Good Communities and Bad Communities: Does membership affect performance?

1. INTRODUCTION

The current generation of Massive Open Online Courses (MOOCs) are designed to leverage student knowledge to augment instructor guidance. Activity in these courses is typically centered on a threaded forum that, while curated by the instructors, is largely student driven. When planning MOOCs, it is commonly hoped that open forums will allow students to interact freely and that better students will help the poorer performers. It has not yet been shown, however, that this occurs in practice.

In our ongoing work, we are investigating the structure of student communities and social interactions within online and blended courses [1]. Our focus in this poster is on the structure of student communities in a MOOC and the connection between those communities and students' performance in the course. Our goal was to determine whether students in the course form strong sub-communities and whether a student's community membership is correlated with their performance. If students do form strong communities and community membership is a predictor of performance, then it would suggest either that students are forming strong relationships that help to improve their performance or that they are clustering by performance. If they do not, then it suggests that they may be able to connect freely in the forums at the expense of persistent and beneficial relationships.

2. BACKGROUND

Course-level relationships have been shown to influence students' performance and the overall success of a course. Fire et al. examined the impact of immediate peers in a traditional class and found that students' performance was significantly correlated with that of their closest peer [4]. Eckles and Stradley analyzed dropout rates and found that students with strong relationships with students who dropped out were more likely to do so themselves [3].

Rosé et al. [7] examined students' evolving social interactions in MOOCs using a Mixed-Membership Stochastic Block model which seeks to detect partially overlapping communities. They found that dropout likelihood was strongly correlated with community membership. Students who actively participated in forums early in the course were less likely to drop out later. Dawson [2] studied blended courses and found that students in the higher grade percentiles tended to have larger social networks within the course and were more likely to be connected to the instructor.

3. METHODS

Big Data in Education is a MOOC offered by Dr. Ryan Baker through the Teacher's College at Columbia University [8]. This is a 3-month long course composed of lecture videos, forum interactions, and 8 weekly assignments. All of the assignments were structured as numeric or multiple-choice exams and were graded automatically. Students were required to complete assignments within two weeks of their release and were given three attempts to do so, with the best score being used. 48,000 students enrolled in the course with 13,314 watching at least one video, 1,380 completing at least one assignment and 778 posting in the forums. Of that 778, 426 completed at least one assignment. 638 students completed the course, some managed to do so without posting in the forums.

We extracted a social network from the forums, each student, instructor, and TA was represented by a node. Each student node was annotated with their final grade. Forum users could: start new threads, add to existing threads, or add comments below existing posts. We added directed edges from the author of each item to the author of the parent post, if any, and to the author of the items that preceded it in the current thread. We then elim-
nated all self-loops and collapsed all parallel edges to form a simple weighted graph for analysis. We extracted two different classes of graphs. The ALL graphs include everyone who participated in the forums while the Student graphs omit the instructor and TAs. We produced two versions of each graph: one containing all participants and one that excluded students with a course grade of 0.

We identified communities using the Girvan-Newman Edge Betweenness Algorithm [5]. This algorithm takes as input an undirected graph and a desired number of communities. It operates by identifying the edge with the highest edge-betweenness score: the edge that sits on the shortest path between the most nodes. It then removes that edge and repeats until the desired number of disjoint graphs have been made. We applied exploratory modularity analysis to identify the natural number of communities [1].

Having generated the graphs and determined the natural cluster numbers, we clustered the students into communities. We treated the cluster assignment as a categorical variable and tested its correlation with final course grades. An examination of the grade distributions showed that they were non-normal, so we applied the Kruskal-Wallis (KW) test to evaluate the relationship [6].

The KW test is a non-parametric analogue of the ANOVA test.

## 4. RESULTS AND DISCUSSION

The raw graph contained 754 nodes and 49,896 edges. After collapsing the parallel arcs and removing self-loops we retained a total of 17,004 edges. Of the 754 nodes, 751 were students. Of those, 304 obtained a grade of 0 in the course leaving 447 nonzero students. The natural cluster number for each of the graphs is shown in Table 1 along with the result of the KW tests. As Table 1 illustrates, cluster assignment was significantly correlated with performance. While it may be the case that the better performing students are still most connected to others who perform at a similar level. This is consistent with prior work on traditional classrooms and issues such as dropout. It runs counter to the naive assumption that good students will help to improve the others. While it may be the case that the better performing communities contain poorer-performing students who increased their grades through interaction with better students, the presence of so many low-grade clusters suggests that students do fragment into semi-isolated communities that do not perform very well.

More research is required to determine why these communities form, whether it is due to motivational factors or similar incoming characteristics. We present some work along these lines in [1]. We will also examine the stability of the communities over time to determine whether they can be changed or if they are a natural outgrowth of the forums and must be accepted as is.

## 5. ACKNOWLEDGMENTS

Work supported by NSF grant #1148269: “Modeling Social Interaction & Performance in STEM Learning” Yoav Bergner, Ryan Baker, Danielle S. McNamara, & Tiffany Barnes Co-PIs.

## 6. REFERENCES


### Table 1: Community cluster numbers and Kruskal-Wallis test of student grade by community.

<table>
<thead>
<tr>
<th>Users</th>
<th>Zeros</th>
<th>Clusters</th>
<th>$K$</th>
<th>$df$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>Yes</td>
<td>212</td>
<td>349.03</td>
<td>211</td>
<td>&lt; 0.005</td>
</tr>
<tr>
<td>All</td>
<td>No</td>
<td>173</td>
<td>216.15</td>
<td>172</td>
<td>&lt; 0.02</td>
</tr>
<tr>
<td>Students</td>
<td>Yes</td>
<td>184</td>
<td>202.08</td>
<td>78</td>
<td>&lt; 0.005</td>
</tr>
<tr>
<td>Students</td>
<td>No</td>
<td>169</td>
<td>80.93</td>
<td>51</td>
<td>&lt; 0.005</td>
</tr>
</tbody>
</table>

Figure 1: Student communities with edges of weight 1 removed. Nodes represent communities. Size indicates number of students. Color indicates mean grade.