2018 IEEE International Conference on Data Mining Workshops (ICDMW)

Khan Academy: A Social Networking and Community Question Answering Perspective

1st Sneha Mondal
IBM Research-India
Bangalore, India
snmondal@in.ibm.com

2nd Akshay Gugnani
IBM Research-India
Bangalore, India
akgugnan@in.ibm.com

3rd Renuka Sindhgatta
IBM Research-India
Bangalore, India
renuka.sr@in.ibm.com

4th Vinay Kumar Reddy Kasireddy
IBM T J Watson Research Centre
YorkTown Heights, NY, USA
vinay.kasireddy@ibm.com

***Abstract*—This paper studies the social networking and community question answering aspects of Khan Academy, a popular yet largely uninvestigated online educational forum. We start with a brief description of our dataset and data collection methodology. We then proceed to construct the underlying network and study its topology based on degree distribution and degree correlation. We examine the performance of different ranking algorithms vis-a-vis user-provided expertise ranking, and explain the observed high correlation with PageRank. Furthermore, we empirically observe how interactions evolve as a course advances, and note that while the network progressively shrinks because low-performing nodes drop out, it also becomes a more tight-knit community. We infer that users who drop out are possibly novice learners, who ask several questions but lack the required expertise to answer many questions themselves. Throughout our work, we draw parallels with existing studies on other web-based question-answering forums which are primarily targeted towards an adult population.**

***Index Terms*—education, e-learning, community question answering, social networks**

I. Introduction

Massive open online courses (MOOCs) have become an exciting part of e-learning, with a plethora of general as well as specialized video lectures, to complement the traditional education system. Coursera, edX, Udacity and Khan Academy are some of the most popular MOOCs hosting websites. Among them, Khan Academy stands out because all its courses are self-paced, and the platform builds entirely on human expertise, the motivation of individuals to learn, ask questions and share knowledge. While there are efforts to incentivize participation through *badges* which users must earn by contributing to the community, there are no incentives in terms of assignment grading or course-completion certificates.

Khan Academy offers practice exercises, instructional videos, and a personalized learning dashboard. While they offer courses on subjects as diverse as mathematics, science, economics, humanities, and computer programming; most of the content available is for Math. The content spans early math through Class-12 (i.e., high school), and covers topics ranging from arithmetic to multivariate calculus.

Each course is divided into topics, which further consist of multiple concepts. The learning environment involves watching a video explanation of a concept followed by self-assessments in the form of questions. Each video explanation is supplemented by a transcript of the video and a discussion forum (akin to a community Q&A) where students and instructors can interact, ask questions and resolve doubts.

Community Q&A sites are studied similarly to social networks where traditional friendship relationships are replaced with interactions leading to information exchange. User interactions on Khan Academy are initiated by a user asking a question, they typically continue when other users answer the question, and may extend further through the exchange of insightful comments and follow-up questions. Additionally, anyone with an account can provide feedback by up-voting useful questions, comments and answers; or by down-voting factually incorrect and low-quality content or spam. Analysis of the graph emerging from the different types of user interactions provides insight into the activity patterns of users, and in particular, of experts and novice learners.

II. Related Work

There have been studies from the social networking viewpoint of question answering communities such as StackOverflow [1], [2], as well as social coding communities such as GitHub [3]. On the other hand, there have also been studies on examining the social network aspect of YouTube by measuring the subscription graph, comment graph, and video content corpus [4]. YouTube is found to deviate significantly from network characteristics that mark traditional online social networks (such as homophily, reciprocative linking, and assortativity). However, it shows remarkable similarities with another content-driven online social network, Twitter [5]. MOOCs such as Khan Academy are a blend of the video hosting aspect of YouTube and the question answering aspect of StackOverflow, thus making them a unique and interesting topic of study from a social networking viewpoint.

There exists literature on incentivizing student participation in online educational forums [6], [7]. Furthermore, a line of research on community question answer systems focuses on detecting collusive manipulations [8], [9] and preventing them [10], while another line of research focuses on evaluating and predicting answer quality [11]. There have been studies from the sociological viewpoint on the impact of Khan Academy on learning and education [12], [13].

2375-9259/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDMW.2018.00058

To the best of our knowledge, ours is the first study focusing on the social networking and the community question answering aspects of Khan Academy.

III. Data Collection and Description

We decided to focus our initial experiments on a small subset of Khan Academy's vast data source. We take into consideration concepts of Math, specifically, *Trigonometry*, *Probability and Statistics*, and *Linear Algebra*, because a preliminary analysis showed that these three concepts have a relatively higher engagement of users as well as a reasonable number of topics covered under each.

Khan Academy provides open APIs for developers to access and use Khan Academy data[1]. The RESTful API gives developers access to nearly the entirety of data and outputs easy-to-parse JSONs.

At the highest levels, the API provides information about the content covered by Khan Academy. Videos under individual topics are accessible via topic names (e.g. Overview and History of Algebra, Banking and Money, Circulatory and Pulmonary Systems, etc). Besides, one can use the topic names to access topic-level exercises (e.g. Trigonometric Functions, SAT Math Level 1, Tests for Convergence and Divergence, etc). Since exercises and videos are linked, it is also possible to access videos pertaining to a specific exercise.

Khan Academy's *topictree* API allows one to make unauthenticated API calls and mine nearly the entire library, organized by topic, individual videos or exercises. For our purpose, we employ the Topictree API[2] to obtain a hierarchical organization of all topics, along with their videos. Topics are organized into groups and sub-groups/concepts such as "Math" and "Linear Algebra", respectively. The general organization follows the "Jump to Topic" bar of links found on Khan Academy's homepage.

As discussed earlier, a video is accompanied by a transcript and a community Q&A forum. For each desired video, we obtain the transcript, as well as discussion threads (i.e., answers and comments) of up to a hundred questions from the ensuing forum.

For every question, answer or comment, we gather pertinent information such as: the unique identifier of the author, the upvote and downvote count, badges associated with the content; and meta-details such as the timestamp of the content. We leverage all this data in the context of our study of the online community, as detailed in the following sections.

IV. Empirical Study of the Online Community

The content on Math forms a bulk of Khan Academy's offering, with instructional videos and practice exercises on topics as diverse as linear algebra, trigonometry and multivariate calculus. For ease of analysis, in this section, we restrict our study to *Trigonometry* and *Probability and Statistics*, two of the largest topics under Math. Since our observations on both topics are similar, we report results for the Trigonometry dataset alone.

A. Constructing the Social Network

From the videos posted under Trigonometry, we create a directed graph (henceforth referred to as the Trigonometry graph). In the graph $G = (V, E)$, nodes in set V correspond to users who either pose a question, or submit an answer. Edges between these nodes in E indicate a question-answer interaction. The edges are directed from the author of the question to the author of the answer, and are weighted by the number of such interactions between the two users. Hence, the in-degree of a node represents the number of *distinct users* whose questions it has answered, whereas the weighted in-degree represents the total number of such answers, i.e., the *overall answer-contribution*. The Trigonometry graph has **1691** nodes and **2291** directed edges. The sum of edge weights in the graph (i.e. the number of question-answer pairs) is **9655**.

The aforementioned network is based on question-answering only. It is worth noting that, unlike some other case studies, one cannot study the who-upvotes-whom network for Khan Academy owing to its anonymous upvoting system.

B. Characterizing the Social Network

1) Degree Distribution: Typical social networks are observed to follow a power-law degree distribution, which is of the form $f(d) \propto d^{-\alpha}$, where d is the degree, $f(d)$ is the fraction of nodes with degree d, and α is the power-law exponent. Most real networks including social networks have a power-law exponent between 2 and 3 [5].

The power-law degree distribution is a reflection of the highly skewed distribution of participation. However, owing to the heavy tail of such a distribution, nodes with very high degrees are more common than we would expect in a distribution such as Gaussian. That is, there are relatively more common users with very high in-degrees; similarly there exist significant number of users with very high out-degrees as well.

Figure 1 plots the degree-distribution of the Trigonometry network. The power law exponent is 3.43 with respect to in-degrees, and 2.72 with respect to out-degrees. The power law exponents for the studied network are on the *higher* side, implying that there are some extremely active users who answer a lot of questions while a majority of users answer only a few. Likewise, many users pose only a single question, but some ask a dozen or more. This observation is further emphasized by the fact that 899 out of 1691 nodes,i.e. nearly 53% have *zero* out-degree. These nodes have not posed a single question, but only contributed to the community through answers.

Further, as Figure 2 shows, the top 5% (in terms of overall contribution) of answering-nodes contribute an overwhelming 25% to the overall answer-content, with the marginal contribution falling progressively. Thus, the top few contributors are critical in keeping the community Q&A active, and Khan Academy boosts users to become top contributors by incentivizing them through badges of increasing prestige.

[1] https://api-explorer.khanacademy.org/

[2] http://www.khanacademy.org/api/v1/topictree

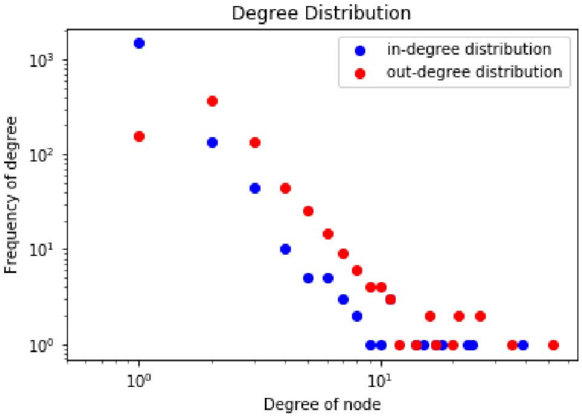

Fig. 1. Degree distribution

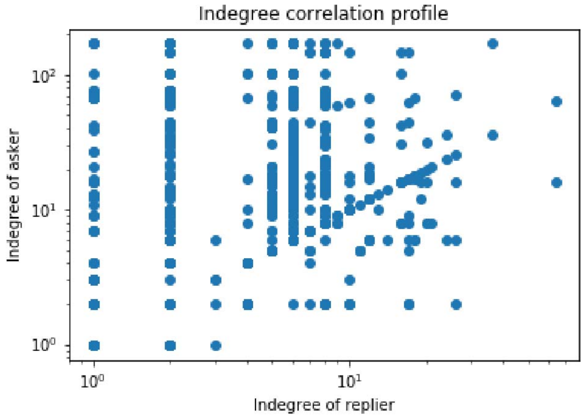

Fig. 3. Assortative mixing of answerers

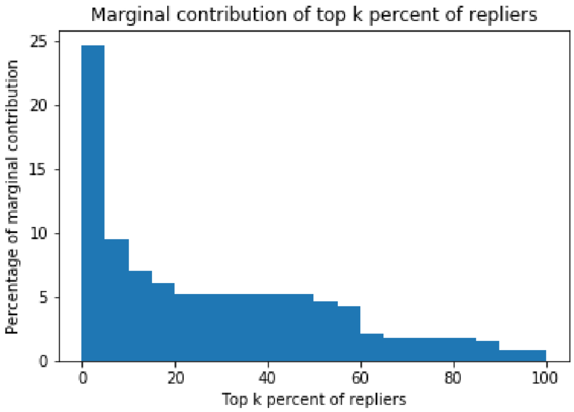

Fig. 2. Marginal percentage contribution of top repliers

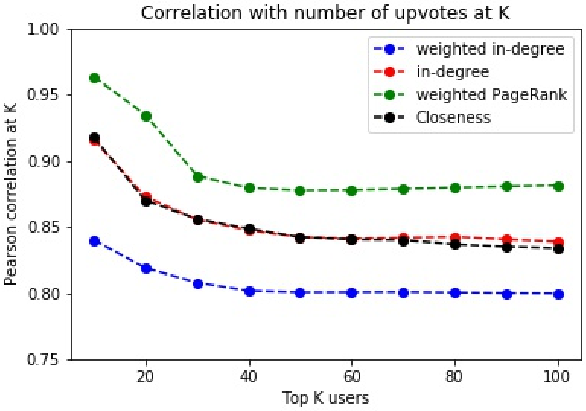

Fig. 4. Correlation of different ranking algorithms with number of upvotes for top K users

2) Degree Correlation: Along with degree distribution, degree-correlation or assortativity provides crucial information about the nature of interactions in a network. For instance, we would like to know whether high-volume repliers only reply to novices, or to other high-volume repliers like themselves.

Figure 3 presents a simplified correlation profile that, for each asker-replier pair counts the indegree of the replier versus the indegree of the asker. While positive assortativity is quite common in social networks, the Trigonometry graph is neither assortative nor dissortative, with a Pearson correlation coefficient of 0.152.

Unexpectedly, Figure 3 seems to suggest that low in-degree users (ones who lack the expertise to answer others' questions) tend to reply to high in-degree (expert) users as well. However, on cross-checking with the Khan Academy website, we observed that several of these "answers" don't address the question, but are actually follow-up questions themselves, or simple comments posted on the same thread.

C. Node Expertise Ranking

There exist multiple algorithms for ranking nodes by importance (or expertise) in a social network. For instance, if a person answers a lot of questions on a topic, it is often the case that they know the topic well. Thus *weighted-indegree* forms a measure of a user's expertise.

A slightly different measure is counting how many other users a user helps. For instance, a user may answer a lot of questions by repeatedly interacting with a specific set of users. On the other hand, a user who posts fewer answers, but in the process helps a greater number of users, could have broader expertise. In our context, this is measured by a node's *in-degree*.

Closeness centrality measures how well-connected a node is. Users with high values for closeness centrality that have better access to information and more direct (i.e., shorter) interaction paths with most nodes in the network.

PageRank is a celebrated ranking algorithm that encodes the peer assessment of the value of node by taking into account

TABLE I
DETAILS OF THE SUBGRAPHS CORRESPONDING TO PROGRESSIVE SUB-TOPICS

Sub-topics	Nodes	Edges	Density
Vectors and Spaces	3058	5169	0.5×10^{-3}
Matrix Transformations	1575	2581	1.04×10^{-3}
Alternate Co-ordinate Systems	881	1366	1.69×10^{-3}

not just the number of nodes linking to it, but also the number of nodes pointing to those nodes, and so on. In our context, an answering node is given higher importance if it answers questions from other high-volume answering nodes.

The number of upvotes on an answer provides explicit feedback on the quality of the content, as assessed by the user-community. Thus, we rank nodes by the average upvotes their answers receive, and use this as a “gold standard” for comparison. Figure 4 illustrates the correlation of different ranking algorithms with number of upvotes for top K users.

We note that PageRank outperforms the other metrics in terms of correlation with the gold standard, and one possible explanation for this observation is as follows : High in-degree nodes (i.e., high-volume repliers) are typically experts, and if at all they pose questions, we’d expect them to be more challenging than a question posted by a newbie. When a node answers questions from an expert (consequently receiving a higher weightage in PageRank), it is also a testament to their own expertise which is reflected in the quality of the answer (consequently garnering significant upvotes from the community), thus leading to the observed high correlation as compared to other ranking algorithms.

D. Characterizing the Temporal Graph

In this section, we examine how user-interaction in Khan Academy’s discussion forum evolves when a course become progressively harder. To suitably study these patterns, we require a topic with an unambiguous ordering of difficulty amongst sub-topics. We therefore restrict ourselves to the topic of *Linear Algebra*, which is further divided into *Vectors and Spaces*, *Matrix Transformations* and *Alternate Co-ordinate Systems*.[3] While Khan Academy itself doesn’t mention that the videos must be viewed in the aforementioned order, we can reliably assume that the listed sub-topics are in increasing order of difficulty, in keeping with the design of a standard classroom course on Linear Algebra.

We construct three subgraphs (each corresponding to a sub-topic), from the Linear Algebra graph. The node, edge and density statistics of the subgraphs are shown in Table I. From these values, we make three important observations :

- As the course advances, there is a distinct fall in participation, indicated by the progressive decrease in the number of interacting nodes. Thus, the graphs *shrink* in size as the course proceeds.
- The shrinkage is caused primarily owing to the departure of low-volume repliers (see Figure 5). For instance, the

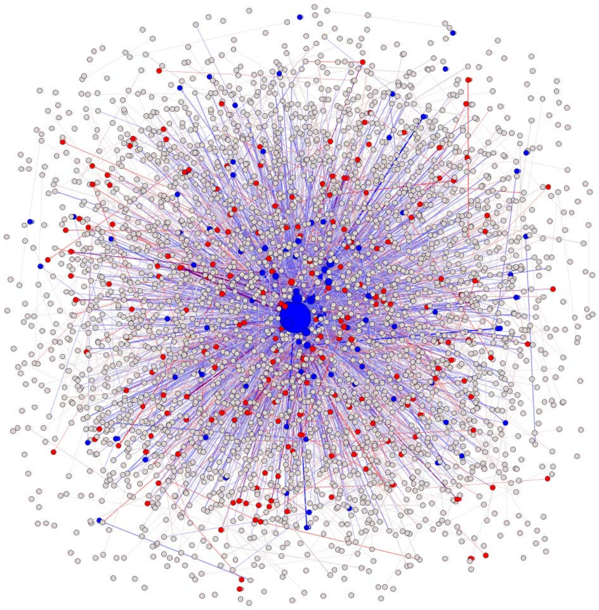

Fig. 5. Persistence of users over the advancement of course (blue = nodes which persist through all three topics, red = nodes which persist for the first two topics, gray = nodes which depart after the first topic) where node sizes correspond to their in-degrees

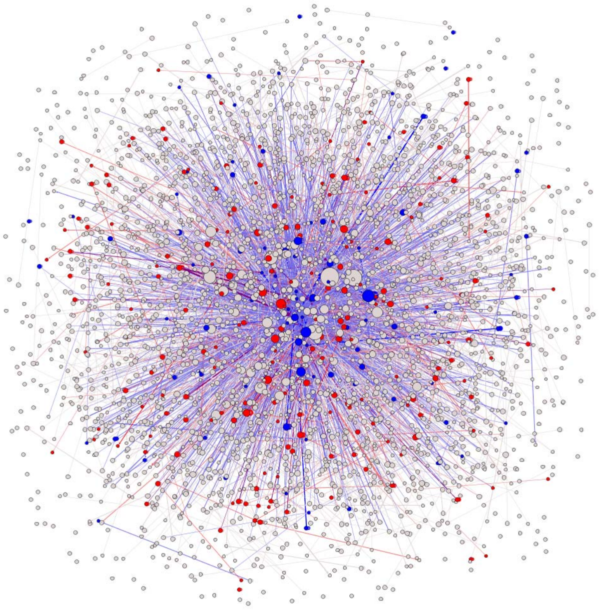

Fig. 6. Persistence of users over the advancement of course (blue = nodes which persist through all three topics, (red = nodes which persist for the first two topics, gray = nodes which depart after the first topic) where node sizes correspond to their out-degrees

[3] https://www.khanacademy.org/math/linear-algebra

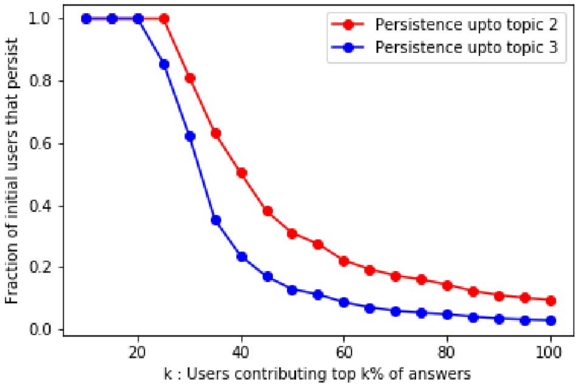

Fig. 7. Persistence of users over the advancement of course

number of users who answer only a single question (i.e., nodes with in-degree = 1) is almost halved at each step. These nodes are typically novice learners, who lack the expertise to contribute significantly in terms of answers, and eventually drop out as the content gets harder.

- Although the sub-graphs shrink in size, they progressively become *more* dense i.e., there are more frequent interactions amongst a smaller set of participants. One possible explanation could be that the users persisting to the end remain committed to the learning process and tend to engage more actively as compared to users who drop out.

Figures 5 and 6 illustrate this point further. We observe that that nodes which drop out are typically the ones who ask several questions (high out-degree in Figure 6) but answer very few questions (low in-degree in Figures 5). Hence they are possibly novice learners who participate primarily to have their doubts clarified on basic topics, but lack the required expertise to answer many questions themselves.

Figure 7 illustrates the persistence of users over the advancement of course, based on how they are positioned as answer contributors. The top 20% of answer-contributors are consistent and contribute in all three sub-topics, but beyond that point, we see a steep decline the fraction of users that persist as the content becomes more challenging.

V. Conclusions & Future Work

In summary, we wanted to examine how users interact on a self-paced learning platform such as Khan Academy, and whether the underlying network provides insights into user behaviour.

To do this, we followed three steps : First, we constructed the network and observed its topology in terms of degree distribution and correlation, and found that a few high-volume users provide a majority of the content, whereas most users ask or answer only a couple of questions. Next, we evaluated node-expertise ranking algorithms against user-provided feedback in the form of upvotes, and discovered that users who satisfactorily answer questions from experts garner significant upvotes from the community. Finally, we observed that as the curriculum becomes more demanding, most low-performing users drop out, whereas high-performing experts tend to persist. Also, although viewership reduces as the course progresses, the smaller community tends to engage more actively than before.

These findings are preliminary results, and barely scratch the surface of the information that Khan Academy provides. As an immediate next step, we would like to study the entire Math dataset, and see if natural communities emerge in the underlying network; and if so, whether these communities correspond to specific topics under math. It would also be interesting to understand if node expertise is limited to certain specific topics, or extends across topics to Math as a whole. We expect that more careful analysis of Khan Academy user activity will help us differentiate between usage patterns of early-stage learners, vis-a-vis high-school students; thereby providing insights into evolving patterns of learning.

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," in *ASONAM*. IEEE/ACM, 2013, pp. 886–893.

[2] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Discovering value from community activity on focused question answering sites: a case study of stack overflow," in *SIGKDD*. ACM, 2012, pp. 850–858.

[3] Y. Yu, G. Yin, H. Wang, and T. Wang, "Exploring the patterns of social behavior in GitHub," in *Proceedings of the 1st international workshop on crowd-based software development methods and technologies*. ACM, 2014, pp. 31–36.

[4] M. Wattenhofer, R. Wattenhofer, and Z. Zhu, "The YouTube social network," in *ICWSM*. AAAI, 2012, pp. 354–361.

[5] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *WWW*. ACM, 2010, pp. 591–600.

[6] A. Ghosh and J. Kleinberg, "Incentivizing participation in online forums for education," in *EC*. ACM, 2013, pp. 525–542.

[7] R. D. Vallam, P. Bhatt, D. Mandal, and Y. Narahari, "A stackelberg game approach for incentivizing participation in online educational forums with heterogeneous student population," in *AAAI*. AAAI, 2015, pp. 1043–1049.

[8] A. Fayazi, K. Lee, J. Caverlee, and A. Squicciarini, "Uncovering crowdsourced manipulation of online reviews," in *SIGIR*. ACM, 2015, pp. 233–242.

[9] B. Li, T. Jin, M. R. Lyu, I. King, and B. Mak, "Analyzing and predicting question quality in community question answering services," in *WWW*. ACM, 2012, pp. 775–782.

[10] G. Ghalme, S. Gujar, A. Kumar, S. Jain, and Y. Narahari, "Design of coalition resistant credit score functions for online discussion forums," in *AAMAS*. IFAAMAS, 2018, pp. 95–103.

[11] C. Shah and J. Pomerantz, "Evaluating and predicting answer quality in community QA," in *SIGIR*. ACM, 2010, pp. 411–418.

[12] C. Thompson, "How Khan Academy is changing the rules of education," *Wired Magazine*, vol. 126, pp. 1–5, 2011.

[13] J. A. Ruipérez-Valiente, P. J. Muñoz-Merino, D. Leony, and C. D. Kloos, "ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform," *Computers in Human Behavior*, vol. 47, pp. 139–148, 2015.