

# Hierarchical Cluster Analysis Heatmaps and Pattern Analysis: An Approach for Visualizing Learning Management System Interaction Data

Ji Eun Lee

Utah State University  
jieun.lee@aggiemail.usu.edu

Mimi Recker

Utah State University  
mimi.recker@usu.edu

Alex J. Bowers

Columbia University  
bowers@tc.columbia.edu

Min Yuan

University of Utah  
min.yuan@eccles.utah.edu

## ABSTRACT

This paper presents a form of visual data analytics to help examine and understand how patterns of student activity – automatically recorded as they interact with course materials while using a Learning Management System (LMS) – are related to their learning outcomes. In particular, we apply a data mining and pattern visualization methodology in which usage patterns are clustered using hierarchical cluster analysis (HCA) then visualized using heatmaps to produce what is called a clustergram. We illustrate the application of this methodology by building two clustergrams in order to explore university students' LMS activity patterns using both semester and weekly summary data. The resulting clustergrams reveal differences in LMS usage between high-achieving and low-achieving/dropout students.

## Keywords

Hierarchical Cluster Analysis, Heatmap, Learning Management System, Visual Data Analytics

## 1. INTRODUCTION

With the explosive growth in the use of LMS to support instructional activities, several recent studies have applied Educational Data Mining (EDM) to analyze the vast datasets collected by LMS. Results from such studies can help identify at-risk learners, monitor student performance, and inform course re-design [4, 5].

Some approaches tend to take a variable-centered approach, examining features and trends in key usage variables. In contrast, a person-centered approach can highlight individual sub-groups of students that share common data patterns [1, 7], that when pattern analyzed, link to important differences in overall course or educational outcomes. In this way, data points are not aggregated, thereby obscuring their individual patterns [6].

This study takes the latter, person-centered approach. As a form of visual data analytics, we describe and apply a data mining and pattern visualization methodology, in which usage patterns are clustered using hierarchical cluster analysis (HCA) then visualized using heatmaps to produce what is called a *clustergram* [1]. We illustrate the application of this methodology by analyzing data collected from a widely used LMS, Canvas. In particular, we address two questions: To what extent do clustergrams help understand patterns of student activity in the course? How do these patterns of activity relate to student learning outcomes?

## 2. BACKGROUND

### 2.1 EDM and LMS

Much prior EDM research applied to LMS data has typically taken a variable-centered approach by examining usage at an aggregated level [3]. While useful, these results aggregate and

average users' behaviors, and thus make it difficult to recognize the diverse patterns displayed by different groups of users [6]. Thus, in the present study, we take a more person-centered approach to visually investigate what sub-groups of students may share common patterns, and how these relate to their learning outcomes.

### 2.2 Hierarchical Cluster Analysis Heatmaps

HCA is a multivariate statistical method for classifying related units in an analysis across high dimensionality data. More recently, HCA has been combined with heatmap visualizations, called a *clustergram* [1]. The clustergrams represent each participant's row of data across each of the columns of variables as a color block, using stronger intensities of one color to represent lower levels of the variable, and increasing intensities of a different color to represent higher levels. We apply cluster analysis heatmap visualizations to Canvas LMS data from a large, online course. In this way, we test the utility of the analysis and visualization technique when applied to the potentially larger data patterning and visualization issues around these types of student interaction data.

## 3. METHODS AND DATA SOURCES

The data are drawn from a larger dataset containing all student recorded by the Canvas LMS at a medium-sized U.S. western university. For the present study, we extracted the student interaction data from a large (N=139) introductory level mathematics online course taught during the fall 2014 semester.

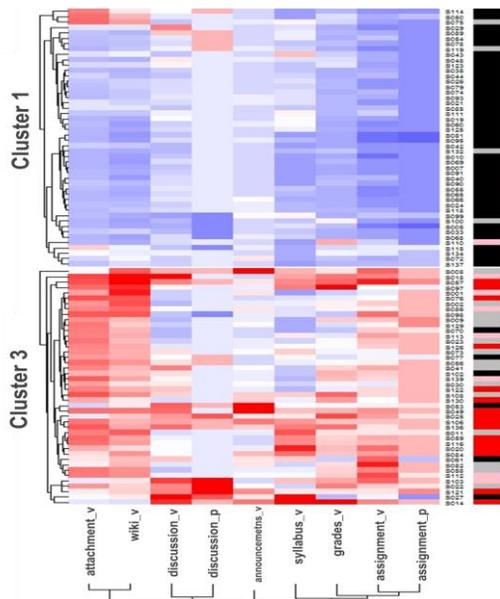
Two clustergrams were built, one with semester summary data and the other with weekly summary data. First, for the clustergram using the semester summary data, all student activity data were transformed to z-scores in order to standardize variance. HCA was applied to cluster both rows and columns. Color gradients ranges from colder blue for -3 SD below the mean to a hotter red for value +3 SD above the mean. Second, for the clustergrams using the weekly summary data, we used raw data and HCA was applied to only the rows. In addition, we applied k-means clustering on the rows for more precise interpretation of clustergrams. Lastly, student final course grade was included as an overall outcome variable in the final column. For all analyses, we used the R studio with the "ComplexHeatmap" packages. Regarding algorithms, the clustergrams were clustered using K-means, then HCA (using average linkage and Euclidean distance) was applied to each row-cluster.

## 4. RESULTS

### 4.1 Clustergram using semester summary

Figure 1 presents a section of the clustergram using the semester summary data. As shown in Figure 1, most students with lower

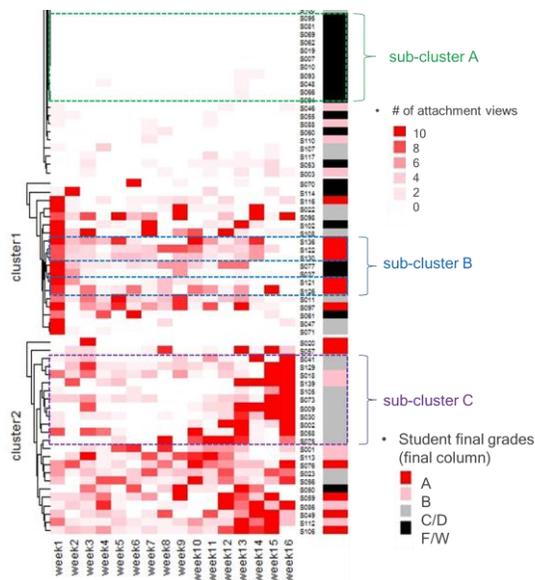
activity (Cluster 1) either received a grade of F or withdrew (W) from the course, whereas many students with higher activity (Cluster 3) received a grade of A.



**Figure 1. Section of the clustergram using the semester summary data (for full image: [goo.gl/Y7VFHJ](http://goo.gl/Y7VFHJ))**

A correlational analysis revealed that the variables related to engaging with ‘assignment’ features had the highest positive correlations with final grades ( $r = .70, p < .05$ ). Variables related to views of grades ( $r = .56, p < .05$ ), wiki ( $r = .42, p < .05$ ), syllabus ( $r = .35, p < .05$ ), and attachments ( $r = .35, p < .05$ ) had the next highest positive correlations with final grades. The remaining variables (views of announcements, participation in discussions) were not significantly correlated with final grades.

#### 4.2 Clustergram using weekly summary



**Figure 2. Section of the clustergram for the number of students' attachment views by week (for full image: [goo.gl/IwcCch](http://goo.gl/IwcCch))**

In order to investigate how student activities changed over the course of the semester, we built a clustergram using the weekly summary data. Figure 2 presents a section of the clustergram for the number of students' ‘attachment’ views by week.

The clustergram shows that the students with a grade of A (sub-cluster B) showed relatively consistent views of attachments over the course of the semester. Interestingly, the students with a grade of A (sub-cluster B) tended to show higher attachment views at the beginning of the course and more consistently throughout the semester. However, the students with grades of C/D (sub-cluster C) tended to have higher attachment views at the end of the course, representing perhaps a less-successful ‘cramming’ strategy.

### 5. CONCLUSION

This study demonstrates the utility of cluster analysis heatmap visualizations as a means to use visual data analytics to examine student patterns of activity at different grain sizes (week vs. semester). Combining this technique with the large sets of LMS provides a unique opportunity to examine the patterns of student activity as they relate to overall student outcomes. This type of visual data analytics expands the number of tools available for instructors and administrators to help identify the features and specific LMS interaction data that are most useful to their students. As recent critiques of LMS interaction data have shown that past analytic methods are insufficient to understand the rich complexity of how students learn through an LMS [2], this study provides an additional means to approach these complex data analytic issues.

### 6. REFERENCES

- [1] Bowers, A.J. 2010. Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. *Practical Assessment, Research & Evaluation*. 15, 7, 1-18.
- [2] Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., and Rosé, C. P. 2015. Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science*. 6, 4 (Jul. 2015), 333-353.
- [3] Macfadyen, L. P., and Dawson, S. 2010. Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers & Education*. 54, 2 (Feb.2010), 588-599.
- [4] Picciano, A. G. 2014. Big data and learning analytics in blended learning environments: Benefits and concerns. *International Journal of Artificial Intelligence and Interactive Multimedia*. 2, 7 (Sep. 2014), 35-43.
- [5] Romero, C., Ventura, S., and García, E. 2008. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*. 51, 1 (Aug. 2008), 368-384.
- [6] Wilkinson, L., and Friendly, M. 2012. The history of the cluster heat map. *The American Statistician*. 63, 2 (Jan, 2012), 179-184. DOI= <http://10.1198/tas.2009.0033>
- [7] Xu, B., and Recker, M. 2011. Understanding teacher users of a digital library service: A clustering approach. *Journal of Educational Data Mining*. 3, 1 (Oct. 2011), 1-28.