Avg. Path Length	3.889
Assortativity	-0.053
Avg. Clustering Coefficient	0.056

TABLE I. The statistics of the sampled network

Indeed, we discovered that there are only 324 reciprocated pairs of users (i, j) in G – i.e., 2.33% of the total edges. A reciprocated pair (i, j) means that i has answered at least one of j's questions and vice versa, j has answered at least one of i's questions. Given such a tiny fraction of reciprocated users, we can conclude that any given user i tends to help another user j no matter what (i.e., independent of if she has been helped by j before). In other words, Stack Overflow is a network of selfless users that benefits the community as a whole such that the notion of *generalized reciprocity* is strongly reflected here. This observation also resonates the recent findings in [6] for the same network of Stack Overflow and in [9] for a medical professional social network.

Furthermore, the assortativity coefficient measures the node similarity in the network. It is the measure of the tendency of nodes in the network to connect preferentially to other nodes that are like or unlike them in some way. The value of the assortativity coefficient ranges from -1 to 1, i.e., from disassortative to assortative. A network with zero assortativity is potentially a random graph with no clusters. Table I shows that the assortativity of our network is negative and close to zero. This means that the network is rather random (with no clusters) and is weakly disassortative. Weak disassortativity suggests that well-connected nodes (with potentially higher expertise) weakly tend to connect with other nodes with few connections (that potentially have lower expertise). Given the special characteristic of the network as a space for users to share and seek expertise, it is reasonable that more experienced and knowledgeable users tend to connect with those with less expertise. This has further confirmed the observation that the generalized reciprocal behaviors are observed among its users. This also strongly agrees with the findings in [9].

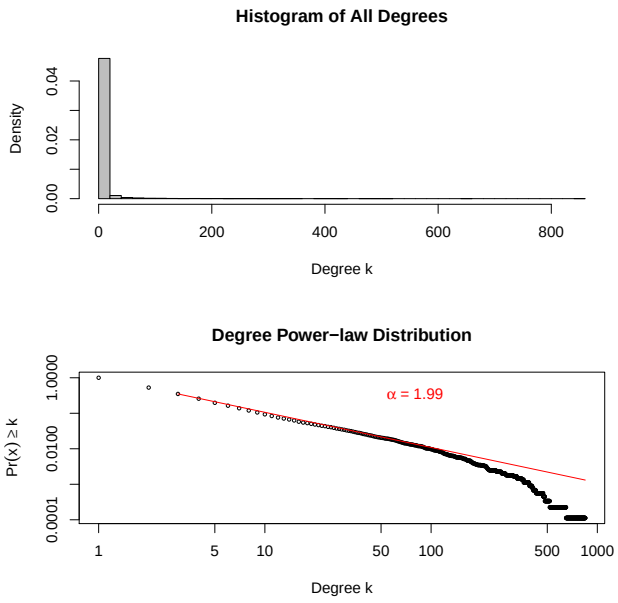

Fig. 1. The overall degree distribution of the sampled network

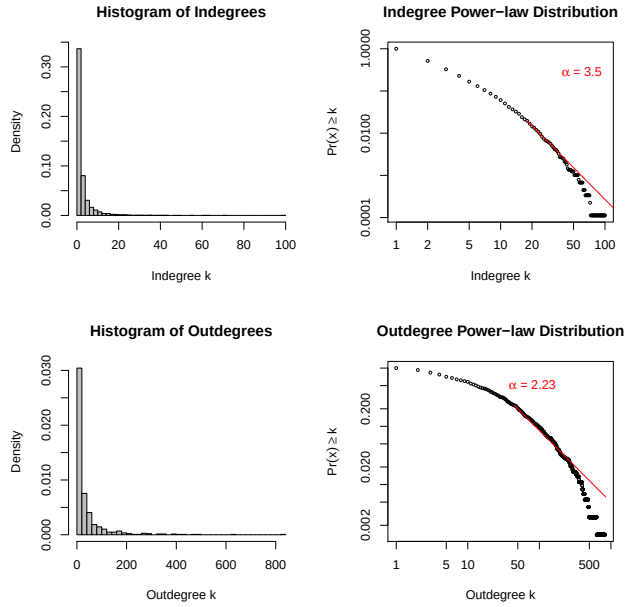

Fig. 2. The in- and outdegree distribution of the sampled network

IV. Exploratory Network Structure Analysis

It is of interest to examine the network structure of G to see how expertise is distributed. In particular, in this section, we will look into the degree distribution of G to see if it obeys the famous power-law distribution that is often found in empirical data of complex networks [10]. A power-law distribution typically features a one-sided long tail with many large-valued outliers.

Given a network $G = (V, E)$ and let x be a random variable indicating the degree of a given node $v \in V$. If the degree distribution of G obeys the power law, then the cumulative degree distribution of G is given by

$$\Pr(x \geq k) = Ck^{-\alpha}, \tag{1}$$

where C is a positive constant and α is the exponent parameter. The exponent α is typically in the range of $(2, 3)$, though not always [10]. A network whose degree distribution obeys the power law is also called a *scale-free* network [11] because α does not scale up as G grows in size. Many networks that occur in the natural world as well as man-made world, such as the World Wide Web, various biological and social networks, have been found to be scale-free [10].

Indeed, examining the degree distribution of G has revealed its scale-freed property. Due to the directed nature of the network, three kinds of degree distribution have been examined: the overall degree distribution, the indegree distribution, and the outdegree distribution. Fig. 1 illustrates the overall degree distribution of G with $\alpha = 1.99$ and Fig. 2 illustrates the indegree ($\alpha = 3.5$) (upper half) and outdegree ($\alpha = 2.23$) distribution (lower half). All the three distributions have characteristic long (right) tails as illustrated in their histograms. In both figures, the cumulative probability distributions (on the right hand side) are plotted on a log-log scale.

We further performed the non-parametric Kolmogorov-Smirnov (KS) test on the distributional difference between the empirical distribution and the parameterized power law for each kind of degree distribution. The null hypothesis is that there is no difference and the alternative is that they are different. The results are shown in Table II. In the table, $k_{\min}$ denotes the lower bound of the degree to which we wish to fit the power-law distribution $\Pr(x \geq k) \sim k^{-\alpha}$ for $k \geq k_{\min}$, α is the exponent parameter, and D is the D-statistic of the goodness of fit of the KS test. If $D > 0.05$, we reject the null hypothesis; otherwise, we accept it. Thus, the overall and indegree distribution are significantly the same as the parameterized power law at the 5% level. Whereas the outdegree distribution is different from the power law at the 5% level. This agrees well with earlier findings in [2], [3] on the power-law degree distributions of the networks constructed from the Sun Java Forum and Yahoo! Answers respectively.

	Overall	Indegree	Outdegree
$k_{\min}$	3	18	44
α	1.99	3.50	2.23
D	0.014^*	0.028^*	0.061

TABLE II. KS tests of the power-law degree distributions of the sampled network

The power-law degree distribution of the network gives us a valuable insight that if we associate a node's degree (specifically its indegree) with expertise – which is the main task of the next section, then the distribution of expert users is such that there are very few of them (characterizing by having large indegrees) and the majority of the nodes in the network are novice users trying to seek expertise from those few expert nodes. Thus, experts are rare and are highly sought after.

V. Expertise Network Analysis

A. Current Reputation System

Stack Overflow currently implements a reputation system that awards its users with points (or punishes them by sub-

tracting points) when other users vote up (or down) on their questions or answers. A minimum reputation of 1 is always maintained for all users. Apart from being an indicator of expertise, according to the website, "reputation is a rough measurement of how much the community trusts you" [12]. A highly reputed user with high reputation points have elevated privileges in moderating the website. Table III summarizes the current reputation system in use on Stack Overflow according to [5]. Refer to [5] for a more detailed description of the reputation mechanism.

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

TABLE III. STACK OVERFLOW'S CURRENT REPUTATION SYSTEM [5]

For our main task of expertise analysis, we use the reputation points (whose mechanism is given in Table III) as a direct proxy for measuring a user's expertise. This may be biased and not truthfully reflect a user's expertise in its strictest sense as there may exist strategies to improve one's reputation (as outlined in [5]). However, otherwise, we would not have any measurable means to quantify one's expertise that can be easily obtained. Hence, we resort to using reputation as the sole indicator of expertise.

B. Network Centrality Measures

From the constructed network G, for each user, we take the following seven (7) centrality measures as the explanatory variables to model the user's expertise on Stack Overflow.

Degree centralities. We define "AnsNum" as the total number of answers a user v has given. This corresponds to the total weighted indegree of node v. "Indeg" is the indegree of v in the network. This corresponds to the total number of *other* users that v has helped (i.e., answered their questions).

While replying to many questions implies that one has high expertise, asking a lot of questions is usually indicative of one's lack of expertise on some topic. Hence, we adopt the **Z-scores** defined in [2] as a centrality measure that combines both one's asking and replying patterns. Let q be the number of questions and a the number of answers a given user posts respectively, if the user makes $n = q + a$ posts, we want to be able to measure how different this behavior is from a "random" user who is equally likely to ask and answer with probability $p = 0.5$. Thus, we would expect such a random user to post $n/2$ answers with a standard deviation of $\sqrt{n}/2$. The Z-score measures how many standard deviations above or below the expected "random" value a user lies:

$$z = \frac{a - n/2}{\sqrt{n}/2} = \frac{a - q}{\sqrt{a + q}}. \tag{2}$$

If a user asks and answers about equally likely, the Z-score will be close to 0. If they answer more than ask, the Z-score will be positive, otherwise, it is negative. We calculate the Z-score for both the total number of questions one asks and answers and the number of other users one has helped and received replies from, denoted as "Z_num" and "Z_deg" respectively.

There is a potential problem in counting the number of answers one posted or the number of other users one has helped. A user who has answered, say, 100 "easy" questions will be ranked as equally expert as another who has answered 100 "advanced" questions. Apparently, the latter often has greater expertise than the former. Therefore, we further make use of **HITS Authority** (or "Auth") [13] and **PageRank** (or "PR") [14] centralities for our expertise analysis. In a nutshell, these centrality measures give more weights to those who are pointed to by other high-degree or influential users. Their differences lie in their implementations. PageRank provides a kind of peer assessment of one's centrality by taking into account not just the number of inlinks one receives from their pointers, but also the number of inlinks those pointers receive. While HITS Authority makes use of the notion of hubs – i.e., an authoritative user is pointed to by many good "hubs", e.g., those users who ask a large number and a variety of questions. These measures can be easily obtained by standard algorithms proposed by their respective inventors.

We finally use the **betweenness** centrality as the last explanatory variable. We shorthand it as "BTW" in this paper. BTW of node v measures the number of shortest paths from i to j that have to go through v for all $i, j \in V$ and $i, j \neq v$. Metaphorically, a high betweenness indicates that v has an advantageous "brokerage" position in the network such that other nodes have to go through it in order to reach one another or a "bridge" between different communities of users.

Fig. 3 shows the Pearson correlation coefficients of these 7 explanatory variables with reputation. It can be seen from the figure that AnsNum and Indeg correlate quite well with reputation at both more than 0.60, next are the PR and Auth centralities at both more than 0.50. BTW also correlates, but weakly, with reputation at about 0.26. What is surprising is that the Z-scores do not seem to correlate (well) with reputation, especially with Z_deg weakly anti-correlates with reputation (having negative coefficient). The next sections will try to model and explain these relationships.

C. Linear Regression Analysis

For the analysis in this subsection and the next, we further reduced the sample to include only users who had answered at least 10 questions in order to obtain a set of high reputation users. This drastically reduced the cardinality of the set to $n = 642$ "expert" users because of the power-law degree distribution. The reputation of these ranges from 122 (min) to $465,200$ (max). We then examined the empirical distribution of reputation and found that it is highly right-skewed with a very long (right) tail. That is, there exist a few very large-valued outliers with extremely high reputation that could significantly bias the analysis. Therefore, we manually removed those extremely highly reputed users to aid the analysis. Specifically, we removed those whose reputation exceeds $250,000$ points – there are only three of them. Thus, the final dataset contains $n = 639$ users. Moreover, all the explanatory variables in this section and the next are standardized (i.e., subtracted from the mean and then divided by the standard deviation) to account

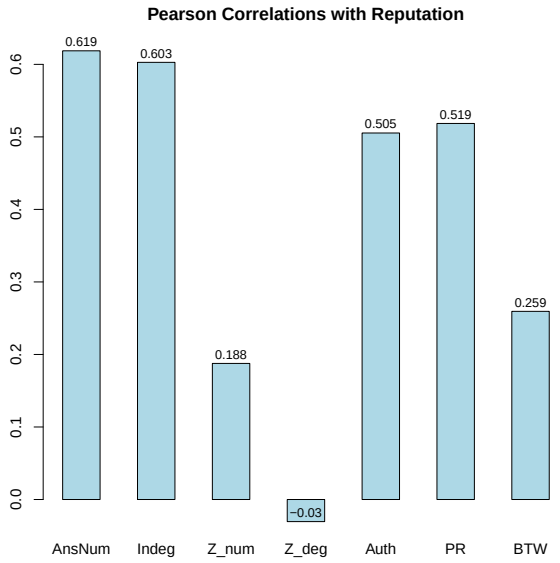

Fig. 3. Pearson correlations of the explanatory variables with with Reputation

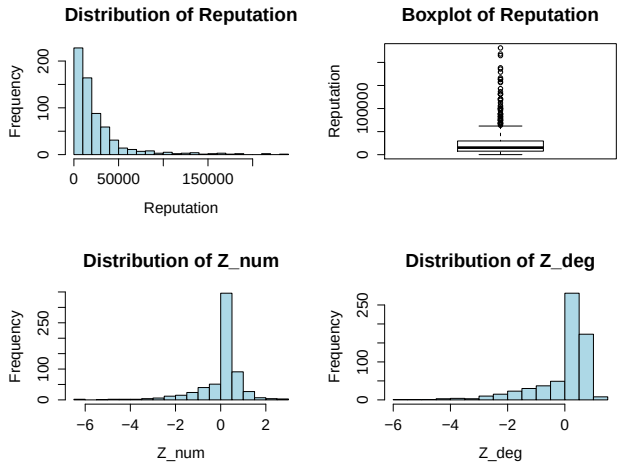

Fig. 4. Distributions of the sampled Reputation (upper) and Z_scores (lower)

for the discrepancies in units. They thus all have mean 0 and standard deviation 1.

Fig. 4 (upper half) illustrates the distribution of the re-sampled reputation. The figure shows that the distribution is still right-skewed with a long (right) tail, but it is much less extreme than otherwise. The boxplot particularly shows that there still exist numerous "outliers" on the right tail. The lower half of the figure shows the distributions of Z-scores for the (resampled) set of users. Of noteworthy is that both distributions of Z-scores show that these users are both more prone to answering questions and helping other users than randomly. This is shown by the left skews in both distributions, and especially for Z_deg.

To understand the relationship between reputation and its explanatory variables, we first performed a linear regression analysis using reputation as the response variable. Table IV shows the results of this regression. The table shows that AnsNum, Z_num, Z_deg, and PR are the significant (at the 5% level) explanatory variables for reputation, and the linear model is able to account for over 40% of the total variance. However, before attempting to interpret these results, we would like to check an important assumption of our linear model, i.e., the explanatory variables are *linearly independent* vectors. If the assumption does not hold, regression results are biased and relationships interpreted are spurious. This is often referred to as the problem of *multicollinearity* in linear models.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25079.68	970.52	25.84	< 0.01
AnsNum	27215.79	5925.51	4.59	< 0.01
Indeg	-4696.27	4668.07	-1.01	0.31
Z_num	-9275.57	3395.52	-2.73	0.01
Z_deg	9012.50	3455.33	2.61	0.01
Auth	-2569.99	2332.49	-1.10	0.27
PR	4592.86	1570.26	2.92	< 0.01
BTW	-1381.90	2257.63	-0.61	0.54

TABLE IV. LINEAR REGRESSION ON ALL PREDICTORS, $R^2 = 40.28\%$

We next look at the correlations across the explanatory variables (or *predictors*). Table V shows the Pearson correlation matrix of the seven predictors. Note that the correlation matrix is symmetrical about the diagonal with each variable perfectly correlates with itself. From the table, we see that there *is* indeed a problem of multicollinearity here if all the predictors are included into a linear model. For example, AnsNum and Indeg are highly correlated with each other (at 97%) and AnsNum and Auth also highly correlate together (at 87%). This suggests that we should only take either one of the highly correlated variables and not both to avoid multicollinearity and information redundancy. Interestingly, the table also shows that Z_deg tends to be negatively correlated with the other predictors, except Z_num. The next subsection will try to explain this phenomenon.

	AnsNum	Indeg	Z_num	Z_deg	Auth	PR	BTW
AnsNum	1.00	0.97	0.28	-0.11	0.87	0.73	0.47
Indeg	0.97	1.00	0.24	-0.13	0.78	0.76	0.50
Z_num	0.28	0.24	1.00	0.88	0.29	0.15	-0.59
Z_deg	-0.11	-0.13	0.88	1.00	-0.07	-0.15	-0.80
Auth	0.87	0.78	0.29	-0.07	1.00	0.51	0.33
PR	0.73	0.76	0.15	-0.15	0.51	1.00	0.42
BTW	0.47	0.50	-0.59	-0.80	0.33	0.42	1.00

TABLE V. PEARSON CORRELATION MATRIX OF ALL PREDICTORS

D. Principal Component Analysis

In this section, we try to solve the multicollinearity problem observed in the previous one, we also ask the question if we could further reduce the dimensionality of the dataset (i.e., reduce redundant information) to arrive at a more *parsimonious* model. We use principal component analysis (PCA) for this purpose. The basic idea of PCA is to reduce a large set of variables to a smaller one that still contains most of the information in the large set. Hence, correlated variables are projected onto a principal component (or dimension). A principal component (PC) is thus a linear combination of *optimally weighted* variables into a lower dimensional space while retains the maximal amount of information. Each PC is guaranteed to be uncorrelated with one another; as a result, this solves our multicollinearity problem.

Results of the PCA are given in Table VI with the loadings of the predictors on each PC. A *loading* of a variable on a

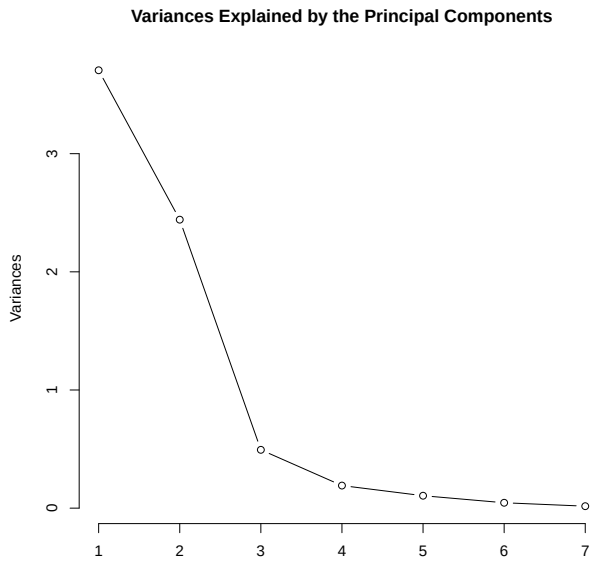

Fig. 5. Total variance accounted for by each of the principal component

PC is its projected weight onto it. In other words, it is the weight by which each standardized original variable should be multiplied to get the component score. Of noteworthy here is all the predictors have positive loadings onto the first PC (PC1), except for Z_deg. Z_num, however, has very weak positive loading onto PC1 at 0.04. The second PC (PC2) has mostly negative loadings, expect for BTW, with the Z-scores having the strongest negative weights. We leave the interpretations of these loadings until later in this section.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
AnsNum	0.50	-0.14	0.13	-0.12	0.30	-0.08	-0.78
Indeg	0.50	-0.11	-0.03	-0.36	0.51	0.16	0.57
Z_num	0.04	-0.63	-0.02	-0.24	-0.28	-0.68	0.13
Z_deg	-0.17	-0.59	-0.09	-0.25	-0.22	0.69	-0.14
Auth	0.43	-0.16	0.61	0.50	-0.32	0.15	0.19
PR	0.42	-0.06	-0.77	0.42	-0.20	0.04	0.01
BTW	0.34	0.44	0.00	-0.56	-0.61	0.05	-0.02

TABLE VI. LOADINGS OF ALL PREDICTORS ON THE PRINCIPAL COMPONENTS

Fig. 5 shows the total variance of the dataset that are explained for by each of the PC. The figure visually suggests that after the third PC, the total variance accounted for are negligible. In fact, the first three PC's combined explains for about 94.87% of the total variance. Therefore, we decided to take the first three PC's (whose loadings are given in Table VI) as the new reduced dimensions of our dataset.

We call "Score1", "Score2", and "Score3" the principal component score of the first, second, and third PC respectively. We use them as the new explanatory variables for our linear model, with reputation as the response. In other words, we perform a principal component regression (PCR) of reputation on these three scores. The results are shown in Table VII. The table shows that only the first two component scores are highly significant at the 5% level, while the third one is not. Moreover, the R^2 coefficient of goodness-of-fit of the model has been slightly reduced to 38.14% at the gain of a simpler and more parsimonious model (with fewer dimensions).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25079.68	984.59	25.47	< 0.01
Score1	9740.53	511.86	19.03	< 0.01
Score2	-3280.88	630.63	-5.20	< 0.01
Score3	-2155.89	1402.35	-1.54	0.12

TABLE VII. PRINCIPAL COMPONENT REGRESSION ON THE THREE PC SCORES, $R^2 = 38.14\%$

The coefficient estimates of Table VII indicate that, in general, having a higher value for each of the predictor (while holding everything else constant) benefits one's reputation. However, the effects are *not* equal and some boost one's reputation much more than the others. Specifically, the PR centrality has the highest positive effect on one's reputation. Next are the degree centralities with about equal strength. The third most effective is the Auth centrality, which is followed by BTW. Finally, the Z-scores also do good to one's reputation but rather moderately, and with the least effect comes from the Z_deg score.

Finally, Fig. 6 shows the biplot of the PCA that visualizes how the original predictors are projected onto the first two PC's together with all the data points (represented by the integer node id's). The figure is basically a two-dimensional plane whose axes are the first two PC's and the predictors (represented by vectors whose directions and magnitudes are given by their corresponding loadings on the axes) that are similar to one another are clustered closely together by having a small angle of separation. The plot visually suggests that there are roughly three clusters of predictors that (positively) correlate with reputation to different degrees. The strongest one is given by the cluster of the degree centralities, PR, and Auth. BTW centrality follows and forms its own cluster. The third cluster that has the weakest effect is the Z-scores with the two scores Z_num and Z_deg quite separated from each other (i.e., having a rather wide angle of separation between them) and Z_deg having the weakest effect on reputation.

In terms of expertise-sharing activity, our findings indicate that while answering more questions or helping more other users as well as being more diverse and flexible in those activities do help one establish their reputation and expertise (the former is reflected in the positive effects of degree centralities, PR, and Auth and the latter in BTW centrality), being *too* diverse may offset those benefits instead. This is represented by the weak effects of the Z-scores (and particularly of Z_deg) in the PCR analysis on reputation and this in turn explains for the notable negative correlation of Z_deg on reputation seen in Fig. 3. Having higher Z_deg represents the tendency to help more other users, which when exceeds a certain threshold might imply diffused and unfocused effort as a very large group of helpees may pose questions from a diversity of subject matters that are possibly not closely related to one another. This in turn suggests a lack of expertise as one doesn't normally have in-depth knowledge in diverse and possible unrelated subject matters. In other words, focus matters when it comes to establishing reputation and expertise in a fact-based QA community such as Stack Overflow.

VI. CONCLUSION

In this paper, we first attempted to model the interactions of users on Stack Overflow – a popular online QA community

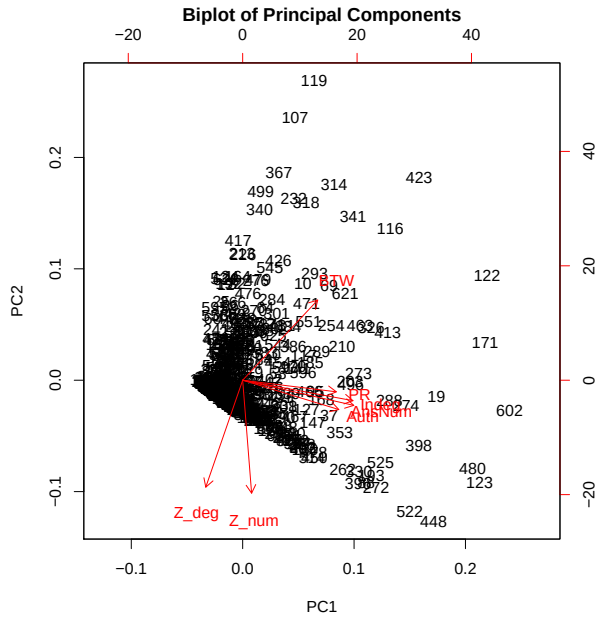

Fig. 6. Biplot of the principal components

that focuses on programming tasks – using a network of interactions between the users. The network has weighed directed edges that represent the directions and strengths of those interactions. We found out they network does not have tightly knit clusters, and that users generally help each other no matter what. This is a testimony to the good standards of the community and how it is supposed to be – to share expertise freely among its members. This result agrees with recent findings in [6] for the sample QA community and in [9] for a different community that focuses on medicine. We then analyzed the network structure. We found out that the network follows the power-law degree distribution for its overall degree and indegree. However, the outdegree distribution does not quite follow the power law. This also agrees with earlier findings in the online QA communities of Sun Java Forum [2] and Yahoo! Answers [3].

However, we were more concerned about the reputation system currently in use on Stack Overflow and how it can be characterized by the various network-centric measures. The underlying problem was how to model and explain for a user's expertise using their network of interactions and if the current reputation system was doing a good job at reflecting the user's expertise. To this end, we made use of linear models and principal component analysis. Specifically, we made use of seven easily obtainable network centrality measures to account for one's reputation on the network. These are the degree centralities, PageRank [14], HITS Authority [13], betweenness centrality, and the Z-scores proposed in [2] that represent one's tendency to answer more questions (than ask) and help more other users (than being helped). We found out that they all reflect reputation but to different degrees. Specifically, the PageRank centrality has the highest predictive power for reputation, which is just how it was designed to do – to predict influential nodes in a network. It is then followed by the degree centralities, which also makes sense as the more questions one answers or the more other users one helps does reflect one's higher level of expertise. To lesser extents are the HITS Authority and the betweenness centrality with the betweenness being less – perhaps because the network does not have distinct communities. At the lower end are the Z-scores. What is most interesting is, between the Z-scores, Z_deg has the lowest predictive power. This centrality measure reflects one's tendency to help many other users and to some extent, one's level of focus on answering the questions posed. Higher Z_deg may reflect one's lack of focus because of one's tendency to help a larger number of other users, who possibly ask questions on a diversity of topics and being less focused on one. This also makes sense because in order to be an expert at a field, one needs to focus on their own field and not being so versatile. This is particularly true for Stack Overflow as most of the questions on it are factual and highly technical. This finding also resonates an earlier finding on Yahoo! Answers [3] in which focus matters when it comes to answering factual questions.

Finally, we since PageRank is the most predictive factor of one's reputation on Stack Overflow, and in turn for one's level of expertise, we conclude that the reputation system in use on Stack Overflow does quite a decent job at reflecting one's expertise – because PageRank is a well-known and effective measure of influential and authoritative nodes on a network. However, our model was able to capture only about 40% of the reputation measure. Perhaps, the other 60% that we could not account for using network centralities comes from one's trustworthiness, which is a major part of the reputation mechanism on Stack Overflow. This in turns comes from its voting mechanism. Thus, an interesting future work on this line of research is how to represent a network of trust for online QA communities and improve the trust mechanisms on those.

ACKNOWLEDGMENT

This research is supported by the Singapore National Research Foundation (NRF) under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office.

REFERENCES

[1] C. Treude, O. Barzilay, and M.-A. Storey, "How do programmers ask and answer questions on the web?: NIER track," in *Software Engineering (ICSE), 2011 33rd International Conference on*. IEEE, 2011, pp. 804–807.

[2] 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.

[3] 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.

[4] 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 27th international conference on Human factors in computing systems*. ACM, 2009, pp. 779–788.

[5] 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.

[6] S. Wang, D. Lo, and L. Jiang, "An empirical study on developer interactions in Stack Overflow," in *Proceedings of the 28th Symposium on Applied Computing*. ACM, 2013, p. to appear.

[7] A. Bacchelli, "Mining Challenge 2013: Stack Overflow," in *The 10th Working Conference on Mining Software Repositories*, 2013, p. to appear.

[8] R. Atkinson and J. Flint, "Accessing hidden and hard-to-reach populations: Snowball research strategies," *Social research update*, vol. 33, no. 1, pp. 1–4, 2001.

[9] G. Hua and D. Haughton, "A network analysis of an online expertise sharing community," *Social Network Analysis and Mining*, vol. 2, no. 4, pp. 291–303, 2012.

[10] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[11] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[12] S. O. FAQ. (2013, Apr.) Frequently Asked Questions – Stack Overflow. http://stackoverflow.com/faq. [Online]. Available: http://stackoverflow.com/faq

[13] J. M. Kleinberg, "Hubs, authorities, and communities," *ACM Computing Surveys (CSUR)*, vol. 31, no. 4es, p. 5, 1999.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web." 1999.