Visualizing Learner Behaviour in MOOCs using Sankey Diagrams

Karsten Lundqvist
Victoria University of Wellington, New Zealand

Michael Godinez
Victoria University of Wellington, New Zealand

Steven Warburton
Victoria University of Wellington, New Zealand

It can be difficult to assess the design of, and learning, within Massive Open Online Courses (MOOCs). It is especially hard when trying to analyse this at the level of the individual learner. This study has developed a tool, inspired by Sankey diagrams, to visualise learners' behaviour and paths through MOOC content. This tool can be used to investigate if learners are interacting with the content as planned when the course was designed. It has been designed iteratively through four stages of rapid prototyping. This paper presents the narrative of the development of the tool with an emphasis on validation via feedback from three user groups at each prototype stage.

Introduction

Massive Open Online Courses (MOOCs) arguably have been one of the most significant disruptive innovations within education in recent years. These free to study courses, offered largely by universities, have attracted millions of learners to a growing catalogue of subject areas. In 2016, more than 6850 MOOCs ran with over 58 million learners (Shah 2016). They are often broadly categorised as either cMOOCs or xMOOCs where cMOOCs utilise a connectivist learning approach (Milligan, Littlejohn, and Margaryan 2013), and xMOOCs, which are commonly more didactic in nature, comprise a mix of video material, textual content and assessments using behaviourist approaches (Daniel 2012).

The success of MOOCs builds on the active engagement of massive numbers of learners (McAuley et al. 2010) who through engagement with other participants self-organise into learning communities where they share skills, objectives, knowledge, and interests, most often by commenting within the MOOC discussion fora and other social networking tools (McAuley et al. 2010). Downes suggests that when looking at the success factors of a MOOC, one should investigate why the course was made the way it was, and if the design has successfully achieved those aims. This should preferably be done at the individual participants level because each person has a different objective or motivation for taking a course and has different needs and objectives (Downes 2015). The analysis of individual user experiences is an important aspect of course evaluation but difficult to achieve when there are thousands of participants (Shi et al. 2014). High learner numbers make it virtually impossible to follow individual progress through material and gain a clear understanding of learner behaviour within the course. This is especially problematic within the more structured xMOOCs where it becomes difficult for the educationalists and learning designers to get an overview of the effectiveness of the structure, and indeed where certain parts of the MOOC might need to be edited to be more effective.

In this paper we present a tool developed at anonymous to visualize user behaviour within courses hosted on the edX MOOC platform. It uses the log files made available to edX partners and creates a specialised Sankey diagram. This has been found to create a useful overview for learning designers working within the MOOC team at anonymous.

The next section is a literature review of visualizing learner behaviour in MOOCs. Following this is the method section which leads into the implementation of the visualizer tool. This section is a narrative of four iterations of prototypes. It includes feedback from a group of users which is used as validation of changes for the following prototype.

Visualizing learner behaviour in MOOCs

Earlier studies that have investigated MOOCs from the perspective of individual learners have mostly used surveys or interviews around the user experience, participant demographics, and metrics of learner progression through course e.g. number of videos viewed or tests taken (Kop and Fournier 2010; Kop 2011; Kop, Fournier, and Mak 2011; Levy 2011; B. Stewart 2010; Breslow et al. 2013). Over time, as the number of MOOCs has increased, participation size and completion rate have become popular metrics to show the relative success of individual MOOCs, or to measure learner satisfaction (Adamopoulos 2013; Jordan 2013; Khalil and Ebner

2014; Jordan 2015; Onah, Sinclair, and Boyatt 2014; Stephens-Martinez, Hearst, and Fox 2014). These metrics are relatively easy to calculate and mirror common metrics already used by universities when evaluating formal education courses. The findings for these studies are mostly visualised using standard aids such as bar and pie charts, box plots and scatter diagrams. This represents a common approach to overcoming the complexity of MOOCs whereby participant behaviour is compressed into simple metrics, or individual experiences are taken as an overall reflection of the course experience. Although these tools and approaches can provide a useful basis for comparing MOOCs they inevitably hide the complexity of behaviour. However, there may be valuable information that becomes hidden during this process that would be helpful, for example, in rectifying problems with the learning design.

Process mining is a technique where event log data is analysed to create a model that can be used to analyse business processes. The model can be created using any type of data mining technique, however it commonly results in a visualisation to help further the understanding of a particular process (W. M. van der Aalst 2011). This technique has been applied in educational settings, for example, a combination of flow charts and process cubes have been used to analyse the video lectures in a business information systems course at Eindhoven University of Technology (W. M. P. van der Aalst, Guo, and Gorissen 2015). Process cubes have been designed to investigate multi-dimensional data, however there are challenges when using them for comparing and visualizing different types of cells (W. M. van der Aalst 2013).

A Sankey-like diagram was first used by Charles Joseph Minard in 1869 to visualise Napoleon's Russian campaign of 1812 (Friendly 2002). They are named after Captain H. R. Sankey who is accredited as using it first in an academic publication, where he used it to illustrate flows within a turbine (Schmidt 2008). Sankey diagrams visualise flows from one state to another by using the width of the arrows to indicate the quantity of flow within the system. They have been used in education for overviews of video consumption within MOOCs. Here, a specialized type of Sankey state transition diagram was used to illustrate the number of users who viewed each part of the videos hosted on the MOOC (Coffrin et al. 2014). Students who had watched all of the videos in succession could be identified and the course could be analysed from the viewpoint of 'qualified' and 'non-qualified' students. Google Analytics includes standard Sankey diagrams and can be used to visualise transitions on websites (Emmons, Light, and Börner 2017; Beaven, Codreanu, and Creuzé 2014; Kay et al. 2013). However, they are session based and therefore the same user will potentially show up many times in the same Sankey diagram with various starting points when the user uses the website many times (Analytics 2017). In MOOCs, participants are expected to access the course multiple times, potentially from many devices, and therefore this approach cannot be used to visualize the full interaction with the course by users.

Method

The application has been developed following a rapid prototyping paradigm (Connell and Shafer 1989). There have been four different prototypes that were used in the feedback sessions. Getting user feedback after each prototype is critical in driving a consistent improvement through the iterations. The edX reference group (ERG) at anonymous provided this feedback with suggestions and improvements for the following iteration. This group comprise learning designers, academics (some with prior MOOC experience), researchers in education and MOOCs and university administrators engaged in the MOOC production at anonymous. This provided access to representative of the four user types identified as key stakeholders for the visualizations, and were thus well placed to provide relevant feedback.
The anonymous edX MOOC was used throughout the development of the visualiser, and all of the Sankey diagram figures are from this course. The various prototypes used the logs available at the time of a particular feedback session, and therefore, some of the visualisations are not consistent over time.

Implementing the visualiser

This section describes four prototype iterations. Each iteration includes the feedback from the ERG, providing validation for the changes within the next iteration of the tool. This provides a coherent narrative for understanding the development and usefulness of the tool.

Initial prototype

The aim of this project was to create a visual tool that provides an overview of participant behaviour within MOOCs. It should use edX log files to provide the data in the first place, but it should be developed to be extendible. The decision was made to base the visualisation on the Sankey diagrams as they are an intuitive and well-understood.

The d3 library is a JavaScript library for producing dynamic, interactive data visualisations in a web browser, the visualization was made using and extending the d3 library's functionality for creating online Sankey diagrams (d3 2017). User data was extracted from database files and it was modified to the json data format needed by the d3 library (Figure 1) The logs were not used in order to speed up the development of the first prototype and the visualization (figure 2) was created with data embedded as an svg image.

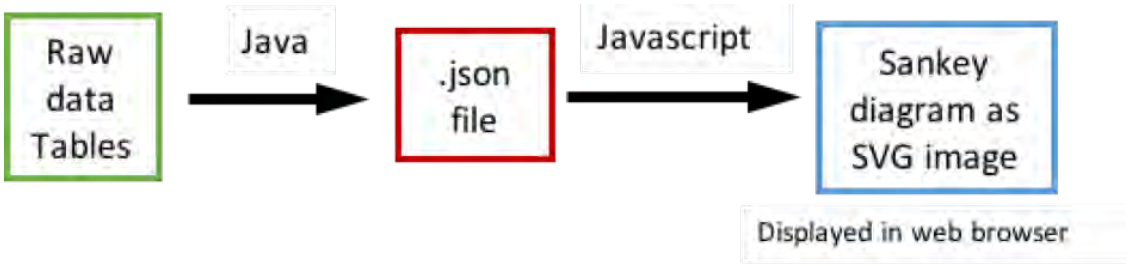

Figure 1: The data collection process for the initial prototype.

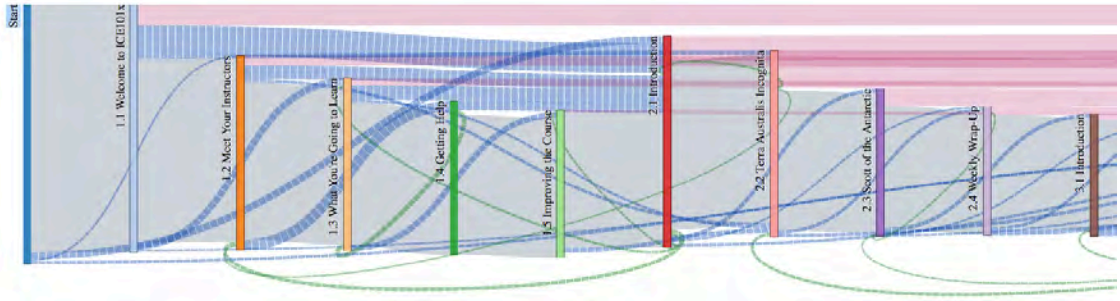

Figure 2: The first prototype showing the first two topics.

Each of the vertical bars are the course content pages laid out in the order provided on edX. The grey vertical paths indicate the participants who are moving from one step to the following step (the expected behaviour), blue paths are forward jumps (skipping the following page/s), green paths show backward steps, and red paths are the students who had their last activity at that given page. The visualization only shows paths with more than 20 students. When hovering over a bar or path it is becomes highlighted and shows the title and the exact number of participants following the path, or entering the associated content page.

In the feedback session with the ERG it was concluded that this approach was interesting and the group was able to use the view to make observations. Therefore, it was concluded that the approach of using Sankey diagrams was intuitive, in terms of presenting the course activity and several interesting participant behaviours were identified:

- It was clear when participants dropped out of the course. 15.1% of participants got to the last page i.e. end of the course. It is however impossible to see in this diagram how many participants in total completed the final quiz (which could also be defined as the end of the course), because the vertical bars included all visits to the associated content page. In other words, if a participant visited the quiz several times that participant would be counted numerous times.
- Introduction pages were skipped by many and therefore potentially essential information for participating in the course would not be seen. This suggested that learning designers should consider introducing important information as embedded in the regular course material.
- Most leavers departed early in the course. This highlighted that the designers should present reasons or hooks to combat attrition in the early sections of the course.
- Figure 3 shows that quizzes have a lot of unusual movement in and out, and reveals that many people skipped the quiz. It was difficult to understand the real impact of this behaviour, as the diagram did not show if this happened before, at, or after the first time seeing the quiz. The diagram was hiding key information about the user behaviour.
- The activity at the end of the course was much more linear than in the earlier sections (figure 4). The quizzes did produce some behavioural changes, but seemingly not to the same degree as the first quiz. Either the students had changed their behaviour or the visualization was hiding something.

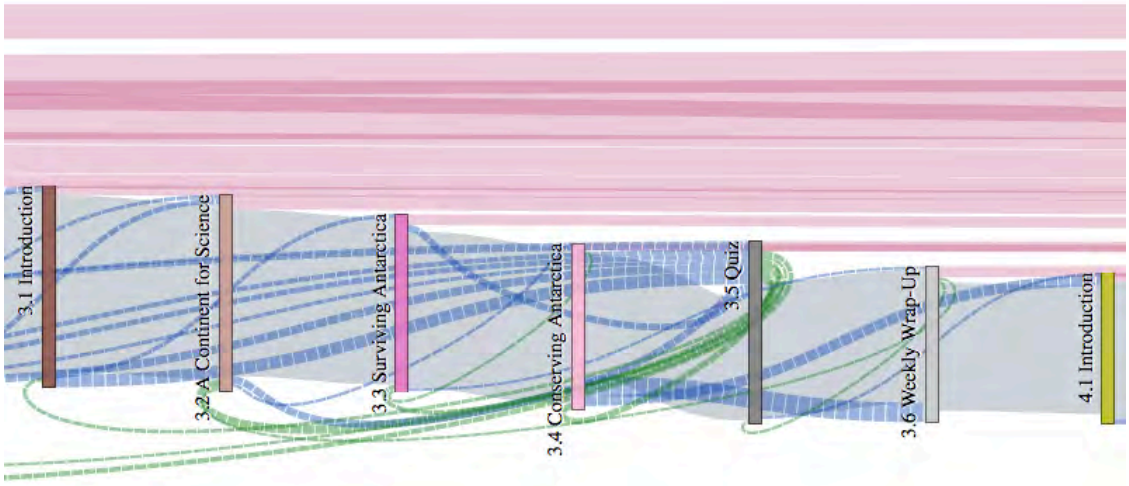

Figure 3: The first prototype showing the third topic.

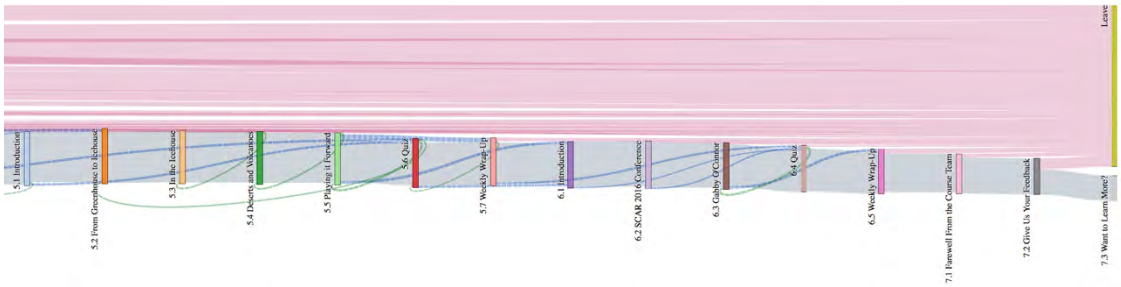

Figure 4: The first prototype showing the end of the course.

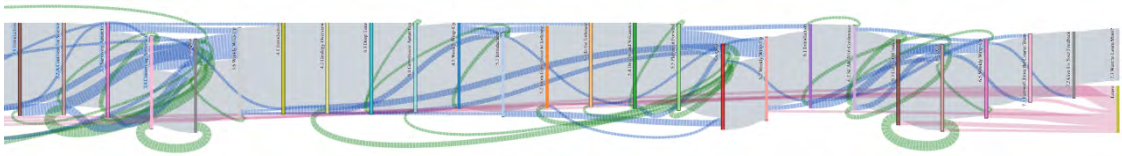

Figure 5: The first prototype showing the behaviour of paying participants.

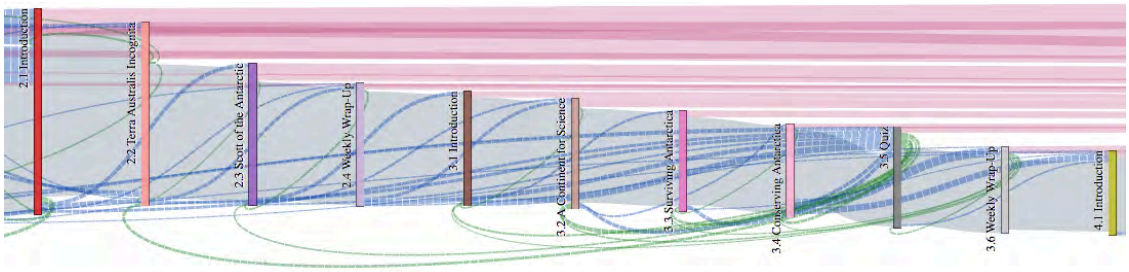

Figure 6: The first prototype showing the behaviour of non-paying participants around the first quiz.

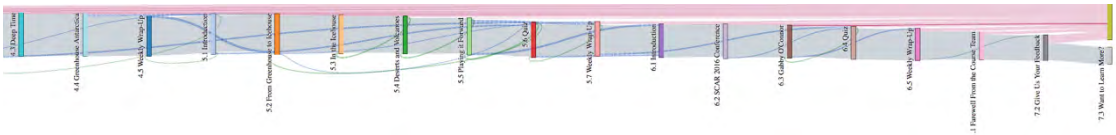

Figure 7: The first prototype showing the behaviour of non-paying participants at the end of the course.

The ERG suggested that it would be an interesting test to first, create differing views based on paying and non-paying participants, and second, activity before and after a major edX platform user interface change to see if this visualization could be used to indicate underlying reasons for the observations.
The observations made from these visualizations:

- Approximately 23% of non-paying participants jumped ahead to the first quiz, with most of the jumps being prior to topic 3.1. 75% of those then jumped back to the content, seemingly using the quiz as a guide on what to learn (figure 6).
- Although 38.7% of the paying customers jumped forward to the first quiz, 65.9% of these were short jumps (topic 3.1 or after), and they seemingly are engaging with the content more, and the quizzes are used to engage with previously viewed material throughout the MOOC (figure 5).
- The non-paying participants followed the end of course linearly with only a few jumps. For example, under 4% jumped ahead to the last quiz and 2.9% jumping back to previous material from that quiz. This indicates a change in behaviour (figure 7).
- The paying participants kept jumping around the material throughout the course. For instance, 19.8% jumped to the last quiz, with 20.4% jumping back to explore the material further, indicating a continued engagement with the material (figure 5).

The positive feedback from the ERG, and the indication that it could be used as a tool to analyse behaviour, led to the continued development of the tool. The following list of changes for the next iteration were based on the feedback from ERG and the initial objectives:

- Use edX logs files instead of database files. Using logs will provide future extendibility as they contain more information than the provided database files.
- To ensure that the vertical bars would show the number of unique visitors to a content page there were two suggestions; split the diagram into two rows with the upper row showing the first visits to a page and the lower row all subsequent visits. Embed a differently coloured bar inside of the vertical bar to indicate the proportion of first visitors to the page. It was decided to use the "two rows" option because this might potentially show a more detailed view of the participant's behaviour with the content.

2nd Prototype

The next prototype was still a visualization with the data embedded inside the view, so therefore not a tool to be used with other edX MOOCs, however it was created as a website instead of an image. The recommended changes had been implemented, while keeping the same colour scheme and style of the previous view. At the data extraction level, the only difference was that the raw logs were used instead of the database files provided by edX.

The following are the observations of the ERG

- The diagram now showed movements of participants to previously viewed content, which provide a more informative picture of their engagement with the material (figure 8). The ERG concluded that the new view therefore was richer and more expressive than the initial prototype. However, some of the extra detail had made the visualization more difficult to understand. They agreed that the expressiveness was more important than the usability, but methods should be sought to make it possible to engage with the visualisation and to increase usability.
- The engagement with the content is still more linear in the latter stages of the course (figure 9).
- The first quiz is still disruptive to learner behaviour (figure 10) The first observation was that 4.6% more learners visit the weekly wrap-up than the quiz, and therefore skip the quiz. The second observation was that the combination of the quiz and the wrap-up seemingly was used by participants to study or review the previous material to enable them to answer the quiz. The quiz had over double as many revisits than first time views, and the previews content had been traversed extensively by these users. It was suggested that this to some extent might be the difference between the behaviour of paying and non-paying participants.
- Later quizzes showed the same impact on behaviour with 7%-8% avoiding them and an increase in revisits of previous content, although the jumps backwards only showed jumps to the closest previous content.

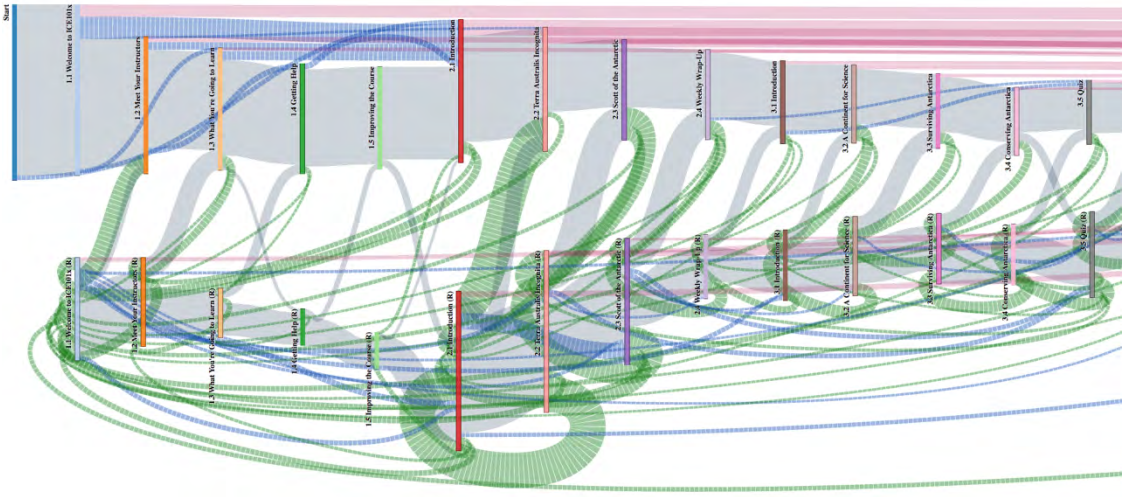

Figure 8: The second prototype showing the first three topics.

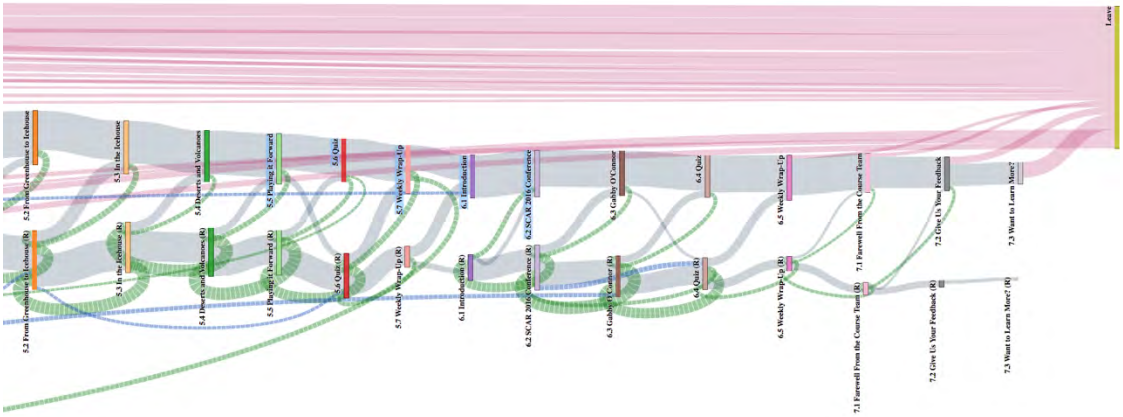

Figure 9: The second prototype showing the end of the course.

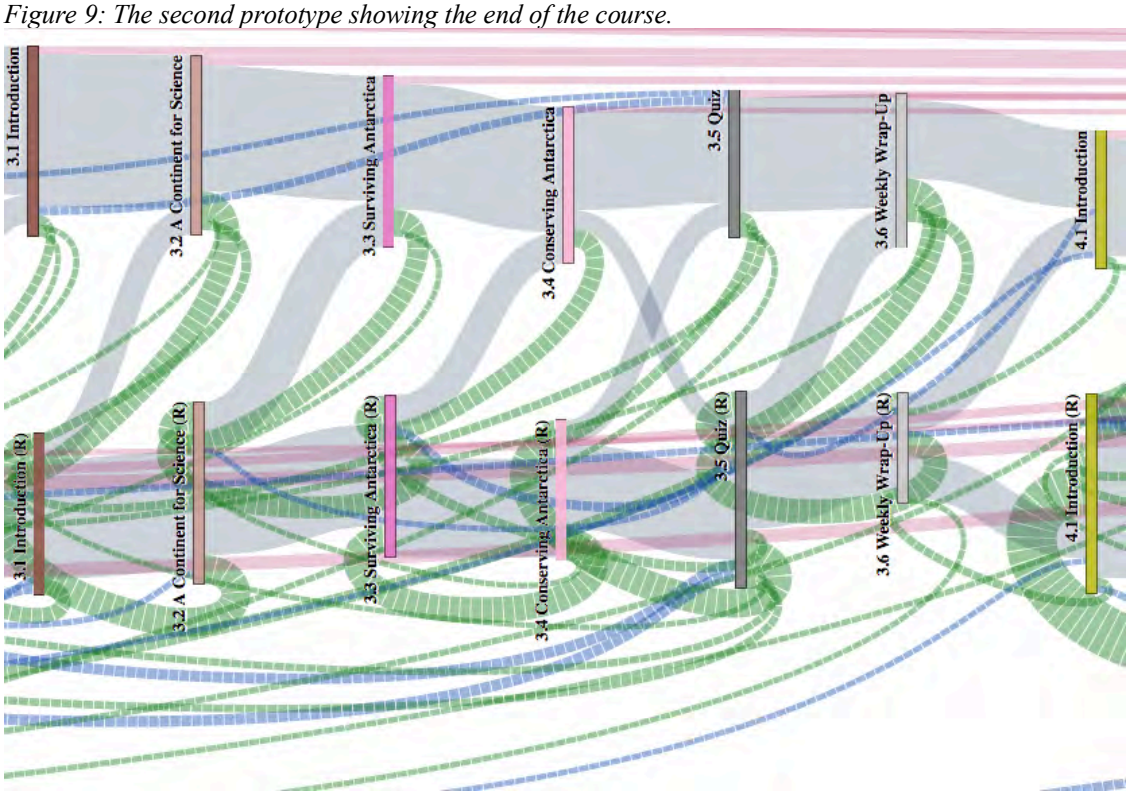

Figure 10: The second prototype showing the end of the course.

Furthermore, the ERG had become aware that a mid-course user interface change by edX (28/04/2017) might have contributed to the observed behavioural changes. The menu of content had been moved from the top of edX's course pages to a less prominent space at the bottom of the screen. Two special views were created to see if this had changed the participants' behaviour. Figure 11 and Figure 12 show the same content pages (vertical bars). The blue paths that can be seen in the top of figure 12 are there because all users were registered as visiting that page for the first time, even if they had been online before the user interface date. Almost all participants are following the expected route with only a few jumps forward and backwards. There is also a noticeable decrease in revisits to content pages.

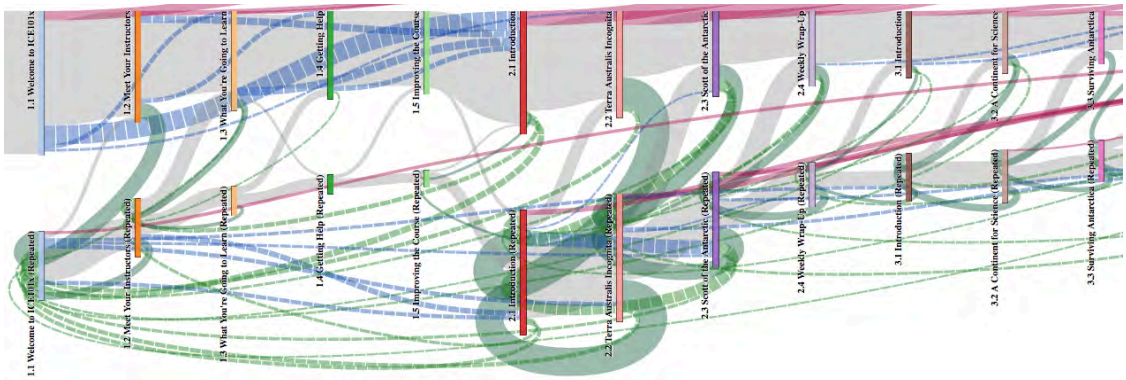

Figure 11: The second prototype showing before the user interface change.

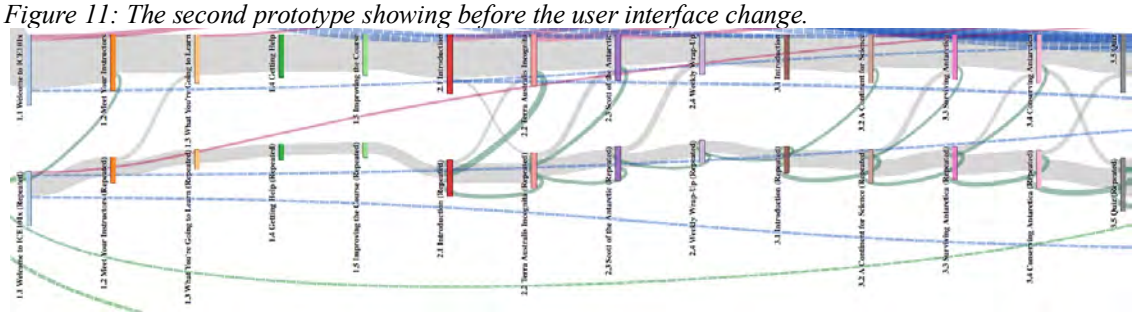

Figure 12: The second prototype showing after the user interface change.

The following changes were suggested for the next prototype:

- Increase usability of the visualization.
- Prepare for new views based on segments of users.
- Create views based on users who followed particular paths or visited certain content sections as represented by the vertical bars.
- Develop it as a tool that can be used with logs from other edX courses.
- The visualisation does not include movements lower than 20 to decrease complexity of the visualization. However, all visits to the content page ought to be shown in the vertical bars. They were excluded in this version. Therefore, the next version should include all visits in vertical bars, but still exclude low paths.
- Provide an option for the user to set the minimum number pf paths that are shown.

3rd Prototype

The biggest changes for this iteration were technical. The data extraction was developed into a tool that makes a webpage of the provided logfiles. This prepares the tool for other MOOCs and also for future expansions to create multiple views based on participant groupings, dates and other feature that can segment into behavioural groupings.

A web server was included to provide the views within browser. This allowed the generation of a new view when the user of the tool double-click a path or vertical bar based on the users who did this activity.
To increase the usability of the visualization the vertical bars were made movable so that the path ways in and out become clearer.

An input box was created to set the minimum number of users for displayed paths. The inclusion of all participants (i.e. include below the minimum number user paths in the vertical bars) did not radically change the view massively, but it now provides the correct number of participants on the vertical bars.

The ERG agreed with the changes. It was observed that the tool was slow at producing the specialized views, and it was agreed to investigate ways to increase speed.

4th Prototype

It was found that the speed issue stemmed from traversing the log files to produce the data for specific views. An intermediate data format was created that extract all user behaviours into on single list. This could be saved as a .ser file in the tool.

A GUI was added to allow a user to input the .ser file or raw log data files themselves. If raw data files are selected, a .ser file will be generated so that the raw data files don't have to be read next time. After selecting the relevant files in the GUI and clicking the start button, the web server is started, and the default browser is opened to show the Sankey diagram.

This approach has increased the speed of the tool significantly. The tool has been released on an open source license at https://github.com/MikeSolvalou/MikeSolvalou.github.io

Conclusion and Future work

The processes described above have helped validate the visualiser as a valuable tool for the MOOC development team at anonymous.
It is still being maintained and further developed. The current plan is to integrate user functionality to segregate the behaviours of various different user grouping and behavioural differences, so that any user will be able to create with views without manually creating the associated logs. This is currently achieved using scripts or by creating bespoke programs whenever a question arises.

A related future feature is to incorporate statistical tests. It would be useful to compare two different diagrams from the same course and be able to see if the two are significantly different from each other. It seems that the data is not normally distributed, so common statistical tools such as t-tests can probably not be used. The plan is to seek advice on this from statistical experts on this.

The tool has been prepared to support data from other MOOC platform, but due to lack of available data this has not been fully implemented yet.

The tool in its current form has already been used to analyse and understand user behaviour. It is being used with the learning design team to reflect and shift course design and pick up on potential loops in learner pathways that can be caused by inappropriate tests or content that is not well matched to the course objective.

References

Aalst, W. M. van der. (2011). Data mining. In *Process mining* (pp. 59–91). Berlin, Heidelberg: Springer.

Aalst, W. M. van der. (2013). Process cubes: Slicing, dicing, rolling up and drilling down event data for process mining. In *Asia-pacific conference on business process management* (pp. 1–22). Springer.

Aalst, W. M. P. van der, Guo, S., & Gorissen, P. (2015). Comparative process mining in education: An approach based on process cubes. In P. Ceravolo, R. Accorsi, & P. Cudre-Mauroux (Eds.), *Data-driven process discovery and analysis: Third ifip wg 2.6, 2.12 international symposium, simpda 2013, riva del garda, italy, august 30, 2013, revised selected papers* (pp. 110–134). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-46436-6_6

Adamopoulos, P. (2013). What makes a great MOOC? An interdisciplinary analysis of student retention in online courses. In *Proceedings of the 34th international conference on information systems* (pp. 1–21). Milan.

Analytics, G. (2017). Using the flow visualization reports.

Beaven, T., Codreanu, T., & Creuzé, A. (2014). Motivation in a language mooc: Issues for course designers. *Language MOOCs: Providing Learning, Transcending Boundaries. Berlin: De Gruyter Open*, 48–66.

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, *8*, 234–241.

Coffrin, C., Corrin, L., Barba, P. de, & Kennedy, G. (2014). Visualizing patterns of student engagement and performance in moocs. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 83–92). New York, NY, USA: ACM. https://doi.org/10.1145/2567574.2567586

Connell, J. L., & Shafer, L. (1989). *Structured rapid prototyping: An evolutionary approach to software development*. Yourdon Press.

d3. (2017). D3-sankey.

Daniel, J. (2012). Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. *Journal of Interactive Media in Education*, *2012*(3). https://doi.org/10.5334/2012-18

Downes, S. (2015). The quality of massive open online courses. In B. H. Khan & M. Ally (Eds.), *International handbook of e-learning volume 1: Theoretical perspectives and research* (pp. 65–77). Abingdon: Routledge.

Emmons, S. R., Light, R. P., & Börner, K. (2017). MOOC visual analytics: Empowering students, teachers, researchers, and platform developers of massively open online courses. *Journal of the Association for Information Science and Technology*, *68*(10), 2350–2363. https://doi.org/10.1002/asi.23852

Friendly, M. (2002). Visions and re-visions of charles joseph minard. *Journal of Educational and Behavioral Statistics*, *27*(1), 31–51.

Jordan, K. (2013). MOOC completion rates: The data. http://www.katyjordan.com/MOOCproject.html Accessed 7 June 2017.

Jordan, K. (2015). Massive open online course completion rates revisited: Assessment, length and attrition. *The International Review of Research in Open and Distributed Learning*, *16*(3), 341–358.

Kay, J., Reimann, P., Diebold, E., & Kummerfeld, B. (2013). MOOCs: So many learners, so much potential... *IEEE Intelligent Systems*, *28*(3), 70–77. https://doi.org/10.1109/MIS.2013.66

Khalil, H., & Ebner, M. (2014). MOOCs completion rates and possible methods to improve retention-a literature review. In *World conference on educational multimedia, hypermedia and telecommunications* (pp. 1305–1313).

Kop, R. (2011). The challenges to connectivist learning on open online networks: Learning experiences during a massive open online course. *The International Review of Research in Open and Distributed Learning*, *12*(3), 19–38. https://doi.org/10.1109/MIS.2013.66

Kop, R., & Fournier, H. (2010). New dimensions to self-directed learning in an open networked learning environment. *International Journal of Self-Directed Learning*, *7*(2), 1–18.

Kop, R., Fournier, H., & Mak, J. S. F. (2011). A pedagogy of abundance or a pedagogy to support human beings? Participant support on massive open online courses. *The International Review of Research in Open and Distributed Learning*, *12*(7), 74–93. https://doi.org/10.19173/irrodl.v12i7.1041

Levy, D. (2011). Lessons learned from participating in a connectivist massive online open course (MOOC). In *Proceedings of the chais conference on instructional technologies research 2011: Learning in the technological era* (pp. 31–36).

McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). The MOOC model for digital practice. http://www.elearnspace.org/Articles/MOOC_Final.pdf Accessed 6 June 2017; Massive Open Online Courses: digital ways of knowing; learning, Charlottetown, Canada.

Milligan, C., Littlejohn, A., & Margaryan, A. (2013). Patterns of engagement in connectivist moocs. *Journal of Online Learning and Teaching*, *9*(2), 149.

Onah, D. F., Sinclair, J., & Boyatt, R. (2014). Dropout rates of massive open online courses: Behavioural patterns. *EDULEARN14 Proceedings*, 5825–5834.

Schmidt, M. (2008). The sankey diagram in energy and material flow management. *Journal of Industrial Ecology*, *12*(1), 82–94. https://doi.org/10.1111/j.1530-9290.2008.00004.x

Shah, D. (2016). By The Numbers: MOOCs in 2016. Class Central; https://www.class-central.com/report/mooc-stats-2016/ Accessed 7 April 2017.

Shi, C., Fu, S., Chen, Q., & Qu, H. (2014). VisMOOC: Visualizing video clickstream data from massive open online courses. In *Visual analytics science and technology (vast), 2014 ieee conference on* (pp. 277–278). IEEE. https://doi.org/10.1109/VAST.2014.7042528

Stephens-Martinez, K., Hearst, M. A., & Fox, A. (2014). Monitoring moocs: Which information sources do instructors value? In *Proceedings of the first acm conference on learning@ scale conference* (pp. 79–88). ACM. https://doi.org/10.1145/2556325.2566246

Stewart, B. (2010). Social media literacies and perceptions of value in open online courses. http://portfolio.cribchronicles.com/wp-content/uploads/2012/11/612701_Social_Media_Literacies_MOOCs.pdf Accessed 7 July 2017.

Please cite as: Lundqvist, K., Godinez, M. & Warburton, S. (2018). Designing personalised, automated feedback to develop students' research writing skills. In M. Campbell, J. Willems, C. Adachi, D. Blake, I. Doherty, S. Krishnan, S. Macfarlane, L. Ngo, M. O'Donnell, S. Palmer, L. Riddell, I. Story, H. Suri & J. Tai (Eds.), Open Oceans: Learning without borders. Proceedings ASCILITE 2018 Geelong (pp. 194-203).

https://doi.org/10.14742/apubs.2018.1910