Communities of Performance & Communities of Preference

Rebecca Brown
North Carolina State University
Raleigh, NC
rabrown7@ncsu.edu

Collin Lynch
North Carolina State University
Raleigh, NC
cflynch@ncsu.edu

Yuan Wang
Teachers College, Columbia University
New York, NY
elle.wang@columbia.edu

Michael Eagle
North Carolina State University
Raleigh, NC
mjeagle@ncsu.edu

Jennifer Albert
North Carolina State University
Raleigh, NC
jennifer_albert@ncsu.edu

Tiffany Barnes
North Carolina State University
Raleigh, NC
tmbarnes@ncsu.edu

Ryan Baker
Teachers College, Columbia University
New York, NY
ryanshaunbaker@gmail.com

Yoav Bergner
Educational Testing Service
Princeton, NJ
ybergner@gmail.com

Danielle McNamara
Arizona State University
Phoenix, AZ
dsmcnamara1@gmail.com

ABSTRACT

The current generation of Massive Open Online Courses (MOOCs) operate under the assumption that good students will help poor students, thus alleviating the burden on instructors and Teaching Assistants (TAs) of having thousands of students to teach. In practice, this may not be the case. In this paper, we examine social network graphs drawn from forum interactions in a MOOC to identify natural student communities and characterize them based on student performance and stated preferences. We examine the community structure of the entire course, students only, and students minus low performers and hubs. The presence of these communities and the fact that they are homogeneous with respect to grade but not motivations has important implications for planning in MOOCs.

Keywords

MOOC, social network, online forum, community detection

1. INTRODUCTION

The current generation of Massive Open Online Courses (MOOCs) is designed to leverage student interactions to augment instructor guidance. The activity in courses on sites such as Coursera and edX is centered around user forums that, while curated and updated by instructors and TAs, are primarily constructed by students. When planning and building these courses, it is hoped that students will help one another through the course and that interacting with stronger students will help to improve the performance of weaker ones. It has not yet been shown, however, that this type of support occurs in practice.

Prior research on social networks has shown that social groups, even those that gather face-to-face, can fragment into disjoint sub-communities [37]. This small-group separation, if it takes place in an online course, can be considered negative or positive, depending on one's perspective. If poor students communicate only with similarly-floundering peers, then they run the risk of perpetuating misunderstandings and of missing insights discussed by better-performing peers and teaching staff. An instructor may wish to avoid this fragmentation to encourage poor students to connect with better ones.

These enduring subgroups may be beneficial, however, by helping students to form enduring supportive relationships. Research by Li et al. has shown that such enduring relationships can enhance students' social commitment to a course [18]. We believe that this social commitment will in turn help to reduce feelings of isolation and alienation among students in a course. Eckles and Stradley [9] have shown that such isolation is a key predictor of student dropout.

We have previously shown that students can form stable communities and that those communities are homogeneous with respect to performance [3]. However that work did not: show whether these results are consistent with prior work on immediate peer relationships; address the impact of hub students on these results; or discuss whether students' varying goals and preferences motivate the community structure. Our goal in this paper is to build upon our prior work by addressing these issues. In the remainder of this paper we will survey prior educational literature on community formation in traditional and online classrooms. We will then build upon our prior work by examining the impact of hub users. And we will look at the impact of user motivations on community formation.

2. RELATED WORK

2.1 MOOCs, Forums, & Student Performance

A survey of the literature on MOOCs shows the beginnings of a research base generating an abundance of data that has not yet been completely analyzed [19]. According to Seaton et al. [29], most of the time students spend on a MOOC is spent in discussion forums, making them a rich and important data source. Stahl et al. [30] illustrates how through this online interaction students collaborate to create knowledge. Thus students' forum activity is good not only for the individual student posting content or receiving answers, but for the class as a whole. Huang et al. [14] investigated the behavior of the highest-volume posters in 44 MOOC-related forums. These "superposters" tended to enroll in more courses and do better in those courses than the average. Their activity also added to the overall volume of forum content and they left fewer questions unanswered in the forums. Huang et al. also found that these superposters did not suppress the activity of less-active users. Rienties et al. [25] examined the way in which user interaction in MOOCs is structured. They found that allowing students to self-select collaborators is more conducive to learning than randomly assigning partners. Further, Van Dijk et al. [31] found that simple peer instruction is significantly less effective in the absence of a group discussion step, pointing again to the importance of a class discussion forum.

More recently Rosé et al. [27] examined students' evolving interactions in MOOCs using a Mixed-Membership Stochastic Block model which seeks to detect partially overlapping communities. They found that the likelihood that students would drop out of the course is strongly correlated with their community membership. Students who actively participated in forums early in the course were less likely to drop out later. Furthermore, they found one forum sub-community that was much more prone to dropout than the rest of the class, suggesting that MOOC communities are made up of students who behave in similar ways. This community can in turn reflect or impact a student's level of motivation and their overall experience in a course much like the "emotional contagion" model used in the Facebook mood manipulation study by Kramer, Guillroy, and Hancock [16].

Yang et al. [36] also notes that unlike traditional courses students can join MOOCs at different times and observed that students who join a course early are more likely to be active and connected in the forums, and less likely to drop out, than those who join later. MOOCs also attract users with a range of individual motivations. In a standard classroom setting students are constrained by availability, convention, and goals. Few students enroll in a traditional course without seeking to complete it and to get formal credit for doing so. MOOCs by virtue of their openness and flexibility attract a wide range of students with unique personal motivations [10]. Some join the course with the intent of completing it. Others may seek only to brush up on existing knowledge, obtain specific skills, or just watch the videos. These distinct motivations in turn lend themselves to different in-class behaviors including assignment viewing and forum access. The impact of user motivations in online courses has been previously discussed by Wang et al. [32, 33]; we will build upon that work here. Thus it is an open question whether these motivations affect students' community behaviors or not.

2.2 Communities, Hubs, & Peers

Kovanovic et al. [15] examined the relationship between social network position or centrality, and social capital formation in courses. Their work is specifically informed by the Community of Inquiry (COI) framework. the COI framework is focused on distance education and is particularly suited to online courses of the type that we study here. The model views course behavior through three *presences* which mediate performance: cognitive, teaching, and social.

This *social presence* considers the nature and persistence of student interactions and the extent to which they reinforce students' behaviors. In their analysis, the authors sought to test whether network relationships, specifically students' centrality in their social graph, is related to their social performance as measured by the nature and type of their interactions. To that end, they examined a set of course logs taken from a series of online courses offered within a public university. They found that students' position within their social graph was positively correlated with the nature and type of their interactions, thus indicating that central players also engaged in more useful social interactions. They did not extend this work to groups, however, focusing solely on individual hub students.

Other authors have also examined the relationship between network centrality, neighbor relationships, network density, and student performance factors. Eckles and Stradley [9] applied network analysis to student attrition, finding that students with strong social relationships with other students who drop out are significantly more likely to drop out themselves. Rizzuto et al. [26] studied the impact of social network density on student performance. Network density is defined as the fraction of possible edges that are present in a given graph. Thus it is a measure of how "clique-like" the graph is. The authors examined self-reported social networks for students in a large traditional undergraduate psychology course. They found that denser social networks were significantly correlated with performance. However, a dominance analysis [1] showed that this factor was less predictive than pure academic ability. These results serve to motivate a focus on the role of social relationships in student behavior. Their analysis is complicated, however, by their reliance on self-report data which will skew the strength and recency of the reported relationships.

Fire et al. [11] studied student interaction in traditional classrooms, constructing a social network based on cooperation on class assignments. Students were linked based on partnership on group work as well as inferred cooperation based on assignment submission times and IP addresses. The authors found that a student's grade was significantly correlated with the grade of the student with the strongest links to that student in the social network. We perform similar analysis in this paper to examine whether the same correlation exists in MOOCs.

Online student interaction in blended courses has also been linked to course performance. Dawson [8] extracted student and instructor social networks from a blended course's online discussion forums and found that students in the 90th grade percentile had larger social networks than those in the 10th percentile. The study also found that high-performing students primarily associated with other high-performing students and were more likely to be connected to the course instructor, while low-performing students tended to associate with other low-performers. In a blended course, this effect may be offset by face-to-face interaction not captured in the online social network, but if the same separation happens in MOOC communities, low-

performing students are less likely to have other chances to learn from high-performing ones.

2.3 Community Detection

One of the primary activities students engage in on forums is question answering. Zhang et al. [38] conducted a social network analysis on an online question-and-answer forum about Java programming. Using vertex in-degree and out-degree, they were able to identify a relatively small number of active users who answered many questions. This allowed the researchers to develop various algorithms for calculating a user's Java expertise. Dedicated question-and-answer forums are more structured than MOOC forums, with question and answer posts identified, but a similar approach might help identify which students in a MOOC ask or answer the most questions.

Choo et al. [5] studied community detection in Amazon product-review forums. Based on which users replied to each other most often, they found communities of book and movie reviewers who had similar tastes in these products. As in MOOC forums, users did not declare any explicit social relationships represented in the system, but they could still be grouped by implicit connections.

In the context of complex networks, a community structure is a subgraph which is more densely connected internally than it is to the rest of the network. We chose to apply the Girvan-Newman edge-betweenness algorithm (GN) [13]. This algorithm takes as input a weighted graph and a target number of communities. It then ranks the edges in the graph by their edge-betweenness value and removes the highest ranking edge. To calculate Edge-betweenness we identify the shortest path $p(a,b)$ between each pair of nodes a and b in the graph. The edge-betweenness of an arc is defined as the number of shortest paths that it participates in. This is one of the centrality measures explored by Kovanovic et al. above [15]. The algorithm then recalculates the edge-betweenness values and iterates until the desired number of disjoint community subgraphs has been produced. Thus the algorithm operates by iteratively finding and removing the highest-value communications channel between communities until the graph is fully segmented. For this analysis, we used the iGraph library [7] implementation of G-N within R [24].

The strength of a candidate community can be estimated by modularity. The *modularity score* of a given subgraph is defined as a ratio of its intra-connectedness (edges within the subgraph) to the inter-connectedness with the rest of the graph minus the fraction of such edges expected if they were distributed at random [13, 35]. A graph with a high modularity score represents a dense sub-community within the graph.

3. DATA SET

This study used data collected from the "Big Data in Education" MOOC hosted on the Coursera platform as one of the inaugural courses offered by Columbia University [32]. It was created in response to the increasing interest in the learning sciences and educational technology communities in using EDM methods with fine-grained log data. The overall goal of this course was to enable students to apply each method to answer education research questions and to drive intervention and improvement in educational software and systems. The course covered roughly the same material as a graduate-level course, Core Methods in Educational Data Mining, at Teachers College Columbia University. The MOOC spanned from October 24, 2013 to December 26, 2013. The weekly course was composed of lecture videos and 8 weekly assignments. Most of the videos contained in-video quizzes (that did not count toward the final grade).

All of the weekly assignments were structured as numeric input or multiple-choice questions. The assignments were graded automatically. In each assignment, students were asked to conduct analyses on a data set provided to them and answer questions about it. In order to receive a grade, students had to complete this assignment within two weeks of its release with up to three attempts for each assignment, and the best score out of the three attempts was counted. The course had a total enrollment of over 48,000, but a much smaller number actively participated. 13,314 students watched at least one video, 1,242 students watched all the videos, 1,380 students completed at least one assignment,and 778 made a post or comment in the weekly discussion sections. Of those with posts, 426 completed at least one class assignment. 638 students completed the online course and received a certificate (meaning that some students could earn a certificate without participating in forums at all).

In addition to the weekly assignments the students were sent a survey that was designed to assess their personal motivations for enrolling in the course. This survey consisted of 3 sets of questions: MOOC-specific motivational items; two PALS (Patterns of Adaptive Learning Survey) sub-scales [21], Academic Efficacy and Mastery-Goal Orientation; and an item focused on confidence in course completion. It was distributed to students through the course's E-mail messaging system to students who enrolled in the course prior to the official start date. Data on whether participants successfully completed the course was downloaded from the same course system after the course concluded. The survey received 2,792 responses; 38% of the participants were female and 62% of the participants were male. All of the respondents were over 18 years of age.

The MOOC-specific items consisted of 10 questions drawn from previous MOOC research studies (cf. [2, 22]) asking respondents to rate their reasons for enrollment. These 10 items address traits of MOOCs as a novel online learning platform. Specifically, these 10 items included questions on both the learning content and features of MOOCs as a new platform. Two PALS Survey scales [21] measuring mastery-goal orientation and academic efficacy were used to study standard motivational constructs. PALS scales have been widely used to investigate the relation between a learning environment and a student's motivation (cf. [6, 20, 28]). Altogether ten items with five under each scale were included. The participants were asked to select a number from 1 to 5 with 1 meaning least relevant and 5 most relevant. Respondents were also asked to self-rate their confidence on a scale of 1 to 10 as to whether they could complete the course according to the pace set by the course instructor. All three groups of items were domain-general.

4. METHODS

For our analysis, we extracted a social network from the online forum associated with the course. We assigned a node to each student, instructor, or TA in the course who added to it. Nodes representing students were labeled with their final course grade out of 100 points. The Coursera forums operate as standard threaded forums. Course participants could start a new thread with an initial post, add a post to an existing thread, and add a comment or child element below an existing post. We added

a directed edge from the author of each post or comment to the parent post and to all posts or comments that preceded it on the thread based upon their timestamp. We made a conscious decision to omit the textual content of the replies with the goal of isolating the impact of the structure alone.

We thus treat each reply or followup in the graph as an implicit social connection and thus a possible relationship. Such implicit social relationships have been explored in the context of recommender systems to detect strong communities of researchers [5]. This is, by design, a permissive definition that is based upon the assumption that individuals generally add to a thread after viewing the prior content within it and that individual threads can be treated as group conversations with each reply being a conscious statement for everyone who has already spoken. The resulting network forms a multigraph with each edge representing a single implicit social interaction. We removed self loops from this graph as they indicate general forum activity but not any meaningful interaction with another person. We also removed vertices with a degree of 0, and collapsed the parallel edges to form a simple weighted graph for analysis.

In the analyses below we will focus on isolating student performance and assessing the impact of the faculty and hub students. We will therefore consider four classes of graphs: *ALL* the complete graph; *Student* the graph with the instructor and TAs removed; *NoHub* the graph with the instructor and hub users removed; and *Survey* which includes only students who completed the motivation survey. We will also consider versions of the above graphs without students who obtained a score of 0, and without the isolated individuals who connect with at most one other person. As we will discuss below, a number of students received a zero grade in the course. Because this is an at-will course, however, we cannot readily determine why these scores were obtained. They may reflect a lack of engagement with the course, differential motivations for taking the course, a desire to see the course materials without assignments, or genuinely poor performance.

4.1 Best-Friend Regression & Assortativity

Fire et al. [11] applied a similar social network approach to traditional classrooms and found a correlation between a student's most highly connected neighbor ("best friend") and the student's grade. The links in that graph included cooperation on assignments as well as partnership on group assignments. To examine whether the same correlation existed in a massive online course in which students were less likely to know each other beforehand and there were no group assignments, we calculated each student's best friend in the same manner and performed a similar correlation.

The simple best friends analysis gives a straightforward mechanism for correlating individual students. However it is also worthwhile to ask about students who are one-step removed from their peers. Therefore we will also calculate the grade assortativity (r_G) of the graphs. Assortativity describes the correlation of values between vertices and their neighbors [23]. The assortativity metric r ranges between -1 and 1, and is essentially the Pearson correlation between vertex and their neighbors [23]. A network with $r=1$ would have each vertex only sharing edges with vertices of the same score. Likewise, if $r=-1$ vertices in the network would only share edges with vertices of different scores. Thus grade assortativity allows us to measure whether individuals are not just connected directly to individuals with similar scores but whether they correlate with individuals who are one step removed.

Several commonly studied classes of networks tend to have patterns in their assortativity. Social networks tend to have high assortativity, while biological and technological networks tend to have negative values (dissortativity) [23]. In a homogeneous course or one where students only form stratified communities we would expect the assortativity to be very high while in a heterogeneous class with no distinct communities we would expect it to be quite low.

4.2 Community Detection

The process of community detection we employed is briefly described here [3]. As noted there we elected to ignore the edge direction when making our graph. Our goal in doing so was to focus on communities of learners who shared the same threads, even when they were not directly replying to one-another. We believe this to be a reasonable assumption given the role of class forums as a knowledge-building environment in which students exchange information with the group. Individuals who participate in a thread generally review prior posts before submitting their contribution and are likely to return to view the followups. Homogeneity in this context would mean that students gathered and communicated primarily with equally-performing peers and thus that they did not consistently draw from better-performing classmates and help lower-performing ones *or* that the at-will communities served to homogenize performance, with the students in a given cluster evening out over time.

While algorithms such as GN are useful for finding clusters they do not, in and of themselves, determine the *right* number of communities. Rather, when given a target number they will seek to identify the *best* possible set of communities. In some implementations the algorithm can be applied to iteratively select the maximum modularity value over a possible range. Determining the correct number of communities to detect, however, is a non-trivial task especially in large and densely connected graphs where changes to smaller communities will have comparatively small effects on the global modularity score. As a consequence we cannot simply optimize for the best modularity score as we would risk missing small but important communities [12].

Therefore, rather than select the clusterings based solely on the highest modularity, we have opted to estimate the correct number of clusters visually. To that end we plotted a series of modularity curves over the set of graphs. For each graph G we applied the GN algorithm iteratively to produce all clusters in the range $(2,|G_N|)$. For each clustering, we then calculated the global modularity score. We examined the resulting scores to identify a *crest* where the modularity gain leveled off or began to decrease thus indicating that future subdivisions added no meaningful information or created schisms in existing high-quality communities. This is a necessarily heuristic process that is similar to the use of Scree plots in Exploratory Factor Analysis [4]. We define the number identified as the *natural* cluster number.

5. RESULTS AND DISCUSSION

Before removing self-loops and collapsing the edges, the network contained 754 nodes and 49,896 edges. The final social network contained 754 nodes and 17,004 edges. 751 of the participants were students, with 1 instructor and 2 TAs. One individual was incorrectly labeled as a student when they were acting as the Chief Community TA. Since this person's posts clearly indicated that he or she was acting in a TA capacity with regard to the forums, we relabeled him/her as a TA. Of the 751 students 304 obtained a zero grade in the course leaving 447 nonzero students. 215 of the 751 students responded to the motivation survey.

There were a total of 55,179 registered users, so the set of 754 forum participants is a small fraction of the entire course audience. However, forum users are not necessarily those who will make an effort or succeed in the course. Forum users did not all participate in the course, and some students who participated in the course did not use the forums: 1,381 students in the course got a grade greater than 0, and 934 of those did not post or comment on the forums, while 304 of the 751 students who did participate in the forums received a grade of 0. Clearly students who go to the trouble of posting forum content are in some respect making an effort in the course beyond those who don't, but this does not necessarily correspond to course success.

5.1 Best-Friend Regression & Assortativity

We followed Fire et al.'s methodology for identifying Best Friends in a weighted graph and calculated a simple linear regression over the pairs. This correlation did not include the instructor or TAs in the analysis. We calculated the correlation between the students' grades to their best friends' grades in the set using Spearman's Rank Correlation Coefficient (ρ) [34]. The two variables were strongly correlated, $\rho(748) = 0.44$, $p < 0.001$. However, the correlation was also affected by the dense clusters of students with 0 grades. After removing the 0 grade students we found an additional moderate correlation, $\rho(444) = 0.29$, $p < 0.001$.

Thus the significant correlation between best-friend grade and grade holds over the transition from the traditional classroom to a MOOC. This suggests that students in a MOOC, excluding the many who drop out or do not submit assignments, behave similarly to those in a traditional classroom in this respect. These results are also consistent with our calculations for assortativity. There we found a small assortative trend for the grades as shown in Table 1. These values reflect that a student was frequently communicating with students who in turn communicated with students at a similar performance level. This in turn supports our belief that homogeneous communities may be found. As Table 1 also illustrates, the zero-score students contribute substantially to the assortativity correlation as well with the correlation dropping by as much as a third when they were removed.

Table 1: The grade assortativity for each network.

Users	Zeros	V	E	r_G
All	Yes	754	17004	0.29
All	No	447	5678	0.20
Students	Yes	751	15989	0.32
Students	No	447	5678	0.20
Non-Hub	Yes	716	9441	0.37
Non-Hub	No	422	3119	0.24

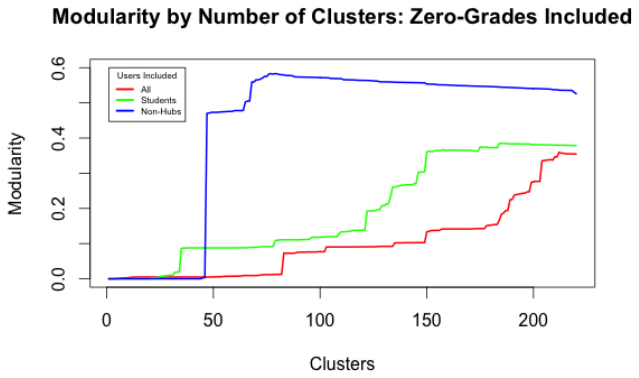

Figure 1: Modularity for each number of clusters, including students with zeros.

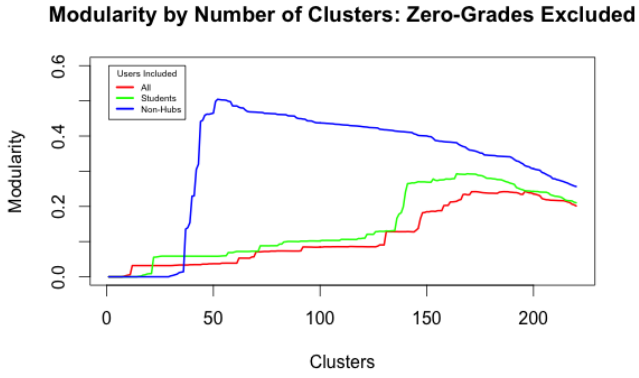

Figure 2: Modularity for each number of clusters, excluding students with zeros.

5.2 Community Structure

The modularity curves for the graphs both with and without zero-score students are shown in Figures 1 and 2. We examined these plots to select the natural cluster numbers which are shown in Table 2. As the values illustrate the instructor, TAs, and hub students have a disproportionate impact on the graph structure. The largest hub student in our graph connects to 444 out of 447 students in the network. The graph with all users had lower modularity and required more clusters than the graphs with only students or only non-hubs (see Table 2), with

Table 2: Graph sizes and natural number of clusters for each graph.

Users	Zeros	V	E	Clusters
All	Yes	754	17004	212
All	No	447	5678	173
Students	Yes	751	15989	184
Students	No	447	5678	169
Non-Hub	Yes	716	9441	79
Non-Hub	No	422	3119	52
Survey	Yes	215	1679	58

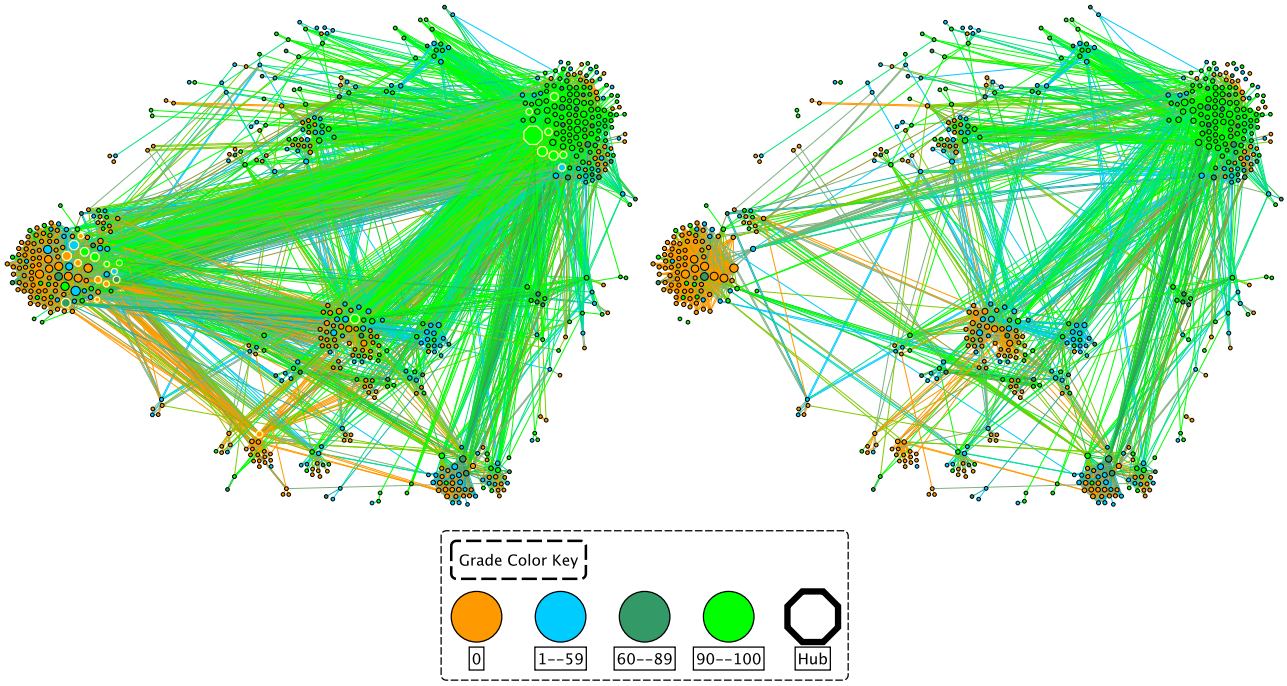

Figure 3: View of the student communities with edges of frequency <2 removed. The Student network with (left) and without (right) hub-students, with each vertex representing a student and grade represented as color.

the non-hub graph having the highest modularity. This suggests that non-hub students formed more isolated communities, while teaching staff and hubs communicated across these communities and connected them.

This largely consistent with the intent of the forums and the active role played by the instructor and TAs in monitoring and replying to all relevant posts in the forums. It is particularly interesting how closely the curves for the ALL and Student graphs mirror one another. This may indicate that the hub students are also those that followed the instructor and TAs closely, thus giving them isomorphic relationships, or it may indicate that they are more connected than even the instructors and thus came to bind the forums together on their own. This impact is further illustrated by the cluster plots shown in Figure 3. Here the absence of the hub students results in a noticeable thinning of the graph which in turn highlights the frequency of communication that can be attributed to this, comparatively small, group.

The difference between the full plots and those with zero values are also notable as the zero grade students were clearly a major factor in community formation. A direct examination of the user graph showed that many of the zero students were only connected to other zero students or were not connected at all. This is also highlighted in Figure 3. In both graphs the bulk of the zero score students are clustered in a tight network of communities on the left-hand side. That super-community consists primarily of zero score students communicating with other zero-score students, a structure we have nick-named the 'deathball.'

5.3 Student Performance & Motivation

As the color coding in Figure 3 illustrates, the students did cluster by performance. Table 3 shows the average grade and standard deviation for a small selection of the communities in the ALL reply network including zero-grades, hub students, and teaching staff. Several of the communities, particularly the larger ones, do show a blend of good and poor students, with a high standard deviation. However many if not most of the communities are more homogeneous with good and poor students sharing a community with similarly-performing peers. These clusters have markedly lower standard deviation.

Table 3: Grade statistics by community, selected to show examples of more and less homogeneous communities.

Members	Average Grade	Standard Deviation
118	21.62	36.58
41	22.00	32.45
34	25.41	40.44
31	56.13	47.69
20	49.05	45.64
16	12.44	31.13
14	88.43	22.47
12	96.08	6.36
11	96.45	7.38
4	3.00	6.00
4	8.50	9.81
4	4.25	8.50
4	96.25	3.50

An examination of the grade distribution for each of the clusters showed that the scores within each cluster were non-normal. Therefore we opted to apply the Kruskal-Wallis (KW) test to assess the correlation between cluster membership and perfor-

Table 4: Kruskal-Wallis test of student grade by community, for each graph.

Users	Zeros	Chi-Squared	df	p-value
All	Yes	349.0273	211	< 0.005
All	No	216.1534	172	< 0.02
Students	Yes	202.0814	78	< 0.005
Students	No	80.93076	51	< 0.005
Non-Hub	Yes	309.8525	183	< 0.005
Non-Hub	No	218.9603	168	< 0.01
Survey	Yes	99.99840	577	< 0.005

mance. The KW test is a nonparametric rank-based analogue to the common Analysis of Variance [17]. Here we tested grade by community number with the community being treated as a categorical variable. The results of this comparison are shown in Table 4. As that illustrates, cluster membership was a significant predictor of student performance for all of the graphs with the non-zero graphs having markedly lower p-values than those with zero students included. These results are consistent with our hypothesis that students would form clusters of equal-performers and we find that those results hold even when the highly-connected instructors, TAs and hub students are included.

We performed a similar KW analysis for the questions on the motivation survey and for a binary variable indicating whether or not the student completed the survey at all. For this analysis we evaluated the clusters on all of the graphs. We found no significant relationship between the community structure on any of the graphs and the survey question results or the survey completion variable. Thus while the clusters may be driven by separate factors they are not reflected in the survey content.

6. CONCLUSIONS AND FUTURE WORK

Our goal in this paper was to expand upon our prior community detection work with the goal of aligning that work with prior research on peer impacts, notably the work of Fire et al. [11]. We also sought to examine the impact of hub students and student motivations on our prior results.

To that end we performed a novel community clustering analysis of student performance data and forum communications taken from a single well-structured MOOC. As part of this analysis we described a novel heuristic method for selecting natural numbers of clusters, and replicated the results of prior studies of both immediate neighbors and second-order assortativity.

Consistent with prior work, we found that students' grades were significantly correlated with their most closely associated peers in the new networks. We also found that this correlation extended out to their second-order neighborhood. This is consistent with our prior work showing that students form stable user communities that are homogeneous by performance. We found that those results were stable even if instructors, hub players, students with 0 scores, and students who did not fill out the survey were removed from consideration. This suggests that either the students are forming communities that are homogeneous or that the effect of those individual and network features on the communities and on performance is minimal.

We also found that community membership was not a significant predictor of whether students would complete the motivation survey or of students' motivations. We were surprised by the fact that even when we focused solely on individuals who had completed the survey, the students did not connect by stated goals. This suggests to us that the students are more likely coalescing around the pragmatic needs of the class or conceptual challenges rather than on the winding paths that brought them there. One limitation of this work is that by relying on the forum data we were focused solely on the comparatively small proportion of enrolled students (6%) who actively participated in the forums. This group is, by definition a smaller set of more actively-involved participants.

In addition to addressing our primary questions this study also raised a number of open issues for further exploration. Firstly, this work focused solely on the final course structure, grades, and motivations. We have not yet addressed whether these communities are stable over time or how they might change as students drop in our out. Secondly, while we ruled out motivations as a basis for the community this work we were not able to identify what mechanisms do support the communities. And finally this study raises the question of generality and whether or not these results can be applied to MOOCs offered on different topics or whether the results apply to traditional and blended courses.

In subsequent studies we plan to examine both the evolution of the networks over time as well as additional demographic data with the goal of assessing both the stability of these networks and the role of other potential latent factors. We will also examine other potential clustering mechanisms that control for other user features such as frequency of involvement and thread structure. We also plan to examine other similar datasets to determine if these features transition across classes and class types. We believe that these results may change somewhat once students can coordinate face to face far more easily than online.

7. ACKNOWLEDGMENTS

This work was supported by NSF grant #1418269: "Modeling Social Interaction & Performance in STEM Learning" Yoav Bergner, Ryan Baker, Danielle S. McNamera, & Tiffany Barnes Co-PIs.

8. REFERENCES

[1] R. Azen and D. Budescu. The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, 8(2):129–48, 2003.

[2] Y. Belanger and J. Thornton. Bioelectricity: A quantitative approach Duke University's first MOOC. *Journal of Learning Analytics*, 2013.

[3] R. Brown, C. F. Lynch, M. Eagle, J. Albert, T. Barnes, R. Baker, Y. Bergner, and D. McNamara. Good communities and bad communities: Does membership affect performance? In C. Romero and M. Pechenizkiy, editors, *Proceedings of the* 8^{th} *International Conference on Educational Data Mining*, 2015. submitted.

[4] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

[5] E. Choo, T. Yu, M. Chi, and Y. Sun. Revealing and incorporating implicit communities to improve recommender systems. In M. Babaioff, V. Conitzer, and D. Easley, editors, *ACM Conference on Economics and Computation, EC '14, Stanford,*

CA, USA, June 8-12, 2014, pages 489–506. ACM, 2014.

[6] K. Clayton, F. Blumberg, and D. P. Auld. The relationship between motivation learning strategies and choice of environment whether traditional or including an online component. *British Journal of Educational Technology*, 41(3):349–364, 2010.

[7] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[8] S. Dawson. 'Seeing' the learning community: An exploration of the development of a resource for monitoring online student networking. *British Journal of Educational Technology*, 41(5):736–752, 2010.

[9] J. Eckles and E. Stradley. A social network analysis of student retention using archival data. *Social Psychology of Education*, 15(2):165–180, 2012.

[10] A. Fini. The technological dimension of a massive open online course: The case of the CCK08 course tools. *The International Review Of Research In Open And Distance Learning*, 10(5), 2009.

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

[12] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proc. of the National Academy of Sciences*, 104(1):36–41, 2007.

[13] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. of the National Academy of Sciences*, 99(12):7821–7826, June 2002.

[14] J. Huang, A. Dasgupta, A. Ghosh, J. Manning, and M. Sanders. Superposter behavior in MOOC forums. In *Proc. of the first ACM conference on Learning@ scale conference*, pages 117–126. ACM, 2014.

[15] V. Kovanovic, S. Joksimovic, D. Gasevic, and M. Hatala. What is the source of social capital? the association between social network position and social presence in communities of inquiry. In S. G. Santos and O. C. Santos, editors, *Proc. of the Workshops held at Educational Data Mining 2014, co-located with 7th International Conference on Educational Data Mining (EDM 2014), London, United Kingdom, July 4-7, 2014.*, volume 1183 of *CEUR Workshop Proc.* CEUR-WS.org, 2014.

[16] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[17] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[18] N. Li, H. Verma, A. Skevi, G. Zufferey, J. Blom, and P. Dillenbourg. Watching MOOCs together: investigating co-located MOOC study groups. *Distance Education*, 35(2):217–233, 2014.

[19] T. R. Liyanagunawardena, A. A. Adams, and S. A. Williams. MOOCs: A systematic study of the published literature 2008-2012. *The International Review of Research in Open and Distributed Learning*, 14(3):202–227, 2013.

[20] J. L. Meece, E. M. Anderman, and L. H. Anderman. Classroom goal structure, student motivation, and academic achievement. *Annual Review of Psychology*, 57:487–503, 2006.

[21] C. Midgley, M. L. Maehr, L. Hruda, E. Anderinan, L. Anderman, and K. E. Freeman. *Manual for the Patterns of Adaptive Learning Scales (PALS)*. University of Michigan, Ann Arbor, 2000.

[22] MOOC @ Edinburgh 2013. MOOC @ Edinburgh 2013 - report #1. *Journal of Learning Analytics*, 2013.

[23] M. E. Newman. Assortative Mixing in Networks. *Physical Review Letters*, 89(20):208701, Oct. 2002.

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.

[25] B. Rienties, P. Alcott, and D. Jindal-Snape. To let students self-select or not: That is the question for teachers of culturally diverse groups. *Journal of Studies in International Education*, 18(1):64–83, 2014.

[26] T. Rizzuto, J. LeDoux, and J. Hatala. It's not just what you know, it's who you know: Testing a model of the relative importance of social networks to academic performance. *Social Psychology of Education*, 12(2):175–189, 2009.

[27] 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 *Proc. of the first ACM conference on Learning@ scale conference*, pages 197–198. ACM, 2014.

[28] A. M. Ryan and H. Patrick. The classroom social environment and changes in adolescents' motivation and engagement during middle school. *American Educational Research Journal*, 38(2):437–460, 2001.

[29] D. Seaton, Y. Bergner, I. Chuang, P. Mitros, and D. Pritchard. Who does what in a massive open online course? *Communications of the ACM*, 57(4):58–65, 2014.

[30] G. Stahl, T. Koschmann, and D. Suthers. Computer-supported collaborative learning: An historical perspective. *Cambridge handbook of the learning sciences*, 2006:409–426, 2006.

[31] L. Van Dijk, G. Van Der Berg, and H. Van Keulen. Interactive lectures in engineering education. *European Journal of Engineering Education*, 26(1):15–28, 2001.

[32] Y. Wang and R. Baker. Content or platform: Why do students complete MOOCs? *MERLOT Journal of Online Learning and Teaching*, 11(1):191–218, 2015.

[33] Y. Wang, L. Paquette, and R. Baker. A longitudinal study on learner career advancement in MOOCs. *Journal of Learning Analytics*, 1(3), 2014.

[34] Wikipedia. Spearman's rank correlation coefficient — Wikipedia, the free encyclopedia, 2013. [Online; accessed 27-February-2013].

[35] Wikipedia. Modularity (networks) — Wikipedia, the free encyclopedia, 2014. [Online; accessed 5-February-2015].

[36] D. Yang, T. Sinha, D. Adamson, and C. P. Rose. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proc. of the 2013 NIPS Data-Driven Education Workshop*, volume 10, page 13, 2013.

[37] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

[38] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *Proc. of the 16th international conference on World Wide Web*, pages 221–230. ACM, 2007.