Chang, M. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

Using Clickstream to Understand Learning Paths and the Network Structure of Learning Resources: Using MOOC as an Example

Ming GAO[a], Jingjing ZHANG[b*], Di SUN[c] & Jiang ZHANG[d]
[a]*Research Centre of Distance Education, Beijing Normal University, China*
[b]*Big Data Centre for Technology-mediated Education, Beijing Normal University, China*
[c]*IDD&E, Syracuse University, USA*
[d]*School of Systems Science, Beijing Normal University, China*
* jingjing.zhang@bnu.edu.cn

Abstract: Massive open online courses (MOOCs) have attracted great attention from the public and learners. However, their high dropout rates have been criticized over the past few years. There has been a body of work dedicated to using learning analytics to provide feedback for instructors and designers to improve learners' engagement and retention rates. However, the open and flexible nature of MOOCs has often been overlooked in these analytics studies. In this work, we used clickstream data to construct a flow network model in order to identify MOOC learners' learning paths and the network structure of available learning resources from an open system perspective. We found that learners tend to adopt linear learning paths within chapters and continue to watch video lectures in next chapters instead of taking quizzes at the end of the chapters, and they rarely review previous chapters during studies. We also found that learning paths of all learners have formed a centralised network structure which implies that certain learning resources in such a network structure dominate the way in which learners learn in MOOCs.

Keywords: MOOCs, learning behaviour, learning paths, network structure, learning analytics

1. Introduction

Massive open online courses (MOOCs) play an important role in higher education (Li & Stephen, 2013). MOOCs gather global education resources and provide opportunities for learners to take online courses from prestigious universities. In the early stages of MOOC development, MOOCs on Udacity, Coursera, and edX platforms attracted more than 100,000 registrants per course (Seaton, Bergner, Chuang, Mitros, & Pritchard, 2014). However, there have also been high dropout rates, with only a very small numbers of registrants completing most MOOCs successfully (Koller, Ng, Do, & Chen, 2013). It is critical to find out more about how to keep learners engaged and how to improve the dismal retention rates. A great deal of recent work has been done using machine learning methods to predict the learners who are most likely to drop out (e.g., Halawa, Greene, & Mitchell, 2014; Moreno-Marcos, Alario-Hoyos, Munoz-Merino, & Delgado Kloos, 2018; Xing & Du, 2018). While the accuracy rate of dropout prediction is high, the explanatory power of such studies is low. The selected variables, which have a strong impact on prediction, may not provide enough information for course designers and instructors to adapt MOOC design or content responsibly. Some research has addressed this by highlighting the importance of analysing learners' learning paths to provide more practical feedback for instructors and designers (Davis, Chen, Hauff, & Houben, 2016). For example, Wen and Rose (2014) investigated learners' habits through their sequential use of learning activities. Nevertheless, in analysing the learning paths of individuals, only a few studies have taken the open and flexible nature of MOOCs into consideration (e.g., Zhang, Lou, Zhang, & Zhang, 2019). For instance, a learner's use of time is not always continuous once they start online learning: they can stop watching a video or pause it, then restart or close it whenever they like.

In this work, we attempt to understand MOOC learners' learning paths and to explore the network structure of learning resources by modelling a flow network from learners' clickstream data. Specifically, we focus on analysing learners' learning behaviour in relation to learning resources

consisting of lecture videos and assignments, attempting to discover learners' learning behaviour or habits and to measure the network structure of learning resources from an open perspective.

2. Methodology

2.1 MOOCs Dataset

The dataset used in this work was gathered from a course run in XuetangX, titled "Introduction to Psychology" (2015 autumn semester). It contained 12 chapters (68 learning resources, including videos and assignments) and a final exam. The design of this course implied a traditional linear trajectory through the learning material (see Figure 1). Each chapter commenced with a video lecture, followed by an assignment to evaluate what the learners had attained from this chapter. Nearly 20,000 learners had registered for the course but only 10,359 actually accessed it.

In the dataset, learner ID, timestamp (the time at which the page was opened/closed), URLs, page title, stay time, and page type were saved in each log. Learners and learning resources could be identified uniquely through the learner ID and URLs. To take into account the flexibility of the learners' learning time, the timestamp was used to extract their learning sessions through sorting and segmenting their clickstream data. As a rule of thumb, online behaviours that occurred over 25.5 minutes apart were determined as separate sessions (Catledge & Pitkow, 1995). The 30-minute threshold was used in this study to delineate a sequence as a new session (see Figure 2). In other words, in cases where there was 30 minutes between the use of two different resources, we assumed that the learners had stopped learning from the former resource. Stay time was used to filter invalid click access, and page titles and types provided the basic descriptions of a web page.

Some of these course learners used mobile devices to access the course. Their data were not included in the analysis due to incomplete formats. Some logs associated with zero seconds (0s) of stay time on websites and other resources (except video lectures and assignments) were also excluded from the data analysis. After finishing some pre-processing work, 5,506 users with at least 1 second of log data who used web browsers to visit the course (63,060 logs) were included.

Figure 1. Screenshot of "Introduction to Psychology"

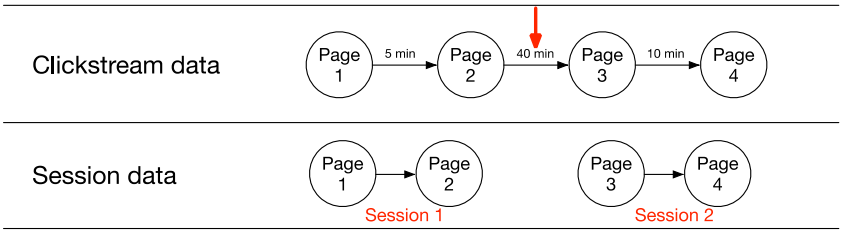


Figure 2. Transformation from clickstream data to session data

2.2 Flow Network Model

This work used the flow network model proposed by Zhang et al. (Zhang & Wu, 2013; Wu, Zhang, & Zhao, 2014) to model the openness of MOOCs. This model has been found useful in many different contexts, including assessing the impact of websites (Wu & Zhang, 2011), the "stickiness" of a web forum (Wu et al., 2014), and the geometric representation of Internet ecology (Shi et al., 2015). The flow network model is an open, directed, and weighted network in which nodes represent the websites and weighted edges represent traffics (the amount of users' transition) between two websites. It can be constructed from the users' clickstream data and its structure can be seen in Figure 3. The openness of the flow network means that it can connect the online and offline world by adding two special nodes, "source" and "sink". The direction from the "source" to other nodes represents the point at which users go online; when users go offline, this is marked by the nodes with "sink". In a balanced network, moreover, the amount of inflow of a node equals its outflow, except for the "source" and "sink. As mentioned above, 68 nodes (the number of resources, including all video lectures and assignments) can be used to construct the flow network.

Based on the flow matrix hidden in the flow network, three indexes – A_i, C_i, and η – can be used to describe the attributes of this network. A_i measures the total flux of node i, and C_i measures the impact of node i on the whole network. After getting A_i and C_i for every node i, we can explore the existence of the allometric scaling law ($C_i \sim A_i^{\eta}$) (Jiang Zhang & Guo, 2010), a universal relationship found in river systems (Rodriguez-Iturbe & Rinaldo, 2001), living organisms (West, Brown, & Enquist., 1999), and international trade systems (Shi, Luo, Wang, & Zhang, 2013). The exponent η reflects the flow structure of the network. An example provided by Shi et al. (2013) is two networks with the same distribution of A_i having a different value of η. A_i = [1, 2, 3, 4, 5], $C_i(1)=A_i^{0.5}$={1, 1.4, 1.7, 2, 2.2}, $C_i(2)=A_i^{2}$={1, 4, 9, 16, 25}. For the network with η=2, the largest two nodes – the fifth and fourth nodes (25/(1+4+9+16+25)=45%, 16/(1+4+9+16+25)=29% of impacts, respectively) – almost control the whole network. For the network with η=0.5, the largest node only dominates 2.2/(1+1.4+1.7+2+2.2) = 27% of impacts. Thus, if $\eta > 1$, the network structure is centralised. In other words, some nodes control the flows of the network. If $\eta <= 1$, that means the network structure is more decentralised. Each node plays the more equal role in the network. In the learning resources network, we can interpret A_i as the total amount of viewing frequency through resource i; C_i reflects the influence of resource i on the entire network (both direct and indirect) and η reflects the flow structure of the resource network. The larger the η, the more centralised the flow structure (Wu & Zhang, 2013). More detail on these three variables is given in Zhang et al. (2010).

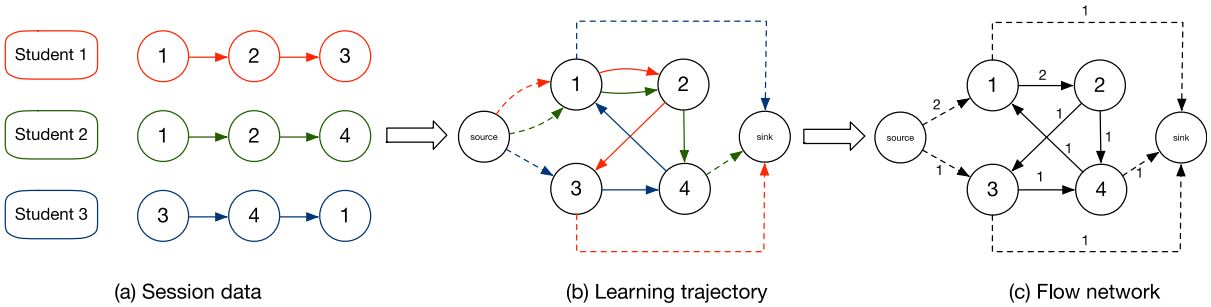


Figure 3. Process of flow network construction

3. Findings

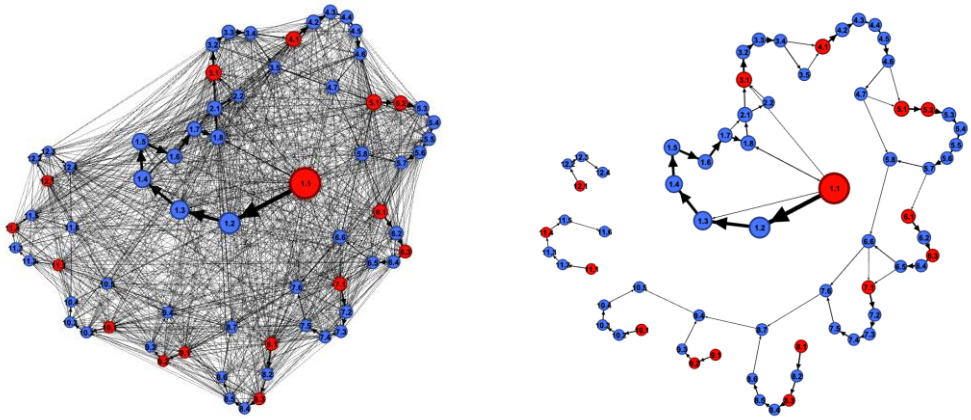
The learning resource network was constructed based on students' clickstream data and visualised using visualisation software called Gephi (https://gephi.org/) (Figure 4(a)). The size of the nodes represents their importance (C_i). The colour of the nodes represents their state of flux. A red node indicates that the flux of outflow (excluding the flow to sink) was larger than the flux of inflow (excluding the flow from source), while a blue node indicates that its flux of inflow (excluding the flow from source) was larger than the flux of outflow (excluded the flux to sink). That is, red nodes show that there are more learners continuing to learn after they have completed this resource, and the blue nodes show that some of the learners leave these resources. The arrows between two nodes represent a learner's learning direction (learning path) between resources. The thickness of the edges indicates the viewing frequency.

3.1 Learning Paths

We almost found that red nodes (the flow of the outflow is larger than the inflow) are the first resource of each chapter (see Figure 4). This means that some learners might keep on learning after the first video lecture, while some may leave during any video lecture or assignment within the same chapter.

Given the non-formal nature of MOOCs, learners can determine their way of learning and do not have to follow the predesigned learning sequence, for example, employing nonlinear navigation (Guo & Reinecke, 2014). The complex links between resources show the nonlinear trajectories to some extent (see Figure 4(a)). However, almost all the thicker directed edges appear within the chapter. This indicates that most of learners still follow the predesigned learning path – a linear learning path – especially within each chapter.

To highlight the relationship between resources, we filtered some edges and left the backbone edges based on learners' frequent transitions between resources (shown in Figure 4(b)). No isolated point was left when the lower limiting value of edges was reduced to 133.344. Two separate learning resource communities (Chapters 11 and 12) were found with no links to other chapters and with thin edges. This depicts the learning paths of learners who finish the course successfully; these are the last two chapters of the course, and the topics were broken up into weeks and released one chapter at a time per week. On the other hand, this may reflect that some learners were just interested in the topic of these two chapters. We also found links between the last nodes of some chapters, for example between Chapter 4 and Chapter 10. The last node (resource) of a chapter was an assignment aiming to help learners evaluate their learning performance in this chapter. It indicates that some learners intend to continue watching the next chapter's video lectures instead of testing their learning performance at the end of a chapter. This was also verified by the links between the last-but-one resource of former chapters and the first resource of later chapters. It is interesting to see that no links return to former nodes. This indicates that fewer learners reviewed former content. Some activities and assignments that embed or mix previous content can be designed to help learners review and enhance what they have learned.



(a) Before filter (b) After filter

Figure 4. The flow structure of resources. The size of a node stands for the impact of the resource (C_i). The width of the line means the amount of flux. The colour represents the difference between inflow

and outflow: red represents that the flux of its outflow is bigger than the inflow; blue represents that the flux of the outflow is smaller than the inflow

3.2 The Network Structure of Learning Resources

As shown in Figure 4(b), some of the largest nodes (larger value of C_i), which reflect their importance, were almost all located in the first chapter, which gave an introduction to psychology. This indicates that the introduction may play an important role in the course learning resources network.

Further, we explored the impact of nodes on the flux of whole flow network. In other words, we examined whether some resources play an important role in connecting other resources to affect a learner's learning or understanding by exploring the relationship between the size of a node and its impact. As shown in Figure 5, it shows the A_i and C_i values of all resources on the log–log plot, and they have been fitted with a line. The minimum square error (R^2) was used to evaluate the performance of fitness. The value of R^2 is 0.93, which shows a high degree of fit. The exponent (η) is 1.08. This shows that the flow network structure is centralised that some resources, represented above the red line in Figure 5, have a larger impact (C_i) on the whole network than estimated. Namely, some resources affect a learner's learning or understanding through connecting related or necessary resources.

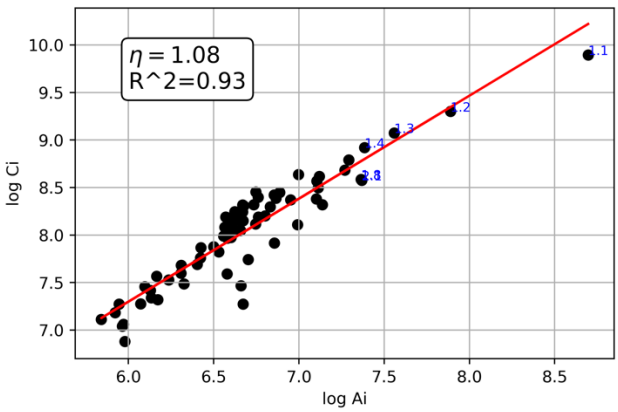


Figure 5. The power–law relationship between A_i and C_i for courseware

4. Conclusion and Future Work

In this study, we attempted to use a flow network model to improve our understanding of MOOC learners' learning behaviour and the relationship or structure of learning resources from an open perspective. We conducted an in-depth analysis on learners' learning paths and the relationship between the course resources. Some interesting results emerged from the learning path analysis, for example that learners tend to adopt a linear learning path within chapter learning; few learners review or enhance what they have learned. We also found that some resources play a role in connecting other resources, with an effect on learners' learning and understanding. In future, we will further explore learners' learning paths through different learning achievement groups and explore the learning resources structure in different subject categories to provide more practical feedback for course instructors and designers.

Acknowledgements

The authors would like to thank XuetangX for providing us with the clickstream data for this research. This project is funded by Student Research Foundation of Faculty of Education, Beijing Normal University [Project No. 1812201].

References

Catledge, L. D., & Pitkow, J. E. (1995). Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems*, *27*(6), 1065–1073. https://doi.org/10.1016/0169-7552(95)00043-7

Davis, D., Chen, G., Hauff, C., & Houben, G.-J. (2016). Gauging MOOC learners' adherence to the designed learning path. In *Proceedings of the 9th International Conference on Educational Data Mining* (pp. 54–61). Retrieved from http://www.derecho.unam.mx/cultura-juridica/eventos/programa-2-Curso-DyL.pdf

Guo, P. J., & Reinecke, K. (2014). Demographic differences in how students navigate through MOOCs. In *Proceedings of the First ACM Conference on Learning @ Scale Conference - L@S '14* (pp. 21–30). https://doi.org/10.1145/2556325.2566247

Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. In *Proceedings of the European MOOC Stakeholder Summit 2014* (pp. 58–65). Retrieved from http://www.emoocs2014.eu/sites/default/files/Proceedings-Moocs-Summit-2014.pdf

Koller, D., Ng, A., Do, C., & Chen, Z. (2013). Retention and intention in Massive Open Online Courses: in depth. *Educause Review*, *48*(3), 62–63.

Li, Y., & Stephen, P. (2013). *MOOCs and Open Education: Implications for Higher Education*. *Cetis*. https://doi.org/10.4324/9781315751108-1

Moreno-Marcos, P. M., Alario-Hoyos, C., Munoz-Merino, P. J., & Delgado Kloos, C. (2018). Prediction in MOOCs: a review and future research directions. *IEEE Transactions on Learning Technologies*, https://doi.org/10.1109/TLT.2018.2856808

Rodriguez-Iturbe, I., & Rinaldo, A. (2001). *Fractal River Basins: Chance and Self-Organization*. Cambridge: Cambridge University Press.

Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2014). Who does what in a Massive Open Online Course? *Communications of the ACM*, *57*(4), 58–65. https://doi.org/10.1145/2500876

Shi, P., Huang, X., Wang, J., Zhang, J., Deng, S., & Wu, Y. (2015). A geometric representation of collective attention flows. *PLoS ONE*, *10*(9), 1–21. https://doi.org/10.1371/journal.pone.0136243

Shi, P., Luo, J., Wang, P., & Zhang, J. (2013). Centralized flow structure of international trade networks for different products. In *2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings* (pp. 91–99). IEEE. https://doi.org/10.1109/ICMSE.2013.6586267

Wen, M., & Rose, C. P. (2014). Identifying latent study habits by mining learner behavior patterns in Massive Open Online Courses. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 1983–1986). New York: ACM. https://doi.org/10.1145/2661829.2662033

West, G. B., Brown, J. H., & Enquist., B. J. (1999). The fourth dimension of life: fractal geometry and allometric scaling of organisms. *Science*, *284*(5420), 1677–1679.

Wu, L., & Zhang, J. (2011). The decentralized structure of collective attention on the Web. *ArXiv Preprint ArXiv:1110.6097*, 1–12.

Wu, L., & Zhang, J. (2013). The decentralized flow structure of clickstreams on the web. *European Physical Journal B*, *86*(6), 1–6. https://doi.org/10.1140/epjb/e2013-40132-2

Wu, L., Zhang, J., & Zhao, M. (2014). The metabolism and growth of web forums. *PLoS ONE*, *9*(8), 1–11. https://doi.org/10.1371/journal.pone.0102646

Xing, W., & Du, D. (2018). Dropout Prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research*. *57*(3), 547-570. https://doi.org/10.1177/0735633118757015

Zhang, J., Lou, X., Zhang, H., & Zhang, J. (2019). Modeling collective attention in online and flexible learning environments. *Distance Education*, *40*(2), 278-301. https://doi.org/10.1080/01587919.2019.1600368

Zhang, J., & Guo, L. (2010). Scaling behaviors of weighted food webs as energy transportation networks. *Journal of Theoretical Biology*, *264*(3), 760–770. https://doi.org/10.1016/j.jtbi.2010.03.024

Zhang, J., & Wu, L. (2013). Allometry and dissipation of ecological flow networks. *PLoS ONE*, *8*(9), 1–8. https://doi.org/10.1371/journal.pone.0072525