Examining MOOC superposter behavior using social network analysis

Mandira Hegde
Whiting School of Engineering
Johns Hopkins University
Baltimore, MD, USA
mhegde1@jhu.edu

Ian McCulloh
Accenture
Washington, DC, USA
ian.mcculloh@accenturefederal.com

John Piorkowski
Applied Physics Lab
Johns Hopkins University
Laurel, MD, USA
jpiorko2@jhu.edu

***Abstract*— This paper examines quantity and quality superposter value creation within Coursera Massive Open Online Courses (MOOC) forums using a social network analysis (SNA) approach. The value of quantity superposters (i.e. students who post significantly more often than the majority of students) and quality superposters (i.e. students who receive significantly more upvotes than the majority of students) is assessed using Stochastic Actor-Oriented Modeling (SAOM) and network centrality calculations. Overall, quantity and quality superposting was found to have a significant effect on tie formation within the discussion networks. In addition, quantity and quality superposters were found to have higher-than-average information brokerage capital within their networks.**

Keywords— Stochastic Actor-Oriented Model, SAOM, social network, MOOC, Coursera, online learning, collaborative learning, centrality

I. INTRODUCTION

The popularity of online education has exploded in the last decade [1]. A significant downside to online education is the lack of interaction among students and instructors which would organically occur as a major component of a live, in-person course. Many online courses rely on online discussion forums to compensate for this lack of interaction. In particular, Massive Open Online Courses (MOOCs) use discussion forums as the primary mode of interaction among instructors and students [2]. Within these forums, students are able to share their opinions and feedback, discuss course material, and seek help from both students and instructors alike.

Given the significance of online discussion forums in online classes, understanding the behavior of students within these forums is of critical importance to course designers and instructors. This paper examines two particular forms of student behavior which may play an outsized role in online discussion forums: quantity superposting and quality superposting. We define quantity superposters as students who post significantly more frequently than their peers. Additionally, we define quality superposters as students who receive significantly more upvotes on their discussion contributions than their peers.

In particular, we focus our research on superposting behavior using the following research questions (RQs):

- RQ1: Do the quality superposting behavior and quantity superposting behavior of a student's connections within a network affect that student's tie formation throughout the duration of the course?
- RQ2: To what extent do quantity and quality superposters serve as information brokers within discussion forums?

These research questions are explored by performing statistical network modeling and social network centrality calculations on discussion data collected from five Coursera MOOCs.

II. BACKGROUND

Previous work performed by Huang et al. [2] sought to analyze whether high-volume contributors ("superposters") contribute value to MOOC discussion forums or negatively impact contribution from the large remainder of students in the course [2]. This work found high superposter activity to correlate positively and significantly with higher overall activity and forum health. Huang et al. measured forum health by assessing the total contribution volume within the forums, the perceived utility of superposters in terms of votes they received, and the number of orphaned threads in the forums. Huang et al. found high superposter value was correlated with higher average perceived utility in terms of received votes, and a smaller fraction of orphaned threads.

This paper aims to extend this work by examining superposter value creation using a social network analysis (SNA) lens. Using a SNA approach, superposters are viewed as embedded within a network and as such, their value is assessed considering their structural influence within the network. Specifically, Stochastic Actor-Oriented Modeling (SAOM) is performed to assess the effect that superposters have over ties formed in the network during the duration of the courses analyzed. In addition, the average brokerage capital of superposters is assessed using betweenness centrality calculations and compared to that of the average student user. Furthermore, this paper extends Huang et al.'s definition of a superposter to include two types of superposters: 1) quantity superposters (as measured by the number of posts made), and 2) quality superposters (as measured by the number of votes received).

ASONAM '19, August 27–30, 2019, Vancouver, BC, Canada
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6868-1/19/08 $15.00
https://doi.org/10.1145/3341161.3345310

III. DATASET

The dataset used for this analysis contains the anonymized versions of discussion threads from the forums of 60 Coursera MOOCs. This dataset was downloaded directly from https://github.com/elleros/courseraforums. This data was originally collected and by Rossi and Gnawali in [3] to study the evolution of forum activities and investigate language-independent features to classify the discussion threads based on the types of the interactions among the users using a machine learning approach [3].

Each Coursera MOOC contains sub-forums in which students discuss various material related to the class. Sub-forums are comprised of threads which are 2 levels deeps, containing an ordered sequence of posts with additional comments attached to the posts. For each thread, post, and comment created in these courses, the dataset contains a user ID for the user author, the user type of the author (student, instructor, staff, or anonymous), the timestamp for when the user created the thread, post, or comment, the number of votes each thread, post, and comment received, and the course ID for each thread, post, and comment that was created.

Data for five of these courses was analyzed. Table 1 provides a summary of the analyzed course datasets.

Table 1: Dataset summary for courses analyzed.

Course Name	Number Users	Number Students	Number Threads	Course Duration (weeks)
Computational Method for Data Analysis	268	266	188	10
Concepts and Tools for University Physics	186	179	109	9
Data Structures and Algorithms	250	239	284	14
Networks: Friends, Money, and Bytes	202	200	103	13
Numerical Analysis for Engineers	102	100	119	9

IV. METHODOLOGY

A. Social Network Creation

Social networks were extracted from the discussion thread data provided for each of the five courses analyzed. There are a variety of ways in which a social network can be extracted from online discussion forum data. For example, a simple approach would be to create an undirected network graph by creating links among all users within a particular forum who contribute to the same discussion threads. A more complex approach to network creation was taken, similar to the approach taken by Sinha [4], in order to attribute importance to discussion initiators. In this approach, users who created a new thread in the forum were directly connected with an outward link to all users who posted a response to their thread. In addition, users whose post within a thread generated three or more comments were directly connected with an outward link to users who commented on their post; this was done to informally identify the formation of substantive sub-threads within each thread. To form the final social network graph, the union of these two networks was taken, anonymous users were removed, and duplicate ties within the network were removed, resulting in a directed, unweighted network graph. These networks were created using the statnet package in R and saved as adjacency matrices for further analysis [5].

The following user attributes for each user node in the network were extracted from the discussion thread data for each of the five courses analyzed:

1. Type
2. Quantity superposter status
3. Vote score
4. Popularity score

A user's type defines whether the user is a student, instructor, or staff member. A user's quantity superposter status is a Boolean value which is true if the user is a student and is within the set of users who belong to the top 5% of forum participants in the course with respect to the number of threads, posts, or comments they created. In other words, quantity superposters are the students who posted the most frequently throughout the course. The vote score for each user was calculated by summing the number of votes each user's created thread, post, or comment received and subtracting from that total the number of downvotes each of that user's created thread, post, or comment received. A popularity score for each user was then assigned based on the user's type and vote score: student users with a vote score greater than 10 were assigned a popularity score of 5, student users with a vote score between 6 and 10 inclusive were assigned a popularity score of 4, student users with a vote score between 0 and 5 inclusive were assigned a popularity score of 3, student users with a vote score between -5 and -1 inclusive were assigned a popularity score of 2, and student users with a vote score less than -5 were assigned a popularity score of 1. All non-student users (i.e. course instructors and staff) were assigned a popularity score of 3. A single attribute file containing the attributes for all users within a given course was created for each course for further analysis.

For all five courses analyzed, social networks and user attributes were extracted at the end of the course such that all threads, posts, and comments created at any point throughout the course contributed to the network adjacency matrices and attribute files that were created. In addition, for four of the five courses analyzed – 1) Concepts and Tools for University Physics, 2) Data Structures and Algorithms, 3)Networks: Friends, Money and Bytes, and 4) Numerical Analysis for Engineers – additional social networks and user attributes were extracted at earlier points during the course. Specifically, social networks and user attributes were extracted as described in IV.A. using the data from the time the class started until one third of the class duration had elapsed ("wave one"), and again using data from the beginning of the class until two thirds of the class duration had elapsed ("wave 2"). Thus, a total of 13 social networks and 13 attribute files were extracted.

Fig. 1 shows the social network that was extracted during wave 1 of the Numerical Analysis for Engineers course; student users who were identified as quantity superposters are shown in red; of note, users who had not yet created any threads, posts, or comments or responded to any threads or posts (i.e. not participated in the discussion threads) are shown in the periphery as disconnected from the rest of the network.

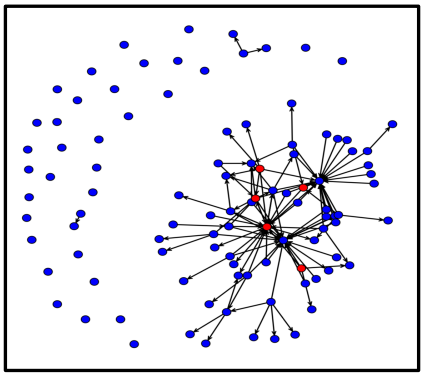


Fig 1: Numerical Analysis for Engineers wave 1 network.

It is interesting to note in Fig. 1 that the top five superposters are not necessarily the top five nodes in terms of betweenness centrality. While these may be correlated, they measure different constructs in the social group.

B. Stochastic Actor-Oriented Modeling (SAOM)

In order to answer RQ1, Stochastic Actor-Oriented Model (SAOM) simulations were performed to determine the influence that quantity superposters and quality superposters had over tie formation within the course network as the course progressed.

SAOMs have historically been used for the analysis of longitudinal social network data, modeling change from the perspective of an actor within an evolving network. Ties between actors are defined as going from "ego" (self) to "alter" (other). The estimation procedure uses repeated simulations of network evolution from each wave to the next, resulting in a large set of networks that could have potentially brought the observed networks from one to the other [6].

The specification of the network objective function takes the form

$$f_i^{net}(x) = \sum_k \beta_k^{net} s_{ik}^{net}(x) \tag{1}$$

where $f_i^{net}(x)$ is the value of the network objective function for actor i depending on the state x of the network. Functions $s_{ik}^{net}(x)$ are network state-based effects, and the β_k^{net} are the values to be estimated. Similarly, the behavioral objective function is defined as

$$f_i^{beh}(x, z) = \sum_k \beta_k^{beh} s_{ik}^{beh}(x, z) \tag{2}$$

with dependence on behavior value z [6].

Of note, SAOMs view network evolution as a series of choices made by actors to create, maintain, or cut ties to other actors [6]. Therefore, a key assumption was made that students within the analyzed course forums were given the choice of who to respond to within the forums (i.e. students were given the choice of with whom to form ties) throughout the duration of the courses.

The RSiena package in R was used to implement the SAOM simulation [7]. To statistically model the Concepts and Tools for University Physics, Data Structures and Algorithms, Networks: Friends, Money and Bytes, and Numerical Analysis for Engineers courses, three waves of network data and user attribute data were used. Specifically, the quantity superposter status and popularity score attributes were considered in this analysis, while controlling for node outdegree. The first wave of network and attribute data for each course consisted of the adjacency matrix and attributes extracted from the time the class started until one third of the class duration had elapsed. The second wave of network and attribute data for each course consisted of the adjacency matrix and attributes extracted from the beginning of the class until two thirds of the class duration had elapsed, and the third wave of network and attribute data for each course consisted of the adjacency matrix and attributes extracted from the beginning of the class until two thirds of the class duration had elapsed. Using this longitudinal data, the effect that the quantity superposting status of a node's connection ("superposting alter") as well as the popularity score of a node's connection ("popularity alter") had on that node's tie formation was tested.

C. Betweenness Centrality Calculation

In order to answer RQ2, between centrality scores for each node in each extracted network were calculated. A node's betweenness centrality score is considered a measure of gate keeping and brokering ability within a network. Users with a high betweenness centrality scores are likely to control the flow of information within a network [8]. In the context of MOOC discussion forums, forum participants with high betweenness centrality scores, and therefore high brokerage capital, can be assumed to be creative and prominent in spreading information and ideas in a network and therefore an important asset in value creation within the course network [10].

The betweenness centrality of a node v is given by the expression

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through [9].

The 'sna' package in R was used to calculate the betweenness centrality score of each user in the course networks formed at the end of each course [5]. Additionally, the following average (i.e. arithmetic mean) betweenness centrality scores were computed for the following user subgroups:

1. All student users within each course
2. Student who belong to the top 5% of forum participants within each course with respect to the number of threads, posts, or comments created (i.e. most frequent student posters or "quantity superposters")
3. Students who belong to the top 5% of forum participants within each course with respect to the number of upvotes received (i.e. most upvoted students or "quality superposters")

V. Results and Analysis

A. Stochastic Actor-Oriented Modeling (SAOM) Results

Table 2 provides the SAOM results for four courses. As shown, the quantity superposting status of a given node's

connection ("superposting alter") was statistically significant in all courses, except for the "the Networks: Friends, Money, and Bytes" course. The popularity score of that node's connection ("popularity alter") was a statistically significant effect on that node's tie formation in the "Numerical Analysis for Engineers and Concepts" and "Tools for University Physics" courses, but not the other two.

B. Betweenness Centrality Calculation Results

The set of charts in Fig. 2 shows the average betweenness centrality scores of students in each of the five courses analyzed compared to the average betweenness centrality scores of the most upvoted students and the average between centrality scores of the most frequent student posters (i.e. quantity superposters) in those courses. As shown in all five courses, the average betweenness centrality scores of the most frequent student posters and the most upvoted student posters

Table 2: SAOM Analysis Results

Course	RSiena Query	Estimate	Standard Error	Overall Maximum Convergence Ratio
Concepts and Tools for University Physics	Superposting alter	1.0305 ***	0.2206	0.0481
	Popularity alter	0.7311 ***	0.1516	0.0968
Networks: Friends, Money, and Bytes	Superposting alter	-4.8302	7.0599	0.1435
	Popularity alter	0.8149	0.8538	0.0449
Numerical Analysis for Engineers)	Superposting alter	1.3255 ***	0.3127	0.0308
	Popularity alter	1.3859 ***	0.2364	0.0454
Data Structures and Algorithms	Superposting alter	0.8693 ***	0.2789	0.0455
	Popularity alter	-0.2383	0.3285	0.0350

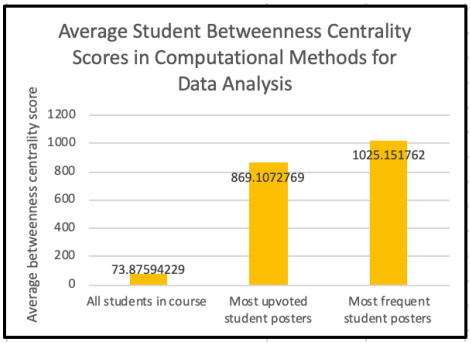
*** *statistically significant finding at the < 0.01 level*

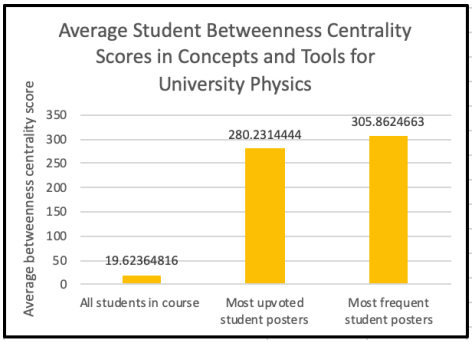
far exceed the average betweenness centrality score of the average student in the course, suggesting that students who post the most frequently as well as the most upvoted students possess much higher-than-average brokerage capital within their respective networks.

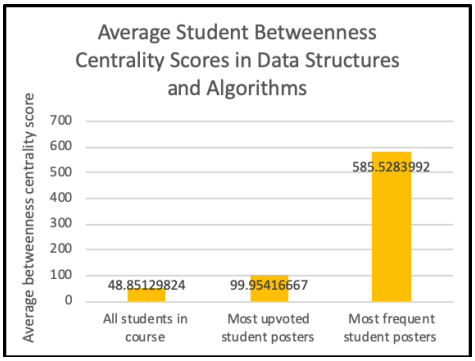
For three of the five MOOC courses analyzed, the most upvoted students were much higher in betweenness centrality compared with all students in the course. For each of these three networks, the popularity alter effect was statistically significant in the SAOM. When the most upvoted students betweenness centrality was similar to all students in the course, there was no statistically significant relationship for popularity alter. This makes intuitive sense in that the popularity effect only contributes to network tie formation when it is notably more popular in terms of betweenness centrality than other students in the course.

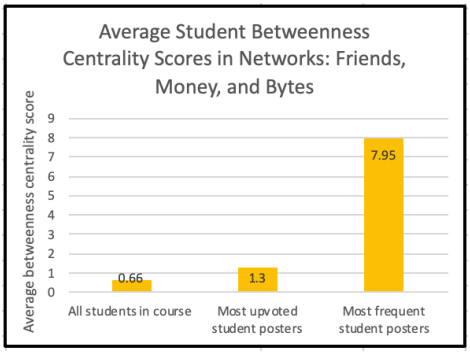
VI. Conclusion

This research presented in this paper aimed to build upon the work performed by Huang et al. [2] by examining superposter value creation using structural social network analysis approaches. In particular, Stochastic Actor-Oriented









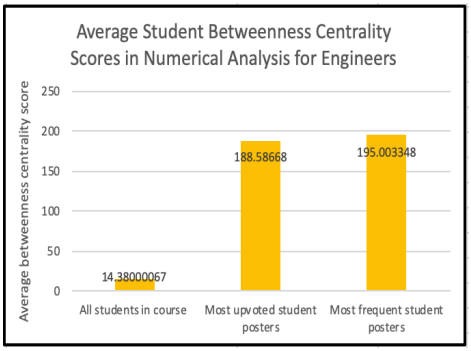


Fig 2: Betweenness centrality scores.

Modeling (SAOM) was performed to assess the effect that superposters have over links created within their networks. The quantity superposting behavior of a given node's connection ("superposting alter") as well as the popularity score of that node's connection ("popularity alter") did have a statistically significant effect on that node's tie formation. In addition, the average brokerage capital of superposters was assessed using betweenness centrality calculations and compared to that of the average student user. While this research took a different approach from Huang et al [2] our findings corroborate the conclusion made by Huang et al. that superposters have a largely positive impact on the health of MOOC discussion forums. Specifically, superposters possess a higher-than-average level of betweenness centrality, making them important contributors and influencers in the spread of information throughout course discussion networks. Superposters likely play a significant role in coordinating interactions amongotherwise unconnected students in the network, creating ties to those in the network and spread information without the presence of a mediator or intermediary [10].

In categorizing students as superposters, this work did not consider the content of a student's discussion thread contributions. To continue the study of the effect of superposter behavior in MOOC discussion forums, a combined structural and content-based social network analysis could be employed in which Natural Language Processing (NLP) techniques are used to assess the quality of a student's discussion contribution. Various other student behaviors, such as the tendency to positing inflammatory messages, or the tendency to post encouraging messages, could also be studied to help further inform course instructors and designers to improve the quality of online discussion forums.

REFERENCES

[1] J. Dijsselbloem, "The Rise of MOOCs: Can Online Distance Learning Replace Traditional Education", Diggit Magazine, November 2018.

[2] J. Huang, A. Dasgupta, A. Ghost, J. Manning, and M. Sanders, "Superposter behavior in MOOC forums," Proceedings of the First ACM Conference on Learning at Scale Conference, pp. 117-126, March 2014.

[3] L. Rossi and O. Gnawali, "Language independent analysis and classification of discussion threads in Coursera MOOC forums," IEEE International Conference on Information Reuse and Integration (IRI), August 2014.

[4] T. Sinha, "Supporting MOOC instruction with social network analysis," ACM, January 2014.

[5] R. Acton, L. Jasny, "An Introduction to Network Analysis with R and statnet," Subbelt XXXIV Workshop Series, February 2014.

[6] "Stochastic actor-oriented modeling for studying homophily and social influence in OSS projects," Empirical Software Engineering an International Journal, vol. 22, February 2017.

[7] R. Ripley, R. Snijders, Z. Boda, A. Voros, P. Preciado, Manual for RSiena. University of Oxford, April 2019.

[8] I. McCulloh, H. Armstrong, and A. Johnson, Social Network Analysis with Applications. Wiley, July 2013.

[9] "Betweenness Centrality (Centrality Measure)," GeeksforGeeks: A computer science portal for geeks.

[10] M. Saqr, U. Fors, M. Tedre, and J. Nouri, "How social network analysis can be used to monitor online collaborative learning and guide an informed intervention," PLoS One, vol. 13(3), March 2018