Predicting Programming Community Popularity on StackOverflow from Affiliation Networks

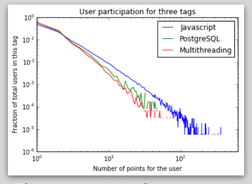
Brie Bunge, Melissa Johnson, & Sophia Westwood

overview

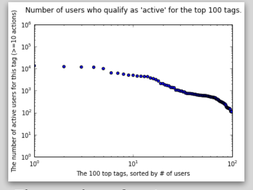
- ***Task:*** Predict tag rank in Nov 2013 based on first four weeks of a tag's lifetime
- ***Results:*** **Initial affiliation network around a tag is more indicative of later success than metrics on initial activity**
- ***Important features:*** Clustering coefficient and average shortest path
- ***Classifiers:*** Applied Random Forest Classifiers, Linear SVCs, Logistic Regression, and AdaBoost Classifiers to predict tag success

background

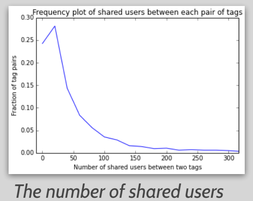
- StackOverflow is a question-and-answer site for programmers, where users post questions labelled with tags for certain technologies, such as c#, java, javascript, php, android, jquery, c++, and python.
- Users label questions with tags
- Answers and comments inherit their question's tag
- Question can have multiple tags



The user activity for Javascript, PostgreSQL, and multithreading tags.



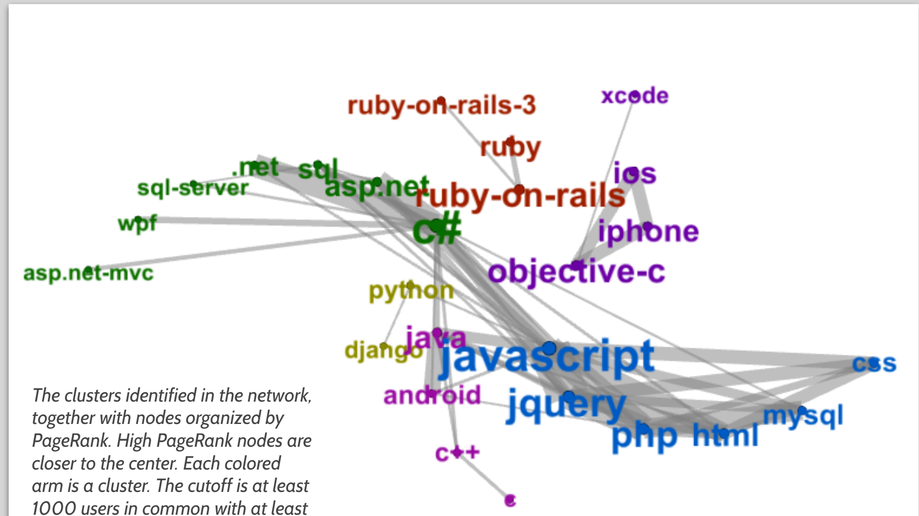
The number of active users per tag follows a power-law distribution.



The number of shared users between two tags follows a power-law distribution.

network models

- Analyzed data from StackOverflow
  - "Communities" based on tags for specific technologies
  - Examined activity for the top 1000 tags, as determined by post count
- Graph for a tag represents a snapshot of the StackOverflow tag network during the first 28 days of the tag's existence
- Two types of 1000 weighted, undirected networks
  - Tag co-occurrence affiliation network
    - tags as nodes and weighted edges representing the number of posts tagged with both tags
  - User activity affiliation network
    - tags as nodes and edges with weights representing the number of users that posted in both communities
    - only considered users that have made at least 5 contributions to each tag



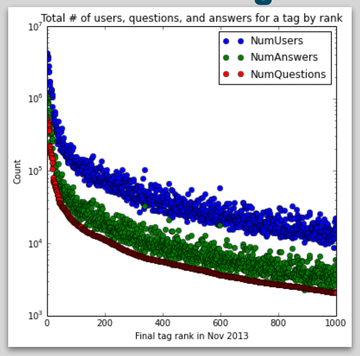
The clusters identified in the network, together with nodes organized by PageRank. High PageRank nodes are closer to the center. Each colored arm is a cluster. The cutoff is at least 1000 users in common with at least 10 contributions.

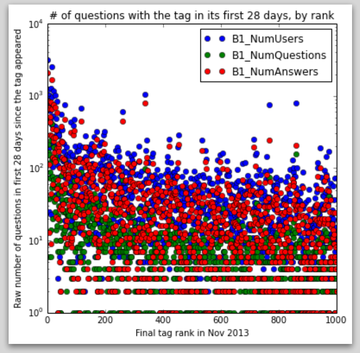
features

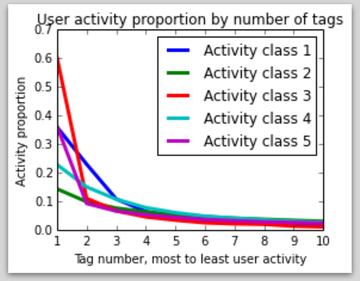
All of the following are for the first four weeks after the creation date of the first post with our tag.

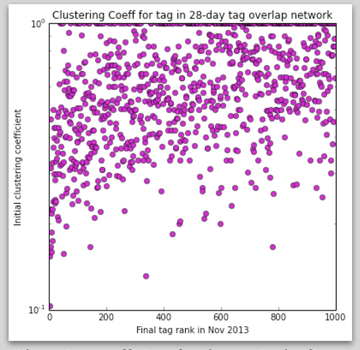
- Number of tag's posts also tagged with a top ten tag
- Percentage of tag's posts also tagged with a top ten tag
- Number of tag's posts also tagged with another tag
- Percent of tag's posts also tagged with another tag
- Number of tag's answers
- Number of tag's questions
- Number of tag's comments
- Aggregate number of users that have contributed to a tag's questions, answers, and comments
- Features for tag co-occurrence and user activity affiliation networks:
  - Degree centrality
  - Closeness centrality
  - Average shortest path
  - PageRank
  - Clustering coefficient

results & analysis









Clustering coefficient for the tag in the first 28-day affiliation network where edges between tag nodes are weighted by # of posts where the tags co-occur.

Classifying future popularity only from # Answers, # Questions, # Comments, and # Users during first 28 days of tag.

	Accuracy:		Less Popular:		More Popular:	
	Test	Train	Precision	Recall	Precision	Recall
Random Forest	.55	.93	.51	.66	.61	.46
Linear SVC	.60	.60	.55	.80	.72	.44
Logistic Regression	.58	.59	.53	.80	.70	.40
AdaBoost	.62	.65	.56	.77	.71	.49

Classifying future popularity only from % and # of tag posts also with a top ten tag during first 28 days of the tag.

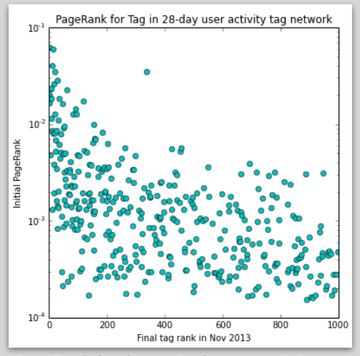
	Accuracy:		Less Popular:		More Popular:	
	Test	Train	Precision	Recall	Precision	Recall
Random Forest	.65	.65	.65	.50	.65	.77
Linear SVC	.63	.62	.58	.66	.68	.60
Logistic Regression	.63	.62	.58	.66	.68	.60
AdaBoost	.64	.63	.75	.31	.61	.91

Classifying future popularity based on degree center, closeness center, average shortest path, PageRank, and the clustering coefficient of the tag T in the co-occurring tag network (nodes are tags, edges are weighted by the number of posts tagged with both tags during the first 28 days of tag T.

	Accuracy:		Less Popular:		More Popular:	
	Test	Train	Precision	Recall	Precision	Recall
Random Forest	.63	.98	.57	.70	.69	.56
Linear SVC	.65	.66	.60	.70	.71	.61
Logistic Regression	.65	.65	.60	.72	.71	.60
AdaBoost	.68	.70	.69	.53	.67	.80

Classifying future popularity based on Avg Shortest Path, Page Rank, and Clustering Coeff from the user activity tag affiliation network during first 28 days of tag.

	Accuracy:		Less Popular:		More Popular:	
	Test	Train	Precision	Recall	Precision	Recall
Random Forest	.64	.72	.71	.37	.62	.87
Linear SVC	.65	.59	.65	.52	.65	.77
Logistic Regression	.65	.59	.65	.52	.65	.77
AdaBoost	.68	.63	.82	.39	.64	.93



PageRank for the tag in the user activity affiliation network. Note that this plot is sparser, as tags that did not meet the user activity thresholds for edges were given a PageRank of 0.