Using Pre-course Survey Responses to Predict Sporadic Learner Behaviors in Advanced STEM MOOCs

Work-in-Progress

Harsh Wardhan Aggarwal^,
School of Industrial Engineering
Purdue University, West Lafayette, IN 47907, USA
^haggarwa@purdue.edu

Peter Bermel*,
School of Electrical & Computer Engineering
Purdue University, West Lafayette, IN 47907, USA
*pbermel@purdue.edu

Nathan M Hicks,
School of Engineering Education
Purdue University, West Lafayette, IN 47907, USA

Kerrie A. Douglas, Heidi A. Diefes-Dux, Krishna Madhavan
School of Engineering Education
Purdue University, West Lafayette, IN 47907, USA

***Abstract*—Massive Open Online Courses (MOOCs) attract learners with different learner intentions, background knowledge, and skills as compared to traditional, closed enrollment settings. This level of diversity in learners introduces new factors that impact student persistence and engagement. Previous research has analyzed MOOC learners who are either fully engaged in a course in its entirety or at least consistently accessing course materials. Sporadic users, those who access course content randomly, have not been studied as in depth, even though they comprise the largest number of learners in a highly advanced engineering MOOC. In this study, we use clickstream data and pre-survey data to understand sporadic learners in a highly advanced nanoelectronics MOOC. We used a MOOC on nanotechnology with a total enrolment of close of 10,000 users offered for a duration of 8-weeks. We identified that academic preparedness, learner intentions, and expected time commitment could be used to predict sporadic users. Finally, an effect size analysis was performed.**

Keywords— MOOCs; highly technical courses; effect size analysis

I. INTRODUCTION

Massive open online courses (MOOCs) provide learners from different educational backgrounds access to highly technical content that was once reserved for university students meeting a specific set of prerequisites [1]. The fact that MOOCs are typically free and lacking in a requirements in terms of participation and deadlines make MOOCs a popular resource to educate oneself in a self-paced environment [1]. There is a very clear difference in how different users access MOOCs; some use MOOCs as textbooks and select their course content based on need, while others consider it more like a university course and participate fully in the course[2].

However, this flexibility in terms of time convenience, self-paced learning and tuition free access to highly technical content is reserved only for the users, as it has been found that substantial resources in terms of cost, time, and labor are required to produce MOOCs [3]. Since the foundation of any MOOC is dependent on instructors' dedication of time, knowledge, and effort, mostly in excess of their regular work load [4], it is vital that we work towards understanding the real impact and utilization of MOOCs.

The first step towards measuring utilization of MOOCs is to understand different user behavior. There have been numerous studies that have looked at the fully engaged users, the ones who access all the course materials and regularly complete the course assignments. Researchers have tried to understand their motivation [5], demographics, and intentions to understand their behavior in a course [6]. However, limited research is available on users who do not access all of the course materials and are sporadic in their access of course resources.

We believe that it is equally important to understand the behavior of sporadic users. Since the number of sporadic users is usually much larger than that of fully engaged users [6], it is important to understand how academic preparedness, intentions, and expected time commitment can help us identify sporadic learners so as to better design MOOCs that cater not just to fully engaged but also to sporadic users.

Thus, the purpose of this paper is to understand the effect of academic preparedness and intention to fully utilize course materials on learner usage, specifically sporadic users, in an advanced engineering MOOC.

II. BACKGROUND

NanoHUB-U offers courses on groundbreaking nanotechnology and nanoengineering topics. It was developed

978-1-5090-5920-1/17/$31.00 ©2017 IEEE

to extend the reach of the Center for Computational Nanotechnology [7] and currently offers courses which range from basic nanoelectronics to organic electronic devices. The majority of the concepts taught through NanoHUB-U courses are very new which make it a one-stop source for many researchers and scientists to access those materials, as they are not covered in textbooks [2]. Such content requires a deep understanding of nanotechnology and a strong foundation in mathematics which is uncommon for general MOOC learners. Still, the courses are made available through the edX platform and have no prerequisite requirements for enrolment. This has resulted in a huge enrolment in such courses. The course focused in this study is titled Fundamentals of Nanoelectronics: Basic Concepts and had 9,888 learners enrolled in its live version.

III. Hypothesis

The following hypothesis were tested in this study.

H1: Learners are less likely to be sporadic users if they plan to spend more than 6 hours in the course

H2: Learners are less likely to be sporadic users if they have a goal of achieving a high grade in the course.

H3: Learners are more likely to be sporadic users if they intend to dedicate less than 3 hours a week.

H4: Learners who intend to watch all of the videos are less likely to be sporadic users.

H5: Learners who intend to participate in all aspects of the course are less likely to be sporadic.

IV. Research Methods

Predicting user behavior using pre-survey questions, is a multi-step process. First, we developed a pre-survey for an advanced engineering MOOC. This was followed by data collection in the form of clickstream data for the complete 8-week duration of the course to categorize users into different learner groups based on their course behavior using a clustering method. Finally, the pre-survey results were analyzed alongside different learner groups to detect predictive behavior from pre-survey results. This section explains these different steps in more detail along with rationales for choosing the course, the pre-survey questions, and the clustering method.

A. Course

This study analyzes user groups and pre-survey questions in a highly advanced engineering course titled Principles of Nanoelectronics: Basic Concepts, offered through edX [6]. The course is specifically aimed at scientists and engineers who work in the field of nanotechnology and nanoelectronics. The course consisted of four units which were divided across 8 weeks. A total of 9,888 users enrolled in the course during the 8-week offering. The course was open for enrollment even after the 8-weeks, but this study only considers the users who participated in the course during the live version.

B. Pre-Survey: Development

A survey was included in the course as part of the first week assignment. The survey gathered information on learners' demographic, prerequisite background (especially mathematics), reasons for enrolling in the course, expected time commitment in the course, employment status, expectations for course participation, personal goals for the course, intrinsic and extrinsic motivation for learning, and English language skills. The survey included both forced-choice and Likert-style items. More information on the survey items can be found in [5], [6].

C. Clustering Technique

We used a K-means cluster analysis to analyze clickstream data [8], [9]. The cluster analysis was used to find similarity in learner usage patterns in accessing the course. We collected a total of 196,836 clicking incidents or records, but only used a subset of 36,000 records for clustering. Since, this subset consisted of all access records to "sequential" or "chapter" modules, which are the only possible way to access "video" or "problem" modules, and the clusters were created to understand learner behavior in terms of their access of videos and assignments, this subset was used to create clusters. For more detail about how the clustering technique has been applied to this course please refer to [6], [10].

D. Cluster Comparisons

We performed a Chi-square test of proportion for every pre-survey question related to our research questions to determine the likelihood that two variables are statistically different from each other (with a significant *p*-value of less than 0.05) [11]. The normal approximation assumption to use a Z-table was justified as $np>5$, where n is equal to the number of samples and p is equal to the proportion. However, the sample was not randomly selected. We included all the users who completed the pre-surveys. As the pre-surveys were optional, those included in the study might represent a self-selection bias.

V. Results

A. Clustering: Description and Comparisons

A total of five clusters were identified: fully engaged learners, consistent learners, two-week engaged learners, one-week engaged learners, and sporadic learners [6]. As the cluster labels suggest, fully engaged learners access the study materials regularly and attempt most of the assessments in the course; consistent viewers access content over the entire course duration but do not regularly complete quizzes and exams. As for the 2-week engaged and 1-week engaged learners, they actively accessed the course materials for the initial 2 and 1 week of the course respectively, but faded out afterwards. Sporadic users, the users focused on in this study, appear to randomly access the course materials over time and can be further divided into two types: (1) those who attempted all five exams without accessing any study material presented in the course and (2) those who did not access any videos but did access the course assignments[10]. From our initial clustering,

going from fully engaged all the way down to sporadic users, we identified 297, 182, 299, 543, and, 1,435 users respectively.

B. Completed Surveys and Questions

While 2,756 users were clustered, only 1,451 users initiated the pre-survey and only 969 users completed the survey. Out of 969 users, 218 users who completed the survey were not part of any clusters so they had to be removed from our dataset. So, we were left with 751 users, of which 172 were fully engaged, 102 were consistent, 121 were 2-week engaged, 196 were 1-week engaged and 160 were sporadic users. An analysis of the fully engaged learners and a cluster comparison has already been presented by [6].

Table 1 lists the questions from the pre-survey that were analyzed for this study. The survey had 25 questions in total, however, only 6 questions were analyzed in this study to test the hypotheses.

C. Sporadic Learner Statistics

Given that only sporadic learners were the focus of this study, we only present statistics related to sporadic learners. Table 2 highlights the number of sporadic learners versus other learners (which are fully engaged, consistent, 1-week and 2-week learners) for all six questions listed in Table 1. The Z test of proportion on a single sample for all six hypotheses showed significant differences (**p < 0.05). Hence, we rejected all six null hypotheses in this study.

TABLE I. SURVEY QUESTION DESCRIPTION

Question Description		
Questions		***Options***[a]
1	How much calculus have you taken?	1. None 2. High School level **3. 1 semester of college level calculus** **4. 2 semesters of college level calculus** **5. More than 2 semesters of college level calculus**
2	Have you taken any courses in differential equations?	**1. Yes** 2. No
3	Have you had prior experience in semiconductor device courses?	**1. Yes** 2. No
4	How much time do you plan to dedicate each week on this course?	1. 1-3 hours 2. 3-6 hours **3. 6-9 hours** **4. 9-12 hours** 5. 12-15 hours 6. uncertain
5	Which of the following best describes your learning goals for this course?	**1. I wish to achieve a high final grade.** 2. I wish to achieve a passing grade. 3. I am not concerned with grade outcomes.
6	Which of the following best describes what you hope to grain from the Nanoelectronics course?	1. I want to learn a broad overview of what Nanoelectronics is. 2. I want to become well acquainted with foundational principles of nanoelectronics. **3. I want to be able to apply material from this course in furutre projects related to nanoelectronics**.

[a.] The users who selected the options in bold were included for hypothesis testing

It is important to note that the total number of respondents for each question (listed in Table 1) were exactly 751, however, the number of users considered for each hypothesis is different due to the framing of the hypothesis. For example, for H2 (Table 2), the total number of users who took any courses in differential equations totaled 620 (out of 751). Out of these 620 users, 120 were sporadic learners (21%) and 500 were learners from other clusters. Therefore, all the hypothesis were scaled to the proportion of sporadic learners. Table 2, mentions the total number of considered for each hypothesis.

TABLE II. SPORADIC LEARNERS STATISTICS

Learner Number, Percentage and Selected Users			
Hypothesis	***Sporadic***	***Others (All other learner clusters)***	***Selected Users***[b]
H1	28 (16%)	147 (84%)	175
H2	45 (15.9%)	237 (84.1%)	282
H3	57 (28%)	142 (72%)	199
H4	37(27%)	99(73%)	136
H5	48(15%)	267(85%)	315

[b.] Total users were 751 for all questions.

D. Effect Size Calculation

Table 3 lists the effect sizes for all the hypothesis. The effect sizes were calculated using r = Z/(n^0.5) where N was the sample size and Z was the z-score found from the Chi square test (**p <0.05).

TABLE III. EFFECT SIZE

Question Description	
Hypothesis	***Effect Size (r)***
H1	-0.13
H2	-0.13
H3	0.18
H4	0.14
H5	-0.15

VI. DISCUSSION

As is evident from the results of the hypothesis tests and Table 3, learners who intended to watch all videos in the course were less likely to be sporadic users. This suggests that having course prerequisites and some background related to the course content could be a good filter for identifying sporadic users from the other cluster groups.

In addition, we found that users who planned to dedicate more than 6 hours on average to the course and aimed for achieving a higher grade in the course were less likely to be sporadic users. This was consistent with the idea that users who are dedicated in the course and plan to spend more time on the course (right from the start) are more likely to access the course resources consistently.

Finally, learners who intended to participate in all aspects of the course were also less likely to be sporadic in their use of course material. This result is particularly interesting as it can be used to filter out sporadic users at the start of the course by identifying the intent of the learners.

VII. Conclusions

This study highlights two important points. First, sporadic learners make up a large number of MOOC users and therefore should be studied in depth. Second, pre-course survey responses to questions related to academic preparedness, intentions, and time commitment can be used to identify this large sporadic learner group.

VIII. Future Work

Although this paper only discusses results about the sporadic learners, we plan to extend our analysis for learners from other categories in our future studies. It is also important to find the differences and similarities between different learner groups so that a cross sectional analysis can be conducted within different courses and different MOOC platforms.

Acknowledgment

This study was funded by a grant from the National Science Foundation (NSF DGE 1544259). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

[1] T. R. Liyanagunawardena, A. A. Adams, and S. A. Williams, "MOOCs: A systematic study of the published literature 2008-2012," *Int. Rev. Res. Open Distrib. Learn.*, vol. 14, no. 3, pp. 202–227, 2013.

[2] K. A. Douglas, P. Bermal, M. M. Alam, and K. Madhavan, "Smart data characterization of a highly technical MOOC-like engineering course," *Submitt. Publ.*

[3] P. M. Nissenson and A. C. Shih, "MOOC on a budget: Development and implementation of a low-cost MOOC at a state university," in 2015 ASEE Annual Conference and Exposition, 2015

[4] S. Kolowich, "The Professors behind the MOOC Hype.," *Chron. High. Educ.*, 2013.

[5] B. Mihalec-Adkins, N. Hicks, K. A. Douglas, H. Diefes-Dux, P. Bermel, and K. Madhavan, "Surveying the motivations of groups of learners in highly-technical STEM MOOCs," in *2016 IEEE Frontiers in Education Conference (FIE)*, 2016, pp. 1–6.

[6] N. M. Hicks *et al.*, "Integrating analytics and surveys to understand fully engaged learners in a highly-technical STEM MOOC," in *2016 IEEE Frontiers in Education Conference (FIE)*, 2016, pp. 1–9.

[7] K. Madhavan, M. Zentner, and G. Klimeck, "Learning and research in the cloud," *Nat. Nanotechnol.*, vol. 8, no. 11, pp. 786–789, 2013.

[8] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," in *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on*, 2010, pp. 63–67.

[9] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 63, no. 2, pp. 411–423, 2001.

[10] K. A. Douglas, P. Bermel, M. M. Alam, and K. Madhavan, "Big Data Characterization of Learner Behaviour in a Highly Technical MOOC Engineering Course," *J. Learn. Anal.*, vol. 3, no. 3, pp. 170–192, 2016.

[11] M. L. McHugh, "The chi-square test of independence," *Biochem. Medica*, vol. 23, no. 2, pp. 143–149, 2013.