JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

Shapes of Educational Data in an Online Calculus Course

Olga Caprotti
Florida State University, USA
olga.caprotti@gmail.com

ABSTRACT. This paper describes investigations in visualizing logpaths of students in an online calculus course held at Florida State University in 2014. The clickstreams making up the logpaths can be used to visualize student progress in the information space of a course as a graph. We consider the graded activities as nodes of the graph, while information extracted from the logpaths between the graded activities label the edges of the graph. We show that this graph is associated to a Markov Chain in which the states are the graded activities and the weight of the edge is proportional to the probability of that transition. When we visualize such a graph, it becomes apparent that most students follow the course sequentially, section after section. This model allows us to study how different groups of students employ the learning resources using sequence analysis on information buried in their clickstreams.

Keywords: Online calculus, Markov chain, clickstream, sequence analysis

1 INTRODUCTION

The amount and breadth of data being collected on student learning is growing quickly in an effort to improve education and learning. In this work, we have investigated student activities as they happen in the virtual learning environment of the World Education Portals (WEPS[1]) with the ultimate goal of being able to build a recommendation system to help students successfully attain their learning goals. WEPS was initiated by Dr. Mika Seppälä to disseminate good practice and innovative learning technologies for the STEM subject areas, with a specific focus on mathematics. As observed in Seppälä (2014), the multidimensional nature of the educational data seems to require visualization approaches that can display complex information. As such, he proposed using a surface model to help navigate and visualize the complexity in a natural way and, like the graph representation, to give the overall picture. In the mathematical model for online courses envisioned by Seppälä (2013), students advance along a graph "in which the vertices are quizzes, workshops and examinations, and the edges correspond to essentially different ways of using the course resources."

In this paper, we show that this mathematical model is a Markov Chain by constructing it from educational data harvested from an online calculus course hosted at WEPS.

[1] Currently hosted at https://geom.mathstat.helsinki.fi/moodle/

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

WEPS is based on the Moodle system, one of the most popular open source learning management systems, approaching 80 million users worldwide. Moodle automatically logs data about students' online activities. Each Moodle installation holds the data about the courses, access details, learning sessions, grades, and clickstreams of each student. The database in the background holds information implicitly, namely data that can be computed from the explicit stored data. To obtain this derived data, it is necessary to study how the system stores the information about each student session, and to query the underlying database tables to retrieve it. This educational data mining in Moodle is widely researched, in particular Romero, Gutiérrez, Freire, and Ventura (2008) review classification methods applied to log data, and Casey, Dublin, and Gibson (2010) carried out studies of the log trails of students analyzing views, logins, and daily activities for a variety of courses.

In this paper, we use educational log data from students enrolled in courses in mathematics, here specifically a calculus course in which traditional teaching was combined with online student-centred learning and peer instruction. In particular, we focus on using visual learning analytics: how to visualize student actions in a course in a way that informs the instructors and guides the students to a better learning strategy for this online calculus class. The ultimate goal of this effort is to be able to recommend to students how to study and suggest which learning activities and resources among those available are more likely to maximize successful attainment of their immediate learning goal.

Here we have used data collected during an online Calculus II course held at Florida State University in 2014 by Mika Seppälä. This online course in mathematics is the result of many iterations over the last decade, documented, for instance, in Seppälä, Caprotti, and Xambó (2006), Caprotti, Seppälä, and Xambó (2007), Ojalainen and Pauna (2013), which resulted in the course structure and methodology adopted currently and described in Pauna (2017). The course comprises heterogeneous learning activities, with the intension to accommodate a variety of study strategies. Contrary to the strict sequential presentation of materials typical of the larger MOOCs, students could access all the study resources of the whole course apart from the graded activities made available at successive times.

We harvested the log files from 140 students enrolled in the online Calculus II course. From the original 133,570 lines, the log file, after curation, was 1,400 lines long and contained data records. Therefore, this course is more consistent with the definition of a small private online course (SPOC), in the sense of Fox (2013), rather than that of a Massively Open Online Course (MOOC), where these methods are more typically used. Hence, these data are far from being as big and broad as intended these days, both in terms of the number of individuals and in terms of heterogeneity of the characteristics we are able to study. However, it is large enough to make the analysis interesting without having to deal with size-induced hardware or software limitations.

The standard log data recorded during the student online sessions is composed of the time stamped clicks that show the student activities across the online course. The clickstream makes up the learning path followed by a student; however, we must keep in mind that this path walks along higher dimensions when we consider all variables still unknown that contribute to learning. The literature on

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

the subject is vast, and at the beginning of this project we decided to include measures of attitudinal and cognitive skills at the beginning of the course, similar to what was suggested by Niemi (2012a; 2012b), along with some demographics data. The results of the analysis of these psychological predictors correlated to the log data of the online calculus course and performance in the course are described in Hart, Daucourt, and Ganley (2017). To further our understanding of student learning in the course, a continuous task has been deciding and mining data to derive additional indicators for describing the student activities from the information implicitly stored across multiple tables in the underlining database. Such information included, for instance, the number of attempts and related scores for a given quiz, the time passed since a graded task was assigned, the grade for that specific activity, and the final grade.

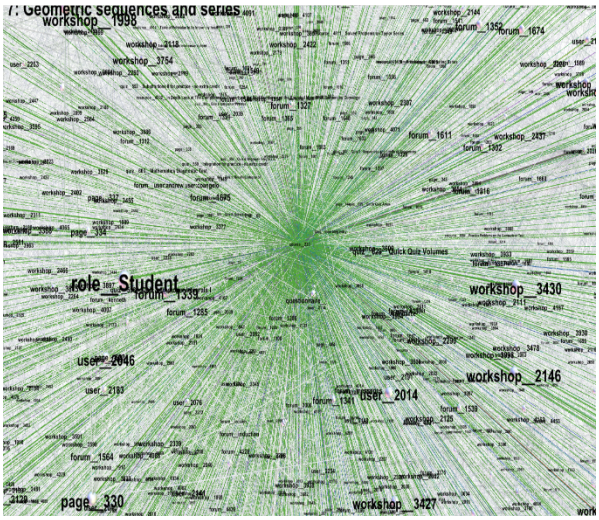


Figure 1: Full graph of the course log.

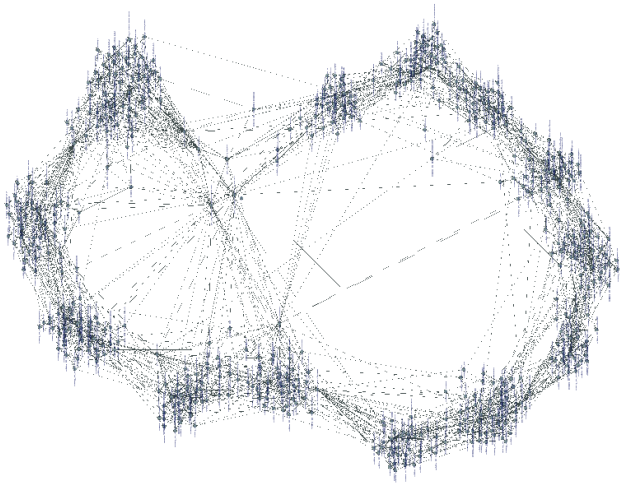


Figure 2: Reduced graph of the course log.

For this paper, we have purposely discarded data regarding clicks associated with activities like participation in the forum discussion in order to concentrate on aspects related only to the usage of the instructor-provided course resources: quizzes, peer-assessed workshops (i.e., homework problem sets), graded exams, and instructor-produced videos of course content. The initial complexity of the log data is visualized well in Figure 1: the graph is obtained by displaying every resource in the log file as a node and by adding an edge between resources that have been visited sequentially by at least one student in the course. This figure is almost uninterpretable. This is because most resources are made available by the system's graphical user interface from the course top page, the node for the course top page has a very high degree and centrality: students mainly access intended activities from the course top page. Therefore, ignoring this extra click, namely considering it as noise induced by the user interface, was the initial step in polishing the data. The resulting graph is already remarkably different: Figure 2 shows that several nodes with high degree appear. It became clear that student activities clustered around specific resources, namely the graded activities that contributed to the final course grade. This also confirmed

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

the intuitive course graph model by Seppälä (2013) that we will show how to formalize as a Markov Chain and further use for analysis of student log data.

The structure of the paper is as follows. Section 2 describes how we obtained the course Markov graph using a Markov Chain model of the course log data, while Section 3 describes how the course Markov graph defines logpaths and indices of usage of learning resources. Finally, in Section 4 we show how to derive a study recommendation system from the log data analysis.

2 THE COURSE MARKOV GRAPH

We start by noting that the clickstreams can be used to visualize student progress in the information space of a course as a graph. One way to do this is to consider graded activities as vertices of the graph, while different ways of using the instructional materials and other activities to prepare for graded activities label the edges of the graph. If a course uses formative assessments, then it is every instructor's natural interpretation that students proceed in their learning by focusing the studying in order to solve the most immediate task. In other words, it is assumed that students are working towards the most immediate graded task; for example, if an exam is coming up, then all student interaction with the course is related to their studying the content for that exam. Given this, abstracting away the granularity from the graph in Figure 2 is done by considering as nodes of the graph only the graded activities (and not every learning resource on which an action is logged) and labelling the edge by information derived from the clickstream that links the two graded activities, corresponding to two adjacent nodes. Basically, we encode the different ways to prepare for a graded activity as labelled edges of the course graph.

To give an idea of the kind of information available, a fragment of the log data is shown in Figure 3 where the Source and Target columns contains nodes of the graph (which represent previous and next graded activity); Time.Prev and Time.Next are the time stamps related to the traversal day for accessing the Source and Target respectively; and Label, the clickstream of learning activities leading from the Source to the Target, contains information that will be used to label the edges of the graph. An example of a clickstream is shown in the grey box in Figure 3 connecting the peer-assessed workshop in one part of the course to a different peer-assessed workshop. The user identification codes have been obfuscated, and it is enough to say that the log is ordered chronologically by User so that it is possible to read each student's progression between the "_START_" and the "_END_" nodes from the Target column. The student in line 50 completed the peer-assessed workshop in Section 13 after doing those right before, in Sections 11 and 12, however the student on line 59, skipped them both. Students, in fact, were not obligated to take part in the peer-assessed workshops, even if these activities contributed to a fraction of the final grade, which explains why some students did not follow a sequential progression in how they completed the graded portions of the class.

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

	User	id	Source	Target	Time.Prev	Time.Next	Label	Source.Section	Target.Section
49	XXXX	50	11-W-aa9352	12-W-aa9357	104	109	13AVwo9360-11AVwo9352-11AVwo9352-12AVwo9...	11	12
50	XXXX	51	12-W-aa9357	13-W-aa9360	109	116	12AVwo9357-12AVwo9357-12AVwo9357-12AVwo9...	12	13
51	XXXX	52	13-W-aa9360	_END_	116	116	7AVwo9338-13AVwo9360-13AVwo9360-7AVwo933...	13	14
52	XXXX	53	_START_	2-W-aa9308	2	19	0PVre4797-0PVre4799-1PVur1301-14PVpa357-2AV...	0	2
53	XXXX	54	2-W-aa9308	2-W-aa9315	19	35	2AVwo9308-2AVwo9308-2AVwo9308-2AVwo9308-...	2	2
54	XXXX	55	2-W-aa9315	6-W-aa9334	35	62	2AVwo9315-2AVwo9315-2AVwo9315-0PVre4797-0...	2	6
55	XXXX	56	6-W-aa9334	7-W-aa9338	62	80	6AVwo9334-6AVwo9334-7AVwo9338-7AVwo9338-...	6	7
56	XXXX	57	7-W-aa9338	8-W-aa9340	80	88	8AVwo9340-8AVwo	7	8
57	XXXX	58	8-W-aa9340	9-W-aa9343	88	92	8AVwo9340-8AVwo	8	9
58	XXXX	59	9-W-aa9343	10-W-aa9348	92	99	9AVwo9343-9AVwo	9	10
59	XXXX	60	10-W-aa9348	13-W-aa9360	99	119	10AVwo9348-10AVwo9348-10AVwo9348-10AVwo9...	10	13
60	XXXX	61	13-W-aa9360	_END_	119	119	13AVwo9360-13AVwo9360-14PVre5440-14PVre481...	13	14
61	XXXX	62	_START_	2-W-aa9308	8	17	1PVur1301-1PVpa330-2AVwo9308-2AVwo9314-1PV...	1	2
62	XXXX	63	2-W-aa9308	2-W-aa9314	18	25	2AVwo9308-2AVwo9308-2AVwo9308-2AVwo9308-...	2	2
63	XXXX	64	2-W-aa9314	2-W-aa9315	25	33	2AVwo9314-2AVwo9314-3AVau9318-3AVau9318-3...	2	2

6AVwo9334-6AVwo9334-7AVwo9338-7AVwo9338-7AVwo9338-7AVwo9338-8AVwo9340-7GAwo9338

Figure 3: Fragment of the log data.

Bearing in mind this interpretation of the log data, it becomes possible to visualize the clickstreams of the students as a Markov Chain in which the states are the graded activities (in this example we have only considered the peer-assessed workshops) and the thickness of the edge connecting the graded activities between states γ_i and γ_j is proportional to the probability of the transition from γ_i to γ_j. Markov chains have been a popular tool in Web path analysis since Sarukkai (2000). In particular, they have been used in the Moodle environment by Marques and Belo (2011) to carry out student profiling. In contrast to their work, we consider how students utilize specifically the resources of a course, interpreting their use of course resources as study strategies driven by the course graded assignments.

Figure 4 shows such a chain of order 1 in the actual sample course. From the probabilities of the transitions, listed in Figure 5, it is apparent from the values on the diagonal that the natural progression followed by most students corresponds to the sequential section-based structure of the course. Some students might skip three or four assignments but these are usually students with low attendance rates. In general, assignments close to the course exams (Sections 4 and 13) have a higher probability of being skipped. Interestingly, the non-zero probability of the trivial path from "_START_" to "_END_" indicates students who have not taken part in any of the formative assessment activities.

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

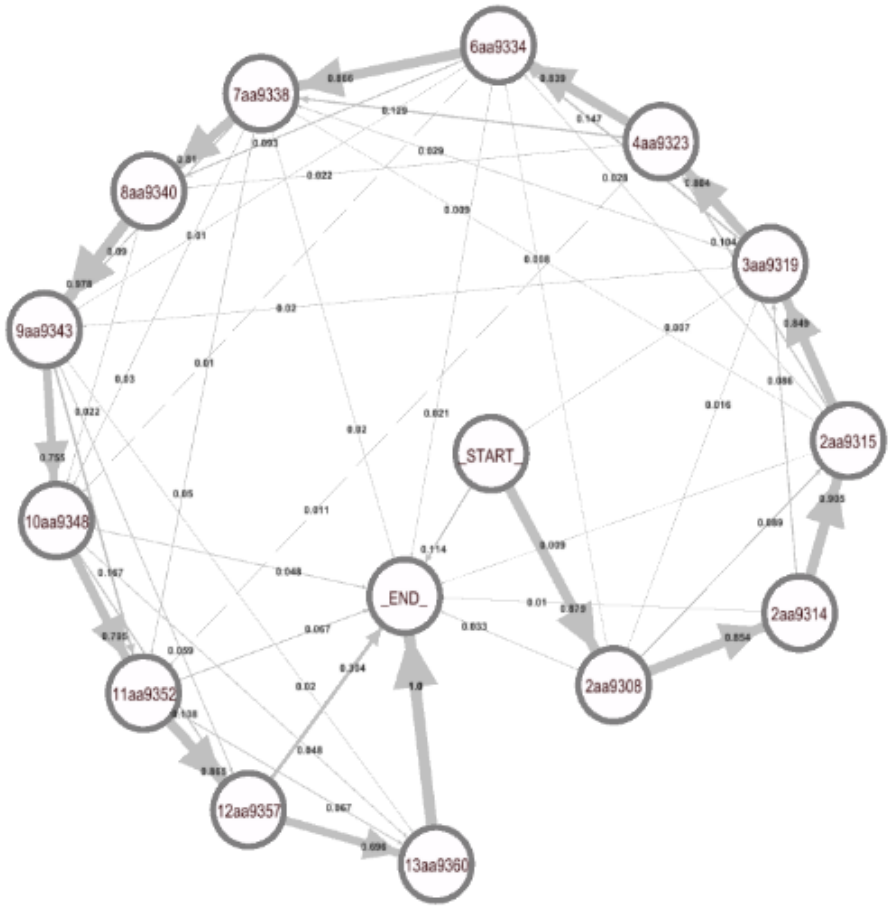


Figure 4: Markov graph of order 1 of the online course.

	2aa9308	2aa9314	2aa9315	3aa9319	4aa9323	6aa9334	7aa9338	8aa9340	9aa9343	10aa9348	11aa9352	12aa9357	13aa9360	_END_
START	0.879	0.000	0.000	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.114
2aa9308	0.000	0.854	0.089	0.016	0.000	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.033
2aa9314	0.000	0.000	0.905	0.086	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010
2aa9315	0.000	0.000	0.000	0.849	0.104	0.028	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.009
3aa9319	0.000	0.000	0.000	0.000	0.804	0.147	0.029	0.000	0.020	0.000	0.000	0.000	0.000	0.000
4aa9323	0.000	0.000	0.000	0.000	0.000	0.839	0.129	0.022	0.000	0.000	0.011	0.000	0.000	0.000
6aa9334	0.000	0.000	0.000	0.000	0.000	0.000	0.866	0.093	0.010	0.010	0.000	0.000	0.000	0.021
7aa9338	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.810	0.090	0.030	0.050	0.000	0.000	0.020
8aa9340	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.978	0.022	0.000	0.000	0.000	0.000
9aa9343	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.755	0.167	0.059	0.020	0.000
10aa9348	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.795	0.108	0.048	0.048
11aa9352	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.865	0.067	0.067
12aa9357	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.696	0.304
13aa9360	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000

Figure 5: Transition matrix.

For the computation of the Markov chain we processed the log data using the system R by the R Core Team (2015), with the library TraMineR by Gabadinho, Ritschard, Müller, and Studer (2011), to extract student sequences of the target nodes from the dataframe, `user_tprev_tnext_target.230` (230

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

was the course number), consisting of the columns User, Time.Prev, Time.Next, and Target from the log data:

```
> usertarget230.seq <-
seqdef(user_tprev_tnext_target.230,var=c("User","Tprev","Tnext","Target"),
informat="SPELL",states=seqstatl(user_tprev_tnext_target.230,var=4),process=FALSE,
left="DEL")
> usertarget230.dss <- seqdss(usertarget230.seq)
```

With minor editing of the discrete state sequence, we produced the input data to the R library clickstream by Scholz (2005) that finally computed the Markov graph of order 1 and the transition matrix. The course Markov graph, and the log data in Figures 1 and 2, were produced by the software Gephi by Bastian, Heymann, and Jacomy (2009).

The visualization shown in Figure 4 includes data related to every student, unfiltered. It is also possible to study how the course graph changes by filtering data based on specific characteristics of students, as will be shown later in Figure 8.

This model of the log data of the course naturally leads to investigating study strategies of students in relation to how they completed the sequence of assignments. To do that, we profiled students by creating a measure of "diligence," which represents the number of assignments they completed, as well as a measure of how many assignments they skipped in a row. The first question to ask is whether the students who followed the sequential path of the course scored higher than those who did not. The correlation between final grade and student diligence is .645. This is not higher because oftentimes the better students did not complete some of the assignments, because these contributed to only a small percentage of the final grade. Further inspection of the average final grade versus diligence in Table 1 indicates that generally diligent students (diligence greater or equal to 9) score higher (if we disregard the *one* good student with diligence 6). This unsurprising observation aligns well with the relevance of the "Daily Course Views" indicator reported, for instance, by Casey et al. (2010).

Table 1: Final Grade Versus Diligence

Diligence	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Final grade avg.	4.25	7	41.5	37.5	56	93	69	55	76.7	83.2	78.6	81.7	80.4	88.8

In the remaining sections of the paper, we study how students employ the learning resources. More specifically, we look into how much students rely on earlier learning materials when completing an assignment, a so-called "look-back," which we interpret as a possible indicator of poor (initial) learning, possibly due to lack of "diligence."

3 LOGPATHS AND LOOK-BACK

"Looking back," as defined by Polya (1973), is the reflective step in the mathematical problem solving process in which the solution is examined. We wanted to study how much students look-back at learning resources studied earlier while preparing the online assignments. In the present course,

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

students had to work out solutions of workshops assignments and also grade the workshop assignments of their peers by examining and evaluating the solutions presented by their peers in a critical problem solving process. During this task, looking-back to learning resources studied in earlier sections of the course occurred. Lee (2012) investigated looking-back in relation to student performance. Lee examined, as look-back indicator cues, the verbs "forgot," "remember," and "repeat," from transcripts of eighth grade students. Although we do not analyze any transcripts here (e.g., the course forum interactions), we are able to investigate look-backs by inspecting the clickstreams with respect to the course sections in the following way.

Assume that every graded activity γ_i belongs to a section of a WEPS course, denoted by $S(\gamma_i)$. This is usually true. Moreover, note that different courses have different configurations of graded activities. The online course under study follows a sequential schedule: students are usually expected to hand in graded activities section after section. Hence, we assume $S(\gamma_i) \leq S(\gamma_j)$ if the graded activity γ_i occurs before γ_j, that also implies (because it is induced by the order of creation of the resources) that the indices for the activities are ordered, so that $i \leq j$. There is usually one graded activity per section, but that is not an assumption. In the specific course, Section 2 had three workshop assignments, and Section 5 did not have any because of the midterm exam. Moreover, if we were to also study quizzes as graded activities, then most sections would have more than one such activity. Given a student A in an online course with graded activities Γ, we call $p_A = [\gamma_i, \gamma_{i+1} \dots, \gamma_k]$, $S(\gamma_i) \leq S(\gamma_j)$, $i \leq j \leq k$, $\gamma_j \in \Gamma$ the *logpath* of A in Γ. Namely, we order the graded activities completed by a student by section and by id; for example, the logpath of the student whose log data is recorded in rows 52–59 is the sequence of activities listed in the Target column in Figure 3. Furthermore, assume sections do not share resources, so that all resources L in a course with t sections can be partitioned by section: $L = \bigcup_{0 \leq i \leq t} L_i$ where L_i are the resources in section i, with L_0 denoting resources at course top level. In the log data in Figure 3, the value appearing in the Label column is the dash-separated concatenation of the names of resources, each prefixed by the section number it belongs to.

We then can also talk of L_1^n resources as the set of resources belonging to the union of the sections 1 to n. Let P_A be a logpath with a graded activity γ_h. We define the hop with target γ_h as the sequence of (actions on) learning resources recorded in the log between the source, γ_{h-1}, of γ_h and γ_h:

$$hop(\gamma_h) = [\lambda_1, \dots, \lambda_e] \subset 2^L, \gamma_h \in P_A.$$

This corresponds to the label of the edge going into $\gamma_h$in the course graph corresponding to the logpath P_A. In the hop towards γ_h, its look-back degree is defined as the number of (clicks of) learning resources belonging to sections below h in the hop towards γ_h:

$$lbd(\gamma_h) = \left| hop(\gamma_h) \cap L_0^{h-1} \right|.$$

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

Similarly, the look-ahead degree is the number of learning resources belonging to sections above the target section h:

$$lad(\gamma_h) = |hop(\gamma_h) \cap L^t_{h+1}|.$$

Finally, the in-section degree is the number of learning resources belonging to the target section h:

$$isd(\gamma_h) = |hop(\gamma_h) \cap L_h|.$$

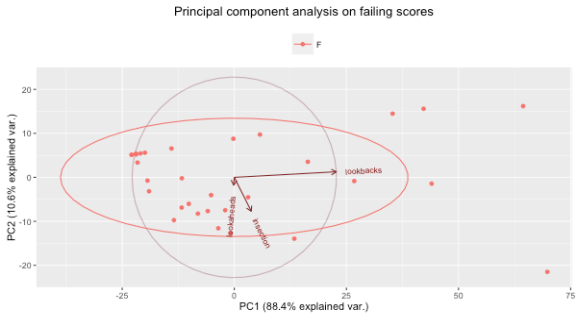
With respect to failing and passing students, principal component analysis of these three indices (look-back degree, look-ahead degree, in-section degree), summed for each hop, for each student and then averaged over the diligence, returned the information in Figure 6.

```
Standard deviations for Fail:
[1] 25.7  8.9  2.7
Rotation:
               PC1   PC2    PC3
look-backs   0.9864  0.16 -0.042
look-aheads -0.0062 -0.22 -0.975
in-section   0.1640 -0.96  0.217

Importance of components:
                          PC1   PC2     PC3
Standard deviation     25.677 8.912 2.66811
Proportion of Variance  0.884 0.106 0.00954
Cumulative Proportion   0.884 0.990 1.00000
```

```
Standard deviations for Pass:
[1] 7.2 3.8 2.5
Rotation:
               PC1   PC2   PC3
look-backs  -0.73  0.67  0.11
look-aheads -0.30 -0.47  0.83
in-section  -0.61 -0.58 -0.54

Importance of components:
                          PC1   PC2    PC3
Standard deviation     7.203 3.845 2.4795
Proportion of Variance 0.713 0.203 0.0844
Cumulative Proportion  0.713 0.916 1.0000
```



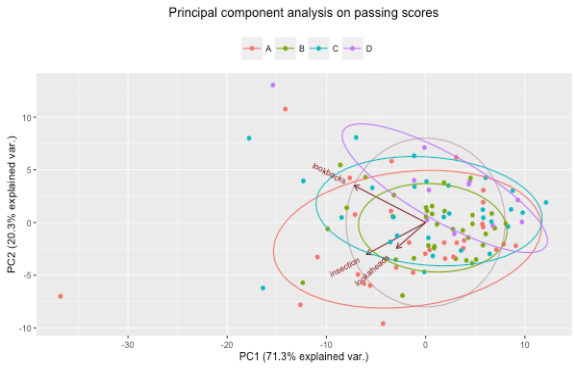


Figure 6: Principal component analysis of Fail and Pass course grades.

In particular, these indices seem to be good indicators for students at the risk of failing. In both cases, fail or pass, look-backs play a bigger role than in-section and look-ahead.

We also carried out detailed analyses of the hops of student cohorts, defined according to criteria related to diligence and to final grades using sequence analysis to try to identify successful studying patterns in accessing the resources.

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

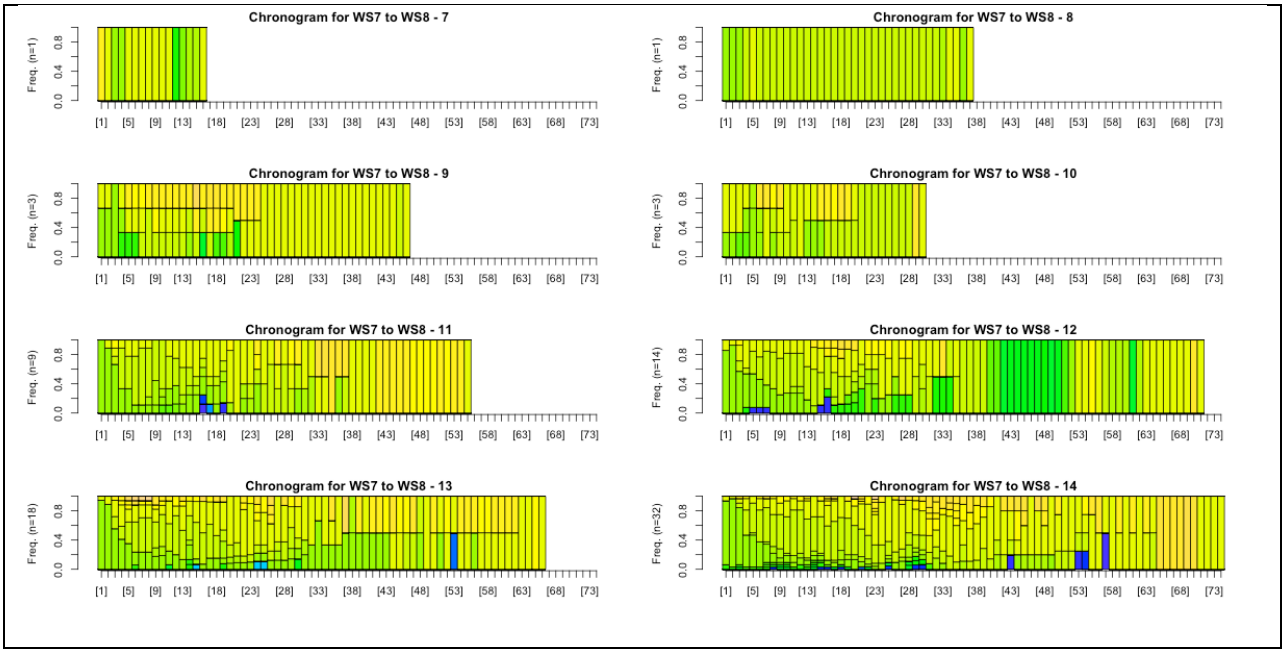


Figure 7: Sequence chronograms of hops between WS7–WS8 by diligence.

Figure 7 shows sequence chronograms for hops between Workshops 7 and 8, filtered by student cohorts based on diligence. Because we are interested in how resources are used with respect to the section, the shades assigned to each resource depend on which index they contribute to: blue represents look-back, green represents in-section, and yellow represents look-ahead. It is apparent that students with higher diligence seem to include more look-ahead resources, possibly an indication of being able to work in parallel on future assignments while preparing for the assignment due. It is also clear how the visualization of the entire graph changes if we try to portray information related to the details of the logpaths.

For example, we show how the look-back degree index can be made apparent in the graph visualization. Seppälä suggested that a correlation distance between two graded activities γ_i and γ_j in the space of graded activities is given by

$$d(\gamma_i, \gamma_j) = \log \frac{1}{\mathrm{Corr}(\gamma_i, \gamma_j)}$$

where $\mathrm{Corr}(\gamma_i, \gamma_j)$ is a measure of correlation between γ_i and γ_j that depends on which aspect of the learning has to be modelled.

Our candidate for this correlation is the average look-back degree: $\mathrm{Corr}(\gamma_{i-1}, \gamma_i) = \mathrm{mean}(lbd_{P_x}(\gamma_i))$ where the mean is taken over all students passing between γ_i and its source. For the visualization, we think of this correlation as related to the weight of the edge joining γ_i and its source, so that the circular layout of Markov course graph will not be affected but the edges will be thicker when the average look-back degree is larger for the paths of the students being visualized. Note that the visualization depends

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

on the cohort of students being considered: the edges grow heavier or thinner and could disappear. The graphs in Figure 8 visualize the difference between students with diligence below 9 (larger correlation in terms of look-back degree) and the rest of the students with higher diligence. Also noticeable at first glance is how less diligent students have dropped out (edges going into the node _END_) much earlier.

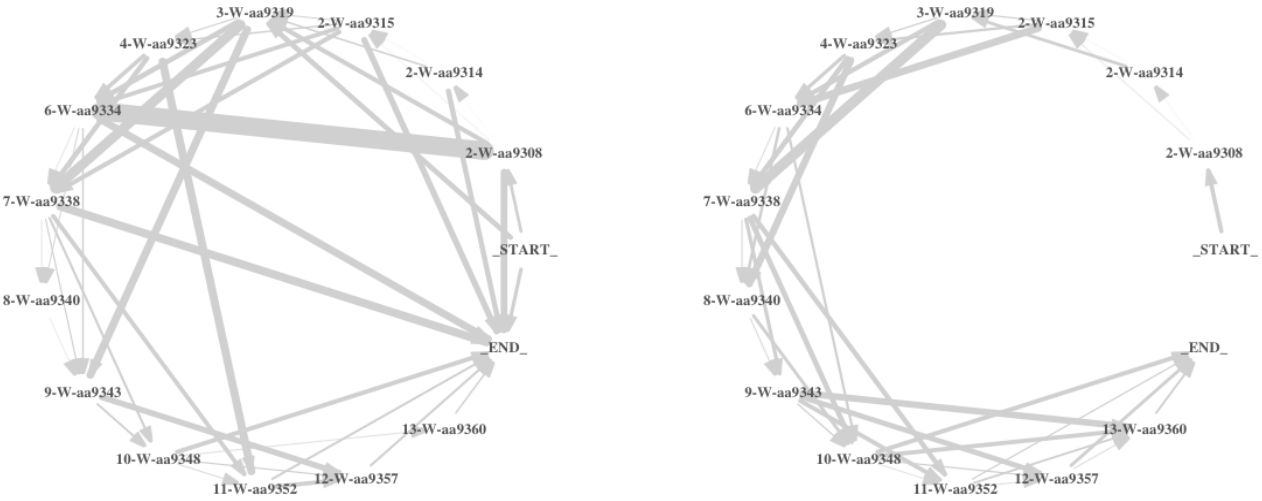


Figure 8: Course graphs of students with diligence <9 and >= 9, weights proportional to the mean of look-back degrees.

Seppälä envisioned that Riemann surfaces could be a very useful exploratory tool for carrying out visual learning analytics tasks on complex, multidimensional, educational data, in a way addressing the issues already pointed out by Hadwin, Nesbit, Jamieson-Noel, Code, and Winne (2007). They suggested a way to construct a surface from a mathematical model of an online course as a graph that we have shown can be formalized as a Markov Chain induced by the graded activities in the logpaths. This interpretation of the log data allowed for the definition of indices that furthered our understanding of how students utilize the online resources. The next section will show how these insights can be used to guide students' study paths.

4 A STUDY RECOMMENDATION SYSTEM FOR ONLINE CALCULUS

The ultimate purpose of analysing the way students complete the online course is to be able to suggest how to best proceed through the learning resources with suggestions for a study strategy. Based on the data collected and clustered according to all graded assignments (quizzes, workshops submissions, and assessments), we are able to construct such a recommendation system as a course browser, inspired by the concept network browser,[2] indicating the learning resources used in the course by students targeting a certain activity. Figure 9 shows an example where the learning resources listed in the middle column are highlighted if they have been used by past students to tackle the workshop assignment on

[2] http://www.findtheconversation.com/concept-map

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

Improper Integrals. Clicking a graded activity will show all related resources with a measure of how much they have been used according to the log data. Furthermore, it is possible to personalize such a system based on additional student profiling data; for instance, just taking into account the strategies employed by top scoring, by most diligent, or by highly motivated students. Based on a few of these investigations for several different courses, we observed that diligent students consult fewer resources and carry out a more focused study activity, thus resulting in a more targeted set of recommendations. However, the resulting recommendation system is not necessarily the best option for the generic student and the personalization must be done after careful analysis of several instances of the same course. Towards this end then, the diligence of a student is an example of a real-time (in terms of skipped assignments) classification of students useful in filtering data collected in past instances of the course for building a personalized course browser.

Ideally, learning resources and activities would be associated more generally with learning goals, which now are only implicitly defined by belonging to specific sections of the course. Using learning goals would add a layer of freedom to the course designer who could replace/change the resources within each goal while keeping the learning goals unaltered. This would allow one to conduct the analysis and construct the recommendation system independently of the instantiation of resources also in terms of version, type of media, or even language. At the time of this investigation, there was only limited support in Moodle for assigning learning goals metadata (as defined, for instance, in the Common Core) to activities and resources.

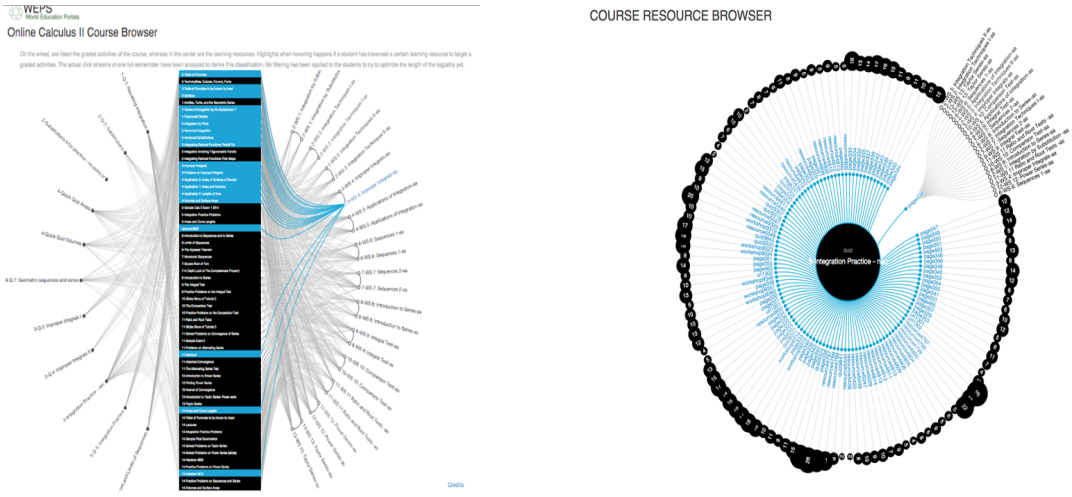


Figure 9: Course browser based on log data.

5 FINAL REMARKS AND FUTURE WORK

We have presented how the intuitive interpretation of the progression of student work in an online calculus course can be formally interpreted as stepping through the graph associated to the Markov Chain induced by the graded activities in the course. This, in turn, gives rise to several possible ways to

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

analyze the study strategies of the students and derive measures that allow us to construct a recommendation engine able to suggest the best learning resources in real-time to target completion of a certain learning task.

The online classroom scenario in which the data was collected was not an a-priori controlled experimental environment but relied on a course that had evolved over several years, structured as a progression of topics organized in sections. While the course activities were kept roughly uniform in each section, their number and composition varied. In particular, some of the final sections did not contain any quizzes. As it turned out, all sections contained one workshop graded activity, except for the first section which had three. Study resources were available to students during the entire course time, whereas graded activities were opened at subsequent times and had to be completed by due dates. More relevant, however, to the goal of deducing study strategies from clickstreams, is the fact that in some weeks, multiple graded activities were open and students could have studied for multiple assignments, in several upcoming sections. While it might be true that study strategies are driven by the most urgent assignment, it is also possible that farther away from the deadline, behaviour is more exploratory. From the point of view of the experienced instructor, this could be a way to 1) motivate students, by advancing to concepts that lie ahead in the course, and 2) train students in thinking along multiple pathways, even if unconsciously. For these reasons, such a course setup cannot be disregarded, but is not an ideal setup for a controlled study.

To improve our model, we need to deal with the issue of *endogeneity*, which we have yet to define exactly in our specific case. For example, in some sections the log data did not record any video being watched or resource being looked at because another activity overshadowed it. In fact, Moodle allows the instructor to arrange resources arbitrarily, and to group them under a general "Page," and this negatively impacted the uniformity of the data collected. Even if every section contained the same kinds of learning resources, these were presented differently by the graphical user interface. The lack of a uniform structure for every course section had the drawback of imposing a cognitive load on students, who had to learn to navigate a different interface in every section; this consequently hindered the possibility of carrying out an unbiased learner profile analysis. The issue of collecting data from resources hosted on third-party servers (e.g., YouTube) is also crucial in obtaining a complete picture of learner online activity. While standards exist to support re-usage of open learning resources, and we successfully experimented with the Tin Can[3] plugin for Moodle, it is still very disrupting to re-design and re-package the whole course. One of the impacts of this research, however, is the insights that will guide design and structuring of the WEPS online courses in the future.

Moreover, we are aware of the fact that we certainly are looking at a very small data set because the broad data landscape influencing learning is extremely varied, ranging from societal background to infrastructure, from well-being to health related conditions, all data which we have not been able to collect so far but might become available in the future.

[3] http://tincanapi.com and http://scorm.com/

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

To tackle the landscape of educational data for online calculus we chose the strategy of understanding smaller portions that contribute to the bigger picture. Because our teaching is online, it makes sense to start by understanding the shape and the geometry of the log data collected by our own online course. This in turn will inform our own future work, that of designers of online learning environments on which actions to track, and that of students on how best to organize their study activity.

ACKNOWLEDGEMENTS

This research was supported by the Shape of Educational Data grant (NSF DRL 1450501) and carried out as a visiting scholar to Florida State University. Much of the work described here was done as a contribution to the research of Yahya Almalki who was attempting to model the logpaths on a Riemann surface. I hope he eventually is able to finish what he started. The mining of educational data has improved because of several discussions and requests stemming from the psychological point of view offered by Colleen Ganley and Sara Hart. To them I also owe the opportunity to carry out these investigations, originally planned by Seppälä, who passed away before we could begin. The colleagues and staff at Florida State University were precious in easing my stay. Finally, I would like to thank the anonymous reviewers for their insightful comments.

REFERENCES

Bastian, M., Heymann, S., & Jacomy M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media* (ICWSM '09) 17–20 May 2009, San Jose, California, USA. Palo Alto, CA: AAAI Press. Retrieved from https://gephi.org/publications/gephi-bastian-feb09.pdf

Caprotti, O., Seppälä, M., & Xambó, S. (2007). Novel aspects of the use of ICT in mathematics education. In M. Iskander (Ed.), *Innovations in e-learning, instruction technology, assessment, and engineering education* (pp. 295–299). Springer: Netherlands. http://dx.doi.org/10.1007/978-1-4020-6262-9_51

Casey, K., Dublin, G. C., & Gibson, P. (2010). (m)Oodles of data mining Moodle to understand student behaviour. International Conference on Engaging Pedagogy (ICEP10), 2 September 2010, Maynooth, Ireland. Retrieved from http://www-public.tem-tsp.eu/~gibson/Research/Publications/E-Copies/CaseyGibson10.pdf

Fox, A. (2013). From MOOCs to SPOCs. *Communications of the ACM, 56*(12), 38–40. http://dx.doi.org/10.1145/2535918

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011), Analyzing and visualizing state sequences in R with TraMineR, *Journal of Statistical Software*, *40*(4), 1–37. http://dx.doi.org/10.18637/jss.v040.i04

Hart, S. A., Daucourt, M., & Ganley, C. M. (2017). Individual differences related to college students' course performance in calculus II. *Journal of Learning Analytics, 4*(2), 129–153. http://dx.doi.org/10.18608/jla.2017.42.11

(2017). Shapes of educational data in an online calculus course. *Journal of Learning Analytics*, *4*(2), 76–90.
http://dx.doi.org/10.18608/jla.2017.42.8

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, *2*(2–3), 107–124. http://dx.doi.org/10.1007/s11409-007-9016-7

Lee, Shin-Yi. (2012) Analysis of "look-back" strategies in mathematical problem solving. *Pre-Proceedings of the 12th International Congress on Mathematical Education: Intellectual and attitudinal challenges* (ICME-12), 8–15 July 2012, Seoul, Korea (pp. 4519–4526).

Marques, A., & Belo, O. (2011) Discovering student web usage profiles using Markov chains. *The Electronic Journal of e-Learning, 9*(1), 63–74. Retrieved from http://www.ejel.org/

Niemi, H. (2012a). How to find motivation: Learning math is taking a new course through self-efficacy. Personal communication.

Niemi, H. (2012b). Self-efficacy and peer support in web based learning. *Assessment and Effective Teaching of Calculus.* University of Helsinki, October 11–13, 2012. NSF–SAVI Collaboration meeting.

Ojalainen, J., & Pauna, M. (2013). Web-based mathematics exercises and their effect on students' achievement and confidence. In R. McBride & M. Searson (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference (SITE 2013), New Orleans, LA, USA* (pp. 2335–2340). Chesapeake, VA: Association for the Advancement of Computing in Education (AACE). Retrieved from https://www.learntechlib.org/p/48453.

Pauna, M. (2017). Calculus courses' assessment data. *Journal of Learning Analytics, 4*(2), 12–21. http://dx.doi.org/10.18608/jla.2017.42.3

Polya, G. (1973). *How to solve it*. Princeton, NJ: Princeton University Press.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Romero, C., Gutiérrez, S., Freire, M., & Ventura, S. (2008). Mining and visualizing visited trails in web-based educational systems. In R. S. J. d. Baker et al. (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (EDM'08), 20–21 June 2008, Montreal, QC, Canada (pp. 182–186). International Educational Data Mining Society.

Sarukkai, R. R. (2000). Link prediction and path analysis using Markov chains. *Computer Networks*, 33, (1–6), 377–386. http://dx.doi.org/10.1016/S1389-1286(00)00044-X

Scholz, M. (2005). Clickstream: An R package for analyzing clickstreams. Retrieved from https://www.jstatsoft.org/article/view/v074i04/v74i04.pdf

Seppälä, M. (2013). Learning analytics, Riemann surfaces, and quadratic differentials. *Riemann and Klein Surfaces, Symmetries and Moduli Spaces*. Linköping, Sweden, 24–28 June 2013. Retrieved from http://youtu.be/qbS_Cum07xg.

Seppälä, M. (2014). Shape of educational data. NSF 1450501 Standard Grant.

Seppälä, M., Caprotti, O., & Xambó, S. (2006). Using web technologies to teach mathematics. In C. Crawford, R. Carlsen, K. McFerrin, J. Price, R. Weber, & D. Willis (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference (SITE 2006), Orlando FL, USA* (pp. 2679–2684). Chesapeake, VA: Association for the Advancement of Computing in Education (AACE).