What You Say is Relevant to How You Make Friends: Measuring the Effect of Content on Social Connection

Yiqiao Xu
Department of Computer Science
NCSU, Raleigh, NC, USA
yxu35@ncsu.edu

Niki Gitinabard
Department of Computer Science
NCSU, Raleigh, NC, USA
ngitina@ncsu.edu

Collin F. Lynch
Department of Computer Science
NCSU, Raleigh, NC, USA
cflynch@ncsu.edu

Tiffany Barnes
Department of Computer Science
NCSU, Raleigh, NC, USA
tmbarnes@ncsu.edu

ABSTRACT

Discussion forums are the primary channel for social interaction and knowledge sharing in Massive Open Online Courses (MOOCs). Many researchers have analyzed social connections on MOOC discussion forums. However, to the best of our knowledge, there is little research that distinguishes between the types of connections students make based upon the content of their forum posts. We analyze this effect by distinguishing on- and off-topic posts and comparing their respective social networks. We then analyze how these types of posts and their social connections can be used to predict the students' final course performance. Pursuant to this work we developed a binary classifier to identify on- and off-topic posts and applied our analysis with the hand-coded and predicted labels. We conclude that the post type does affect the relationship between the students and their closest neighbors or community members clustered communities and their closest neighbor to their learning outcomes.

Keywords

MOOC, social network analysis, forum participation

1. INTRODUCTION & BACKGROUND

Social interactions are an essential component of learning. Peer collaborators in courses provide support by engaging in informal advising, sharing "institutional knowledge", and engaging in the co-construction of knowledge [6]. Students who lack strong social connections are more prone to feeling lost or discouraged in a course and are more likely to drop out[12]. This issue is of particular interest in Massive Open Online Courses (MOOCs), which seek to scale classroom instruction to hundreds or even thousands of students supported by a single instructor. Prior researchers have shown that students in MOOCs do form community structures and that those structures are correlated with their learning [7, 4, 1, 11]. Brown et al. [1], for example, found that students connected to peers with a similar course performance.

While prior research has shown that students form stable social structures in MOOCs, the impact of those connections on students' performance has not always been consistent. Jiang et al. analyzed an algebra MOOC, and found that attirbutes of the students' social network were correlated with the students' final grades, but they found no relationship between the same variables in a different MOOC on finance [4]. Houston et al. likewise examined the forum activities that were most strongly associated with final grades in three different MOOCs and found that the addition of social centrality and similar features had no impact on their predictive models [3]. This inconsistency may be explained by the fact that the types of discussions students have and their relevance can change from class to class and that prior analyses have focused primarily on the overall social structure and not the content of the discussions. Though the peer communication features in MOOCs are intended to foster content engagement, many of the most active discussion-topics are often social conversations, critiques of the class videos, or exchanges of career advice [8]. Prior researchers have developed automated detectors to classify these posts into on- and off-topic comments and to evaluate the relative proportion of relevant discussions to learning outcomes [5, 10].

Our goal in this study is to examine how the topical content of forum discussions affects students' social relationships, as well as how they connect to their learning outcomes by addressing the following question: *Does the type (on- or off-topic) of conversation affect the relevance of students' social networks to their learning outcomes?*

2. DATA

For our analysis, we used data from a MOOC on "Big Data in Education" provided by The Teachers College at Columbia University and hosted on the Coursera (BDE 2013) and EdX (BDE 2015) platforms in 2013 and 2015 respectively. This

Yiqiao Xu, Niki Gitinabard, Collin Lynch and Tiffany Barnes "What You Say is Relevant to How You Make Friends: Measuring the Effect of Content on Social Connection" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 679 - 682

was offered as an 8 week course that includes material from a graduate-level course on educational data mining and the analysis of big data in education. This curriculum introduces students to basic data collection and data analysis methods such as visualization and clustering. The course offers weekly lectures in the form of videos and individual assignments or quizzes which contribute to students' final grade (grade scale from 0 to 1). The students learn how and when to do educational data mining and learning analytics on data. The course was structured around weekly lecture videos and individual quizzes. The students' final course grade is a composite score based upon these quizzes. In the 2013 class, 1380 students completed at least one quiz, 778 students made at least one post or comment on the discussion forum producing a total of 603 discussion threads consisting of 4259 posts in total. In 2015, 320 students completed at least one quiz, 519 students produced 625 discussion threads with a total of 2056 posts. We manually annotated all of the posts and comments separating them into on- and off-topic entries after removing non-English posts from the dataset. Table 1 includes summary information about the number of students who received 0 or non-0 grades and posted on- or off-topic posts from both offerings of the course. Table 2 shows the number of posts and threads (defined by the starter post) of each type in each course.

	Content	0 grade	non-0 grade	Total
BDE 2013	on-topic	156	377	533
	off-topic	187	220	407
BDE 2015	on-topic	58	83	141
	off-topic	51	44	95

Table 1: Demographic of students grade and content

	Content	Post	Thread
BDE 2013	on-topic	2845	405
	off-topic	1388	380
BDE 2015	on-topic	1050	151
	off-topic	1006	367

Table 2: Number of on- and off-topic posts and threads

Tables 1 and 2 shows that in both courses, the students were more likely to take part in the on-topic discussions than the off-topic ones, especially for those who received non-0 grades. However, in 2015, we observed that though the on- & off-topic post counts are close to each other, the number of threads started with off-topic posts were smaller. This may be due to the fact that the EdX platform includes an optional private chat room for users, logs which we did not have access to.

3. METHODS

Prior researchers have shown that students' final grades are strongly related to those of their closest classmate or 'Best Friend' (BF) in traditional classroom [2]. This relationship also holds in online courses [11] and is also true for students' neighbors in a community structure [1]. In contrast to general assumptions the students are not always connecting with others who need help or people who share a goals or background but with people at their same level of performance, in this study, we used exactly the same network generation approach as in our prior work [11] to build the social network graph and to evaluate the relationship between student communities and their final grades. And applied the Girvan-Newman algorithm with the "natural cluster number" approach described in [1] to identify coherent communities. We then applied the Kruskal-Walls(KW) test to evaluate the correlation between clusters and performance.

In this approach, we treat forum participants as nodes, and we construct arcs between the individuals as weighted edges based upon their individual communications. In this approach we add a directed arc from the author's node to nodes representing the authors of all the comments that precede it in the thread. All of the forum contributors in the thread will be connected to one another. As the average thread length of our two datasets are 6.8 and 3.1 respectively, we developed this approach based upon the assumption that participants read the whole thread before they post any comments. Thus, we consider each reply to be an indication of an implicit social connection between forum participants. Once the raw directed graph has been constructed we modify the graph by eliminating all isolated nodes and merging the parallel edges to get a weighted undirected graph.

4. RESULTS

Table 3 shows the order (number of nodes) and size (number of edges) of the graphs that we obtained for the different content types and student groups (with/without 0 grade students) at the end of week-2. As we have established in prior work, the second week is considered the most important cut-off point for students to stay in or drop out of a course and to form their social networks [11]. We can observe that, in both courses, including the 0 grade students, students preferred to participate more in off-topic conversation than on-topic. According to the number of nodes and edges by the end of week-2, for an on-topic network, the edges still increase in frequency after week 2; however, for the off-topic posts, the social network has already formed at this time point. This suggests that the off-topic discussions may have been confined to a stable set of threads that only grew longer, or were confined to the same stable set of chatty people. Furthermore, for the non-0 students, they were more likely to start conversations for on-topic content, than the 0 students. One potential explanation for this is that students who did not plan to obtain a certificate, and who registered for free, participated in conversations such as introducing themselves to each other at the beginning of the course and then lost interest in the course. Another interpretation is that some of the students worked in spurts at the beginning but dropped out because the course did not fit their schedule over time. Our ongoing analysis of the forum content has shown that a number of the posts are also about early issues, such as course logistics and software. These issues became less relevant as the course progressed. Irrespective of the cause, the social structures are well established for off-topic discussions early enough that instructors should be able to provide advice early enough to the students who have lost interest early on.

Table 4 shows the number of students and the average final grade for each group of students with/without 0 grade. We found that students who participated in both the on-

	Content	Node*	Edge*	Grade	Node	Edge
2013	on-topic	392	2185	with 0	508	3778
				non-0	367	2713
	off-topic	356	9483	with 0	429	10389
				non-0	234	1884
2015	on-topic	182	637	with 0	199	721
				non-0	111	332
	off-topic	370	1044	with 0	392	1112
				non-0	98	95

Node*: Number of nodes at the end of week 2
Edge*: Number of edges at the end of week 2

Table 3: Graph order and size

and off-topic discussions received the highest average grade in the course. Students that only participated in off-topic discussions received the lowest final grade when compared to others. These results indicate that not only is sharing knowledge about course content is important, but participating in non-course content also has an impact on their learning process.

We also examined the growth rates of the number of posts, users and new threads as time progressed. We defined a new thread as being on- or off-topic based upon the head post that initiated it. For BDE 2013, on-topic posts and new threads increased over the whole course period, while off-topic posts appeared primarily at the beginning of the course before declining sharply. As for new users, they came in at the beginning to talk primarily on off-topic content, but new users were more likely to make on-topic conversation afterwards. However, for BDE2015, we observed that all three elements in the on-/off-topic social networks grew monotonically at a similar rate and that the number of new off-topic threads and users was always more than the on-topic ones. One of the potential reasons could be that students discussed course content topics in the private chat-room, rather than in the public forum. As one example, at the end of the BDE2013 course, there were 55,179 registered users, yet the final course social interaction graph contained only 778 participants, including 1 instructor and 2 teaching assistants. Some of the forum participants did not complete any quizzes, or even attempt to obtain the certificate, but still chose to engage in on- and off-topic discussions with others. On the other hand, some of the students who worked hard on the course did not contribute to the forum at all. There were 1,381 students who received a non-0 final grade; 934 of which did not post in the forum at all, while 304 zero final grade students did. It is conceivable students only posted when they faced particular difficulties, or that they sought help elsewhere as participation in the course forum was not a necessary condition for completion.

4.1 Social Interaction Analysis

As part of our analysis we also replicated the Best-Friends comparison as used by Brown et al. Here we identified each student's closest neighbor in the course, excluding members of the teaching staff, and then calculated a direct correlation between their grades and those of their best friends. Since the data was non-normal, we used Spearman's Rank Correlation Coefficient as a non-parametric test for association [9]. Our results are shown in Table 5.

	Content	Grade	BF	Comm
BDE 2013	on-topic	with 0	0.96	<0.05
		non-0	<0.05	<0.05
	off-topic	with 0	0.98	<0.05
		non-0	<0.05	0.08
BDE 2015	on-topic	with 0	0.24	<0.05
		non-0	<0.05	<0.05
	off-topic	with 0	0.06	<0.05
		non-0	0.38	0.30

Comm: KW test for community - grade

Table 5: Social connection correlation with content

Table 5 shows that the relationships between students' grades and those of their best friends were consistent between the traditional courses studied by Fire et al. [2] and MOOCs, but not immediately. Our results show that MOOC students, except those who did not submit any assignments, performed similarly to their closest peers.

5. CONCLUSION & DISCUSSION

In this paper, we distinguished posts and comments into on-topic (course relevant) and off-topic(non-course relevant) before analyzing students' social activities and their final grades. Interestingly, we drew a different conclusion from our previous work. In the BDE2013 dataset, including 0 grade students, we found no correlation between the students' closest 'best friend' and their performance, while the clustered community structures were significant related to their to performance. When we break this down into on- and off-topic networks respectively we found that there was a significant correlation with the community structure and grade for on-topic posts. For the off-topic, by contrast, only a moderate relationship was observed. For the BDE2015 non-0 students, the off-topic connection was not relevant to their performance. This is also shown by the average cluster grades for each group. This supports our original argument that the off-topic discussions may be confusing the social network analyses.

Additionally, Students who showed up in both the on- and off-topic discussions received the highest grade, higher than those that focused on the on-topic discussions alone. Students only made off-topic discussion received the lowest grades overall. Thus although participation in the on-topic discussion facilitated the students' learning, chatting with their peers on random topics was also relevant to their learning, albeit weakly. Additionally, according to Table 3, for all of the students, the off-topic graph was much bigger than the on-topic graph, while for the non-0 grade students, they were more likely to post course content topics than random chat. Thus, we conclude that, for the non-0 grade students who focused their efforts on finishing quizzes, passing the course, and receiving a certificate, participation on the forum helped them improve their grade and keep them from dropping out. By contrast, when we consider all of the students, including those with a 0 grade, the behavior of their closest peers does not seem to have affected them consistently. One possible explanation for this may be that 97% of the students received a 0 grade.

			on-topic	off-topic	only in on-topic	only in off-topic	in both on/off-topic
2013	with 0	students	508	429	299	220	209
		grade	0.51	0.38	0.44	0.16	0.61
	non-0	students	367	234	195	62	17
		grade	0.70	0.69	0.67	0.58	0.74
2015	with 0	students	199	392	100	293	99
		grade	0.33	0.14	0.26	0.05	0.40
	non-0	students	111	98	49	36	62
		grade	0.59	0.54	0.53	0.38	0.64

Table 4: Average grades for students in different social interaction

Moreover, as prior studies have shown the students formed the bulk of the social structures by the end of week 2. We found that 91% of the the off-topic connections had been formed by then. For the on-topic social network by contrast, only 57% of the connections had been formed by week 2. This highlights one of the limitations of prior work that conflated these social structures and it highlights the crucial importance of distinguishing posts by content. We observed different results for the BDE2015 dataset. This class had less proportion of on-topic discussions than BDE2013. This may be due in part to the fact that the edX platform provides support for private a chatrooms which students and instructors may use for side discussions. We were unable to access that data and it may be the case that much of the relevant communications were carried on there.

Discussion forums are widely used in MOOCs to support knowledge co-construction, but the connections between online social interaction and learning outcomes is still subject of some debate. As our study has shown the social network structures can be used for information provided we focus on the important on-topic discussions. And, as we have also shown it is possible to use trained models to support this classification, even ones that are trained in part across course offerings. Thus these findings highlight the critical importance of analyzing forum post content when exploring the relationship with learning outcomes; and to draw conclusions carefully when we work with datasets of this type.

In future work we plan to evaluate the impact of automatic classifiers and guidance, both for students and instructors, on students' course performance, topic comprehension, dropout, and their final grades. This kind of help, we argue, will help students and instructors to manage the course content more effectively and will thus increase student engagement, reduce dropout, and improve other student outcomes.

6. ACKNOWLEDGEMENTS

This research was supported by NSF #1821475 "Concert: Coordinating Educational Interactions for Student Engagement" Collin F. Lynch, Tiffany Barnes, and Sarah Heckman (Co-PIs).

References

[1] R. Brown, C. Lynch, Y. Wang, M. Eagle, J. Albert, T. Barnes, R. Baker, Y. Bergner, and D. S. McNamara. Communities of performance & communities of preference. In *EDM (Workshops)*, 2015.

[2] M. Fire, G. Katz, Y. Elovici, B. Shapira, and L. Rokach. Predicting student exam's scores by analyzing social network data. In *International Conference on Active Media Technology*, pages 584–595. Springer, 2012.

[3] S. L. Houston II, K. Brady, G. Narasimham, and D. Fisher. Pass the idea please: The relationship between network position, direct engagement, and course performance in moocs. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 295–298. ACM, 2017.

[4] S. Jiang, S. M. Fitzhugh, and M. Warschauer. Social positioning and performance in moocs. In *Workshop on Graph-Based Educational Data Mining*, volume 14, 2014.

[5] F.-R. Lin, L.-S. Hsieh, and F.-T. Chuang. Discovering genres of online discussion threads via text mining. *Computers & Education*, 52(2):481–495, 2009.

[6] K. Reusser and C. Pauli. Co-constructivism in educational theory and practice. *International encyclopedia of the social & behavioral sciences*, 3:913–917, 2015.

[7] C. P. Rosé, R. Carlson, D. Yang, M. Wen, L. Resnick, P. Goldman, and J. Sherer. Social factors that contribute to attrition in moocs. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 197–198. ACM, 2014.

[8] D. T. Seaton, Y. Bergner, I. Chuang, P. Mitros, and D. E. Pritchard. Who does what in a massive open online course? *Commun. ACM*, 57(4):58–65, Apr. 2014.

[9] P. Sedgwick. Spearman's rank correlation coefficient. *BMJ: British Medical Journal (Online)*, 349, 2014.

[10] A. F. Wise, Y. Cui, and J. Vytasek. Bringing order to chaos in mooc discussion forums with content-related thread identification. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 188–197. ACM, 2016.

[11] Y. Xu, C. F. Lynch, and T. Barnes. How many friends can you make in a week?: evolving social relationships in moocs over time.

[12] D. Yang, T. Sinha, D. Adamson, and C. P. Rosé. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*, volume 11, page 14, 2013.