British Journal of Educational Technology (2015)
doi:10.1111/bjet.12312

Massively Open Online Course for Educators (MOOC-Ed) network dataset

Shaun Kellogg and Achim Edelmann

Shaun Kellogg is a research scholar at the Friday Institute for Educational Innovation and an adjunct assistant professor in the College of Education at North Carolina State University. His research interests center on the application of learning analytics to improve of K-12 educational contexts and educator professional development. Achim Edelmann is a postdoctoral associate at the Duke Network Analysis Center. He received a PhD in sociology from the University of Cambridge. He specializes in the sociology of culture, network analysis, and social theory. His current research focuses on bridging the structure-culture link in theory and practice, and on contemporary developments in the sociology of science. Address for correspondence: Dr Shaun Kellogg, Friday Institute for Educational Innovation, North Carolina State University, 1890 Main Campus Drive, Campus Box 7249, Raleigh, NC 27606, USA. Email: sbkellog@ncsu.edu

Abstract

This paper presents the Massively Open Online Course for Educators (MOOC-Ed) network dataset. It entails information on two online communication networks resulting from two consecutive offerings of the MOOC called *The Digital Learning Transition in K-12 Schools* in spring and fall 2013. The courses were offered to educators from the USA and abroad. Though based on the same course, minor controlled variations were made to both MOOCs in terms of the course length, discussion prompts and group size. The dataset provides opportunities to examine how participants leverage online communication forums to support their learning. In particular, it allows modeling network mechanisms to better understand factors that facilitate or impede the exchange of information among educators.

Dataset

This paper references the following MOOC-Ed social network analysis dataset:
Title: Massively Open Online Course for Educators (MOOC-Ed) network dataset
Location: http://dx.doi.org/10.7910/DVN/ZZH3UB
Creator: Shaun Kellogg, Achim Edelmann
Date: October 30, 2014
Format: Comma Separated Values
Restrictions to use (if any): CC0—"Public Domain Dedication"—Good scientific practices expect that proper credit for use of this data is given via citation.

Introduction

As an emerging model for online learning, Massively Open Online Courses (MOOCs) present new opportunities and challenges for facilitating learning. In particular, they offer novel forms and new dimensions of social interaction that need to be integrated in our understanding of learning. Social interaction has been identified as a crucial factor in the educational process. Prominent theoretical approaches to learning such as Social Cognitive Theory or Social Constructivism more broadly argue that social interaction and dialogue are critical to successful learning (Bandura & McClelland, 1977; Glassman, 2001; Grusec, 1992; Hung & Der-Thanq, 2001; Johnson & Johnson, 2009). However, it remains an open question as to what extent and in what ways

MOOCs support interactions in educational contexts. Assessing the implications of MOOCs for the digital economy, McAuley, Stewart, Siemens and Cormier (2010) report that MOOCs have the potential to "model and build collaborative networks of unprecedented size that transcend time and space" and the "network ties created between people during a MOOC have the potential to continue as sustainable and relevant personal and professional connections beyond the boundaries of the course itself" (p 35).

The dataset described below provides detailed information on the communications taking place between learners in two such MOOCs, along with relevant characteristics of the participants such as their professional roles or their experience in education. The primary use of this dataset is to enable social network analyses (SNAs) of these communications (Rivera, Soderstrom & Uzzi, 2010). SNA provides a set of analytical tools and statistical techniques to account for patterns in how actors connect with each other (Scott, 2000; Wasserman & Faust, 1994). Although this has opened up novel ways of examining theories and questions in the social sciences more broadly, application of SNA in educational research has remained surprisingly limited despite obvious advantages.

As McFarland, Diehl and Rawlings (2011) point out, the statistical methods educational researchers have historically used to study the context of learning rest on assumptions such as statistical independence and normality that run counter to available evidence that "presents classrooms as complex interdependent social environments" (p 90). In contrast, SNA offers statistical methods to adequately account for these complex interdependencies in interaction processes. For example, exponential random graph models (Snijders, Pattison, Robins & Handcock, 2006; Wasserman & Pattison, 1996) can be used to appropriately model various social selection processes such as homophily, reciprocity and triadic closure that shape the structure of social networks (Hunter, Handcock, Butts, Goodreau & Morris, 2008).

However, Carolan (2013) notes, only few scholars have applied such techniques in research on educational processes, partly because we lack appropriate data. This is especially the case with regard to new forms of online education. Being able to apply network analytical techniques to such settings requires a particular form of empirical data. In particular, data must sufficiently capture the structure of social interactions by identifying who communicated with whom and how often. The dataset presented here is intended to help fill this gap. It was specifically created for describing communication patterns and modeling network processes in MOOCs created for educators in order to better understand factors the facilitate or impede the exchange of information and resources among participants (Kellogg, Booth & Oliver, 2014). However, it is suitable for academics and researchers interested in network processes in online learning communities more generally. For example, it allows tracing the ebb and flow of communications over time, analyze the impact of offline proximity on online communication dynamics or the building of cohesive subgroups within the student community. This might help to improve the design of future courses for online learning.

Context of dataset

In spring 2013, the Friday Institute for Educational Innovation at North Carolina State University launched the Massive Open Online Course for Educators (MOOC-Ed) Initiative. MOOC-Eds are designed specifically for professional educators and follow the guidelines for effective professional learning and a special set of design principles: multiple voices, self-directed learning, peer-supported learning and job-connected learning (Kleiman & Wolf, 2015). MOOC-Eds have similarities with and differences from both the structured learning experience of xMOOCs, and the community-oriented aspects of cMOOCs.

xMOOCs are typically online versions of traditional university course formats in which knowledge is delivered from the teacher to the students in the form of lectures, curated resources and

assignments (Haggard, 2013). Like xMOOCs, MOOC-Eds provide a curated set of resources around predefined topics in a centralized learning platform. They also provide structured activities for engaging with this content including assignments and projects.

In contrast, cMOOCs follow a "connectivist" design principle and their core pedagogical model is peer learning. They allow and actively promote communication between teachers and students in various forms. Learning takes place in communities of students and experts, and may be distributed across multiple learning platforms such as blogs, social networking sites and discussion forums. Like cMOOCs, MOOC-Eds are purposefully designed around multiple voices and the exchange of multiple perspectives among educators, students and other experts. Crucial to MOOC-Eds, thus, is the inclusion of activities and online discussion that facilitate the exchange of ideas and experiences among peers.

A typical unit in a MOOC-Ed consists of a short video introduction to the topic, followed by a set of core and supplementary resources (eg, videos from other perspectives, case studies, articles, reports, etc) with embedded opportunities for peer interaction to explore the topic further. To stimulate interaction, the two course instructors regularly engaged with participants. Group spaces within the forums were created based on participants' locations or professional roles to enable focused discussions. The course also includes a small project and peer feedback on submitted projects.

The MOOC-ED initiative was launched with a 6-week online course called *Planning for the Digital Learning Transition in K-12 Schools*. The aim of this course was to help school and district leaders plan and implement K-12 digital learning initiatives. Specifically, the goals of this course were to help educators understand the potential of digital learning in K-12 schools; assess progress and set future goals for their school or district; and begin to develop a plan to achieve your digital learning goals. The course was divided into five units: (1) Envisioning Schools of the Future; (2) Changing the Culture of Teaching and Learning; (3) Elements of a Successful Digital Learning Transition; (4) Leading the Transitions; and (5) Crowdsourcing Feedback, Lessons and Resources.

In autumn 2013, the course was offered a second time in a slightly modified version. The course was expanded to 8 weeks to allow more time to complete assignments and the course project, and modifications were made to the forums to facilitate discussion. Specific changes include: the addition of an online "role-call" asking participants to introduce themselves to their peers at the beginning of the course, replacing low-interest discussion questions, a slight change to initial groups and simplified guidelines for providing feedback on participant projects shared in the forums. The dataset was created from the first two iterations of this course, and is described in more detail below.

Description of dataset

The dataset presented here includes individual characteristics of participants who posted at least once in the online forums as well as information on the online communications between those participants. Individual characteristics were provided by participants on the course registration form. Communications were constructed by a query of the MySQL database that recorded participants' user-logs of activity in the online forums. In particular, the initiator of a discussion thread was identified as the default recipient of all responses in that thread unless comments were explicitly directed at other comments/participants in the thread. We also identified postings that were directed at comments/participants other than the originator of the thread, based on two criteria: First, whether someone other than the originator of the thread was referenced using the forum's "quote" feature, and second whether the comment explicitly entailed the name of participants previously commenting in the thread. To ensure the accuracy of this process, all comments identified as entailing either quotes or the names of previous commentators were visually

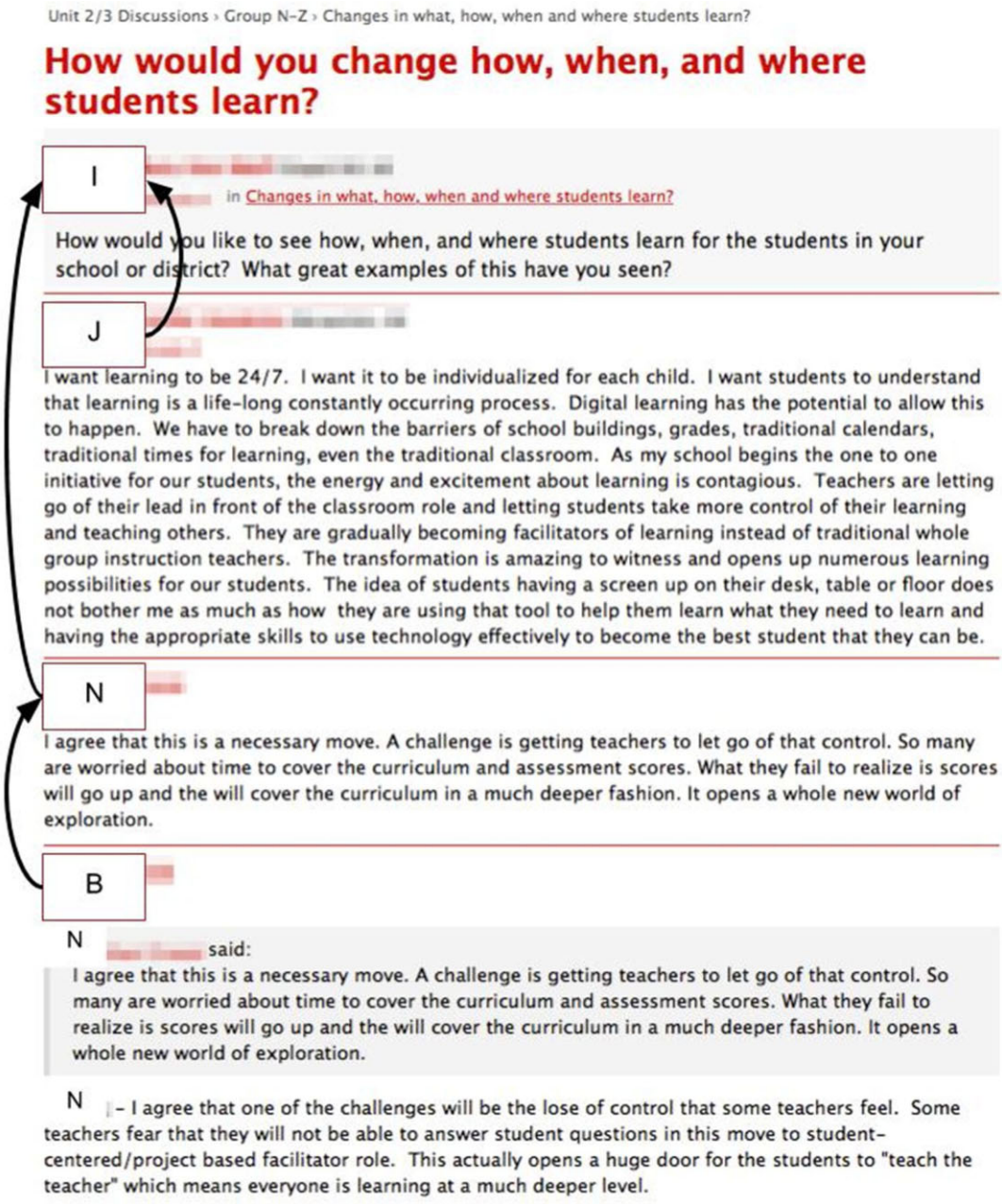

Figure 1: Reply to a discussion comment

inspected. To illustrate this process, consider the beginning of a discussion thread between I, J, N and B in Figure 1. The thread was initiated by I, which thus represents the default recipient of all comments in this thread. Accordingly, the comments by J and N are considered to be directed at I. The comment of B, however, is a direct response to the comment by N as signaled by the use of the quote-feature as well as the explicit mentioning of N's name within B's comment.

Table 1: List of node and edge attribute data for both networks

Variable name	*Description*
Node attributes	
UID	Unique identification number of course participants
Instructor	Identification of course facilitator (1 = instructor)
Role	Professional role (eg, teacher, librarian, administrator)
Experience	Years of experience as an educator
Grades	Works with elementary, middle, and/or high school students; note: "generalists" signals a combination of two or more grade levels
Location	US state or non-US country
Group	Initial assignment of discussion group
Gender	Male, female
Expert	Identifier of "expert panelists" invited to course to share experience through recorded Q&A; (1 = expert panelist)
Connect	Dummy whether participants listed networking/collaboration with others as one of their course goals on the registration form
Edge attributes	
Sender	Unique identifier of author of comment
Receiver	Unique identifier of identified recipient of comment
Timestamp	Time comment was posted
Parent	Primary category or topic of thread
Category	Subcategory or subtopic of thread
Thread_id	Unique identifier of a thread
Comment_id	Unique identifier of a comment

Table 2: Basic descriptive characteristics of both networks

	Network 1	*Network 2*
Directed	Yes	Yes
Nodes	445	492
Isolates	3	66
Density	0.98%	0.85%
Mutual dyads	212	285
Asymmetric dyads	1512	1492
In-/Outdegree	min = 0/0, max = 273/74, mean = 4.35	min = 0/0, max =184/132, mean = 4.19
Tie-frequency	min = 0, max = 15, mean = 1.28	min = 0, max = 12, mean = 1.24
Network centralization	0.38	0.27

The data are stored in two files for each of the MOOCs: a node attribute file and a network edgelist. Both files are in csv-format, with variable names in the headers. The node attribute file contains individual characteristics of participants (nodes). The second file entails the network data in an edgelist format. Each row signifies an online communication (edge) with the corresponding "sender," "receiver" and further attributes of the communication listed across the columns. A description of variable names, as well as basic descriptive statistics of each network, is provided in Tables 1 and 2 respectively.

Limitations

The dataset provided is limited in several ways. First, the actual text of the online communications is not included due to the extensive cleaning that would be required to ensure participants' anonymity. Second, the dataset includes only participants who posted publicly in the course

forums and excludes those who viewed discussions but did not post themselves, or only sent or received messages privately. Finally, due to our coding decision, comments might erroneously be treated as directed to the originator of a thread if participants did not explicitly name or quote the intended recipients.

Statements on open data, ethics and conflict of interest

The collection, processing and analysis of this data was sponsored by Athabasca University aspart of the MOOC Research Initiative and supported through funding from the Bill and Malinda Gates foundations (award number: CO0009135V3). This data is hosted on the Harvard Dataverse Repository and made available under the Creative Commons CC0 Public Domain Dedication. For more information on the CC0 waiver, please visit the Creative Commons website at http://creativecommons.org/about/cc0. Dataverse Community Norms (http://best-practices.dataverse.org/harvard-policies/community-norms.html) as well as good scientific practices expect that proper credit is given via citation for use of this data.

This study and the dissemination of data and findings from this study were approved by Institutional Review Board for the Use of Human Subjects in Research at North Carolina State University (IRB Number: 3413). In addition, the MOOC-Ed Privacy and Terms of Use (https://place.fi.ncsu.edu/mod/page/view.php?id=395) made clear to participants that participation was completely voluntary; that personal information such as name or email address provided would not be released, sold or rented to any entities or individuals outside of the MOOC-Ed project; and that they granted the Friday Institute nonexclusive, royaltyfree, license to use, reproduce, distribute and display content they created in connection with the MOOC-Ed. There is no conflict of interested in the dissemination of data presented in this paper.

References

Bandura, A. & McClelland, D. (1977). *Social learning theory*. Retrieved 13 April 2013, from http://www.jku.at/org/content/e54521/e54528/e54529/e178059/Bandura_SocialLearningTheory_ger.pdf.

Carolan, B. V. (2013). *Social network analysis and education: theory, methods & applications*. Thousand Oaks, CA: Sage Publications, Inc. Retrieved 15 June 2013, from http://www.amazon.com/Social-Network-Analysis-Education-Applications/dp/1412999472).

Glassman, M. (2001). Dewey and Vygotsky: society, experience, and inquiry in educational practice. *Educational Researcher*, *30*, 4, 3–14. Retrieved 13 November 2012, from http://edr.sagepub.com/content/30/4/3.short.

Grusec, J. E. J. (1992). Social learning theory and developmental psychology: the legacies of Robert Sears and Albert Bandura. *Developmental Psychology*, *28*, 5, 776–786. doi: 10.1037//0012-1649.28.5.776.

Haggard, S. (2013). *The maturing of the MOOC* (No. 130). London, UK.

Hung, D. W. L. & Der-Thanq, C. (2001). Situated cognition, Vygotskian thought and learning from the communities of practice perspective: implications for the design of web-based e-learning. *Educational Media International*, *38*, 1, 3–12. doi: 10.1080/09523980110037525.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. & Morris, M. (2008). ergm: a package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, *24*, 3, nihpa54860. Retrieved 4 November 2013, from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2743438&tool=pmcentrez&rendertype=abstract.

Johnson, D. W. & Johnson, R. T. (2009). An educational psychology success story: social interdependence theory and cooperative learning. *Educational Researcher*, *38*, 5, 365–379. doi: 10.3102/0013189X09339057.

Kellogg, S. B., Booth, S. & Oliver, K. M. (2014). A social network perspective on peer support learning in MOOCs for educators. *International Review of Research in Open and Distance Learning*, *15*, 5, 263–289.

Kleiman, G. & Wolf, M. A. (2015). *Going to scale with online professional development: the Friday Institute MOOCs for Educators (MOOC-Ed) Initiative*. Retrieved 5 February 2015, from http://www.mooc-ed.org/wp-content/uploads/2015/02/going-to-scale-with-oTPD.pdf.

McAuley, A., Stewart, B., Siemens, G. & Cormier, D. (2010). *The MOOC model for digital practice*. Retrieved 27 March 2015, from http://www.elearnspace.org/Articles/MOOC_Final.pdf.

McFarland, D. A., Diehl, D. & Rawlings, C. (2011). Methodological transactionalism and the sociology of education. In M. T. Hallinan (Ed.), *Frontiers in sociology of education* (pp. 87–109). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-007-1576-9.

Rivera, M. T., Soderstrom, S. B. & Uzzi, B. (2010). Dynamics of dyads in social networks: assortative, relational, and proximity mechanisms. *Annual Review of Sociology*, *36*, 1, 91–115. doi: 10.1146/annurev.soc.34.040507.134743.

Scott, J. P. (2000). *Social network analysis: a handbook*. Thousand Oaks, CA: Sage Publications Ltd. Retrieved 19 May 2013, from http://www.amazon.com/Social-Network-Analysis-A-Handbook/dp/0761963391).

Snijders, T. A. B., Pattison, P. E., Robins, G. L. & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, *36*, 1, 99–153. doi: 10.1111/j.1467-9531.2006.00176.x.

Wasserman, S. & Faust, K. (1994). H. Johnston & B. Klandermans (Eds), *Social network analysis: methods and applications* Vol. 8. New York, New York: Cambridge University Press. Retrieved 8 December 2012, from http://books.google.com/books?hl=en&lr=&id=CAm2DpIqRUIC&oi=fnd&pg=PR21&dq=social+network+analysis+wasserman&ots=HuKqrf1FS9&sig=TvzSk1RxqfY26gpOuS5LgMWvyVc.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: An introduction to Markov graphs and *p**. *Psychometrika*, *61*, 3, 401–425. Retrieved from http://link.springer.com/article/10.1007/BF02294547