VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SOCIAL AND INFORMATION NETWORKS (CSE3021)

J-COMPONENT

PROJECT REPORT

On

Social Network Analysis of Quora's success

By

SHAUN OOMMEN ALEXANDER (18BCE0610)

KRATU SHARMA (18BCE0642)

Under the Guidance of

PROF. MEENAKSHI S P

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

1. ABSTRACT

Recently, on internet number of question and answer (Q&A) sites have successfully built large growing knowledge repositories, each driven by a wide range of questions and answers from its user's community. While sites like Yahoo Answers have stalled and begun to shrink, one site still going strong is Quora, a rapidly growing service that augments a regular Q&A system with social links between users. Despite its success, however, little is known about what drives Quora's growth, and how it continues to connect visitors and experts to the right questions as it grows. In this, we present results of a detailed analysis of Quora using measurements. We shed light on the impact of three different connection networks (or graphs) inside Quora, a graph connecting topics to users, a social graph connecting users, and a graph connecting related questions. Our results show that heterogeneity in the user and question graphs are significant contributors to the quality of Quora's knowledge base. One drives the attention and activity of users, and the other directs them to a small set of popular and interesting questions.

2. INTRODUCTON

In the last few years, community question-and-answer (Q&A) sites have provided a new way for users to crowdsource the search for specific detailed information, much of which involves getting first-hand answers of specific questions from domain experts. While these sites have exploded in popularity, their growth has come at a cost. For example, the first and still largest of these sites, Yahoo Answers, is showing clear signs of stalling user growth and stagnation, with traffic dropping 63% in last 6 years. In addition, the Google Answers service launched in 2001 was already shut down by 2006. Why is this the case? One of the prevailing opinions is that as sites grow, a vast number of low-value questions overwhelm the system and make it extremely difficult for users to find useful or interesting content. For example, ridiculous questions and answers are so prevalent on Yahoo Answers that a quick Google search for "Yahoo Answers Fail" turns up more than 8 million results, most of which are sites or blogs dedicated to documenting them. Bucking the trend thus far is Quora, an innovative Q&A site with a rapidly growing user community that differs from its competitors by integrating a social network into its basic structure. Various estimates of user growth include numbers such as 150% growth in one month, and nearly 900% growth in one year.

Despite its short history (Quora exited beta status in January 2010), Quora seems to have achieved where its competitors have failed, i.e. successfully drawing the participation of both a rapidly growing user population and specific domain experts that generate invaluable content in response to questions. For example, founders of Instagram and Yelp answered questions about their companies, Stephen Fry and Ashton Kutcher answered questions about actors, and domain-specific answers come from experts such as Navy Seals sharpshooters and San Quentin inmates. So how does Quora succeed in directing the attention of its users to the appropriate content, either to questions they are uniquely qualified to answer, or to entertaining or informative answers of interest? [1]. This is a difficult question to answer, given Quora's own lack of transparency on its inner workings. While it is public knowledge that Quora differs from its competitors in its use of social networks and real identities, few additional details or quantitative measures are known about

its operations. A simple search on Quora about how it works produces numerous unanswered questions about Quora's size, mechanisms, algorithms, and user behavior.

3. OBJECTIVE

The internal structure of question-and-answer sites are often a complex mix of questions, answers, question topics, and users. We will summarize the relationships between different entities. Users can follow individual topics and other users for news and events; questions are connected to other "related" questions, and each question can be tagged with multiple topics. Finally, for each question in the system, there is a user who asked that question (the asker), users who answered that question (answerers), and users who voted on an answer (voters). Quora's internal structure is dominated by three graphs that act as channels that guide user interest and deliver information to users.

1. User-Topic Graph: Quora users follow different topics, and receive updates about questions under topics they follow.

2. Social Graph: Quora users follow each other to form a Twitterlike social graph. Users receive newsfeed about questions their friends participated in.

3. Question Graph: Each question has a list of related questions used by users to browse related questions. The "related" relationship is considered symmetric.

We believe these three graphs are largely responsible for guiding the attention of Quora users. In this project, we will perform detailed analysis on these graphs to understand how they impact user activities, especially how they help users separate a small subset of interesting questions from the larger number of less interesting questions/answers.

4. RELATED WORK

Researchers have studied community-based Q&A (CQA) sites such as Yahoo Answers, MSN QnA, Stack Overflow, Math Over- flow from different perspectives. One perspective focuses on managing questions and topics in CQA sites. Some studies look at question archiving and tagging. Others focus on classifying factual questions with conversional questions, or reusing the knowledge collected from old questions to answer new similar questions. Finally, others evaluate the quality of user generated content, including answer quality and question quality.

A second group of work studies user communities in CQA sites. These projects aim to develop algorithms to identify users with high expertise. One direction is to rank users based on expertise measures generated from user history data (*e.g.* questions, answers, votes). Another direction is modeling user interaction to design network-based ranking algorithms to identify experts. Finally, other works study user community from perspectives such as answering speed and user incentives in CQA sites.

Our work differs from prior art, since we are the first to analyze a social network-based Q&A site using large-scale data measurement and analysis. Instead of treating all users as one big community, we explore the impact of a built-in social network as well other graph structures on the Q&A activities.

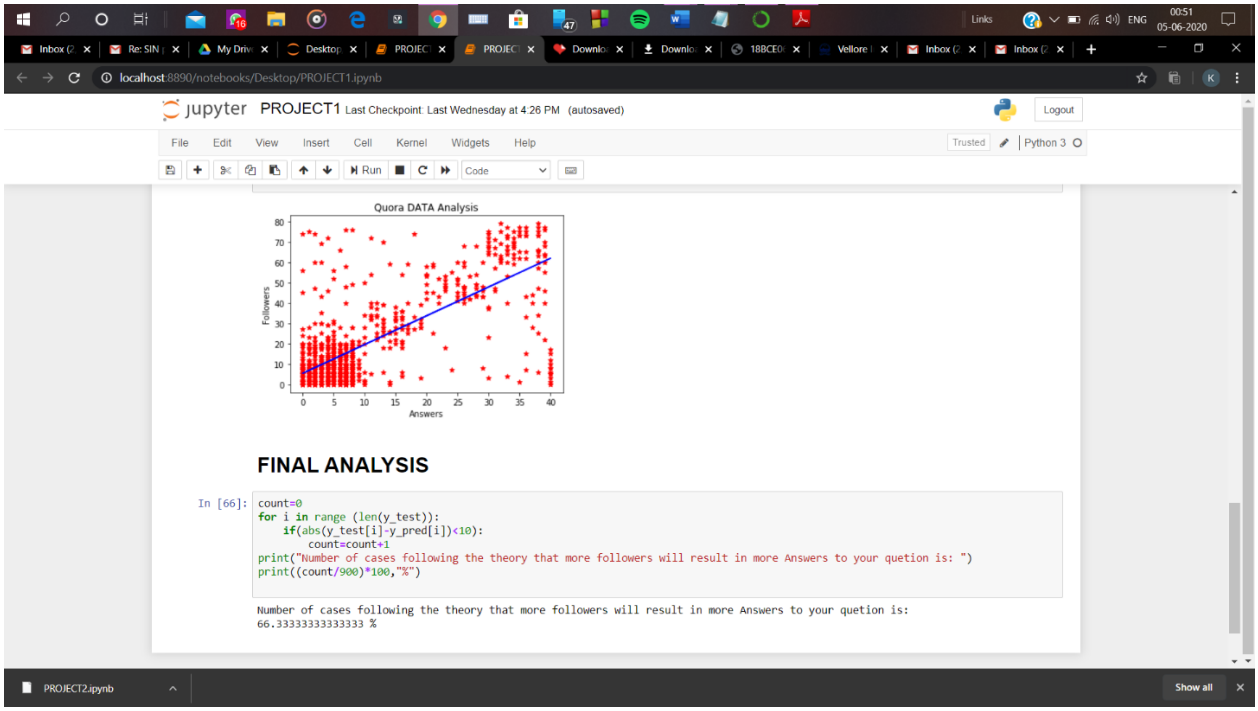
Studies have also looked into the question and answering behaviors in existing online social networks. Users can ask their friends questions by posting tweets in Twitter or updating status in Facebook. These studies answer high-level questions like what types of questions are suitable to ask in social networks, and whether strong ties (close friends) provide better answers than weak ties.

5. SOURCE CODE

5.1 ANSWERS VS FOLLOWERS

```
# Polynomial Regression

# Importing the libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split

# Importing the dataset
dataset = pd.read_csv("kaggle-dataset-7.csv")
X = dataset.iloc[:, [2]].values
y = dataset.iloc[:, [3]].values


# Plotting points as a scatter plot
x = df["answers"]
y = df["followers"]
plt.scatter(x, y, label= "stars", color= "m",
          marker= "*", s=30)
# x-axis label
plt.xlabel('answers')
# frequency label
plt.ylabel('followers')
# function to show the plot
plt.show()


# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.9, random_state = 0)



# Fitting Simple Linear Regression to the Training set
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)

# Predicting the Test set results
y_pred = regressor.predict(X_test)
```

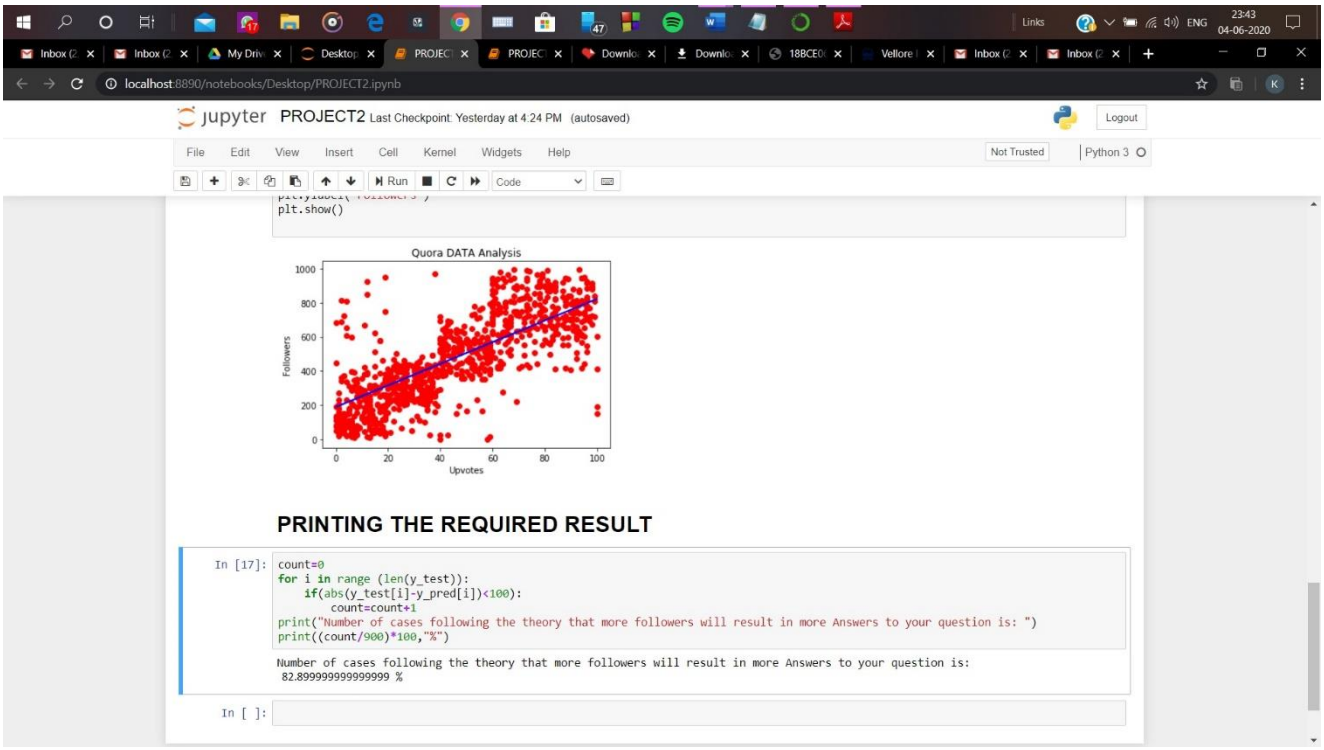
```
# Visualising the Training set results
plt.scatter(X_test, y_test, color = 'red', marker= "*")
plt.plot(X_test, regressor.predict(X_test), color = 'blue')
plt.title('Quora DATA Analysis')
plt.xlabel('Answers')
plt.ylabel('Followers')
plt.show()

count=0
for i in range (len(y_test)):
    if(abs(y_test[i]-y_pred[i])<10):
        count=count+1
print("Number of cases following the theory that more followers will result in more Answers
to your question is: ")
print((count/900)*100,"%")
```



5.2 UPVOTES VS FOLLOWERS

```
# Polynomial Regression

# Importing the libraries
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

```
# Importing the dataset
dataset = pd.read_csv("kaggle-dataset-10.csv")
X = dataset.iloc[:, [2]].values
y = dataset.iloc[:, [5]].values

# Plotting points as a scatter plot
x = df["Upvotes"]
y = df["followers"]
plt.scatter(x, y, label= "stars", color= "m",
           marker= "*", s=30)
# x-axis label
plt.xlabel('Upvotes')
# frequency label
plt.ylabel('followers')
# function to show the plot
plt.show()

# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.9, random_state = 0)



# Fitting Simple Linear Regression to the Training set
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)

# Predicting the Test set results
y_pred = regressor.predict(X_test)

# Visualising the Training set results
plt.scatter(X_test, y_test, color = 'red', marker= "*")
plt.plot(X_test, regressor.predict(X_test), color = 'blue')
plt.title('Quora DATA Analysis')
plt.xlabel('Upvotes')
plt.ylabel('Followers')
plt.show()

count=0
for i in range (len(y_test)):
    if(abs(y_test[i]-y_pred[i])<200):
        count=count+1
print("Number of cases following the theory that more followers will result in more upvotes to
your answers is: ")
print((count/900)*100,"%")
```



6. INITIAL ANALYSIS

6.1 TOPICS

Quora is a general Q&A site with a very broad range of topics. We observed 56K topics in our dataset, which is twice more than that of Stack Overflow, even though Quora is smaller by question count. Table 1[5] lists the top 10 topics with the greatest number of questions in each site. In Quora, the top 10 includes topics in various areas including technology, food, entertainment, health, etc. "Startups" is the most popular one which takes 3.7% of the questions. While all topics in Stack Overflow are different, they are all related to programming [4]. The most popular topic is "C#," which represents roughly 10% of all questions. That's why Quora is used for general questions also that are not related to coding.

Topics in Quora	Number of questions	Topic in Stack Overflow	Number of questions
Technology	5.3M	C#	3.33M
Movies	3.5M	Java	2.77M
Music	3.2M	PHP	2.57M
Writing	2.2M	Javascript	2.42M
Computer Science	1.3M	Android	1.81M
Computer programming	960.8K	Jquery	907K
Venture-Capital	938.5K	iPhone	843K
Google	855.4K	C++	639K
Psychology	705.2K	ASP.net	532K
Startup Advice	378.2K	.net	505K

Table 1: Top 10 topics based on number of questions

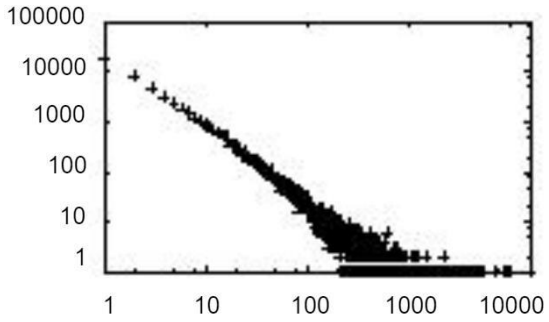


Fig 1: Number of questions per topic

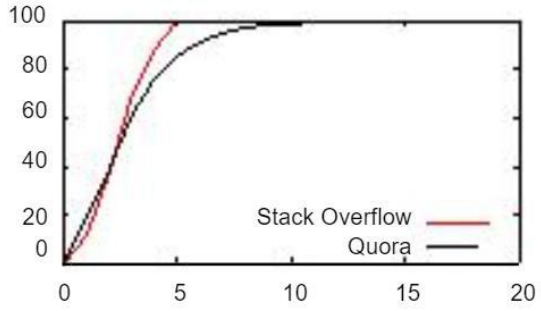


Fig 2: Number of topics per question

Figure 1 plots the distribution of number of questions per topic in Quora in a log-log grid. It shows that for most topics, each topic contains only a handful of questions, while a few popular topics are responsible for most of all questions. The distribution of number of questions per topic mirrors a power-law distribution.

6.2 QUESTION AND ANSWERS

Figure 2 shows the number of topics per question. Stack Overflow requires a minimum of 1 topic and a maximum of 5 topics per question, and the results are evenly distributed between 1 and 5. Although Quora does not have such requirements, a majority (85%) of questions have no more than 5 topics. Very few (<1%) of questions end up with more than 10 topics, which might be an attempt to draw more attention to the question.

6.3 USER ACTIVITY

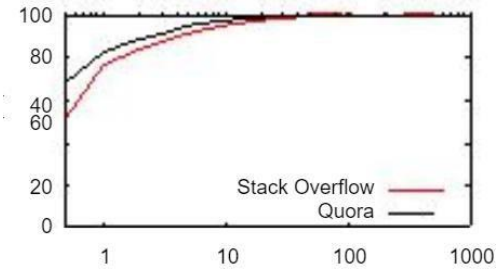


Fig 3: Number of questions per User

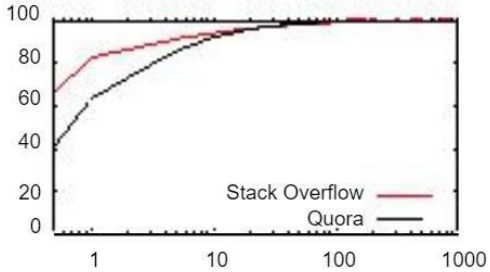


Fig 4:Number of Answers per User

we compare levels of user activity in Quora and Stack Overflow. Figure 3 and Figure 4 show the total number of questions and answers posted by each user. 60% of Stack Overflow users did not post any questions (or answers), while less than 1% of active users post more than 1000 questions (or answers). We observe similar trends in Quora. 40% of the users in our dataset did not post any answers, and 70% of the users have not asked any questions, indicating that a small portion of users have contributed most of the content.

7. WORK DONE AND IMPLEMENTATION

7.1 THE SOCIAL GRAPH

Quora users also follow each other to form a Twitter-like directed social graph. Questions that a user interact with are disseminated to their followers in the form of events in their newsfeed. Therefore, social relationships clearly affect Q&A activities, and serve as a mechanism to lead users to valuable information [3].
In this section, we analyze the Quora social graph to understand the interplay between user social ties and Q&A activities. Specifically, we seek to answer three key questions. First, what triggers Quora users to form social ties? Second, does the presence of popular users correlate with high quality questions or answers? That is, do questions raised by "super-users" with many followers receive more and/or better answers from her followers? Finally, do strong social ties contribute to higher ratings on answers to questions? In other words, do questions answered by super-users get more votes because of the sheer number of their followers?

7.1.1 IMPACT OF FOLLOWERS ON UPVOTES

We analyze answers offered by super users (most followed users). Results in Figure 5 show that for 83% of questions, super users' answers received the highest votes, and for 90% of cases, their answers are among the top-2 most votes. This implies that regardless of the quality of their answers, super users can often get more votes over other users.

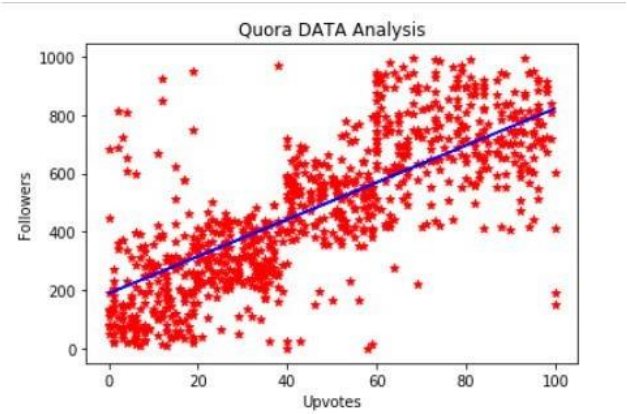


Figure 5: Followers vs upvotes graph

7.1.2 IMPACT OF FOLLOWERS ON QUESTION ANSWERING

We also analyze the number of answers received the user based on their followers. Result in figure 6 shows that for 66% of chances that if a user has a greater number of followers than he/she will receive more number of answers. While 34% of results shows that this is not the case.

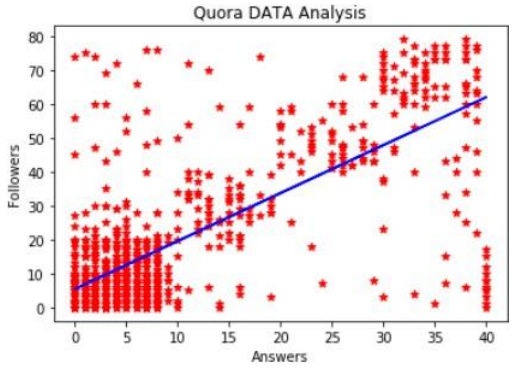


Figure 6: Followers vs Answers graph

7.2 THE USER TOPIC GRAPH

Quora allows users to track specific fields by following the corresponding topics, such as "Startups," "Facebook," and "Technology." This also directly connects users to questions (and associated answers). A question, once created or updated under a topic, will be pushed to the newsfeeds of users who follow the topic. In this section, we model the interaction between Quora users and topics using a user-topic graph and examine the impact of such interactions on question answering and viewing activities.

we list in Table 2 the top 10 topics with the most followers. Clearly, users were highly biased towards certain topics. For example, "Startups" was followed by nearly 18% of users, and "Venture-Capital" by 5% of users. More interestingly, when compared to Table 1 ranking topics by number of questions, only 4 topics ("Startups", "Facebook", "Google", and "Music") are in the top-10 of both rankings. This shows that a high level of interest in a topic, i.e. more followers, does not necessarily produce more questions.

Topic	Number of Followers	Topic	Number of Followers
Startups	474,084	Google	188,867
Facebook	258,569	Science	179,669
Twiter	230,034	TechCrunch	133,310
Technology	211,852	Music	131,084
Entrepreneurship	200,661	Venture-Capital	123,863

Table 2: Top 10 topics in Quora based on number of followers

7.3 THE RELATED QUESTION GRAPH

We used the dataset present in Kaggle (fig 7) to find the related question graph and seek to determine if the structure of the graph plays a role in helping users to find top questions. Intuitively, a similarity-based question graph would produce large clusters of questions around popular topics, with less popular questions relegated to sparse regions of the graph. Thus, users following related question links could encounter popular questions with a higher probability.

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0

Fig 7: Dataset used for related question graph

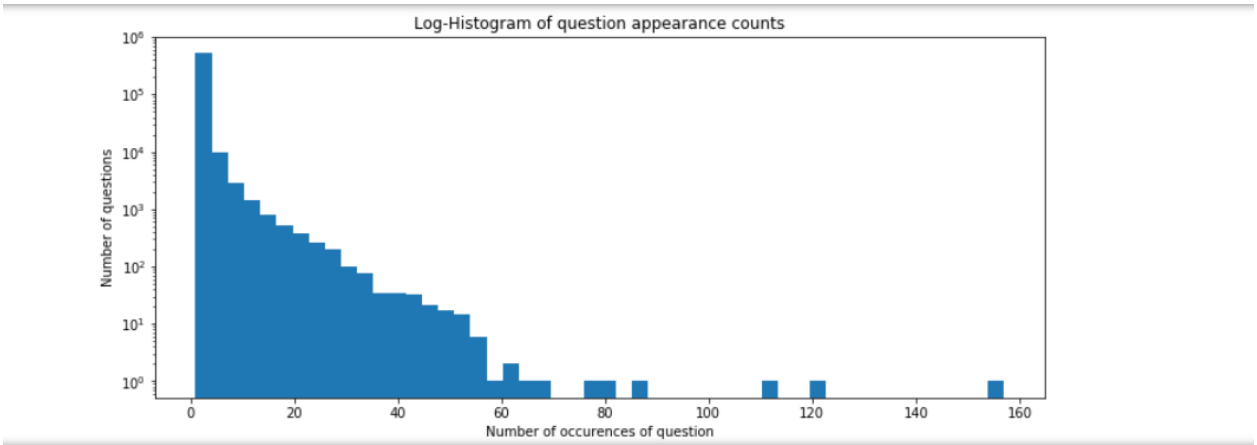


Fig 8: Log histogram of question appearance count

Figure 9 shows the normalized histogram of word count in questions. After getting this graph we can conclude that the average number of words in a question will be 10

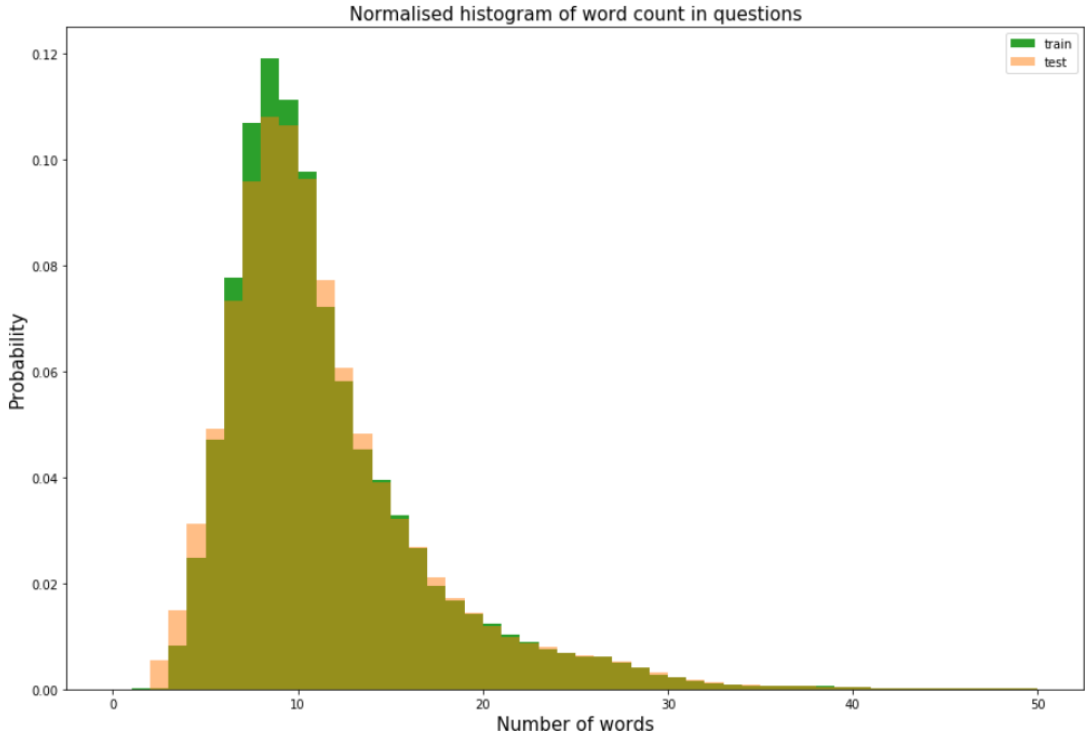


Fig 9: Histogram for word count in questions

Figure 10 shows the general properties shown by how much percentage of questions in Quora. We have analyzed the question with question marks, full stops, capitalized first letter, capital letters and numbers.

```
Questions with question marks: 99.87%
Questions with [math] tags: 0.12%
Questions with full stops: 6.31%
Questions with capitalised first letters: 99.81%
Questions with capital letters: 99.95%
Questions with numbers: 11.83%
```

Fig 10: General analysis of questions in Quora

8. CONCLUSION

Community question and answer sites provide a unique and in- valuable service to its users. Yet as these services grow, they face a common challenge of keeping their content relevant and making it easy for users to "find the signal in the noise," i.e. find questions and content that are interesting and valuable, while avoiding an increasing volume of less relevant content.
We use a data-driven study to analyze the impact of Quora's internal mechanisms that address this challenge. We find that all three of its internal graphs, a user-topic follow graph, a user- to- user social graph, and a related question graph, serve complementary roles in improving effective content discovery on Quora. While it is difficult to prove causal relationships, our data analysis shows strong correlative relationships between Quora's internal structures and user behavior. Our data suggests that the user-topic follow graph generates user interest in browsing and answering general questions, while the related question graph helps concentrate user attention on the most relevant topics. Finally, the user- to-user social network attracts views and leverages social ties to encourage votes and additional high-quality answers. As Quora and its repository of data continues to grow and mature, our results suggest that these unique features will help Quora users continue find valuable and relevant content.

9. REFERENCES

1) What is quora algorithm/formula for determining the ordering or ranking of answers on a question? Quora, Feb. 2016. http://www.quora.com/Quora- product/What-is-Quoras-algorithm- formula-for-determining-the-ordering- ranking-of-answers-on-a-question.
2) How many questions are on quora, answered or not? Quora, Mar. 2015. http://www.quora.com/Quora- Usage-Data-and-Analysis/How-many- questions-are-on-Quora-answered-or-not.
3) On what topics does quora have the best questions and answers? Quora, Apr 2016. http://www.quora.com/Lists-of-Top-Quora- Content/On-what-topics-does-Quora-have- the-best-questions-and-answers.
4) What is a closed question? Stack Overflow,Oct 2014.http://meta.stackoverflow.com/questions /10582/what-is-a-closed-question
5) https://www.quora.com/What-are-the-most-viewed-topics-on-Quora-1