Expertise and Network Properties in StackExchange Sites

Dan Guo (dguo1113), Vani Khosla (vkhosla), Vignesh Venkataraman (viggy)

I. Introduction

This project aims to look at the network properties as they pertain to the StackExchange family of online forums, including StackOverflow. Specifically, the goal of this project is to identify, understand, and designate experts and expertise within StackExchange and StackOverflow data. Because the various subdomains of StackExchange cover a wide range of material, experts can be analyzed both globally and over several different subsets of data to see if there are general patterns or trends that can be extracted. Better defining who experts are and what it means to be an expert, especially in a particularly specific subfield, would allow knowledge to be managed and distributed more judiciously throughout the community. Rating the reputation of users through the use of network properties would create a credibility system that supersedes the StackExchange sites' generic user-specific reputation scores.

II. Literature Review

There have been several papers that have already looked at classifying experts within different online communities.

"Analysis of the Reputation System and User Contributions on a Question Answering Website: StackOverflow" looked to correlate graph centrality metrics to user expertise, a hand coded label. The also paper built a classifier for classifying a user as "expert" and "non-expert" based on simple features extracted from the first three months of the user on SO. The paper also applied PageRank and SVD to understand patterns of high scorers.

The paper "Expertise Networks in Online Communities: Structure and Algorithms" studied the Java Forum question and answer network. The key modeling step in this paper's analysis was taking a set of forum posts and decomposing them into a graph - the asker of a question has outgoing edges to all of the responders to his question, thus giving users who answer lots of questions high in-degree and users who ask lots of questions high out-degree. Examining the degree distribution of the Java Forum graph shows that it is highly skewed, showing that there are some users who ask/answer tons of questions, but also a lot of users who ask/answer very few questions. The authors also applied a variant of the PageRank algorithm (dubbed ExpertiseRank) and the HITS algorithm to the Java forum graph, with mixed results as compared to basic degree and structure analyses.

The paper "Knowledge Sharing and Yahoo Answers: Everyone Knows Something" used a variety of metrics to predict if a given answer will be selected as a best answer; the metrics used to determine this relied on the category clustering, user focus, and user entropy to define what an expert user and response looks like in a particular category. In order to classify which categories were the most active, k-means clustering was applied on three different metrics. A network was then created for each category cluster by connecting users who asked a question to users who answered the question, allowing the graph to easily show which users tended to answer more questions and which users tended to ask more questions.The overall conclusion found that users who tended to be more focused on specific categories generally had better answers within the more technical categories. In addition, the results indicated that users who answered questions and users who asked questions were distinct.

III. Data

The data for this project is from StackExchange, as found here. StackExchange is a network of question and answer websites on topics in varied fields, each site covering a specific topic, where questions, answers, and users are subject to a reputation award process. Each site is modeled directly off of StackOverflow, a programming question and answer site, which was the first site in the network. The entire StackExchange platform's data, in anonymized form, is periodically posted to the Internet Archive for research purposes under a general Wiki license. The data is segregated by StackExchange website, with each subdomain (e.g. sports.stackexchange.com, english.stackexchange.com, etc.) publishing large XML data files containing Posts, Users, Votes, Comments, PostHistory and PostLinks respectively. Posts are obviously the core of the StackExchange platform - these are either posted questions or direct top-level responses to a question on the site. Users are the entities who create Posts - they have relevant metadata (i.e. email address, username) and personal reputation scores that correspond to their perceived intelligence on the platform. Votes are upvotes or downvotes on a particular post - questions and answers can both receive votes (questions for being insightful, answers for being correct and readily comprehendible. Comments are lower-level remarks pertaining to a particular Post - on the website, they are listed as sub-objects of a particular Post. They too can receive upvotes and downvotes. PostHistory tracks the edits and changes made to Posts, and PostLinks show how posts link to each other (especially duplicates, related questions, or points of reference). Owing to the massive scale of the StackOverflow website (~30 GB of data), for the purposes of rapid iteration, we focus our attention on the math subdomain (math.stackexchange.com), which is the largest (~10 GB of data) site in the StackExchange family, after StackOverflow.

The math subdomain of StackExchange has 63.2 MB of Badges, 567.1 MB worth of Comments, 2.39 GB of PostHistory, 11.9 MB of PostLinks, 1.33 GB of Posts, 113 KB of Tags, 62.9 MB of Users, and 375.2 MB of Votes. This corresponds to 558,813 Badges, 2,137,567 Comments, 3,344,471 PostHistory edits, 101,022 PostLinks, 1,169,070 Posts, 1212 Tags, 199,660 Users, and 3,946,589 Votes.

One of the primary difficulties with this project thus far has been the data scale of our intended target StackExchange site, StackOverflow. Because each StackExchange site's data is structured in exactly the same manner, we have chosen to limit the effects of this problem into the future by doing all our initial analysis and exploration on the Math StackExchange site, which is 3x smaller in size. We also limited some of our analyses to just a single question tag (geometry) within this site, in order to understand our original problem of expertise in a particular topic domain. To put this in perspective, the geometry subdomain has 13,993 posts and 10,707 Users.

IV. Model, Algorithm, Method

4.1 Analysis of the PostLinks network

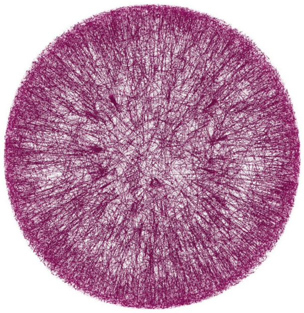

Figure 1: Math StackExchange Links Graph. Directed graph of outgoing links for each post.

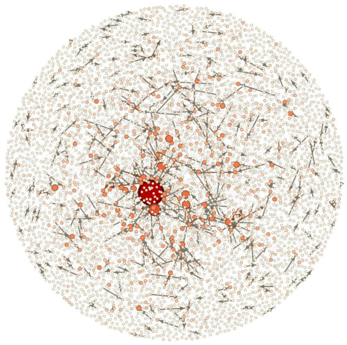

Figure 2: Math Branch: Geometry Tag--Post Links Graph. Directed graph of outgoing links for each post.

As explained earlier, we first looked at the post-to-post link network of the Math StackExchange site. Using the provided PostLinks.xml file, a network was created using the posts as nodes and links between posts as directed edges, from the source to the destination. This graph contains 1169072 posts and 101024 edges between the posts. Note there are many posts that do not have links, or have only one link to a preceding Post. This renders the graph not terribly interesting - nodes in this network have an average outdegree of 0.83 (see Figure 1).

Since the first graph did not exhibit interesting results, a smaller subset of the data was explored. A new network was built using the <geometry> tag within the math StackExchange site (see Figure 2). The subcommunity has 13,993 posts and 10,707 Users. While this network had similar overall characteristics to the first graph (nodes have average outdegree 0.706), this network revealed some interesting results, had much more variety in degree, and, from a purely visual perspective, had more easily visible structure. The most central node in the graph had 53 outgoing links. The top three posts with the most response links are as follows (author's note: these questions will not be unfamiliar to avid Internet users, as all of these questions went viral over the last two years):

1. Is value of pi equal to 4?
2. Is this Batman equation for real?
3. What is the most elegant proof of the Pythagorean theorem?

From both Figure 1 and Figure 2, it can be seen that there is very little linking between posts. However it is important to note that neither graph accounted for duplicate questions, and thus each duplicate question is currently represented by a different node. Accounting for duplicates and removing totally isolated nodes might make this analysis more informative in the future.

4.2 Analysis of the User Network

Moving away from the post network, we examine a user centric network. When creating the user network, we focus on the smaller subset of the Math domain, looking at users who participate in posts with the <geometry> tag. This network is created by linking users who ask questions to users who answer those questions; the edges are added as a directed link, going from the user who asks the question to the user who provides an answer. This network is used for most of the later analyses, and plays an important role in helping to achieve one of the major goals of this paper: to derive different user profiles.

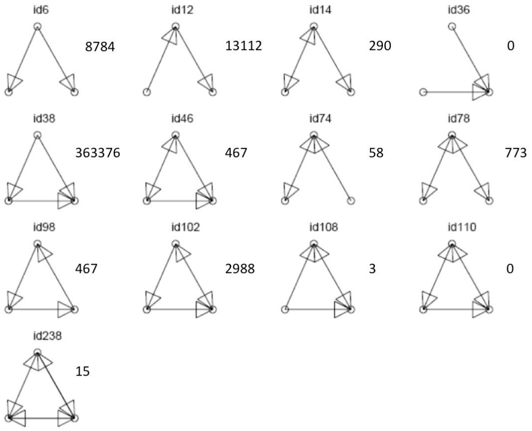

Figure 3: All possible 3-triads and their frequencies in the geometry subgraph.

There are two main ways that we have derived different user profiles: categorizing various interpersonal structures in the network and examining if there are network statistics that would be representative of a majority. We proceed with a motif analysis on the Math StackExchange question-answer graph, modeled like the Java Forum paper with directed edges from an author of a post to a responder to that post. We count the various directed 3-triads in the network and their respective frequencies as in Figure 3. The main purpose for counting triads is to see if there are particular interactions that are common in the community. More importantly is that these interactions may highlight nodes that are strong participants in the network who may be highly reputable in their fields. The most common triad is id38, where there are nodes A,B, C where A and B reposted in response to C, and A reposted in response to B. Another very common triad is id12, is similar to triad id38 except A does not re-post directly to B. These shed light into the structure of expertise in the geometry subgraph. Looking at particular instances of Triad id38, it is usually the case that A is a highly reputable user that shedding wisdom on a comment that B has made in response to C. For example, there are many Triads with A like the user bubba, who is top 3% overall in reputation on Math StackExchange, while B and C have comparably lower reputations.

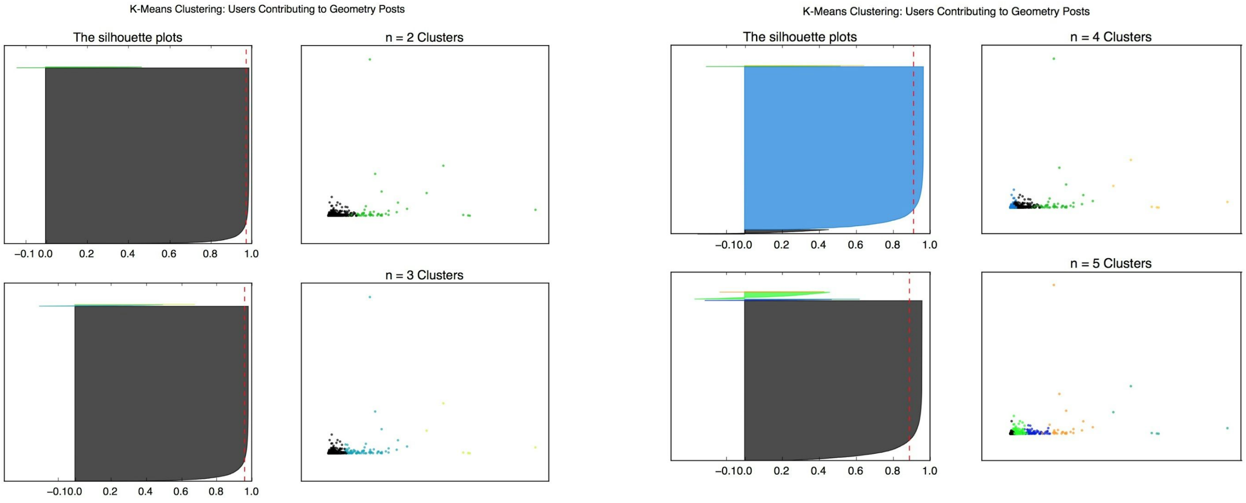

Figure 4: K-Means Clustering for Users contributing to geometry posts for [2,3,4,5] centroids with Silhouette Plots

In the second method of deriving different user profiles, to examine if there are network statistics that would be representative of a majority, feature vectors were extracted for each individual from the Math <geometry> community. The individual vectors for each user included the following features: views, down votes, up votes, outdegree, indegree, closeness centrality, eigenvector centrality, and PageRank centrality. These feature vectors were then used to do *k*-means clustering to associate particular network traits to different personas of people in StackOverflow. The implied physical interpretation of the clustering results is interesting (see Figure 4). From the silhouette plots (Figure 4), it is apparent that for each cluster group, the silhouette values are mostly high, with averages of 0.973 for 2 clusters, 0.959 for 3 clusters, 0.911 for 4 clusters, and 0.887 for 5 clusters. These silhouettes averages are high enough to indicate strong clustering, with a peak at 2 clusters. Visually it is apparent that most users are clustered together, and regardless of the number of clusters, the majority of users occupy a single cluster and very few users occupy the remaining clusters. This makes sense given a large portion of users have a similar interaction profile, of very little participation across multiple graphs. The user who has answered the most questions has answered 535 questions, followed by 445, followed by a few users who have answered between 400 and 200 questions, followed by many users who rarely participate in answering questions. One thing to note is that it's unclear if these users are answering many different types of questions, or all answering many questions that belong to one branch. This would be an interesting result to explore in future research.

While there appear to be some diversity in the number of questions asked by users, there is a much smaller spread in the users who ask questions. The user with the most questions has asked 61 questions, while there are many users who have not asked a question at all (users who mostly answer questions). Both these groups of users (users who ask a lot of questions and users who answer a lot of questions) have high upvotes in the range of 600-1200, but the users who answer many questions have a larger number of down votes than users who ask a lot of questions. For example, one user who has answered many questions has 176 downvotes, while the top three users that have asked questions have 1, 3, and 6 downvotes respectively. This suggests that people don't generally downvote a question, but may consistently downvote answers. But interestingly, avid question askers and question answers both have a large number of views, which indicates the population's interest in them. This perhaps is revealing that we as curious beings do appreciate and attribute value towards thought provoking questions as well as clearly explained, accurate answers.

4.3 Heuristic Approaches to Rank Users by Expertise

Another goal of our analysis is to evaluate and uncover domain expertise. A common dilemma that has come up in our own experiences as students is the quandary of whether or not something posted on the Internet is trustworthy; in the case of StackExchange and StackOverflow, especially when an answer is not readily evaluable, one can often be at the mercy of a user's reputation score. While StackExchange's reputation system awards points to the authors of quality answers, it is less transparent about what areas those quality answers lie in - a user with a high reputation may be an expert in Python, but may not know anything about neural networks, and this fact is inaccessible to a neural networks novice. This can lead to the frustrating situation of having an incorrect or misleading answer accepted as correct, making it the top result when navigating to the page, and ultimately engendering confusion for visiting users. We attempt to solve the problem of evaluating domain expertise with two approaches - a heuristically grounded approach with inputs culled from question-answer graph properties, and a statistical/machine-learning approach using features again culled from question-answer graph properties.

Our heuristic approach ranks users based on a custom function approximating domain expertise numerically, based on a series of inputs. We wrote scripts to fetch in the entire question-answer graph for the <geometry> tag in Math StackExchange and also loaded in all user, post, and badge data. We then wrote a series of increasingly complex heuristic functions - the first just returned a user's reputation (this is the StackExchange baseline), the second returned a user's total degree (indegree minus outdegree) in the question-answer graph, and the third returned a user's total degree times his or her net vote count (upvotes minus downvotes). Moving into computed measures, we also used Betweenness centrality, Eigenvector centrality, and PageRank centrality weighted by badge count as additional heuristics. The top 10 users as ranked by each of these heuristics are displayed in the tables below:

	1. Reputation	**2. Degree**	**3. Degree * Votes**
1	André Nicolas	MvG	Ross Millikan
2	Brian M. Scott	Ross Millikan	André Nicolas
3	Asaf Karagila	André Nicolas	amWhy
4	Did	Blue	Michael Hardy
5	Arturo Magidin	Christian Blatter	lhf

	4. Betweenness	**5. Eigenvector**	**6. PageRank**
1	MvG	MvG	Henning Makholm
2	Ross Millikan	Ross Millikan	André Nicolas
3	Blue	Blue	Ross Millikan
4	André Nicolas	Christian Blatter	Hagen von Eitzen
5	Christian Blatter	André Nicolas	robjohn

6	Qiaochu Yuan	Hagen von Eitzen	robjohn
7	Robert Israel	Jack D'Aurizio	Hagen von Eitzen
8	robjohn	joriki	Gerry Myerson
9	Ross Millikan	Mick	Guess who it is.
10	Bill Dubuque	lab bhattacharjee	Christian Blatter

6	Mick	Mick	joriki
7	Hagen von Eitzen	Jack D'Aurizio	Blue
8	Joseph O'Rourke	Hagen von Eitzen	Jack D'Aurizio
9	bubba	bubba	MvG
10	Jack D'Aurizio	Narasimham	Robert Israel

Table 1: Top 10 Users rated by various heuristic metrics.

Closely examining these rankings yields very promising results. The baseline reputation rankings reveal the “best” overall users on math.stackexchange.com, as expected - looking at the current-day user profiles of André Nicolas and Brian M. Scott on StackExchange, for instance, reveal that they are in the top 0.01% of users all-time and this year respectively. Brian M. Scott is also revealed to be a professor emeritus of math at Cleveland State University, which matches his very high reputation. However, going down the list of profiles reveals that these users are *not* geometry experts per sé - André Nicolas only has a silver badge[1] in geometry (and has only answered 395 questions there, compared to over 1500 questions on calculus and 2300 questions on probability) and Brian M. Scott is a set-theoretic and general topologist with an interest in combinatorics (he has only answered 87 questions related to geometry, compared to 2285 in the general topology tag).

Using just the degree from the question-answer graph also yields expected results - now we arrive at users who spend most of their time answering questions related to the geometry tag. User MvG has 78% of his posts in the geometry tag for a score of 866 (it is, as one would expect, his top tag), and user Blue has 342 posts for a total score of 932. However, these users are *not* the highest rated people on math.stackexchange.com - while they are still in the top 2% of all users, there is a huge difference (~200x) in reputation between the top 0.01% and the top 2%.

The key idea is that we want to isolate experts in geometry, but want to ensure that they are also quite reputable in the general community. The combined degree * vote heuristic attempts to blend the results of the two aforementioned heuristics together by modulating geometric focus (degree) and overall score (vote differential). This produces exceptional results: user Ross Millikan has a gold badge in geometry with 493 posts while also sitting in the top 0.08% of global users, and users André Nicolas and Michael Hardy sit in the top 0.1% of global users with silver badges in geometry (user amWhy’s account is currently suspended on StackExchange, but he/she also had over 500 questions answered in geometry). These basic, direct-data-derived heuristics offer very promising results, which is promising in our goal of understanding domain expertise.

The first computed metric we use in our heuristic is Betweenness centrality, which is roughly equal to the number of shortest paths from all vertices to all others that pass through the measured node. Using this as our expertise measure yields results that are quite similar to the degree heuristic, which makes sense, as degree and connectivity are the key components involved in computing betweenness centrality. Ranked highly are perennial stalwarts MvG, Ross Millikan, and André Nicolas, while the rest of the top 10 are other names that have appeared on our other metric lists. Very similar to the Betweenness metric is the Eigenvector centrality measure, which produces results that are nearly identical to the Betweenness metric. The final (and most complicated) heuristic we consider is a weighted

[1] The silver badge is strictly less reputable than a gold badge. These badges are related to total scores in user answers. Check http://stackoverflow.com/help/badges?tab=general&filter=silver for more details.

combination of PageRank centrality in the geometry question-answer graph and the net number of badges that a user has received, which balances geometry specific knowledge with overall competence. The recursive nature of PageRank seemed appropriate - a reputable user supporting another user would suggest the latter is reputable as well. Under this metric, we shuffle the results of our previous analyses, but the overall results are clearly reflective of what we have seen before. Top ranked user Henning Makholm is in the top 0.06% of users this year and has a silver (approaching gold) badge in geometry, while the rest of the top 5 have had their attributes analyzed in earlier sections. This result, which is typical of all our heuristics, is exactly what we hoped to extract - the ability to determine who the credible experts in a field are.

It is also interesting to compare the Spearman rank correlation coefficient of each of the 5 heuristics (#2-6) versus the baseline (#1): they are [-0.066667, 0.090909, -0.442424, -0.090909, 0.236364] respectively. The indication is that heuristics #2-5 have very little correlation to the baseline, while heuristic #6 has only a slight correlation with the baseline. Logically, this actually makes sense, as the baseline is just pure reputation while what we care about is geometry expertise. The Spearman rank correlation coefficient of heuristics #3-6 compared to a new geometric baseline, #2, yields outputs [0.115152, 0.587879, 0.684848, -0.260606]; this shows significantly stronger association between the outputs of our heuristic analyses, and illustrates that all of our heuristics, on the whole, are statistically correlated with each other.

4.4 Machine Learning Approach to Rank Users by Expertise

Per our statistical approach for domain expertise, we also apply regression models to our data, with a slightly modified target of predicting reputation scores using domain-specific network features. Features that we experimented with include some StackExchange specific constructs such as Views, DownVotes, UpVotes, days since account creation, and days since last login, as well as more advanced, topological metrics such as indegree (number of posts user has responded to), outdegree (number of posts user has had responses about), betweenness centrality, closeness centrality, eigenvector centrality, and PageRank. All regression training and testing were done on a dataset of 10k examples, with 10-fold cross validation, where for 10 iterations, a different tenth of the dataset was withheld for testing, while the remaining 9/10 of the dataset were used for training. Our evaluation metric is root mean squared (RMS) error between the model's prediction and the ground truth reputation.

Models	**Training RMS Error**	**Testing RMS Error**
Linear	7.20	8.74
Ridge Linear	7.21	8.59
SVR (RBF kernel)	2.25	9.02

Table 2:The Performance of different Regression Models in training and testing.

We employ a straightforward linear regression model as a first iteration of reputation prediction. Linear Regression attempts to fit a line that minimizes the squared loss among all training examples. The model is extremely simple, however Figure 2 shows that it does not perform terribly in testing. Along the same vein, we invoke a Ridge Linear Regression; this model includes an L2 regularization term, a penalty for weight vectors that are very large. This reduces the feature space, thus preventing overfitting. The Ridge Linear model has the best, lowest, testing error, perhaps because of its attempt to increase generalizability. The third model we utilize is Epsilon-Support Vector

Regression with a Radial Basis Function kernel. During training, this model fits a linear function that will achieve within epsilon of the truth value. The Support Vector Regression performs even better than both linear regression models on the training data, achieving a total root mean squared deviation of 2.25. However the SVR model generalizes less well considering its test set error is much higher. This gives insight that the Support Vector Regression model is overfitting the training data. There is a possibility of making the model more general if we include some regularization parameter or using fewer features, thus decreasing the feature space and reducing overfitting. In lieu of the above analysis of the three model, it is worthwhile to note that reputation scores on Math StackExchange vary from 0 to 100,000+ so a regressor of RMS error <10 is fairly powerful. Note that in our milestone, we were getting RMS error of a magnitude larger (~100) for a previous linear model, call it A. Given that the training error was quite high for that model A, it was indicative of a high bias situation where our model A was too simple to capture the complexities necessary for reputation prediction. Thus, we increased the feature space of our regression model, including network features beyond degree, such as betweenness, closeness centrality, eigenvalue centrality, and PageRank. This drastically improved our regression results.

A correlation coefficient analysis of these newly added features shed light into how useful they are in prediction. There are two figures below, one examining the correlation of topological features to reputation (see Figure 5), because reputation is the quantity that we are interested in. The second is a correlation matrix among a group of features (see Figure 6).

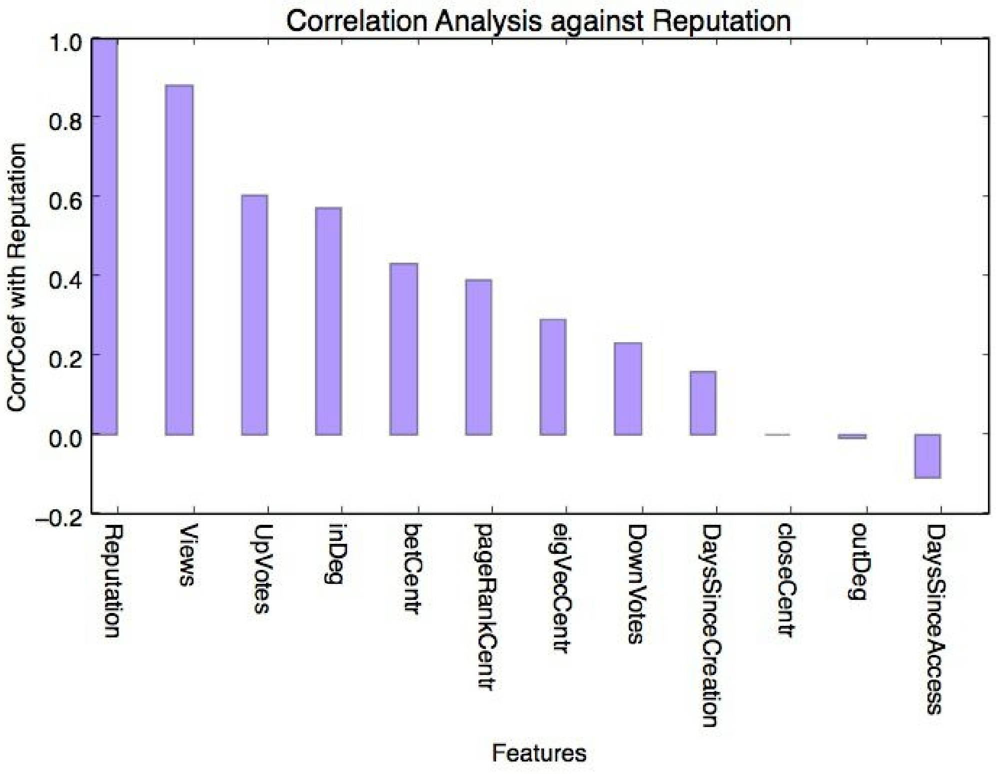

Figure 5: the correlation analysis of topological features to reputation

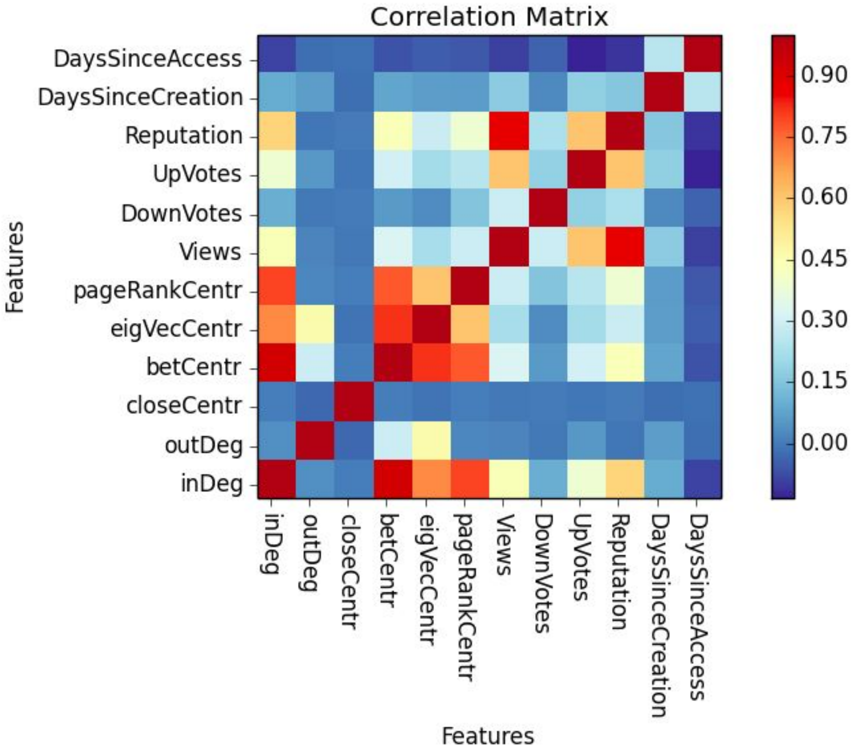

Figure 6: Correlation matrix among a group of features.

The reputation of a user is strongly correlated with his/her view count and number of posts they have responded to (inDeg). As well, reputation is negatively correlated with time since last access account. This perhaps hints that those who use the StackExchange websites more will acquire more knowledge and develop better reputations. Or this may suggest that the websites are more engaging for people of higher reputation so highly reputable people are inclined to login more frequently. Two quantities that seem to have insignificant correlation with reputation are closeness centrality and outdegree. Note that we used the undirected incarnation of the graph for closeness calculations, which nullifies any directionality of question asker and question answerer; intuitively people answering questions may be different from those responding, and hence closeness may not be a great feature for reputation prediction. Outdegree, mapped into our context, is the number of questions a user has asked and gotten responses for. A negligible correlation suggests that users of varying reputations will ask questions, as opposed to a proposed hypothesis that highly reputable users are more inclined to answer than ask questions.

V. Results and Findings

The analysis done in this project consisted of basic graph structure plotting and analysis, triad analysis to look at experts versus learners, *k*-means clustering to cluster different groups of users, using heuristics to retrieve expert users, and applying machine learning to predict user reputation and expertise.

Triad analysis sheds light into how interactions occur at the micro-level. Triad frequency analysis gives a broad, but granular, view on how people are related to one another. On the whole, the Triads that appear most frequently do not have bidirectional edges, suggesting that in StackExchange, communication is usually singular, with a particular question asker and question answer. As opposed to a proposed hypothesis that there is a back and forth in discussion among users, the Triad analysis suggests that the question asker usually reads the answer without any further followup or discussion. This is insightful because it raises questions of why it is the case that communication is usually unidirectional and if the forum should be amended to create more open discussion among users.

The results from *k*-means clustering indicate that the strongest clustering schema is with 2 clusters; in addition to the silhouette plots, it is clear that most users belong to one group, with a small portion of users belonging to the second. This indicates that there is a strong contingent of one type of user (the causal question asker/answerer), and a small contingent of more gregarious users.

The heuristic analysis that we conducted showed that it is in fact possible to come up with various approximation functions that suitably represent domain expertise. As seen in the previous section, experiments conducted on the geometry tag of the math StackExchange subdomain showed that heuristics based on the user interaction question-answer graph, including properties like degree, centrality, badge count, and vote count, corresponded to output rankings that prioritized expertise within the realm of geometry while also lending weight to a user's overall reputation on the website. The tuning of these functions is a seemingly limitless follow-up task; one could doubly emphasize a user's original reputation, perhaps, or even neglect it altogether in favor of a user's PageRank. However, the base results that we produced using heuristics indicate that evaluating or predicting domain expertise through a data-driven approach is not only feasible, but readily accomplishable.

Supervised learning demonstrates great possibilities in predicting the reputation of a user. Using the newly added in network features, we are able to predict reputation scores with small deviations. This is a major improvement from the milestone report, where because of a poor variety and quantity of features, the linear models had large bias errors; adding more features drastically improved results. We would like to tune the feature space even more, and our correlation analysis highlights features that may be redundant or exigent for the linear regression model. The correlation analysis reveals interesting insights related to reputation and time of access. Specifically to the end of understanding domain experts, the correlation analysis reveals that because there is a negligible correlation between the number of questions asked by a user and her reputation, this suggests that people of various reputations all ask questions; as well, the analysis reveals that the number of questions answered is correlated with reputation for a user. This suggests a "just better" model of discussion, where people who are slightly more informed answer the question of those slightly less informed, as opposed to a "best" model. Coupled with the Triad Analysis, it suggests that such communication is fairly unidirectional, without further discussion, even though the expertise levels of both question asker and answer may not be too different.

VI. References

[1] "1.1. Generalized Linear Models." 1.1. Generalized Linear Models — Scikit-learn 0.17 Documentation. Web. 9 Dec. 2015.

[2] "Badges." - *Stack Overflow*. Web. 9 Dec. 2015.

[3] D. Movshovitz-Attias, Y. Movshovitz-Attias, P. Steenkiste, and C. Faloutsos. “Analysis of the Reputation System and User Contributions on a Question Answering Website: StackOverflow.” ASONAM, 2013, pp. 886–893.

[4] Jun Zhang, Mark S. Ackerman, Lada Adamic. “Expertise Networks in Online Communities: Structure and Algorithms.” 16th International Conference on World Wide Web, May 08-12, 2007, Banff, Alberta, Canada.

[5] Lada Adamic, Jun Zhang, Eytan Bakshy, Mark S. Ackerman. “Knowledge Sharing and Yahoo Answers: Everyone Knows Something.” 17th International Conference on World Wide Web, April 21-25, 2008, Beijing, China**.**

[6] "Spearman's Rank-Order Correlation." - A Guide to When to Use It, What It Does and What the Assumptions Are. Web. 9 Dec. 2015.

[7] StackExchange Data Dump (Published August 18, 2015). Retrieved 12 October 2015. https://archive.org/details/stackexchange.