Analysis and Prediction of Question Topic Popularity in Community Q&A Sites: A Case Study of Quora

Suman Kalyan Maity, Jot Sarup Singh Sahni and Animesh Mukherjee
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur, India - 721302
Email: {sumankalyan.maity, jsahni, animeshm}@cse.iitkgp.ernet.in

Abstract

In the past few years, Quora a community-driven social platform for question and answering, has grown exponentially from a small community of users into one of the largest and reliable source of Q&A on the Internet. Quora has a built-in social structure integrated to its backbone; users can follow each other, follow question, topics etc. Apart from the social connections that Quora provides, it has developed a knowledge base nicely organized via hierarchy and relatedness of topics. In this paper, we consider a massive dataset of more than four years and analyze the dynamics of topical growth over time; how various factors affect the popularity of a topic or its acceptance in Q&A community. We also propose a regression model to predict the popularity of the topics and discuss the important discriminating features. We achieve a high prediction accuracy (correlation coefficient ∼0.773) with low root mean square error (∼1.065). We further categorize the topics into a few broad classes by implementing a simple Latent Dirichlet Allocation (LDA) model on the question texts associated with the topics. In comparison to the data sample with no categorization, this stratification of the topics enhances the prediction accuracies for several categories. However, for certain categories there seems to a slight decrease in the accuracy values and we present an in-depth discussion analyzing the cause for the same pointing out potential ways for improvement. We believe that this thorough measurement study will have a direct application to a service like recommending trending topics in Quora.

Introduction

"It is better to know some of the questions than all of the answers"
—James Thurber

Since its foundation in June, 2009 (publicly available in June, 2010), Quora has grown into one of the largest and popular question-and-answer (Q&A) sites. As of September 2013, most of the Quora traffic (33.3%) comes from India followed by US (26.5%) and UK (3.8%) [1]. Apart from being a Q&A site, Quora has a social network backbone that nicely integrates its user base and is a very unique feature compared to the other Q&A sites. People can tag questions with various topics, follow a question, follow a topic, share questions and its answers apart from the basic features like upvoting/downvoting, commenting etc. These social aspects of question-answering make Quora a unique Q&A site for further investigation.

In Quora's ecosystem of knowledge sharing through question-answering, topics play an important role. People follow topics to get important and valuable content related to a topic of their interest. Similarly, when a user posts a question, he/she can tag it with relevant topics so that the topical experts and people interested in the topics get to know about the question and can provide better answers thus helping to control the content quality in Quora. Further, the users in Quora usually provide compelling answers to the questions in which they are interested. Therefore, topics form an essential organizing tool for Quora's knowledge corpus. In fact, the importance of topical organization have also attracted in-house Quora scientists to investigate the structure of the topical network [2]

In this paper, we plan to study the dynamics of topic growth in Quora over time; in other words, how the Quora knowledge base is changing over time with the influx of new topics, growth or decay of older topics. One of the primary interests of this study is to identify factors that have a direct impact on the growth of popularity of the question topics. Understanding the popularity of topics is important because it helps us identifying trending topics. This study has a direct application in recommendation of the trending topics to various users in Quora.

The major contributions of the paper are three-fold.

- Using automated crawls, we have gathered a massive Q&A dataset spanning a period of over four years (Jan 2010 - May 2014).
- We study the temporal growth of topics in Quora and the inter-topic dynamics to understand stability and migration of topics. We observe that core of the topic network is stable whereas the periphery keeps on changing. We also present some case studies and compare them with Google trends.

[1] http://valleywag.gawker.com/most-of-quoras-traffic-is-now-coming-from-india-1341592714

[2] http://data.quora.com/The-Quora-Topic-Network-1

- As a next step, we propose a prediction framework to predict the popularity of the topics and discuss the important features driving the popularity of a topic. We achieve a high correlation between the predicted value and the ground-truth popularity (correlation coefficient ∼0.773) with low root mean square error (∼1.065). We further categorize the topics into a few broad categories by implementing a Latent Dirichlet Allocation (LDA) model on the question texts associated with the topics. In comparison to the data sample with no categorization, this stratification of the topics helps in better prediction accuracies for several categories.

To the best of our knowledge, this is the first rigorous and in-depth measurement study on a massive dataset spanning over a period of more than four years that focuses on the prediction of popular question topics and can potentially have a significant impact on a service like trending topic recommendation in Quora. The organization of the paper is as follows. The next section surveys related work. In section 3, we describe the dataset preparation techniques. Section 4 is devoted for analysis of topical growth in Quora over time. In section 5, we study the stability aspects of the popular topics over time. Section 6 discusses the inter-topic microdynamics. In section 7, we discuss the prediction framework. Section 8 is devoted to performance evaluation of the proposed model. In section 9, we employ LDA model on the question texts to obtain latent categories of topics and separately learn our prediction model on those topical categories and then discuss the prediction accuracies obtained in each case. In section 10, we draw conclusions pointing to the key contributions of our work and discuss potential future directions.

Related work

There has been a considerable amount of work on various aspects of Q&A sites in the past decade. Most of these studies have been conducted on Yahoo Answers (Adamic et al. 2008; Harper, Moy, and Konstan 2009; Harper et al. 2008; Mendes Rodrigues and Milic-Frayling 2009; Shah and Pomerantz 2010; Shtok et al. 2012) though few works have also been done on other Q&A sites like Stack Overflow (Anderson et al. 2012; Mamykina et al. 2011), MSN QnA (Hsieh and Counts 2009; Rodrigues, Milic-Frayling, and Fortuna 2008) and a few recent studies have also used Quora data (Wang et al. 2013; Paul, Hong, and Chi 2012). One direction of research on these Q&A sites focuses on finding experts. The studies done in (Adamic et al. 2008; Li and King 2010; Pal, Chang, and Konstan 2012) focus on ranking the users from expertise measures based on user's history and activities; on the other hand, there have been a few studies (Jurczyk and Agichtein 2007; Lerman and Galstyan 2008; Zhang, Ackerman, and Adamic 2007) that consider the inherent user interactions to model them as a complex system and design network-based ranking algorithms to rank the users. Another direction of research focus on the quality of the user generated content in Q&A sites that includes quality of questions (Anderson et al. 2012; Li et al. 2012) and quality of answers (Adamic et al. 2008; Jeon et al. 2006; Shah and Pomerantz 2010; Tausczik and Pennebaker 2011).

Apart from the two broad themes of research, there are some research works (Harper, Moy, and Konstan 2009; Mendes Rodrigues and Milic-Frayling 2009; Rodrigues, Milic-Frayling, and Fortuna 2008; Shtok et al. 2012; Hsieh and Counts 2009; Mamykina et al. 2011) that do not fit into a single theme but have addressed many interesting questions. In (Harper, Moy, and Konstan 2009), Harper et al. have proposed a classification framework for classifying factual and conversational questions. Shtok et al. (Shtok et al. 2012) have attempted to reduce the rate of unanswered questions by reusing the knowledge of past resolved questions to answer new unresolved similar questions. Bhat et al. (Bhat et al. 2014) have proposed a framework for predicting the response time (getting the first answer) for a newly posted question. Correa et al. (Correa and Sureka 2014) have studied the deleted questions in Stack Overflow. Other works study user community from the perspectives like speed of answering (Mamykina et al. 2011) and user incentives in community Q&A (Hsieh and Counts 2009). Rodrigues et al. (Rodrigues, Milic-Frayling, and Fortuna 2008) have looked at question managing and tagging in Q&A. Our work is different from the above in the sense that we study the dynamics of topic growth and understand various key factors associated with popularity of topics and, thereby, build a model to predict topics that are going to be popular in future.

Dataset preparation

We obtained our Quora dataset through web-based crawls between June 2014 to August 2014. This crawling exercise has resulted in the accumulation of a massive QA dataset spanning a period of over four years starting from January 2010 to May 2014. We followed crawler etiquettes defined in Quora's robots.txt. We used FireWatir, an open-source Ruby library, to control a PhantomJS (Headless Webkit) browser object simulating clicks and scrolls to load the full page. We initiated crawling with 100 questions randomly selected from different topics so that different genre of questions can be covered as stated in (Wang et al. 2013). The crawling of the questions follow a BFS pattern through the related question links. Each question has information like question content, tags, no. of views, no. of shares, comments, follower count (recently replaced by "want answers"), answer content, answer shares, answer comments etc. Separately each topic's page was crawled to get the follower count of the topic. In addition, we separately crawled the user profiles to get the follower count of the users.

Following the above strategy, we obtained 822,040 unique questions across 80,253 different topics and 1,833,125 answers to these questions. Note that in our dataset, we also have many questions which do not have any answers. The detailed dataset description is presented in table 1.

Temporal aspects of Quora topics

Quora organizes questions and answers via a wide spectrum of topics. We have found 80,253 topics in our dataset, which

Category	Quantity
No. of Questions	822,040
No. of Answers	1,833,125
No. of Topics	80,253
No. of followers per question	13.52
No. of views per question	2490.74
No. of shares per question	1.89
No. of followers per topic having at least 100 questions	30633.015

Table 1: Basic dataset statistics.

is sufficiently large and is growing over time. In this section, we analyze and discuss various temporal aspects of Quora topics. In fig 1(a), we show the topic growth over time. We observe that no. of topics grew initially linearly and then exponentially from around mid 2012. This is also an indicator of Quora's overall growth as a community Q&A site. To understand the influx of topics into the Quora system, we study the no. of new topics per month. Fig 1(b) shows monthly rate of arrival of new topics over time. As fig 1(a), this curve also shows a linear steady increase (Dec '10 being an exception) followed by an exponential rise. In fig 1(c), we observe the penetration of new topics in the top 100 and 500 topics (w.r.t no. of questions) through pairwise comparison of the months. Though there is high overlap of topics in the top zone, the penetration of new topics is not negligible (∼12% in top 100 and ∼20% in top 500). Therefore, not only new topics are getting created, some of them are also becoming increasingly popular with a large no. of questions getting tagged by them. Next we study the question influx over the time. Fig 1(d) shows how no. of questions asked per topic varies monthwise. Over the years, the no. of questions per topic increases suggesting that the volume of questions within a topic is rising on an average.

Stability of Quora topics

In this section, we shall perform stability analysis of the popular topics. We first rank the topics according to no. of questions within that topic in every month. Next, we select n top topics every month. After that we find out the no. of topics that appears in all the months. We perform the experiment for various values of n (n = 50, 100, 150,....500). In table 2, we show the no. of stable topics for different values of n. As we increase n, the no. of stable topics increases linearly. We also performed the same experiment by ranking the topics according to no. of answers given for the questions tagged by the topics. The stable topics increase linearly, similarly as in the previous case of ranking with no. of questions. However, corresponding to each n, the no. of stable topics is less when ranked according to answers rather than questions.

We further study how the relative proportions of questions in top 100 (w.r.t no. of questions) stable topics evolve over time (see fig 2). From the figure, it is observed that for "startup" topic, the relative proportion of questions decreases over time whereas for topics like "life" and "psychology", the relative proportion of questions increases over time. We also show the evolution of relative proportions of answers in various stable topics in top 100. We find that in this case also, relative proportions of answers in "life" and "psychology" topics increases while it decreases for "startup". For other stable topics, relative proportions across various time points do not vary too much. Therefore, in general the stable topics experience persistent growth in terms of questions and answers; however there are interesting exceptions as outlined above.

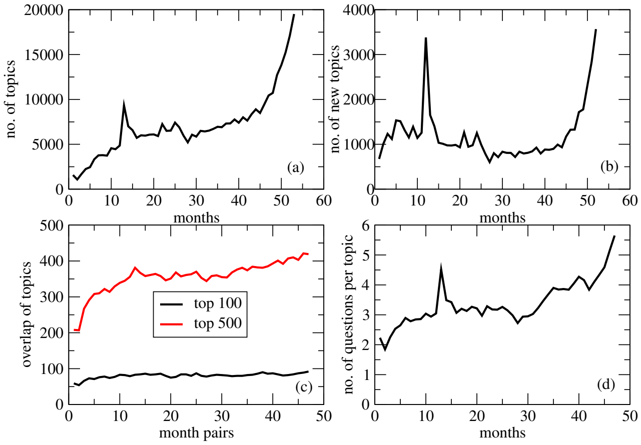


Figure 1: Evolution of a) no. of topics b) no. of new topics over time. 1, 2, ... in x-axes suggest months starting from Jan 2010. c) Consecutive monthwise overlap of topics in top 100 and 500 (w.r.t no. of questions). 1, 2, ... in x-axis suggest consecutive month pairs starting from (Jan '10, Feb'10), (Feb '10, Mar'10), ... d) Evolution of no. of questions per topic over time.

n	no. of stable topics in top (w.r.t no. of questions) n	no. of stable topics in top (w.r.t no. of answers) n
50	11	5
100	26	18
150	39	30
200	50	31
250	54	38
300	68	45
350	74	53
400	79	62
450	86	65
500	94	69

Table 2: Stable topic distribution in top (w.r.t no. of questions as well as answers) n topics for n = 50, 100, ..., 500 across months.

Inter-topic microdynamics

In this section, we shall discuss about the inter-topic microdynamics. The inter-topic network is formed by considering the topics as nodes and an edge between two topics is established if a question is tagged by both the topics. To understand the significance of the topics forming the core of the network and to gather information regarding the temporal evolution of the structure, we perform a k-shell analysis

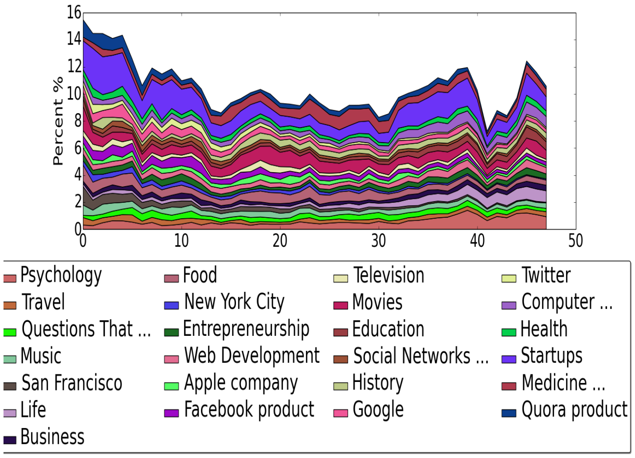


Figure 2: Temporal evolution of percentage of questions for the stable topics in top 100 (w.r.t no. of questions).

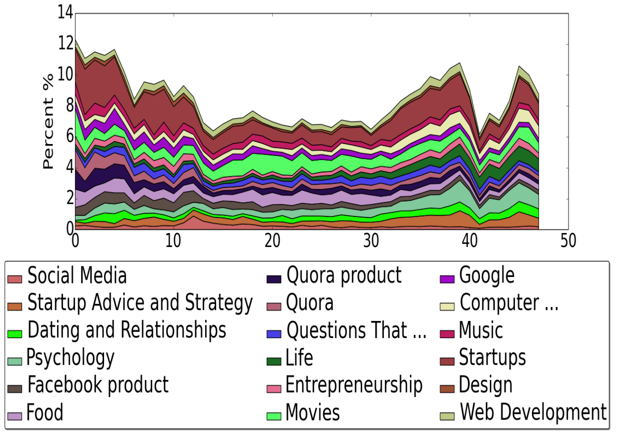


Figure 3: Temporal evolution of percentage of answers for the stable topics in top 100 (w.r.t no. of answers).

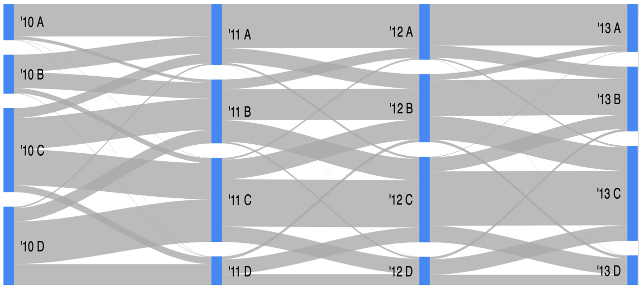


Figure 4: Evolution of the k-shell indices of the topics and the flows between k-shell regions between the years 2010, 2011, 2012 and 2013.

on the topic network. We divide the topics into four regions based on their k-shell indices by dividing the range of k-shell values into four groups of approximately equal sizes. Thus Region A contains words that are in the core of the network ($k \in [\frac{3}{4}k_{max}, k_{max}]$), and Regions B, C, and D contain nodes with increasingly lower k-shell indices. Fig 4 shows the alluvial diagram showing the stability and migration of various topics across different regions of the network for four years ('10 to '14). The height of the blue blocks denote no. of topics in the k-shell region and the shaded areas joining the k-shell regions represent flows of topics between the regions, such that the width of the flow corresponds to the fraction of nodes. We observe that the core of the network, Region A, is remarkably stable compared to the peripheral regions that display a high turnover of nodes. Nodes that are in the core of the network are highly likely to remain so, whereas peripheral nodes frequently either disappear or migrate towards the core.

In table 3, we show the topics that migrated from core to periphery and vice versa. We observe that the no. of migration is higher in the early stage of Quora. However, with time, the migration rate has decreased. There are couple interesting observations we have found here. Pinterest was founded in March 2010 and it has moved from the periphery of 2010 network to core of the 2011 network, similarly the topic "2012 Summer olympics in london" has moved from periphery of 2011 network to core of the 2012 network and moved out from the core of 2012 network to the periphery of 2013 network capturing rise and fall of the topic. Mitt Romney ran for US presidential election in 2012. We observe that the topic "Mitt Romney" has migrated from 2011 C to 2012 A and subsequently migrated from 2012 A to 2013 D. In the same time frame, "The White House" and "Obama Administration" has moved from 2011 C to 2012 A. Thus, the k-shell analysis shows rise and fall of three related event. Another interesting observation is that both "Edward Snowden" and "PRISM NSA Surveillance Program" have migrated from 2012 D to 2013 B. (the NSA Surveillance programme was disclosed in 2013 by Edward Snowden). These above observations are very well supported by Google trends results (see fig 5).

Predicting popularity of topic

In this section, we propose a framework for predicting popularity of a topic. Popularity of a topic is defined by its follower count. Higher the follower count, higher is its popularity. While Quora hosts a large number of topics, and the set is still growing, not all of these are equally popular (in terms of follower count). In fig 4, we present top 20 and bottom 20 topics according to popularity (followers). The top ones in the list include broad and general topics like technology, sports, movies, music, science etc. whereas the bottom topics are more specific. Note that there is little overlap (only 3) between top 10 topics shown in (Wang et al. 2013) with our list of top 20 topics.

Here, we learn the topic popularity features from the evidence of crawled Quora data having information of the top-

[4] https://support.google.com/trends/answer/4355164?hl=en

	’10 to ’11	’11 to ’12	’12 to ’13
high to low core (A to D)	Destiny, Realization, New Atheism, First Impressions	Walking, Travel Startups and Companies	Mitt Romney, 2012 Summer Olympics in London, Quora Features
low to high core (D to A)	Journalists, Pinterest, Marvel Comics, Awkward Situations, Young Entrepreneurs, Viral Videos, Television Writing, Social and Online Music, Occupy Wall Street, Small and Medium Enterprises, Print Media, Volunteering, MacBook Air, Fighting, College and University Students, Streaming Music	2012 Summer Olympics in London	

Table 3: Example topics migrating from core to the periphery of the networks constructed in various years and vice versa.

ics from January 2010 to May 2014 and try to predict the popularity value of the topics 6 months later (December 2014). For this purpose, we separately crawled the follower counts of the topics in December 2014.

Prediction framework:

In this subsection, we describe the experimental framework in detail (see fig 6). Our goal is, given a set of features, to predict the popularity of a topic t at a given time point T. In this work, we have build the features by observing the data from the data from Janauary 2010 to May 2014 and have predicted the popularity of the topics at T = December 2014. Formally, we want to learn a target function $f(X) = \log(n)$, where X is the feature vector of a given topic (t) and n is the follower count of t. The function $f(X)$ is learned from the training examples. We are interested in predicting the magnitude of the acceptance of a topic in a time frame, thus while a topic with 1000 follower is very different from a topic with 5000 follower, 50000 is similar to 55000. Taking logarithm captures this observation. We are trying to learn three aspects in this prediction: (i) what is the feature combination that yields the best prediction? (ii) what are the strongest features and (iii) how do they complement each other? We further categorize the topics by taking all the questions related to the topics as a document and run LDA on the set of documents, each corresponding to a topic. On each of these categories, we separately learn the prediction model and find the prediction accuracies.

Model features

In order to learn our regression model, we consider three types of features for topical popularity, namely context features, content features and user features.

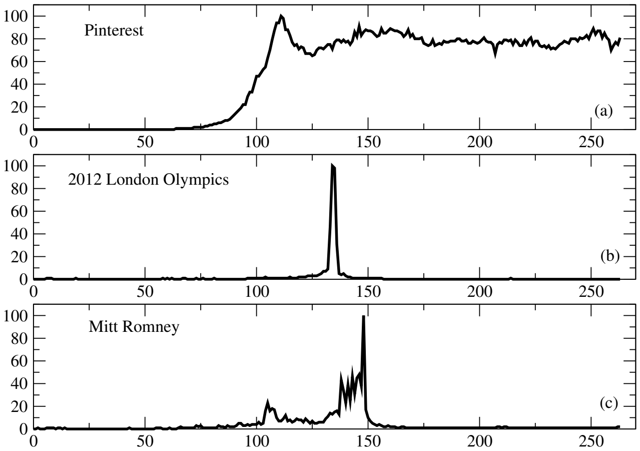


Figure 5: Google trends results for a) Pinterest b) 2012 London Olympics c) Mitt Romney. The y-axis shows the normalized no. of searches that have been done for a particular term, relative to the total number of searches done on Google over time[4]. The numbers (1, 2, 3, ...) on the x-axis represent each week starting from January 2010.

Context features The features in this category are listed below

- Number of questions that have been asked in the topic
- Number of answers per questions in the topic
- Fraction of unanswered questions for the topic
- Average number of question views for the topic
- Average number of question shares for the topic
- Number of comments per questions
- Number of comments per answers
- Average number of answer shares
- Average number of question followers

Content features Content is an important aspect for question and answers. Topics having quality question and answers draw more users and hence gain popularity. The features are mentioned below:-

Topical question diversity: If Q is the document containing all the questions which are tagged by topic i and $p(w|Q)$ is the probability of a word belonging to the document Q then topical question diversity is defined as follows

$$QuesDiv(i) = -\sum_{w \in Q} p(w|Q) \times \log p(w|Q)$$

This feature tells us how much diverse the questions are related to a topic.

Topical answer diversity: Similar to topical question diversity, topical answer diversity is defined as follows

$$AnsDiv(i) = -\sum_{w \in A} p(w|A) \times \log p(w|A)$$

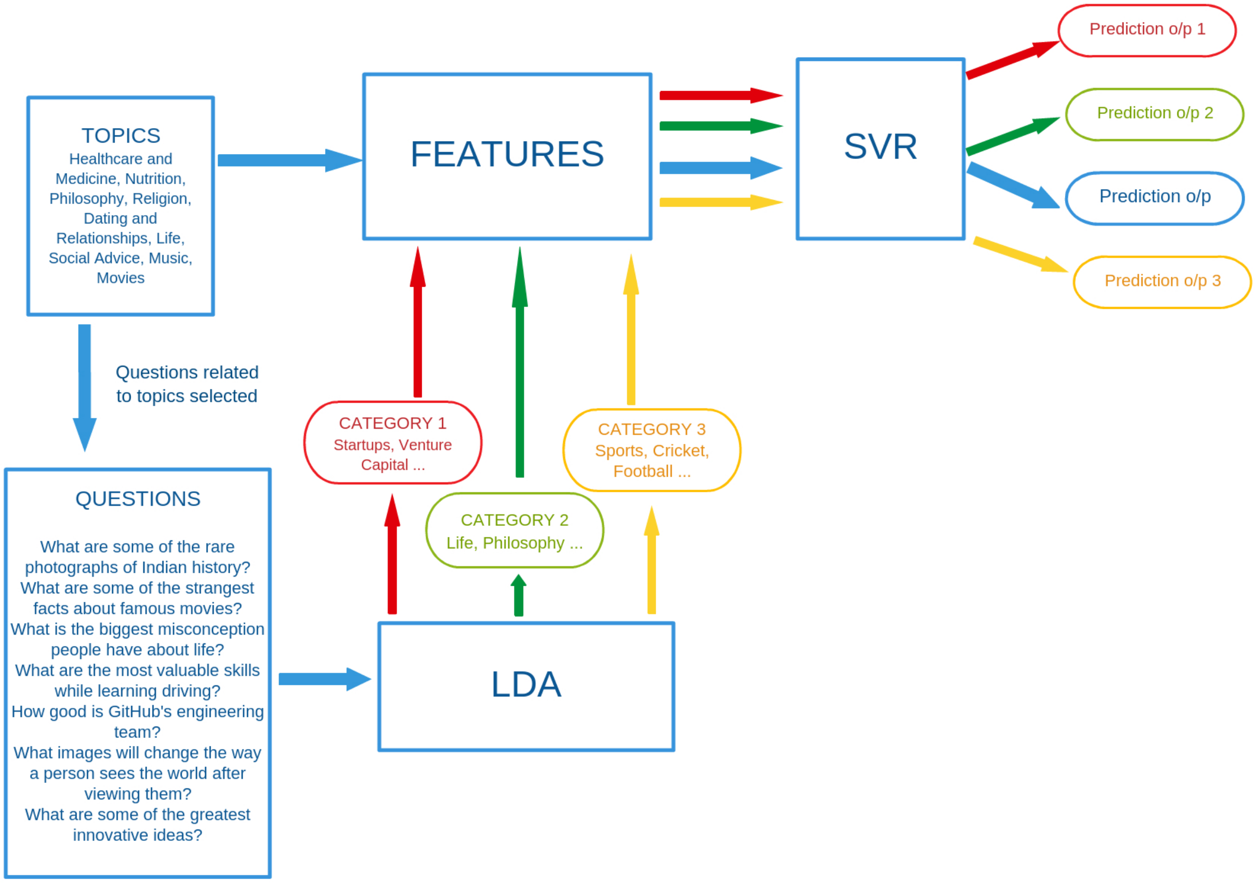

TOPICS
Healthcare and Medicine, Nutrition, Philosophy, Religion, Dating and Relationships, Life, Social Advice, Music, Movies
FEATURES
SVR
Prediction o/p 1
Prediction o/p 2
Prediction o/p
Prediction o/p 3
Questions related to topics selected
CATEGORY 1
Startups, Venture Capital ...
CATEGORY 2
Life, Philosophy ...
CATEGORY 3
Sports, Cricket, Football ...
QUESTIONS
What are some of the rare photographs of Indian history?
What are some of the strangest facts about famous movies?
What is the biggest misconception people have about life?
What are the most valuable skills while learning driving?
How good is GitHub's engineering team?
What images will change the way a person sees the world after viewing them?
What are some of the greatest innovative ideas?
LDA

Figure 6: (Color online) A schematic of our proposed framework. Different color codes (red, green, yellow) are for different categories of topics that are found after running LDA on the question texts related to the topics.

where A is the document containing all the questions which are tagged by topic i and $p(w|A)$ is the probability of a word belonging to the document A.

Topical question clarity: Topical question clarity quantifies the cohesiveness of all questions tagged by the topic. Topical question clarity of topic i ($QuesClarity_i$) is computed as the Kullback-Leibler (KL) divergence between the unigram language model inferred from the document L_Q^i containing all the questions for the i^{th} topic and the background language model from the entire question collection G_Q. This measure is inspired by hashtag clarity measure by (Ma, Sun, and Cong 2012).

$$QuesClarity_i = -\sum_{w \in L_Q^i} p(w|L_Q^i) \times \log \frac{p(w|L_Q^i)}{p(w|G_Q)}$$

Topical answer clarity: Topical answer clarity like topical question clarity quantifies the cohesiveness of all answers to the questions tagged by the topic. Topical answer clarity of topic i ($AnsClarity_i$) is computed as the Kullback-Leibler (KL) divergence between the unigram language model inferred from the document L_A^i containing all the answers for the i^{th} topic and the background language model from the entire answer collection G_A.

$$AnsClarity_i = -\sum_{w \in L_A^i} p(w|L_A^i) \times \log \frac{p(w|L_A^i)}{p(w|G_A)}$$

Frequency of n-grams from the question content in English texts: We search for 2, 3 grams of the words from the question text in the corpus of 1 million contemporary American English words[5]. We use the presence and frequency of bigrams and trigrams, each as a feature for the prediction task. Note that this is one of the very unique features that is introduced by us for the first time in this paper. Our hypothesis is that a popular topic will use more frequent n-grams for better readability of questions.

Frequency of n-grams from the answer content in English texts: We search for 2, 3 grams of the words from

[5] http://www.ngrams.info/samples_coca1.asp

Top		Bottom	
Topics	Followers	Topics	Followers
technology	3.2M	entrance exams	141
science	2.5M	cover songs	129
business	2.3M	record companies	124
books	2.3M	software companies	121
travel	2M	external hard drives	117
movies	2M	playlists	109
music	1.8M	healthcare in the united states	109
health	1.8M	sports injuries	105
food	1.8M	work experience	97
education	1.7M	what scientific evidence exists for x?	95
design	1.6M	atomic molecular and optical physics amo	80
psychology	1.5M	merchant services	78
economics	1.5M	am i too old to do x?	75
history	1.4M	graphic violence	70
entertainment	1.4M	education advice	68
cooking	1.4M	what does x think of y?	61
writing	1.3M	what are the pros & cons of x?	57
sports	1.2M	theists	52
philosophy	1.1M	international festivals and events	46
marketing	1.1M	godzilla 2014 movie	44

Table 4: Top 20 topics and bottom 20 topics in Quora based on number of followers (December 2014).

the answer text in the corpus of 1 million contemporary American English words. Similar to the questions, we use the presence and frequency of bigrams and trigrams, each as a feature for the prediction task.

Question content words From the question content pertaining to a topic, we remove all the function words and Wh-question words; the quantity of the remaining content words is used as a feature for the popularity prediction of a topic.

In-vocabulary words and Out-of-Vocabulary words in question texts: For each of the topic i, we check whether a word appearing in the document (Q_i) consisting of all the questions for the topic, is an In-vocabulary word or an Out-of-vocabulary word. We consider the ratio of In-vocabulary to the Out-of-Vocabulary (OOV) words as a feature of our model.

In-vocabulary words and Out-of-Vocabulary words in answer texts: For each of the topic i, a document (A_i) is created consisting of all the answers for the topic. We then check whether a word is an In-vocabulary word or an Out-of-vocabulary word. The ratio of In-vocabulary to the Out-of-Vocabulary (OOV) words act as a feature.

Cognitive dimension: There could be differences in the cognitive dimension (linguistic and psychological) for different topics (for instance, the cognitive dimension might vary largely between a popular and a less popular topic). To capture the phenomena, we find out the belongingness of a topic's questions to the different cognitive dimensions as regression features. Words from a document containing all the questions are classified into various linguistic (Part-of-speeches of the words, swear words etc.) and psychological categories (physical, social, optimistic, self, anger, positive emotion, negative emotion, sadness etc.) by LIWC software (Pennebaker, Francis, and Booth 2001). We consider 59 such features.

User features Not only contents of the questions and answers, user importance might also be a factor for the popularity prediction. User features include various user related characteristics.

User following: For a topic, we find out the answerers for all the questions in that topic and find out the average number of following of these users. The logarithm of this value act as a feature.

User follower: The user having large number of followers is usually a celebrity. To capture the phenomena that if a celebrity user is answering a question, the topic tagged would eventually get more visibility, we consider logarithm of the average follower count of the answerers as a feature for our model.

User responsiveness: We define the responsiveness of a user as number of questions he/she had answered. If a topic has high responsive answerer, it would have lesser unresolved questions. We consider average responsiveness of the answerers for the questions in a topic as a feature to our prediction model.

Performance of our regression model:

In this section, we analyze the performance of our prediction model. We consider 3222 topics for our prediction task. We train the regression model with 2400 topics and the remaining 822 topics are used for testing. We use Support Vector Regression (SVR) implemented in Weka Toolkit (Hall et al. 2009) for prediction. For evaluating how good the prediction is, we use Pearson correlation coefficient and root mean square error (RMSE). For the above setting, we achieve high correlation coefficient ($\sim$0.773) and low root mean square error ($\sim$1.065). In table 5, we present the contribution of different combinations of feature types, demonstrating how each of these feature types affect to the prediction and whether any feature type is masked by a stronger signal produced by other feature types. We observe that both context and content are strong feature types whereas user features are relatively weak. Among context and content features, content features are more discriminative.

Discriminative features:

In this subsection, we discuss the discriminative power of the individual features. We use $RELIEFF$ feature selection algorithm (Kononenko, Simec, and Robnik-Sikonja

Feature model	Correlation coefficient	RMSE
Context	0.6423	1.2826
Content	0.7708	1.07
User	0.2605	1.6144
Context + Content	0.7723	1.067
Context + User	0.6472	1.2753
Content + User	0.7683	1.074
All	**0.7731**	**1.0653**

Table 5: Performance of various combinations of feature categories.

1997) in Weka Toolkit to rank the attributes. In table 6, we show the rank of the features in terms of their discriminative power for prediction. The rank order clearly indicates the dominance of the content features. In the top 25, only one context feature finds place. Among content features, topical question diversity, topical question clarity, topical answer diversity and topical answer clarity are important factors for popularity prediction. The other important subgroups of content features are the LIWC features and among them Parts-of-Speech categories are more discriminative and featuring in top 25.

Categorization in topics

Rank	Features	Type
1	Topical Question Diversity	Content
2	No of Questions	Context
3	Topical Question Clarity	Content
4	Topical Answer Diversity	Content
5	Topical Answer Clarity	Content
6	fraction of time related words (liwc feature)	Content
7	fraction of third person plural noun (liwc feature)	Content
8	fraction of words with past tense (liwc feature)	Content
9	fraction of first person singular noun (liwc feature)	Content
10	Questions' TriGram Frequency in English texts	Content
11	Answers' InVocabulary to OOV Ratio	Content
12	fraction of relative words (liwc feature)	Content
13	fraction of prepositions (liwc feature)	Content
14	fraction of words with future tense (liwc feature)	Content
15	fraction of articles (liwc feature)	Content
16	fraction of adverbs (liwc feature)	Content
17	Questions' BiGrams Frequency in English texts	Content
18	fraction of second person noun (liwc feature)	Content
19	Questions' BiGrams Presence in English texts	Content
20	Question's Content Words	Content
21	fraction of common verbs (liwc feature)	Content
22	fraction of first person plural noun (liwc feature)	Content
23	Questions' InVocabulary to OOV Ratio	Content
24	fraction of "death" related words (liwc feature)	Content
25	fraction of third person singular noun (liwc feature)	Content

Table 6: Top 25 predictive features and their types.

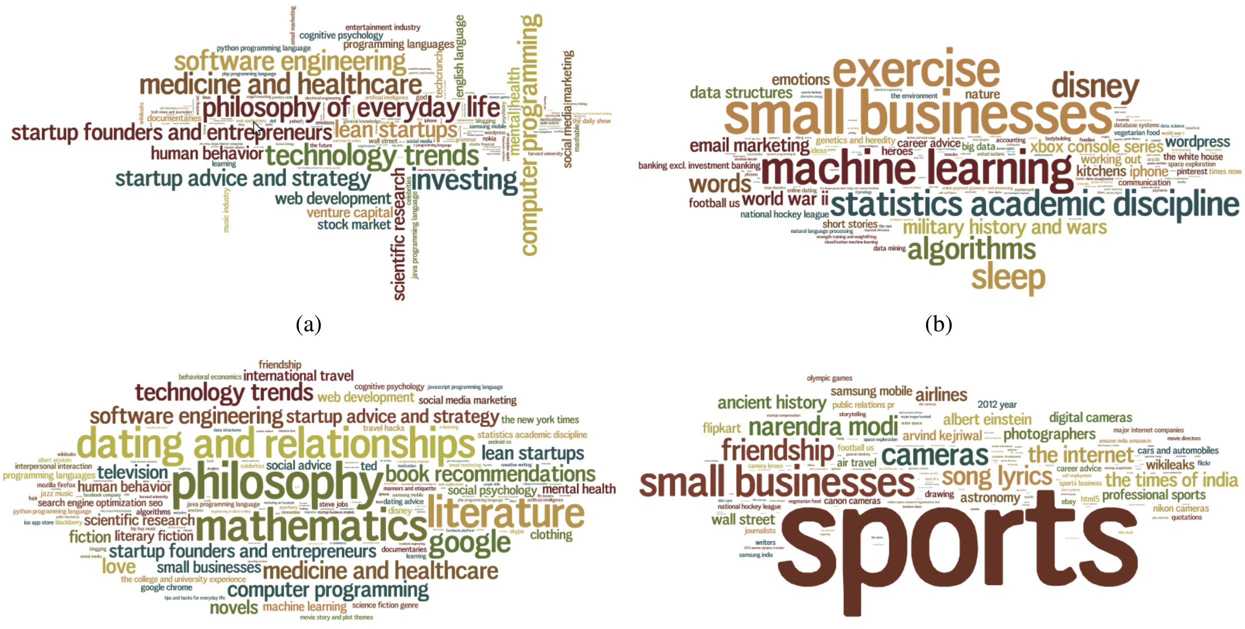
There are inherent broad categories of topics like entertainment, education, health, travel and living, technology etc. that encompass many less broader topics; for example entertainment includes music, movies, tv series, sports etc.; education includes topics like various disciplines of science and arts, topics related to academic institutions etc. Quora manages such kind of topic categories for some of the popular topics [6]. Identifying these categories is difficult because of the overlapping nature of the categories and hence may differ from person to person. In this section, we adopt Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), a well-known generative probabilistic model for discovery of those latent categories. For each topic, we create a document consisting of all the questions related to the topic and then we feed the documents into the LDA model which generates the belongingness probabilities of the topics into different broad categories. Post-categorization, we perform the prediction task with SVR model on each of these categories using 10-fold cross-validation technique. In table 7, we present the prediction accuracies (in terms of correlation coefficient and RMSE) for various values of predefined number of categories K. For each of the different values of K, we obtain at least one category for which the prediction accuracy is higher than the case when we have no categorization. For $K = 12$, we obtain 4 categories of topic for which the correlation coefficient is higher than the case with no categorization. If we consider RMSE as metric of evaluation, we observe that for $K = 4$, in all the topic categories, RMSE values are less than the case with no categorization. This method of stratifying data samples yield better prediction accuracies of certain categories whereas it also yields prediction accuracies for categories which are less than the case with no categorization. One reason for these lower prediction accuracies is the datasize imbalance. However there are other reasons outlined below

In fig 7, we show the topic clouds for the categories for which we achieve best and worst prediction accuracies. The best prediction accuracy is achieved for the category in which topics are mostly related to technology, startups and business are mentioned whereas the worst performing category mostly involve a wide range of various topics starting from cooking to sports, music, songs etc. eventually garbling the topical cohesiveness. In other words, if a category has sufficient data and is thematically well separated then the prediction accuracy is far more than the case where there is either data scarcity or the theme is not well separated.

Conclusions and future works

With increasing popularity and quality control, Quora has developed a rich knowledge base of Q&A. Quora topics play

[6] http://www.quora.com/sitemap



(a)

(b)

(c)

(d)

Figure 7: Topic clouds in various categories corresponding to various values of K. Top row: Topic clouds for the topics in the category that produces a) best prediction accuracy b) worst prediction accuracy for $K = 12$. Bottom row: Topic clouds for the topics in the category that produces a) best prediction accuracy b) worst prediction accuracy $K = 16$. The size of the topic names are proportional to the follower count of the corresponding topic.

vital role in organization of such content. Our study unfolds for the first time the topic dynamics and their popularity. To the best of our knowledge, this work represents the most comprehensive study of topic growth dynamics and understanding of topic popularity in Quora. In this paper, we have analyzed topic evolution over time, the inter-topic dynamics, stability and migration among topics and the factors which affects popularity of a topic.

We proposed a framework for predicting popularity of a topic. Our proposed model achieves a high correlation between the predicted value and actual value (correlation coefficient ∼0.773) with low root mean square error (∼1.065). We observe that the content features are most discriminative compared to others. We further categorize the topics into a set of categories (K) by running a LDA model on the question texts associated with the topics. In comparison to the data sample with no categorization, this stratification of the topics enhances the prediction accuracies for several categories. For $K = 12$, we get many categories performing better in terms of prediction accuracy compared to the topics dataset with no categorization.

There are quite a few other interesting directions that can be explored in future. One such direction could be to study personalized diverse topic recommendation system which will not only recommend trending topics but also interesting topics suiting one's personal requests. We also plan to release the Quora dataset soon for the research community to facilitate further investigations.

Acknowledgments

This work has been supported by Microsoft Corporation and Microsoft Research India under the Microsoft India PhD fellowship Award. SKM would like to thank Google India Pvt. Ltd. and Flipkart Internet Pvt. Ltd. for travel support.

References

Adamic, L. A.; Zhang, J.; Bakshy, E.; and Ackerman, M. S. 2008. Knowledge sharing and yahoo answers: Everyone knows something. WWW '08, 665–674. New York, NY, USA: ACM.

Anderson, A.; Huttenlocher, D.; Kleinberg, J.; and Leskovec, J. 2012. Discovering value from community activity on focused question answering sites: A case study of stack overflow. KDD '12, 850–858. New York, NY, USA: ACM.

Bhat, V.; Gokhale, A.; Jadhav, R.; Pudipeddi, J. S.; and Akoglu, L. 2014. Min(e)d your tags: Analysis of question response time in stackoverflow. ASONAM '14, 328–335.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Correa, D., and Sureka, A. 2014. Chaff from the wheat: Characterization and modeling of deleted questions on stack overflow. WWW '14, 631–642. New York, NY, USA: ACM.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.* 11(1):10–18.

	Categories (K)			
	K = 16	K = 12	K = 8	K = 4
Correlation Coefficient (Root Mean square Error)	0.6755 (1.0769)	0.717 **(0.9967)**	0.7465 (1.0796)	0.7597 **(1.0522)**
	0.7598 (1.0843)	0.6691 (1.1269)	0.7355 (1.0715)	**0.7818** **(1.007)**
	0.7687 **(1.0119)**	**0.7932** **(1.0651)**	0.7339 (1.0646)	0.7675 **(1.0255)**
	0.6747 (1.2437)	0.6846 **(1.0444)**	0.7216 (1.1029)	0.7665 **(1.0572)**
	0.6166 (1.338)	**0.7975** **(1.0605)**	0.7182 (1.0917)	
	0.8053 **(0.954)**	**0.7739** **(1.0277)**	0.7636 (1.0788)	
	0.7555 (1.0799)	**0.7773** **(1.0218)**	**0.7995** **(0.9246)**	
	0.6711 (1.2581)	0.6906 (1.1223)	0.7498 **(1.0571)**	
	0.7198 (1.1412)	0.7054 (1.167)		
	0.682 (1.1764)	0.7584 (1.1315)		
	0.7571 **(1.0622)**	0.69 (1.2072)		
	0.5452 (1.3073)	0.7122 (1.3029)		
	0.6481 (1.187)			
	0.703 (1.1752)			
	0.683 (1.2755)			
	0.751 (1.0796)			

Table 7: Performance of the prediction model for various categories (K = 4, 8, 12, 16) of topics. Bold faces mark the cases where the correlation coefficient is higher than the case for which there was no categorization. It also indicates the cases where RMSE value is less compared to no categorization (see table 5).

Harper, F. M.; Raban, D.; Rafaeli, S.; and Konstan, J. A. 2008. Predictors of answer quality in online q&a sites. CHI '08, 865–874. New York, NY, USA: ACM.

Harper, F. M.; Moy, D.; and Konstan, J. A. 2009. Facts or friends?: Distinguishing informational and conversational questions in social q&a sites. CHI '09, 759–768. New York, NY, USA: ACM.

Hsieh, G., and Counts, S. 2009. Mimir: A market-based real-time question and answer service. CHI '09, 769–778. New York, NY, USA: ACM.

Jeon, J.; Croft, W. B.; Lee, J. H.; and Park, S. 2006. A framework to predict the quality of answers with non-textual features. SIGIR '06, 228–235. New York, NY, USA: ACM.

Jurczyk, P., and Agichtein, E. 2007. Discovering authorities in question answer communities by using link analysis. CIKM '07, 919–922. New York, NY, USA: ACM.

Kononenko, I.; Simec, E.; and Robnik-Sikonja, M. 1997. Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence* 7:39–55.

Lerman, K., and Galstyan, A. 2008. Analysis of social voting patterns on digg. WOSN '08, 7–12. New York, NY, USA: ACM.

Li, B., and King, I. 2010. Routing questions to appropriate answerers in community question answering services. CIKM '10, 1585–1588. New York, NY, USA: ACM.

Li, B.; Jin, T.; Lyu, M. R.; King, I.; and Mak, B. 2012. Analyzing and predicting question quality in community question answering services. WWW '12 Companion, 775–782. New York, NY, USA: ACM.

Ma, Z.; Sun, A.; and Cong, G. 2012. Will this #hashtag be popular tomorrow? SIGIR '12, 1173–1174. New York, NY, USA: ACM.

Mamykina, L.; Manoim, B.; Mittal, M.; Hripcsak, G.; and Hartmann, B. 2011. Design lessons from the fastest q&a site in the west. CHI '11, 2857–2866. New York, NY, USA: ACM.

Mendes Rodrigues, E., and Milic-Frayling, N. 2009. Socializing or knowledge sharing?: Characterizing social intent in community question answering. CIKM '09, 1127–1136. New York, NY, USA: ACM.

Pal, A.; Chang, S.; and Konstan, J. A. 2012. Evolution of experts in question answering communities. In Breslin, J. G.; Ellison, N. B.; Shanahan, J. G.; and Tufekci, Z., eds., *ICWSM*. The AAAI Press.

Paul, S. A.; Hong, L.; and Chi, E. H. 2012. Who is authoritative? understanding reputation mechanisms in quora. *arXiv preprint arXiv:1204.3724*.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. *Linguistic Inquiry and Word Count*. Mahwah, NJ: Lawerence Erlbaum Associates.

Rodrigues, E. M.; Milic-Frayling, N.; and Fortuna, B. 2008. Social tagging behaviour in community-driven question answering. In *Web Intelligence*, 112–119. IEEE.

Shah, C., and Pomerantz, J. 2010. Evaluating and predicting answer quality in community qa. SIGIR '10, 411–418. New York, NY, USA: ACM.

Shtok, A.; Dror, G.; Maarek, Y.; and Szpektor, I. 2012. Learning from the past: Answering new questions with past answers. WWW '12, 759–768. New York, NY, USA: ACM.

Tausczik, Y. R., and Pennebaker, J. W. 2011. Predicting the perceived quality of online mathematics contributions from users' reputations. CHI '11, 1885–1888. New York, NY, USA: ACM.

Wang, G.; Gill, K.; Mohanlal, M.; Zheng, H.; and Zhao, B. Y. 2013. Wisdom in the social crowd: An analysis of quora. WWW '13, 1341–1352.

Zhang, J.; Ackerman, M. S.; and Adamic, L. 2007. Expertise networks in online communities: Structure and algorithms. WWW '07, 221–230. New York, NY, USA: ACM.