Demographical Changes of Student Subgroups in MOOCs: Towards Predicting At-Risk Students

Bokuan Yang
Department of Computer Science, University of Liverpool
Liverpool, UK *b.yang12@student.liverpool.ac.uk*

Lei Shi
Department of Computer Science, University of Liverpool
Liverpool, UK *lei.shi@liverpool.ac.uk*

Armando M. Toda
Institute of Mathematical and Computer Sciences
University of São Paulo, São Carlos, Brazil *armando.toda@usp.br*

Abstract

Past studies have shown that student engagement in Massive Open Online Courses (MOOCs) could be used to identify at-risk students (students with drop-out tendency). Some studies have further considered student diversity by looking into subgroup behavior. Yet, most of them lack consideration of students' behavioral changes along the course. Towards bridging the gap, this study clusters students based on both their interaction with the system and their characteristics and explores how their cluster membership changes along the course. The result shows that students' cluster membership changes significantly in the first half of the course and stabilized in the second half of the course. Our findings provide insight into how students may be engaged in learning on MOOC platforms and suggest the improvement of identifying at-risk students based on their temporal data.

Keywords: FutureLearn, MOOC, subgroup, patterns, learning analytics.

1. Introduction

Massive Open Online Courses (MOOCs) are a unique form of educational information system offering free access to the intellectual holding of universities [8, 25]. It has been spreading in both domestic and international education sectors. Many world-class universities have joined in the MOOC movement, and various MOOCs have been launched across nations in various subjects [18, 30]. Despite the potential and hype associated with MOOCs, the completion rates overall are normally low; some reported that the completion rate could reach below 10% [15]. This has catalyzed considerable studies on identifying dropout possibilities [1, 10, 12, 13].

In comparison to traditional educational methods, MOOCs allow for prediction of when a student may dropout from a course based on their prior voluntary actions logged in the database [3].Most existing studies on identifying at-risk students (students with dropout tendency) in MOOCs heavily rely on the "average/overall" analyses, lacking adequate examination of the potential differences amongst subgroups of students. This approach may produce predictions with potential pitfalls [4, 6, 14]. As such, this study aims at addressing this concern by exploring the diversity of student subgroups and the demographical changes (the percentage of students falling into each subgroup and the subgroup transitional pattern) along the course.

In this study, we combine the previous study on identifying student subgroups, using both students' interaction data with the MOOCs platform and their characteristics to allow for a more accurate clustering [7, 20, 26]. This paper presents the student subgroups clustered from two courses delivered on the FutureLearn MOOCs platform and visualizes demographical changes of these subgroups along the courses progressed to help unmask the behavioral changes at different stages of the course. In particular, this study aims to answer the following three research questions:

RQ1. *How can we subgroup students in online courses?*
RQ2. *How can demographical characteristics of each subgroup change, by weeks?*
RQ3. *Are there transitional patterns amongst subgroups, on a weekly time scale?*

2. Literature Review

2.1. Subgroup Clustering in MOOCs

Some previous studies attempted to classify students based on their interaction with lectures and assignments using a variety of methods and approaches, including bottom-up approaches to identify potential subgroups [14, 19] and top-down approaches to partition students into pre-defined groups [3]. For example, Kizilcec, Piech and Schneider [21] demonstrated a clustering technology based on a derivative single variable for engagement, where they labelled all students either "on track" (took the assessment on time), "behind" (turned in the assessment late), "auditing" (didn't do the assessment but engaged in watching videos) or "out" (didn't participate in the course at all). In their study, four types of engagement trajectories were extracted: "Completing" – students who completed the majority of the assessments; "Auditing" – students who did assessment infrequently if at all and engaged instead of watching video lectures; "Disengage" – students who did assessment at the beginning of the course but then had a marked decrease in engagement; and "Sampling" – students who watched video lectures for only one or two assessment periods [16].

While, in their research, they used the k-means clustering algorithm to categorical data to a certain extent since they just assigned a numerical value to each label (on tack = 3, behind = 2, auditing = 1, out = 0). And converting categorical data into numeric values does not necessarily produce meaningful results in the case where categorical domains are not ordered [17]. Therefore, these approaches have potential problems with converting participation labels, although they still provide a viable way to cluster students based on log data from MOOCs platforms. In our study, we used the one-hot encoding [5] to convert categorical data, thus reducing the impact of the categorical data.

Other studies focused on different approaches to identify students' subgroups, but most of them did not consider behavioral changes over time from the clustering [14, 19, 23, 24]. It is important to explore behavior patterns of subgroup students on a specific time scale, since the characteristics of each subgroup, and the proportion of its total interaction, vary along a course progresses; This can also help the platform adjust the content of the course according to the progress of the course.

In our current study, we apply a bottom-up cluster approach using k-means++ cluster algorithm with students' log data to identify distinct subgroups as well as observe their characteristics changes on a weekly time frame, thus offering a dynamic perspective for students' subgroups.

2.2. Learning Persistence in MOOCs

Considering the problem of the low completion rates in MOOCs, learning persistence was selected as a critical MOOC outcome, which can provide valuable insights into the interactions between the course design and students factors [9, 10, 16]. Several studies have demonstrated possible ways of using learning analytics on interaction and assessment to meaningfully classify student types or subgroups and visually represent patterns of student engagement in different phases of a MOOC. For example, Coffrin et al [7] divided weekly participation into three mutually exclusive student subgroups: Auditors – those who watched videos in a particular week instead of participating assessments; Active learners – those who participated in an assessment in a week; and Qualified learners – those who watched a video or participated in an assessment. The study investigated students' temporal engagement along course progressed. It also showed a way of combining the State-Transition diagram with an analysis of student subgroups to illustrate the students' temporal engagement in courses. Their result indicated that different courses might show similar patterns, although they were different in terms of the curriculum and assessment design.

Similar studies have attempted to compute a description for individual students in terms of how they engaged in each assessment period of a course and then applied clustering techniques to find subgroups in these engagement descriptions [11, 14, 21]. While these studies have successfully concluded the proportion of students in different subgroups by week, they did not attempt to analyze the individual subgroup changes on a specific time scale. Student behavior may change along a course progresses, where they may have been labelled into one subgroup and transit to another in subsequent weeks. It is meaningful to evaluate the transitional pattern for each subgroup on a certain time scale. Therefore, in this study, we measured the proportion of students falling into each subgroup and concluded the transitional pattern for each subgroup on a weekly time frame.

3. Method

3.1. The Course and the Dataset

Our dataset was from the course "Leadership for Healthcare Improvement and Innovation" and "Supply Chains in Practice: How Things Get to you", delivered on the FutureLearn MOOCs Platform, which is freely available for everyone. The course had an official starting week (considered in this study as Week 1) with a duration of six weeks. Each week consisted of multiple steps, which represented lecture streams that students needed to visit. "Registered students" are those who have enrolled in the course; "participated students" are those who have attended at least one steps; "completed students" are those who have completed the courses by the end of Week 6; and "purchased students" are those who have bought the certificate of the course. Table 1 shows the statistics for these two courses.

Table 1. Course design and participants

Course	"Leadership for healthcare improvement and innovation"	"Supply chain in practice: How things get to you"
Duration of the course	6 weeks	6 weeks
Total steps	73	109
Registered students	4,046	5,808
Participated students	2,397	2,924
Completed students	377	318
Purchased students	149	69

The dataset used in this study include:

1. Step record: which student at what time visited which step; when they marked a step as complete.
2. Comment record: which student at what time left what comment on which step; how many "likes" a comment received.
3. Student record: students' demographical information such as age and gender (Table 2).

Students' demographical information was collected using a pre-course survey asking optional questions about their gender, age, etc. (see Table 2, left column). Only 9.5% of the students (506 out of 5,321) answered all the survey questions. Since using incomplete student record would affect the result of data analysis, in this study we only used the records of students who answered all the survey questions.

Table 2. Demographical Information in the Student Record

Variable	Description

User ID	The unique identifier for a student
Gender	The gender of the student
Age group	The age group where the student belongs to
Country	The country where the student belongs to
Highest Educational Level	Student's highest education level
Employment Status	Students' employment status
Employment Area	Students' employment area

3.2. Subgroup Clustering

In previous studies, watching lecture videos and submitting assignments were used for clustering students [14, 19]. Considering the conversational framework of FutureLearn and the course design, two interactive indicators were generated from the *step record* and the *comment record*: 1) steps visited (the proportion of all the steps available visited by a student), and 2) comments submission (the number of comments submitted by a student). Other studies, e.g. [2, 27], used demographic indicators such as gender and age to predict student engagement; [29] focused on the use of learning platform's features in order to analyze learning behavior patterns. Different from previous studies, we selected both students' demographical data and their interaction data for the clustering process. We excluded the highly correlated factors with steps visited or comments submission, leaving gender as an extra factor for the clustering process. The clustering process was based on the k-means ++ algorithm, which could reduce the influence of randomly assigned initial centroids in the k-means algorithm. Similar to previous studies [28], the "Elbow method" was used to calculate a reference K value for the k-means ++ algorithm [22]. Several K values around the reference K was used to cluster subgroups, and we conducted Kruskal-Wallis H tests and Mann-Whitney U tests to examine whether the K value could differentiate subgroups on every clustering factor. Moreover, different from most prior studies, which used cumulative data from the entire course to cluster subgroups, we used cumulative data from each week for subgroups clustering.

3.3. Transitional Pattern for Subgroups

We clustered students into subgroups based on their temporary behavioral data (how they interact with the course including visiting steps and submitting comments from week to week); we used a State-Transition diagram to visualize the weekly transitional patterns amongst subgroups, where the dropped-out students were marked into a separate subgroup. Similar to the subgroup clustering, three indicators were generated, steps visited (the proportional of all the steps available visited by a given student in a given week), comments submitted (the number of comments submitted by a given student in a given week) and the gender (of a given student).

From the State-Transition diagram, we analyzed the proportion for students falling into each of the subgroups by week and generalized the transitional pattern for different subgroups each week.

4. Result

4.1. Subgroup Clustering

In this study, we selected the percentage of steps visited and the number of comments made by students as prime cluster indicators, with additional indicators selected from the student record. From the correlation analysis, we excluded highly correlated indicators. More specifically, we used the $\eta(\mathrm{eta})$ statistics to measure the degree of association between categorical and numeric variables, the independent variable Y(Step and Comment) and dependent variable X(Gender, Country, Age range, Educational level, Employment area and Employment status) (see Table 2 selected as potential cluster indicators.

For the association between categorical variable, the Chi-square test with significant level of .05 was used. The result suggested that there is a strong association between Gender and Employment area ($\chi^2(23) = 39.9, p < 0.05$). Therefore, only one of these two variables will be selected as a cluster indicator. Considering the fact that the MOOCs analyzed in this study are specialized in certain subjects thus may result in special employment distribution, the gender indicator was selected to conclude a general conclusion. As such, our absolute selection of variables included:

1. Steps: the percentage of steps visited by students;
2. Comments: the number of comments submitted by students; and
3. Gender: the gender of students.

Although the FutureLearn platform provides multiple gender options in the pre-course survey, we only considered female and male options, as the others were very underrepresented. Therefore, we considered the gender variable as a dummy variable and we used 0 to represent female and 1 to male.

Using the "elbow method", Mann-Whitney U tests and the K-means ++ cluster algorithm, we successfully clustered 506 students into three distinct subgroups based on the cumulative data. More specifically, we used the "elbow method" to estimate the optimal K-value for the K-means++ algorithm processed in this study, where the result can be seen below in Fig. 1.

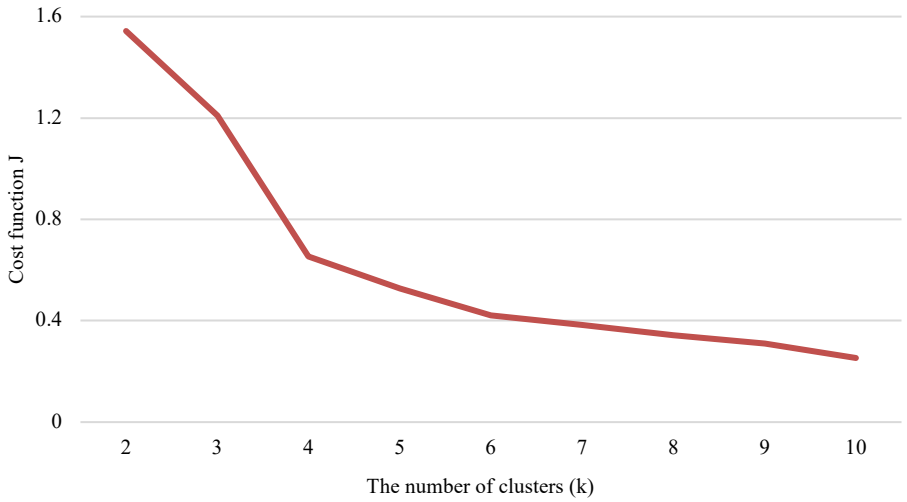

Fig. 1. Cost function J for the dataset

The method believes that one should choose a number of clusters so that adding another cluster does not give much better modelling of data. In Fig. 1, the result cost function J has experienced the most significant decrease in K=4. Hence, the K=4 was chosen as a reference K-value in the subsequent analysis. Based on this reference K-value, several K-values from 2 to 5 has been used to cluster student subgroups. In this case, the Mann-Whitney U test with significant level = 0.05 was chosen to validate whether there was a significant difference among each subgroup, and the results suggested that neither K=4 nor K=5 can differentiate subgroups. Hence K=3 has been chosen this study and the cluster results can be seen in Table 3.

Table 3. Subgroup cluster centroids

	Steps	**Comments**	**Gender**	**N**
Cluster 1 / Samplers	.926	7.16	.360	113
Cluster 2 / Viewers	.107	.91	.353	369
Cluster 3 / All-rounders	.990	67.54	.550	24

Based on the previous work on Anderson et.al, where they labelled students based on lectures watching and assignment submission into three subgroups: Viewer(primary watching lectures handing in few if any assignments), Solvers (primary hand in assignments, viewing few if any lectures) and All-rounders (balance between watching of lectures and handing in

assignments). On the basis of Anderson's work [3], we further clustered students by their positivity. In this study, we did not choose assignment submission as cluster indicators, but we chose the number of comments submitted to replace assignment submission in previous work, where we labelled each subgroup into the following:

- **Viewers** (Cluster 1; 22.33% of the total population). Overall, they visited a very high percentage (92.6%) of steps but submitted very few comments (Mean = 7.16).
- **Samplers** (Cluster 2; 72.92% of the whole population): they made up the largest student group, but they were also the least engaged students – they visited only 10.7% of the steps and on average they left only 0.91 comments.
- **All-rounders** (Cluster 3; 4.74% of the total population). They made up the smallest group, yet they were the most engaged students – they visited 99.0% of the steps and on average they left 67.54 comments.

4.2. Weekly Changes of Cluster Centroid

To further explore the temporal subgroups, we further divided the students into two categories based on the number of steps they visited and the number of comments they made. The students who had partially participated (at least made a comment or visited a step) the courses in a given week were selected and clustered into 3 subgroups, based on the k-means ++ algorithm. Steps, comments and gender were selected as clustering indicators. As shown in Table 4, the cluster centroids stabilized at a certain level across weeks, which suggests that the same subgroup had a similar behavior pattern at different stages of courses.

Table 4. Centroids for weekly subgroups

		Steps	Comments	Gender
Viewer	Week 1	0.964	1.300	0.544
	Week 2	0.979	0.934	0.610
	Week 3	0.988	0.792	0.625
	Week 4	0.936	0.624	0.624
	Week 5	0.986	0.784	0.589
	Week 6	0.971	1.490	0.640
Sampler	Week 1	0.214	0.300	0.428
	Week 2	0.229	0.195	0.507
	Week 3	0.207	0.000	0.467
	Week 4	0.206	0.035	0.517
	Week 5	0.259	0.105	0.526
	Week 6	0.180	0.133	0.467
All-rounder	Week 1	0.986	11.886	0.571
	Week 2	0.998	12.138	0.483
	Week 3	0.990	10.880	0.560
	Week 4	1.000	10.583	0.625
	Week 5	0.998	12.320	0.640
	Week 6	0.952	14.875	0.687

4.3. Weekly Changes of Subgroup

To estimate the subgroup changes along with the courses, the percentage of students labelled in each subgroup per week were also retrieved from the dataset. From Fig. 2 we can see that the first half and the second half of the course had very different demographic characteristics, where the percentage of the students in each subgroup changed significantly in the first half of the courses (Week 1 to Week 3). More specifically, the percentage of Samplers decreased from 50.4% to 17%, which may be caused by a large number of dropout students in the first two

weeks. The proportion of Viewers increased significantly from 42.8% in Week 1 to 68.8% in Week 3 and kept stable at a certain level in the rest weeks. The proportion of All-rounders kept at a relatively stable level, i.e. around 0.1, which suggests that these students were relatively stable, even in the beginning weeks when many students dropped out.

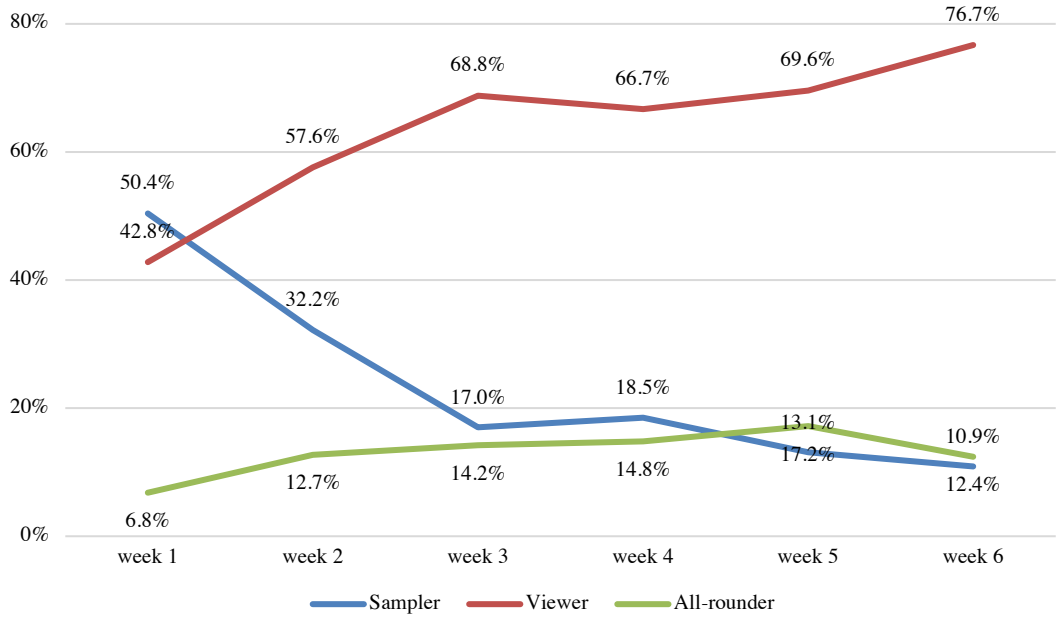

Fig. 2. The percentage of students each subgroup across weeks

Table 5. The number of students each subgroup across weeks

	Sampler	Viewer	All-rounder
Week 1	252	214	34
Week 2	76	136	30
Week 3	30	121	25
Week 4	30	108	24
Week 5	19	101	25
Week 6	14	99	16

Here, we use the State-Transition Diagram to present in detail how the students shifted between subgroups. We assumed possible student subgroups, i.e. Sampler, Viewer, All-rounder and Drop-out, as four possible states each week, and the transitions from one subgroup to another was indicated by the arcs between two states.

Fig. 3 provides a legend to understand the State-Transition diagram used in this study. The legend shows two subgroups A and B; the arcs between circles represent the students transited their subgroup from A to B in a subsequent week. To better visualize the number of students in each subgroup each transition, the circle areas and arc's weight are linearly related to the number of students in subgroups and transitions respectively.

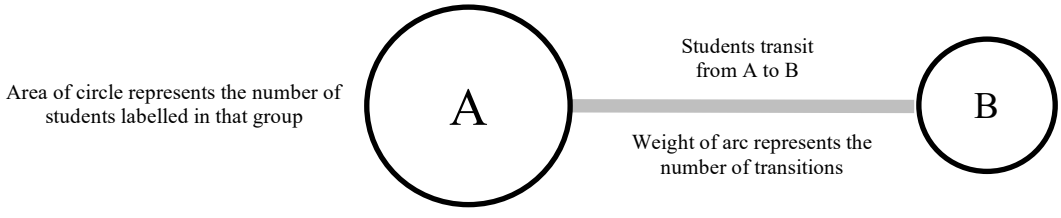

Fig. 3. State-Transition Diagram Legend

Fig. 4 demonstrates the demographical changes for Samplers. A large proportion of students dropped out from the courses in the following week, while a small percentage of them maintained their behavior or transited to become Viewers. However, apart from the first week, no student had transited from Sampler to All-rounders in the following weeks, which suggests that it is unlikely for a highly inactive student to become highly active in a short period.

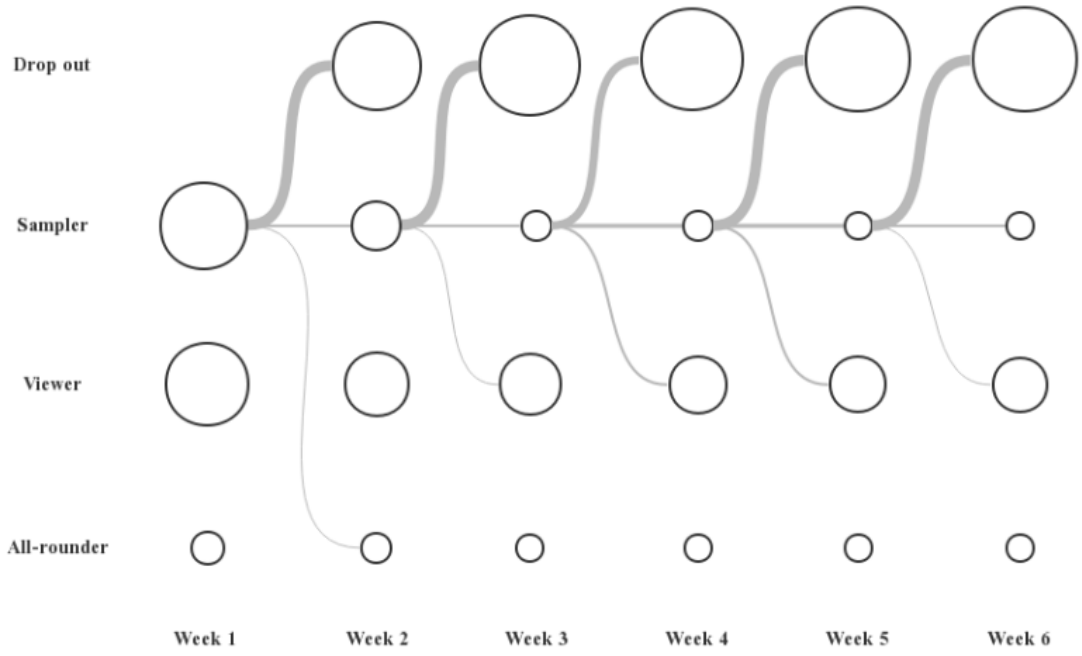

Fig. 4. Samplers' demographical changes across weeks

Fig. 5 emphasizes demographical changes for Viewers, which also indicates that each subgroup had a similar behavioral pattern transition each week. Yet, different from Samplers, most Viewers maintained their behavior patterns in the following week with only a small percentage of them dropped out from the courses or transited to another subgroup.

Similarly, Fig. 6 shows that while All-rounders represented the smallest proportion of the students, they were the most stable subgroup. There was no significant demographical fluctuation event in the first half of the courses, where the number of Samplers and Viewers decreased from 250 to 30 and from 215 to 120, respectively.

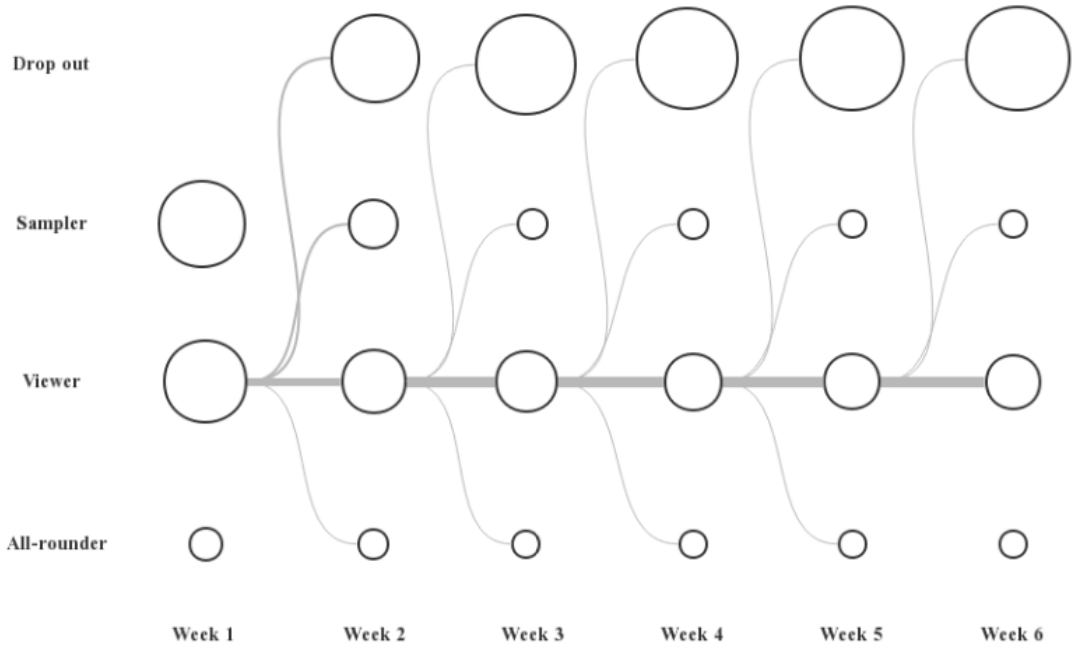

Fig. 5. Viewers' demographical changes across weeks

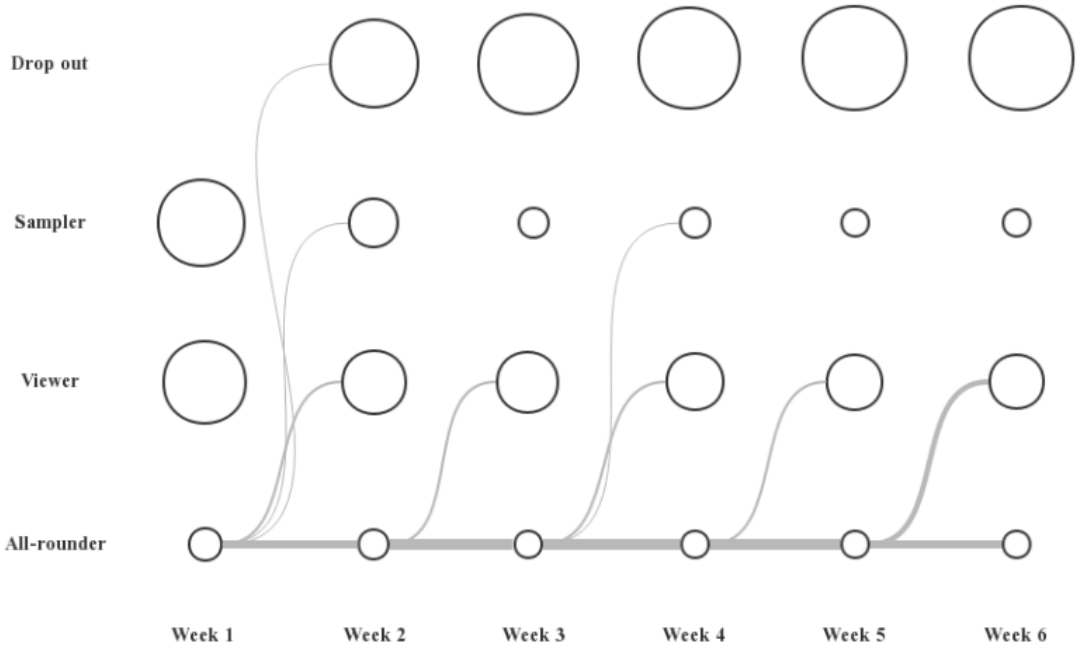

Fig. 6. All-rounders' demographical changes across weeks

5. Discussions and Conclusions

In this study, we have examined demographical changes in student subgroups on a weekly time scale. Different from previous studies that used behavioral data to pre-define or cluster student subgroups, this study used both interaction log data and students' characteristics (gender, in particular). More specifically, to answer Research Question 1, we clustered students into three distinct subgroups using the K-means ++ algorithm and the "Elbow method", as well as the

Mann-Whitney test. Sampler, Viewer and All-rounder were generalized. We have analyzed the differences amongst these subgroups and measured the proportion of students in different subgroups by week. To answer Research Question 2, we examined the behavior changes for students labelled in each subgroup where we found that using similar cluster approaches on weekly accumulated data could generate similar subgroups as the overall clustering result. Most of the subgroup's centroid remained stable within a certain range except All-rounders with the number of comments continuously rising in the second half of the course. To answer Research Question 3, we visualized the demographical changes of subgroups across weeks. Our result suggests that the first half of the course, i.e. Week 1 to Week 3, and the second half of the course, i.e. Week 4 to Week 6, had different demographical characteristic, where the demographics of these subgroups changed significantly from the first half of the former and maintained a certain degree of stability in the latter. More specifically, our study suggests that the less active subgroups took up most of the participants in the early courses, and as the course progressed, the proportion of those subgroups continued shrinking to around 10% (see Fig. 2). This result is opposite to those from previous studies which assume proportion of participants falling into each category keep stable to some extent along courses progress.

For the transition of each subgroup, our result demonstrated that each of them had similar transitional pattern along the courses, where most Samplers dropped out in the subsequent week with small percentage of them kept Sampler's behavior unchanged or transited into Viewer; A large proportional of Viewers maintained the same behavior pattern to subsequent week, and a percentage of students transited to Sampler, All-rounder or dropped out; The All-rounder was the most stable subgroup, where the demographical characteristics stabled from the beginning to the end of the course. Interestingly, the result in Section 4.3 suggests that it was almost impossible for students to switch from highly inactive (Sampler, as in this study) to highly active (All-rounder, as in this study) in a short period of time, and vice-versa.

To conclude, in this study we have analyzed students' data from two MOOCs offered by the FutureLearn platform. The result suggests that the first half and second half of both courses had different demographical characteristics and each student subgroup had their unique behavior and transitional pattern along the courses progressed.

Given the fact that MOOC students have various study behavior, with a very different interaction pattern with the course materials and their peers, when designing MOOCs, there is a strong need for providing personalized support to students that can be labelled into different subgroup at *different stages* of the course. This means the platform should personalize the way they learn, such as adapting learning paths and supporting adaptive intervention for different subgroups. Moreover, the subgroups identified in this study and the weekly demographical changes of those clusters may help inform a range of strategies for the intervention and improvement of MOOCs and MOOC platforms. For example, providing more previews of course materials allows Samplers to make a more informed decision about whether to participate in the first place. Offering more reminders for students who labelled as Sampler on unfinished steps and reduce the incentives for their comment submissions.

This study contributes to the understanding of subgroup clustering and demographical changes in MOOCs. Empirical evidence from this study supports that students' characteristics can also be used as clustering indicators, and the proportion of different subgroups in the total number of students each week may vary along the course. These results highlight the importance of examining subgroup to improve the effectiveness of the identification of at-risk students. In a further study, the same research approach could be applied into courses with more general content where there are more attributes with less association with students' interaction data (Steps visited and Comments submission, as in this study). In this study, the course "Leadership for healthcare improvement and innovation" does not contain any assignment, hence the assessment factor was not considered in subgroup clustering. In a future study, the assignment submission and grade could also be used as cluster indicators.

In terms of limitations, firstly, the dataset available was limited – after removing students with incomplete information, only 506 students' data was retained, and those students might share different characteristics with eliminated students. Besides, the filed involved in the courses used in this study were highly targeted. Therefore, the conclusion drawn from the

analysis of the dataset may be not universally applicable to a course in the other fields. Secondly, the courses that we focused on were unique in duration and structure in which students needed to access both a large number of steps and tools supporting reflection, comment and response.

References

1. Alamri, A., Alshehri, M., Cristea, A., Pereira, F.D., Oliveira, E., Shi, L., Stewart, C.: Predicting MOOCs Dropout Using Only Two Easily Obtainable Features from the First Week's Activities. In: Coy, A., Hayashi, Y., and Chang, M. (eds.) Intelligent Tutoring Systems. pp. 163–173. Springer International Publishing (2019)
2. Alshehri, M., Foss, J., Cristea, A.I., Kayama, M., Shi, L., Alamri, A., Tsakalidis, A.: On the Need for Fine-Grained Analysis of Gender Versus Commenting Behaviour in MOOCs. In: Proceedings of the 2018 The 3rd International Conference on Information and Education Innovations - ICIEI 2018. pp. 73–77. ACM Press, London, United Kingdom (2018)
3. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J.: Engaging with Massive Online Courses. In: Proceedings of the 23rd international conference on World wide web - WWW '14. pp. 687–698. ACM Press, Seoul, Korea (2014)
4. de Barba, P.G., Kennedy, G.E., Ainley, M.D.: The Role of Students' Motivation and Participation in Predicting Performance in a MOOC: Motivation and Participation in MOOCs. Journal of Computer Assisted Learning. 32 (3), 218–231 (2016)
5. Beck, J.E., Woolf, B.P.: High-Level Student Modeling with Machine Learning. In: Gauthier, G., Frasson, C., and VanLehn, K. (eds.) Intelligent Tutoring Systems. pp. 584–593. Springer Berlin Heidelberg, Berlin, Heidelberg (2000)
6. Brinton, C.G., Buccapatnam, S., Chiang, M., Poor, H.V.: Mining MOOC Clickstreams: On the Relationship between Learner Behavior and Performance. arXiv preprint arXiv:1503.06489. (2015)
7. Coffrin, C., Corrin, L., de Barba, P., Kennedy, G.: Visualizing Patterns of Student Engagement and Performance in MOOCs. In: Proceedings of the Fourth International Conference on Learning Analytics and Knowledge - LAK '14. pp. 83–92. ACM Press, Indianapolis, Indiana (2014)
8. Cormier, D., Siemens, G.: Through the Open Door: Open Courses as Research, Learning and Engagement. Educause, 45 (4): 30-39. Retrieved 2014/1/23 from: http://www.educause.edu/EDUCAUSE+ Review (2010)
9. Cristea, A.I., Alamri, A., Alshehri, M., Kayama, M., Foss, J., Shi, L., Stewart, C.D.: Can Learner Characteristics Predict Their Behaviour on MOOCs? In: Proceedings of the 10th International Conference on Education Technology and Computers - ICETC '18. pp. 119–128. ACM Press, Tokyo, Japan (2018)
10. Cristea, A.I., Alamri, A., Kayama, M., Stewart, C., Alshehri, M., Shi, L.: Earliest Predictor of Dropout in MOOCs: A Longitudinal Study of FutureLearn Courses. Presented at the 27th International Conference on Information Systems Development (ISD2018), Lund, Sweden August 22 (2018)
11. Cristea, A.I., Alamri, A., Kayama, M., Stewart, C., Alshehri, M., Shi, L.: How is Learning Fluctuating? FutureLearn MOOCs Fine-grained Temporal Analysis and Feedback to Teachers and Designers. In: 27th International Conference on Information Systems Development (ISD2018). Association for Information Systems, Lund, Sweden. (2018)
12. Davis, D., Jivet, I., Kizilcec, R.F., Chen, G., Hauff, C., Houben, G.-J.: Follow the Successful Crowd: Raising MOOC Completion Rates through Social Comparison at Scale. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17. pp. 454–463. ACM Press, Vancouver, British Columbia, Canada (2017)
13. Dominguez, M., Bernacki, M.L., Uesbeck, P.M.: Predicting STEM Achievement with Learning Management System Data: Prediction Modeling and a Test of an Early Warning System. In: EDM. pp. 589–590. (2016)
14. Ferguson, R., Clow, D.: Examining Engagement: Analysing Learner Subpopulations in Massive Open Online Courses (MOOCs). In: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge - LAK '15. pp. 51–58. ACM Press, Poughkeepsie, New York (2015)

15. Guzdial, M., Adams, J.C.: MOOCs Need More Work; So Do CS Graduates. Communications of the ACM. 57 (1), 18–19 (2014)
16. Halawa, S., Greene, D., Mitchell, J.: Dropout Prediction in MOOCs using Learner Activity Features. Proceedings of the Second European MOOC Stakeholder Summit. 37 (1), 58–65 (2014)
17. Huang, Z.: Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Min. Knowl. Discov. 2 (3), 283–304 (1998)
18. Jung, Y., Lee, J.: Learning Engagement and Persistence in Massive Open Online Courses (MOOCS). Computers & Education. 122 9–22 (2018)
19. Khalil, M., Ebner, M.: Clustering Patterns of Engagement in Massive Open Online Courses (MOOCs): the Use of Learning Analytics to Reveal Student Categories. Journal of Computing in Higher Education. 29 (1), 114–132 (2017)
20. Khalil, M., Ebner, M.: What Massive Open Online Course (MOOC) Stakeholders Can Learn From Learning Analytics? arXiv:1606.02911 [cs]. 1–30 (2016)
21. Kizilcec, R.F., Piech, C., Schneider, E.: Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. In: Proceedings of the third international conference on learning analytics and knowledge. pp. 170–179. ACM (2013)
22. Kodinariya, T.M., Makwana, P.R.: Review on Determining Number of Cluster in K-Means Clustering. International Journal. 1 (6), 90–95 (2013)
23. Kovanović, V., Joksimović, S., Gašević, D., Owers, J., Scott, A.-M., Woodgate, A.: Profiling MOOC Course Returners: How Does Student Behavior Change Between Two Course Enrollments? In: Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16. pp. 269–272. ACM Press, Edinburgh, Scotland, UK (2016)
24. Li, Q., Baker, R.: The Different Relationships between Engagement and Outcomes across Participant Subgroups in Massive Open Online Courses. Computers & Education. 127 41–65 (2018)
25. Rieber, L.P.: Participation Patterns in a Massive Open Online Course (MOOC) about Statistics: MOOC Participation. British Journal of Educational Technology. 48 (6), 1295–1304 (2017)
26. Shi, L., Cristea, A., Toda, A., Oliveira, W.: Revealing the Hidden Patterns: A Comparative Study on Profiling Subpopulations of MOOC Students. In: The 28th International Conference on Information Systems Development (ISD2019). Association for Information Systems, Toulon, France (2019)
27. Shi, L., Cristea, A.I.: Demographic Indicators Influencing Learning Activities in MOOCs: Learning Analytics of FutureLearn Courses. Presented at the The 27th International Conference on Information Systems Development (ISD2018), Lund, Sweden August 22 (2018)
28. Shi, L., Cristea, A.I.: In-depth Exploration of Engagement Patterns in MOOCs. In: Hacid, H., Cellary, W., Wang, H., Paik, H.-Y., and Zhou, R. (eds.) Web Information Systems Engineering – WISE 2018. pp. 395–409. Springer International Publishing, Cham (2018)
29. Shi, L., Cristea, A.I., Awan, M.S., Stewart, C., Hendrix, M.: Towards Understanding Learning Behavior Patterns in Social Adaptive Personalized E-learning Systems. In: The 19th Americas Conference on Information Systems. pp. 1–10. Association for Information Systems, Chicago, Illinois, USA (2013)
30. MOOC completion rates, http://www.katyjordan.com/MOOCproject.html, Accessed: November 03, 2018