# Analyze Diversity in MOOC Participants through Spatial and Temporal Data Mining

Zhongxiu Liu, Abhishek Agrawal, Dhyey Shah
North Carolina State University,Raleigh, NC 27606
{zliu24,akagrawa,dbshah3}@ncsu.edu

***Abstract*—Massive Open Online Courses (MOOC) have provided valuable learning experience for participants from worldwide. However, most prior research analyzed MOOC data without considering the population's diversity and its associated variety in learning behaviors. To address this problem. we applied spatial and temporal data-mining techniques on the interaction data from an 8-week MOOC courses with 24287 participants from 172 countries. We found that participants from different countries have significantly different frequencies in certain activity sequences. We also found different user types and their distributions among spatially clustered US regions. Our work provides valuable information for MOOC designers and analysts to understand and promote the diversity in MOOC population.**

## I. Introduction

Massive Open Online Courses (MOOC) have recently become hot trends in education. According to Pappano[8], Coursera, a MOOC founded in January 2012, attracted 1.7 million users in its first 10 months. A course in MOOC had hundreds of thousands enrolled with over ten thousand participants completed the course. The popularity of MOOC results in rich data and unprecedented challenges for both MOOC educators and data-miners. MOOC Educators, who are unable to observe their students in person, need data-mined knowledge to design and evaluate their courses. Data-miners, who are presented with data from a large worldwide population, need innovative data-mining techniques to address the diversity and the variety in learning behaviors within MOOC participants.

Several topics have drawn great research attention. First, the universally existence of high dropout rate in MOOC courses, as studied by Kim et al. [4] and Kloft et al [6]. Recent research try to understand, predict, and prevent dropout through leveraging data about students' goals, behavioral patterns, educational contents, etc. Second, the participants' engagement during their activities in MOOC, as studied by Kizilcec et al.[5] and Sinha et al[9]. Recent research found participants' engagement is crucial to predicting future drop-out and grades. Prior research have applied innovative techniques to mine participants' engagement pattern from their interaction data in MOOC.

However, there are still topics that need more research attention. Comparing with the majority of educational data, MOOC collects data from a larger and more diverse worldwide population. The scale of MOOC data may help address crucial educational questions related to cultural, geographical and regional difference. However, the diversity of MOOC population is rarely addressed in prior research that analyzed MOOC's interaction data. This can lead to data interpretation that favors groups with dominant numbers of participants. For example, participants from countries with low internet speed usually download MOOC videos first and watch them off-line later. These participants' learning behaviors are not capture by online data collection, so they should not be analyzed together with participants who conduct most of their learning activities online.

In this paper, we applied temporal and spatial data-mining techniques to a 8-weeks long MOOC data participated by 24287 students from 172 countries. For each participant, we summarized their MOOC activities, built temporal activity sequences, and associated these data with the participant's geographical and spatial information. We found that participants from different countries have significantly different frequencies in certain activity sequences, but not participants from different spatially clustered U.S. regions. We found clusters within US regions that shows the existence of similar and different user types. In summary, our research shows the diversity in MOOC's population and its associated differences in learning behaviors. Our research provides valuable information for MOOC designers and analysts to understand and promote the diversity in MOOC population.

## II. Related Work

In regard of temporal data, prior research has yield fruitful results through constructing n-gram models from MOOC's interaction data granulated by defined time-frames. Wen and Rose [10] characterized learning activities as topics consisting of specific start and end activities. This study mined frequent activity sequences within the time-frame of each topics, and generated daily topic distributions for participants of different grade levels. Brooks et al. [1] defined time-frame as 1-day, 3-day, week and month. For each MOOC activity, this study generated 2-grams to 5-grams consist of boolean values indicating whether the activity happened or not, during 2-5 sequences of a selected time-frame. This study then trained decision tree to predict students final grades based on these ngrams temporal data. The study resulted in models that were highly accurate for predicting participants' final grades, but lacked explanatory power.

Spatial information, on the other hand, is barely addressed in association with behavioral patterns. Most prior research limited spatial information to geographical information, which is used to describe the diversity of MOOC population. Nesterko et al. [7] visualized a MOOC's population density, gender

distribution, and achieved final across countries. DeBoer et al. [2] associated geographical information with basic statistics such as average time spent on homework. Both studies found differences between countries in their measured statistics. However, these studies did not address how this diversity transforms into learning behaviors.

In sum, MOOC data contains rich spatial and temporal information about the background of participants and their learning behaviors. While analysis of temporal information has yield fruitful results through the n-gram approach, analysis of spatial information is still limited to geographical information. There is a need to apply more delicate spatial and temporal data-mining techniques to analyze MOOC data.

## III. Methods and Results

### A. Data Description

Our data is collected from an 8-weeks long MOOC from Coursera, taught by Professor Ryan Baker from the Teachers College of Columbia University, on the content of educational data mining. The course started in 2013. Course activities included: weekly lectures taught through video and slides (five lectures + 1 introductory lecture per week); weekly quizzes consists of conceptual knowledge questions and result reports of data-mining assignments; a discussion forum that allows participants to communicate with other participants and the course instructors; and a wiki page with supplementary readings suggestions and course logistics. 24287 participants from 172 countries enrolled in the course, with 638 participants completed the course with a passing grade, calculated from scores of the weekly quizzes'. The completion rate is low but similar to other MOOC courses.

### B. Data Reprocessing

Our raw data are JSON objects containing user id, spatial data in the form of ip address, temporal data in the form of year/month/date/hour/min, and page urls recorded at each time when a participant interact with the MOOC webpage.

We first categorized the urls as one of the following activities: view lecture (VL), attempt/submit quiz (AQ,SQ), read/make a post in forum (RP,MP), and others such as reading wiki pages. We then built activity sequences for each participant in a temporal order, describing the whole time the participant spent on the MOOC. If an activity was conducted on the same content several times consequently, it will only be included once in its sequence. For example, a sequence "viewLecture15, viewLecture15, viewLecture16" will be presented as "viewLecture15, viewLecture16". This is because our data lacks the information about the reasons behind consequently repetitive activities, and our analysis only focuses on the transition between activities.

We then granulated activity sequences with two time-frames: weeks - consecutively 7 days since user's first action in the course; and actions between two specified actions, as in Wen and Rose's work. We generated statistics such as number of each activities and active days within each time-frame .

Lastly, we associated each participant with his/her country, city, and latitude/altitude locations, based on the participant's most frequent ip address.

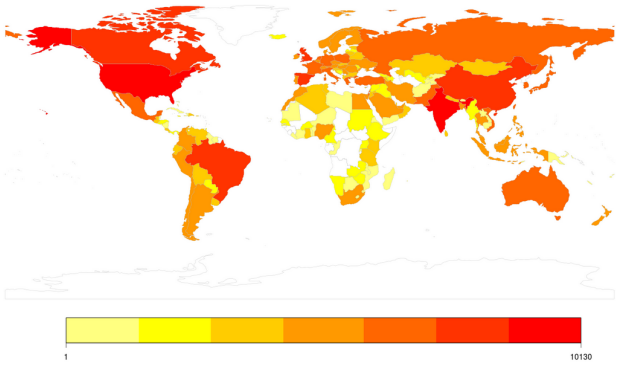


Fig. 1: Heat-Map of User Distribution by Country. The darker the color, the more participants.

| Country | # Partici-pants | active days$\geq$ 7 | Avg Time Span | Completion |
|---|---|---|---|---|
| US | 8385 | 46.8% | 17.2 | 3.17% |
| India | 3682 | 46.3% | 16.2 | 0.81% |
| China | 1038 | 41.6% | 14.9 | 1.54% |
| UK | 805 | 46.8% | 16.2 | 3.00% |
| Brazil | 711 | 48.7% | 17.8 | 1.41% |
| Canada | 689 | 50.0% | 19.0 | 5.08% |
| Spain | 655 | 53.4% | 20.1 | 3.67% |

TABLE I: Stats of MOOC Participants by Country. From left to right: number of participants who started the MOOC, percentage of participants who stayed for at least a week, average number of days the participants stayed (last-day minus first-day), and percentage of participants who completed the course with a grade

### C. Comparing Participants from Countries

Figure 1 shows a heat-map of participants from worldwide. We selected 7 countries with the most participants for further analysis. Table I show statistics about these 7 countries. US has the most participants (8385), followed by India, China, UK, Brazil, Canada and Spain. Spain has the highest percentage of participants who were active for at least a week and the longest average days spent by participants, while China has the lowest in both statistics. Canada has the highest percentage of participants who completed the course with a grade and India has the lowest, with 5.08% and 0.81% respectively.

To investigate the behavioral difference between participants, we mined pair-wise sub-sequences whose frequencies discriminate significantly a country from another, using chi-square test. We applied Bonferroni correction to counteract the problem of multiple comparisons. Notice a frequency is calculated by number of students with a activity sequence, over total number of student who conducted one or more of VL, AQ, SQ, RP or MP activities. We only selected activity sequences with a ¿ 0.12 support.

The result is shown in table II. Participants from china has significantly lower frequencies in activity sequences involving read-post when comparing with other countries. Participants from both China and India have significantly lower frequencies in sequences involving view-lecture, and sequences with view-

| | India | China | Spain |
|---|---|---|---|
| US | ++(VL-RP)<br>++(VL-RP-VL)<br>++(VL-VL)<br>++(RP)<br>++(VL-AQ) | +(RP)<br>+(VL-RP)<br>+(RP-VL)<br>+(VL-RP-VL) | –(AQ)<br>–(VL-AQ)<br>–(RP-AQ) |

| | UK | Brazil | Canada | Spain |
|---|---|---|---|---|
| India | –(VL-RP)<br>–(VL-VL) | –(VL-RP)<br>–(VL-VL) | –(VL-RP)<br>–(RP-VL)<br>–(VL-VL)<br>–(RP) | –(VL-RP)<br>–(VL-VL)<br>–(VL-AQ)<br>–(RP-VL) |
| | UK | Brazil | Canada | Spain |
| China | –(RP)<br>–(VL-RP)<br>–(RP-VL)<br>–(VL-RP-VL)<br>–(VL-VL) | –(RP)<br>–(VL-VL)<br>–(RP-VL)<br>–(VL-RP) | –(RP)<br>–(VL-RP)<br>–(RP-VL)<br>–(VL-RP-VL)<br>–(VL-VL) | –(RP)<br>–(VL-RP)<br>–(VL-RP-VL)<br>–(RP-VL)<br>–(AQ-AQ)<br>–(VL-VL)<br>–(VL-AQ)<br>–(AQ)<br>–(VL-AQ-VL) |

TABLE II: pair-wise activity sequences that are significantly in frequencies when compared between countries. A ++(–) indicates the country on the left most columns has significantly higher(lower) frequency of the sequence compared with country at the top of the corresponding column, with p-value under 0.05;a +(-) indicates significance with p-value between 0.05 and 0.1.

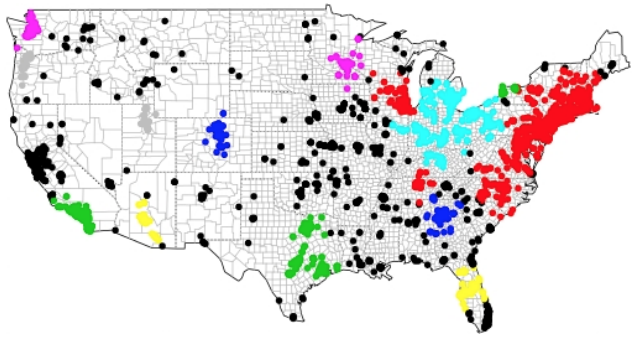

Fig. 2: Spatially Clustered U.S. MOOC Participants. Noise points are in black

| Region | # Participants | active≥ 7 | Avg Time Span | Completion |
|---|---|---|---|---|
| Northeast Coast | 2955 | 46.8% | 17.2 | 2.5% |
| California-Bay | 1133 | 44.7% | 15.7 | 1.68% |
| Michigan-Ohio | 509 | 47.7% | 18.5 | 4.32% |
| Texas | 482 | 48.5% | 19.0 | 4.56% |
| California-LA | 435 | 47.1% | 17.0 | 4.14% |

TABLE III: Stats of MOOC Participants by US Region

lectures in a row. Participants from Spain have significantly higher frequencies in sequences involving attempt-quiz, and combinations of view-lectures and read-posts. No other significantly different frequencies were found between participants from India and China, and between participants from UK, Brazil, Canada when comparing within themselves, and with US or Spain.

### *D. Comparing Participants from Spatially Clustered US Regions*

We applied DBSCAN[3] to identify regions of US participants. DBSCAN is a density-based clustering algorithm designed for spatial data. We choose DBSCAN because we did not know how many clusters exists given the wide distribution of participants across the U.S. We also expected our clusters to be in arbitrary shapes due to the variety of landscapes of the geographical regions where participants are likely to be from. Moreover, we expected large amount of noise points representing participants are from isolated locations. Our result is shown in Figure 2.

We identified some major clusters for further analysis and described them in Table 3. The Northeast Coast cluster has the highest number of participants, followed by California Bay, Michigan-Ohio, Texas and California-Los Angeles clusters. These 5 regions have similar percentage of participants who were active for at least a week, and similar average number of active days. The California Bay cluster has the lowest completion rate of 1.68%, followed by the Northeast Coast cluster with 2.5%. The Michigan-Ohio, Texas and California-LA clusters have higher completion rates above 4%.

We found the shapes of these spatially-clustered clusters resemble the distribution of major universities in the region. A comparison between the Northeast Coast cluster and the Google map of major universities along Northeast Coast is shown in Figure 3.

We applied the same sequential data-mining technique as described in previous section. We did not find any activity sequences whose frequency discriminate significantly one region from another.

To further investigate into the differences between spatially clustered regions, we created a vector for each participant to summarize his/her activeness in MOOC. This vector consists of normalized features selected by principal analysis, including: number lectures viewed, quizzes attempted and all activities during the whole MOOC courses. For each spatially clustered region, we performed k-means clustering based on this vector of activeness. The result is shown in Figure 4. For

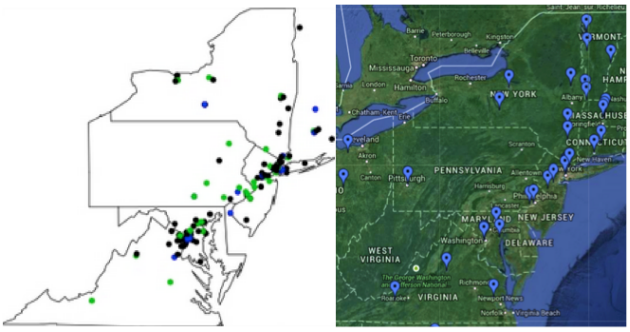


Fig. 3: Locations of Active Participants in the Northeast Coast Cluster (left) and the Major Universities (right) in the Region

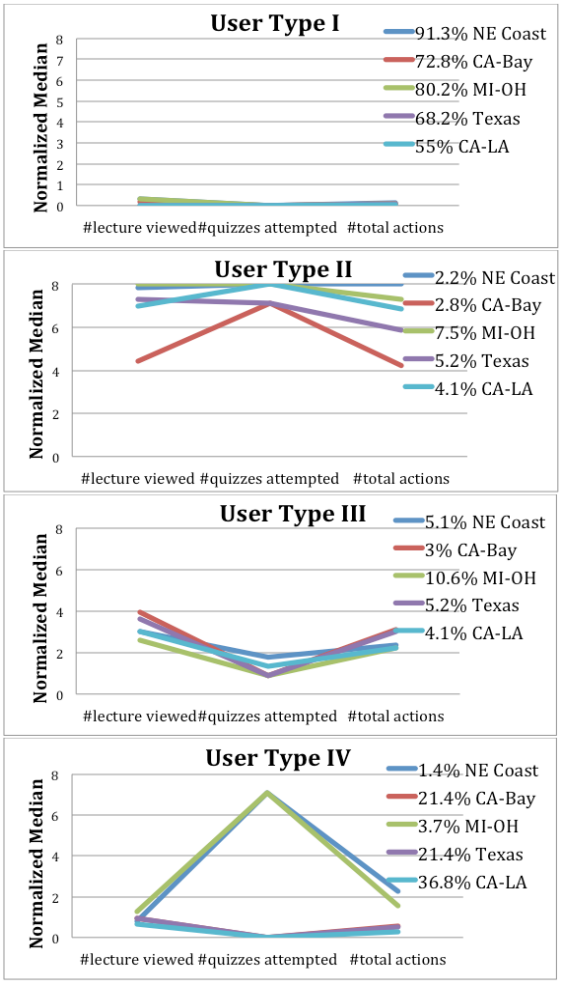


Fig. 4: User Types as clustered from each Region. Each user type plot describes the normalized mean value of total number of lecture viewed, quizzes attempted and actions during the MOOC courses. Each plot contains one out of the four clusters from each region. Percentage of each type of user for each region is marked in legend

each region, we found 4 clusters that categorize its participants. All regions contain three types of participants: I) low number of lectures viewed, quizzes attempted, and total actions II) High number of lectures viewed, quizzes attempted, and total actions III) mid-high number of lectures viewed; mid-low number of quizzes attempted and total actions; VI)mid-low number of lectures viewed. However, for user type VI, Michigan-Ohio and Northeast Coast regions have mid-high number of quizzes attempted whereas California Bay, California LA and Texas regions has low number quizzes attempted.

## IV. Discussion

When comparing between countries, we found significant difference in the frequencies of several activity sequences. These differences can be stemmed from a number of causes, such as differences in how e-learning is viewed in a country, in demographics (age, gender, educational background, first language) of a country's participants, or infrastructure's qualities such as the accessibility and quality of Internet in a country.

When comparing between spatially clustered U.S. regions, we did not find significant difference in the frequency of any activity sequences. This may because these clusters are formed around large US cities and universities, where the causes mentioned above for the differences between countries are of less concerns. However, further analysis on activeness statistics not only revealed the differences between regions, but the diversity within the regions' participants. Some regions have distinct groups of users who attempted high number of quizzes but viewed low number of lectures. This may because some regions have larger career-focus population, who were already familiar with the materials but need the MOOC certificate for career opportunities.

Our results provide helpful information for future MOOC designers and analysts. Future MOOC design should support different learning behaviors through avoiding requirements or measurements that favor certain behavior types. Future MOOC design may also encourage participants from similar regions, countries, or of similar behavioral patterns to collaborate with during learning. On the other hand, MOOC analysts should keep the diversity in mind, and avoid interpretations with biased assumptions.

## V. Conclusion and Future Work (term paper)

Through applying spatial and temporal data-mining techniques, we discovered behavioral differences between participants from different countries and spatially clustered U.S. regions. Our results showed the differences in behavioral sequences and activities activeness between participants from different countries and spatial-clustered U.S. regions. These results provide valuable information for MOOC analysts and designers to understand and promote the diversity in MOOC population.

Future work will be expanded in several directions. In the data-collection process, future research could include online survey to collect richer information from participants, such as their educational backgrounds, their understanding about MOOC, their language proficiencies and their locations' internet speeds. These information will give us more evidences on what caused certain behavior differences during learning. These information will also show how demographic information contributes to the differences between countries and regions. Moreover, Future research should collect more detailed information for each activity. For example, our data contains many consequently interactions with a same video page; some of these interactions may indicate 'forward', 'backward', or 'stop' during a lecture. These detailed activity information will create activity sequences with richer content and engagement patterns.

## References

[1] C. Brooks, C. Thompson, and S. Teasley. A time series interaction analysis method for building predictive models of learners using log data. In *In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 126–135, 2015.

[2] J. DeBoer, G. S. Stump, D. Seaton, and L. Breslow. Diversity in mooc students backgrounds and behaviors in relationship to performance in 6.002 x. In *In Proceedings of the Sixth Learning International Networks Consortium Conference.*, 2013.

[3] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[4] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller. Understanding in-video dropouts and interaction peaks inonline lecture videos. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 31–40, 2014.

[5] R. F. Kizilcec, C. Piech, and E. Schneider. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the 3rd international conference on learning analytics and knowledge*, pages 170–179, 2013.

[6] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart. Predicting mooc dropout over weeks using machine learning methods. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing, Workshop on Modeling Large Scale Social Interaction In Massively Open Online Courses*, page 60, 2014.

[7] S. O. Nesterko, S. Dotsenko, Q. Han, D. Seaton, J. Reich, I. Chuang, and A. D. Ho. Evaluating the geographic data in moocs. In *In Proceedings of the 2013 Conference on Neural Information Processing Systems*, 2013.

[8] L. Pappano. The new york times. *The Year of the MOOC*, ED 26, 2012.

[9] T. Sinha, N. Li, P. Jermann, and P. Dillenbourg. Capturing" attrition intensifying" structural traits from didactic interaction sequences of mooc learners. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing, Workshop on Modeling Large Scale Social Interaction In Massively Open Online Courses*, 2014.

[10] M. Wen and C. P. Rose. Identifying latent study habits by mining learner behavior patterns in massive open online courses. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1983–1986, 2014.