Expertise & Network Properties in StackExchange Sites

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

1. Problem Statement

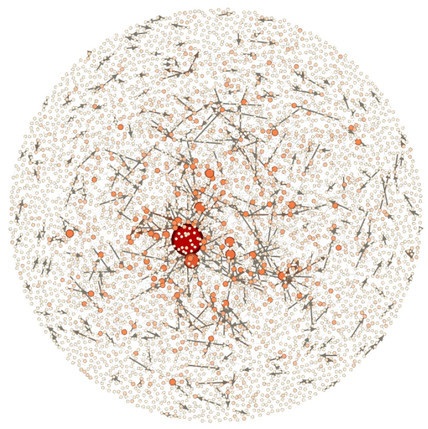
We aim to identify, understand, and designate experts and expertise within StackExchange data. Specifically, our goals:

- Better define and identify experts and expertise, particularly in specific domains, such as geometry.
- Provide structure for knowledge management within an entire online community.
- Create credibility scores that supersede StackExchange's global reputation catch-all.

Prior work in this area include applying network analysis to different question answer forums (Lada), invoking machine learning to predict reputation (Movshovitz), and different perspectives on modeling (Zhang).

2. Datasets

This study looks at the math subdomain of StackExchange, which has 560K Badges, 2M Comments, 1M Posts, 200K Users, and 4M votes.



- Post Graph: Directed graph where nodes are posts and edges are from response to original post.
- User Graph: Directed graph where nodes are users and edges are from commenter to original poster. We use this specific for Geometry tag to focus attention on expertise.

Figure 1. Geometry tag post links.

3. Methods, Models, Findings

3.1. Micro-level Structure

To gain insight in how users interact in a micro setting, we do a Triad Frequency Analysis. This is performed on the User Graph with Geometry tag. The frequency of all possible Triads are tallied up. Triad id38 is very common and maps usually to a highly reputable user shining light on to novice users who are discussing.

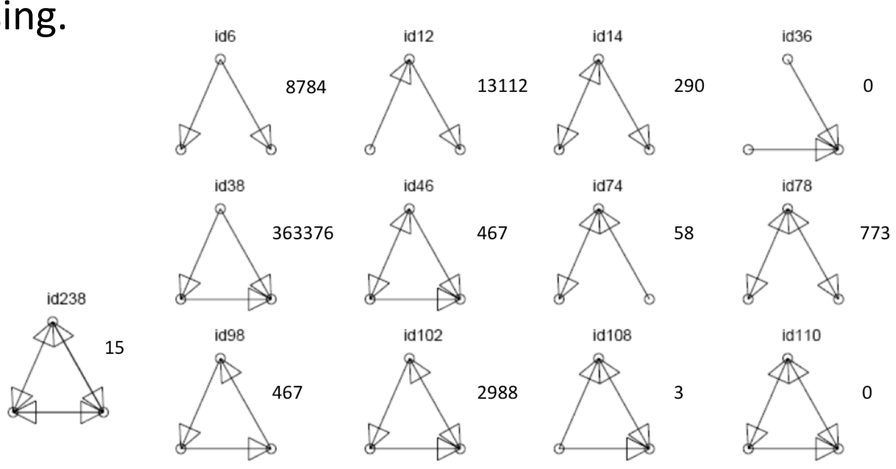


Figure 2. Triads were counted by assign a code each group of linked node sand then mapping those codes to the Triad ids on the left. Note Triad id 38 and id12 occur fairly frequently.

3.2. Clustering

In order to identify particular network traits to different personas of people, *k*-means clustering was performed, on the User Graph with Geometry tag, using the following features: views, up/down votes, in/out degree, and closeness/eigenvector/PageRank centrality.

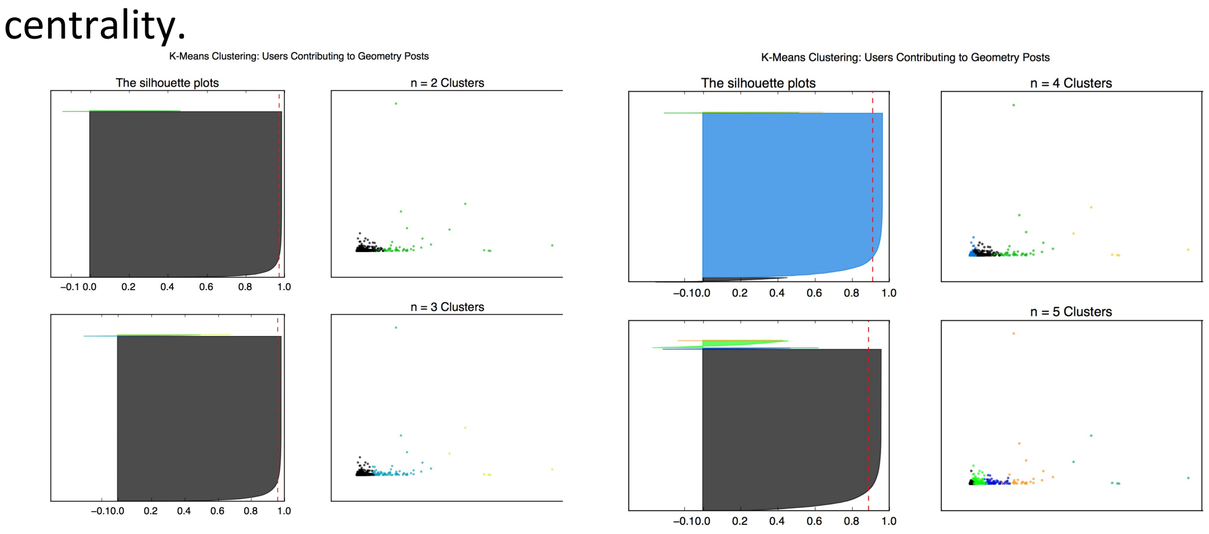


Figure 3. Given the silhouette values of 0.973 for 2 clusters, 0.959 for 3 cluster, 0.911 for 4 clusters, and 0.887 for 5 clusters, the peak clustering number is at 2 clusters.

3.3. Heuristic Approaches to Ranking

A series of increasingly complex heuristic functions successfully prioritize domain expertise over total StackExchange reputation, making these function valid replacements for generic reputation scores.

	Reputation	Degree	Degree * Votes	Betweenness	Eigenvector	PageRank * Badges
1	Andre' Nicolas	MvG	Ross Millikan	MvG	MvG	Henning Makholm
2	Brian M. Scott	Ross Millikan	Andre/ Nicolas	Ross Millikan	Ross Millikan	Andre' Nicolas
3	Asaf Karagila	Andre' Nicolas	amWhy	Blue	Blue	Ross Millikan
4	Did	Blue	Michael Hardy	Andre' Nicolas	Christian Blatter	Hagen von Eitzen
5	Arturo Magidin	Christian Blatter	Ihf	Christian Blatter	Andre' Nicolas	robkohn
6	Qiaochu Yuan	Hagen von Eitzen	Robjohn	Mick	Mick	joriki
7	Robert Isreal	Jack D'Aurizio	Hagen von Eitzen	Hagen von Eitzen	Jack D'Aurizio	Blue
8	Robjohn	Joriki	Gerry Myerson	Joseph O'Rourke	Hagen von Eitzen	Jack D'Aurisio
9	Ross Millikan	Mick	Guess who it is.	Bubba	Bubba	MvG
10	Bill Dubuque	Lab bhatta…	Christian Blatter	Jack D/Aurizio	Harasimham	Robert Isreal

Table 1. These heuristic functions were developed through the User Graph specific to Geometry. This allows us to focus attention on expertise in the Geometry domain. Note these functions had low Spearman Rank correlation to the baseline (0.0545), while they had high Spearman Rank correlation to each other (0.412).

3.4. Supervised Learning

Ten fold cross validation is our training and testing framework. We minimize RMS error. Ridge Linear Regression performed best in test. Given reputation scores 1-100k+, the regressor performs quite well. We also do correlation analysis to examine the most predictive features.

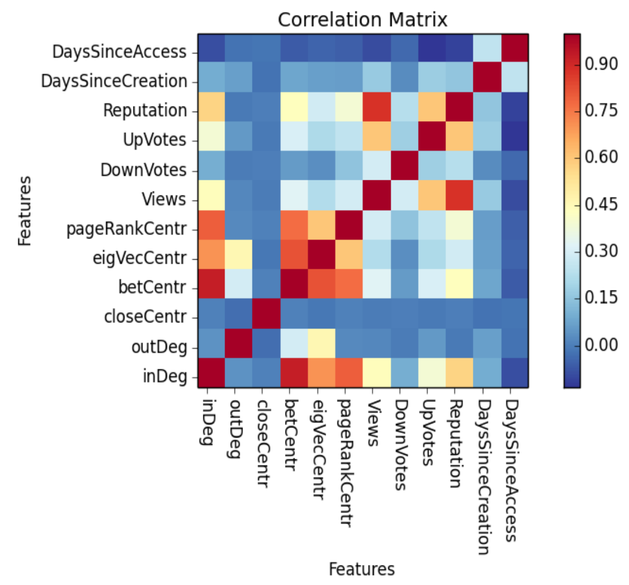


Figure 4. Correlation matrixamong various network level features, StackExchange constructed features, and the reputation score.

Models	Linear	Ridge Linear	Testing Error
Training Error	7.20	7.21	2.25
Testing Error	8.74	8.59	9.02

Table 2. Summary of supervised learning models and their respective training and testing set errors, where we train and test with 10k examples through cross validation.

4. Conclusions & Future Work

- Triad Analysis suggests that the question asker usually reads the answer without any further follow up or discussion, evidenced by the most frequent triads having only unidirectional edges.
- The results from the k-means clustering indicate that there is a strong contingent of one type of user (the causal question asker/answerer), and a small contingent of the expert users.
- Our heuristic analysis shows both feasibility and actual results with regard to the task of predicting and evaluating domain expertise using a data-driven approach that draws from the question-answer graph model.
- Supervised learning demonstrates a low correlation between reputation and number of questions asked, but a positive correlation between reputation and number of questions answered, suggesting a "just better" model of people answering questions of those just slightly less informed.

5. References

[1] 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.

[2] 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.

[3] 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.

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