Jingyuan Zhou
MACS30200
Methods and Initial Results
May.17th

Research question: Do critical users on online social Q&A communities have broader interests in topics compared to others?

Data & Methods

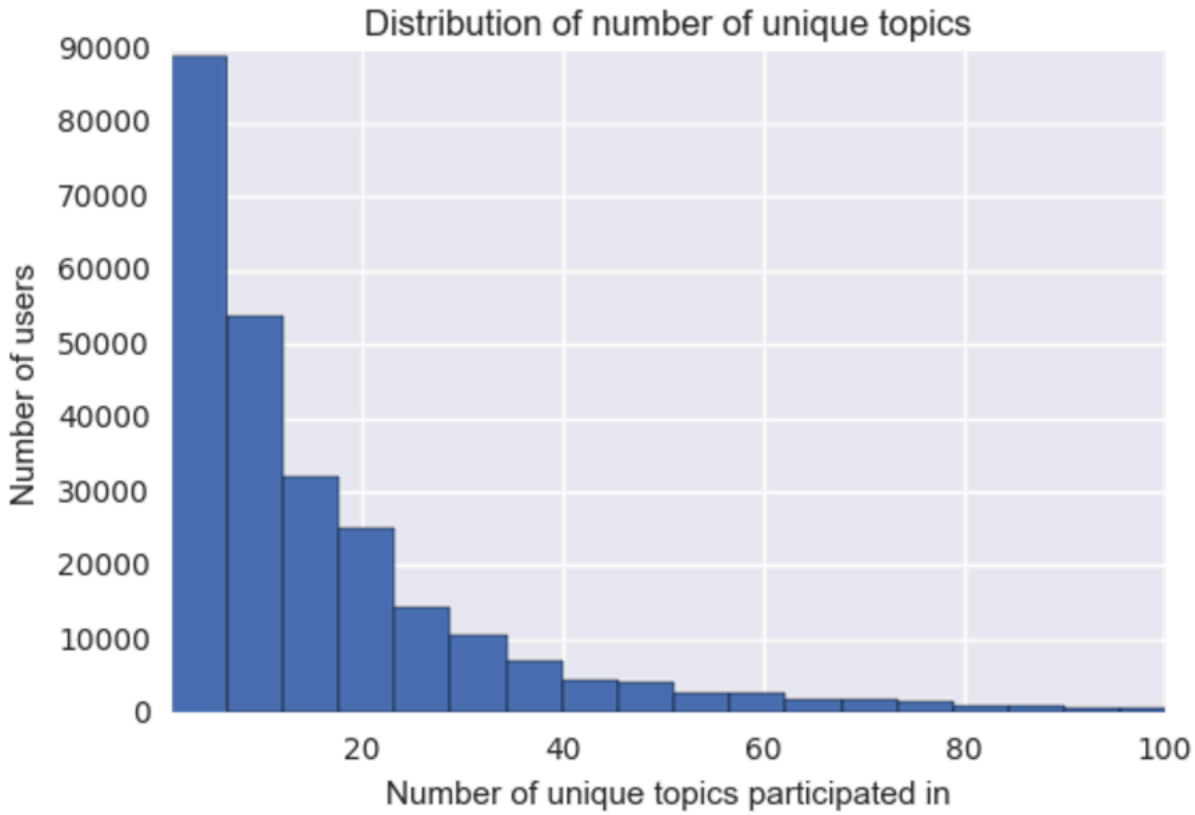
Data is obtained by running scripts from https://github.com/KeithYue/Zhihu_Spider. This collection of data contains 314400 questions and profile information of 261376 users. For each questions, it contains a list of topics that it's tagged with, answers, users related to the answers and a score, which is the number of upvotes of the answer. For each user, it contains a list of users the user of interest is following and a list of followers.

Data is processed so that for each user, we obtain a list of unique topics that it participated in and the frequency associated with each topic. Based on the social network imposed by following, follower lists, we obtained a network of users. Then, with edge betweenness community detection algorithm, community_edge_betweenness() implemented in igraph package in python, we're able to cluster users into communities. Afterwards, the "role" of each user in their community is evaluated by their centrality in the network. For now, we decided to use edge betweenness again because nodes that require the shortest pathways between all other nodes in the network. Semantically, users with a high betweenness centrality may link distinctive groups within their communities. Then, for each community, top 10% users with highest centrality scores are labeled as critical users. Within our dataset, critical users are labeled as 1 and the others are labeled as 0. Consequently, we seem to have obtained all necessary statistics to answer this research question.

Summary of statistics:

	Label	Number of Unique Topics
Count	261376	261376
mean	0.100793	20.932381
std	0.301056	46.995227
25% quantile	0	4
median	0	11
75% quantile	0	22
max	1	5573

As we could observe from the table, distribution of number of unique topics is actually very skewed. Since 75% quantile is 22 and max is 5573, we'll only visualize the number of topics from 0 to 100.



Model and initial results

To get some initial result, we fit a simple linear regression model:

$$p(critical) = \beta_0 + \beta_1 * numberOfTopic$$

OLS Regression Results

Dep. Variable:	label	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.08002
Date:	Wed, 17 May 2017	Prob (F-statistic):	0.777
Time:	07:18:43	Log-Likelihood:	-55694.
No. Observations:	261376	AIC:	1.114e+05
Df Residuals:	261374	BIC:	1.114e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0997	0.001	155.421	0.000	0.098	0.101
n_u_topic	-3.525e-06	1.25e-05	-0.283	0.777	-2.8e-05	2.09e-05

Omnibus:	134367.018	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	600800.380
Skew:	2.675	Prob(JB):	0.00
Kurtosis:	8.153	Cond. No.	56.3

We can observe that p value of number of topics is 0.77, which shows that number of unique topics is not a statistically significant variable. This might be the effect of discarding the hierarchical nature of topics. For example, the topics of a question that asks “how to understand the results of 2016 election of United States” are labeled as “American election”, “American politics”, “Donald Trump” and “Hilary Clinton”. Thus, for further analysis, we would need to look into topics and hand-code them into topics at the same level.