Good Communities and Bad Communities: Does membership affect performance?

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

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

Keywords
MOOC, social network, online forum, community detection

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 authors of the items that preceded it in the current thread. We then elimi-

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 the students' grade performance for all of the graphs. A sample visualization of the student graph is shown in Figure 1.

The students formed detectable communities, and community membership was significantly correlated with performance. While the structure of the communities changed when non-students and zero-students were removed, the significance relationships held. Thus while the specific community structure is not stable under deformations, 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 naïve 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.

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

Users	Zeros	Clusters	K	df	p-value
All	Yes	212	349.03	211	< 0.005
All	No	173	216.15	172	< 0.02
Students	Yes	184	202.08	78	< 0.005
Students	No	169	80.93	51	< 0.005

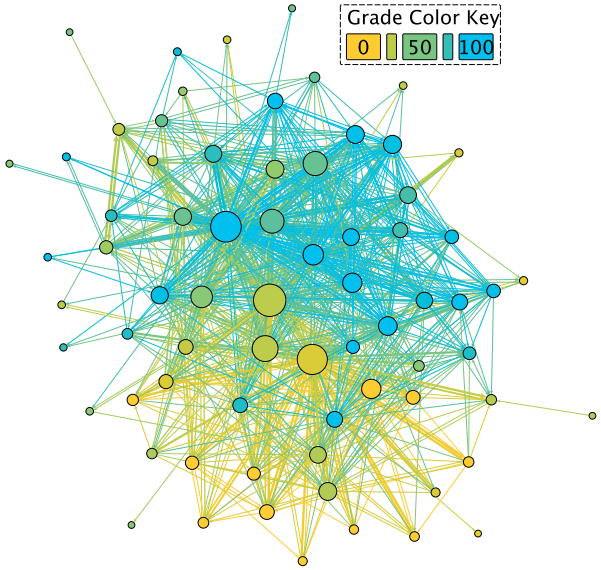

Figure 1: Student communities with edges of weight 1 removed. Nodes represent communities. Size indicates number of students. Color indicates mean grade.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] R. Brown, C. F. Lynch, Y. Wang, M. Eagle, J. Albert, T. Barnes, R. Baker, Y. Bernger, and D. McNamara. Communities of performance & communities of preference. In C. F. Lynch, T. Barnes, J. Albert, and M. Eagle, editors, *Proceedings of the 2nd International Workshop on Graph-Based Educational Data Mining*, 2015. submitted.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

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

[8] Y. Wang, L. Paquette, and R. S. J. D. Baker. A longitudinal study on learner career advancement in moocs. *Journal of Learning Analytics*. (In Press).