

# Exploring Co-studied Massive Open Online Course Subjects via Social Network Analysis

<http://dx.doi.org/10.3991/ijet.v9i8.3581>

K. Jordan

The Open University UK, Milton Keynes, United Kingdom

**Abstract**—Massive Open Online Courses (MOOCs) allow students to study online courses without requiring previous experience or qualifications. This offers students the freedom to study a wide variety of topics, freed from the curriculum of a degree programme for example; however, it also poses a challenge for students in terms of making connections between individual courses. This paper examines the subjects which students at one MOOC platform (Coursera) choose to study. It uses a social network analysis based approach to create a network graph of co-studied subjects. The resulting network demonstrates a good deal of overlap between different disciplinary areas. Communities are identified within the graph and characterised. The results suggests that MOOC students may not be seeking to replicate degree-style courses in one specialist area, which may have implications for the future moves toward ‘MOOCs for credit’.

**Index Terms**—Curricula, Open education, Massive Open Online Courses, MOOCs, Social network analysis.

## I. INTRODUCTION

In the past two years, massive open online courses (MOOCs) have entered the mainstream, attracting several million students [1] and garnering intense media attention.

One of the key characteristics of massive open online courses is the removal of entry pre-requisites to courses [2], allowing students to formulate their own learning pathways, free of the constraints of a modular degree programme. This may be liberating but also potentially problematic for students in order to determine how to fit individual courses together into a coherent whole. Progress is being made on this issue, from bundling individual courses together into ‘specializations’ at Coursera [3] to moves to translate entire subject curricula into the MOOC environment [4]. However, it is not necessarily safe to assume that all MOOC students seek to replicate traditional degree courses in a single subject area through their engagement with MOOCs.

This study seeks to explore the patterns in enrolment of MOOC students on different courses, through social network analysis of courses which Coursera students with public profiles are enrolled upon. The key question is when the entry pre-requisites for courses are removed, do MOOC students stick to courses within a subject area or develop new inter-disciplinary subject areas with their studies?

## II. METHODS

In order to explore which MOOC courses are studied together, a social network analysis approach was taken. Social network analysis conceptualizes individuals as

nodes, which will be connected by edges if a relationship exists between two nodes [5, 6]. In applying this methodology to the question of co-studied MOOC subjects, different courses would be represented as nodes in a one-mode network; an edge is then present between two nodes if one student has enrolled on both courses. This is similar to the approach taken in recommender systems based on purchasing information (for example, book purchases via Amazon [7]). An example of how this would be applied to co-studied courses for three hypothetical students is shown in Figure 1. In scaling up from this to the whole sample, additional courses would be added and edges weighted to reflect the number of students who had enrolled on pairs of courses.

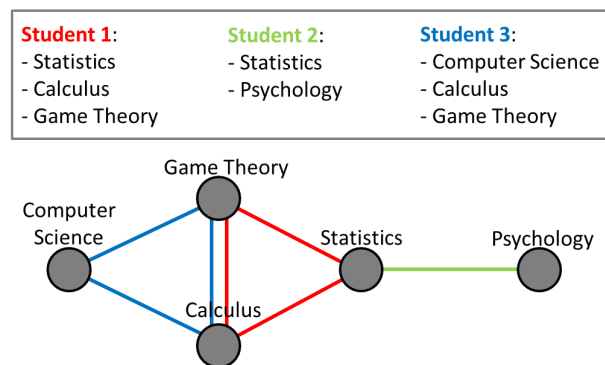


Figure 1. An example of translating the subjects co-studied by three hypothetical students into a social network graph representation.

Data was collected from public Coursera profiles, which list the courses a student had enrolled in. Note that profiles are not public by default; a student must actively opt to make their profile public. As there is no facility within the Coursera website to search for students’ profiles, the sample was identified by internet search. Public profiles were found by searching Google for part of the URL used by profiles, restricted to the Coursera site, using the following search query: "user/i" site:coursera.org . This yielded a total of 287 public profile pages as results. Using public profiles was necessary as it is the only way at present to find this type of data, although it does bring limitations with it. Only a very small proportion of Coursera users appear to have public profiles; at the time of data collection (2nd August 2013), the Coursera website stated a total of 4,262,759 students were registered with the site; 287 public profiles represents a small minority. Students who chose to make their profiles public are not necessarily representative of the whole student body, as their reasons for opting to be public are unknown, and might be self-selecting more active users. Having enrolled on a course is not indicative of whether students actively

engaged with the course materials, although enrolled student numbers is a good predictor of active users (with 50% of enrolled students typically becoming active users) [1].

Since the Coursera Terms of Service prohibit use of web scrapers [8], information about the number of courses and topics a student is enrolled in were collected manually and entered into a spreadsheet. Data was collected on 1st August 2013. In instances where students were enrolled in multiple iterations of the same course, this was only counted as one course. Of the total 287 profiles, three were excluded from further analysis as they belonged to Coursera staff. Distribution of the number of courses studied by the remaining 284 students is shown in Figure 2.

Given the distribution shown in Figure 1, students enrolled in more than 30 courses were excluded from further analysis. Students who were enrolled in zero or a single course were also excluded, as this is insufficient to be able to create an edge in the network. As a result, 201 student profiles were included in the final sample for constructing the network graph. The lists of courses each student is enrolled upon were then rearranged to make pairs of co-studied courses; an undirected link between courses indicating that one person signed up to both courses. Duplicates were allowed in order for a weighed graph to be produced. The spreadsheet was imported into Gephi [9] in

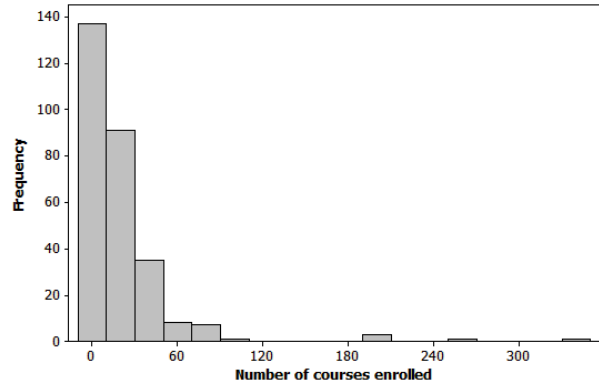


Figure 2. Histogram showing distribution of number of different Coursera courses students are enrolled in.

order to visualise and explore the resulting network, which comprised 301 courses (nodes), and 8175 edges. The modularity algorithm was used in order to detect communities [10]. Categorical data relating to each course was also added in terms of traditional subject classification, in order to examine the extent to which emergent communities follow these classifications. Subject areas used were as defined by the Coursera course list. Where a course fell into multiple areas, a judgment was made as to the primary focus; those which fell into more than four areas were classified as such.

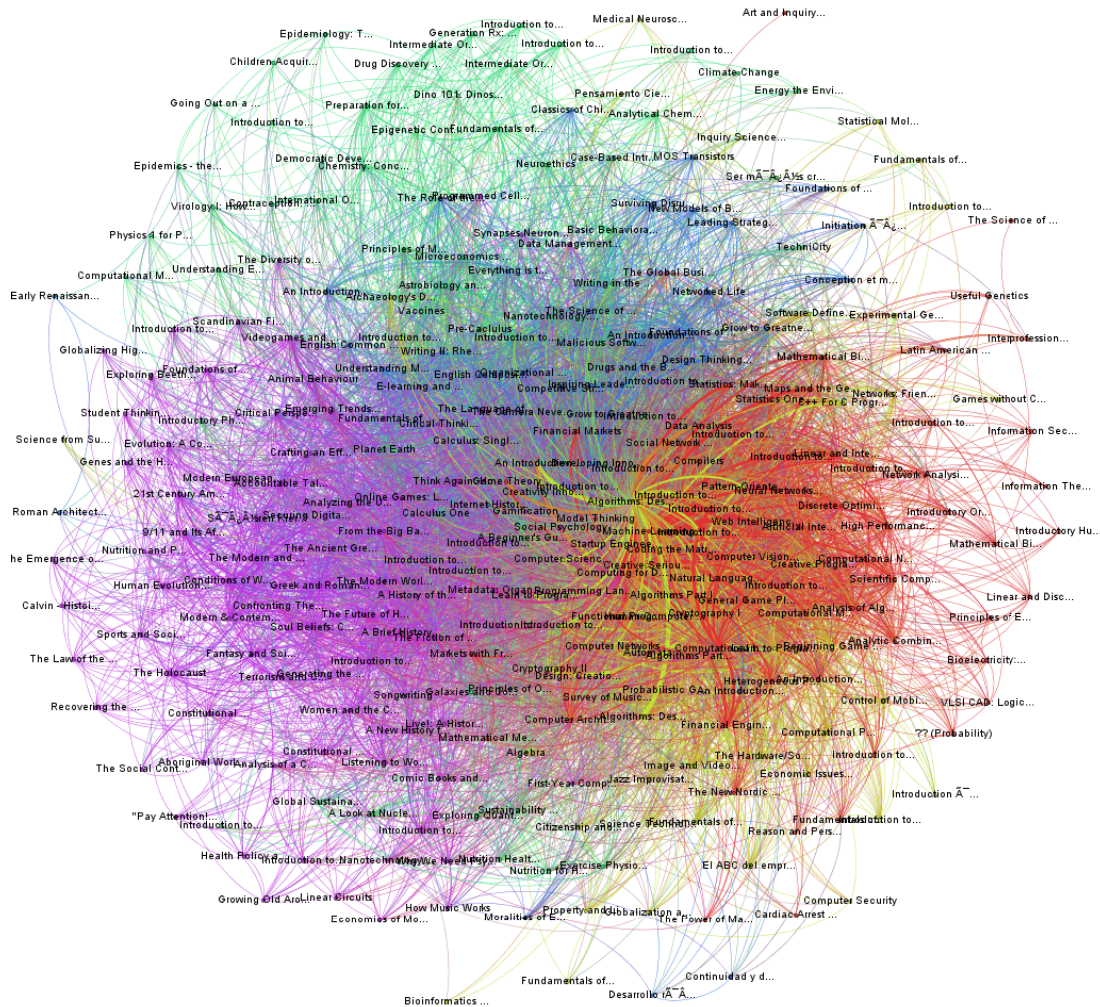


Figure 3. Network graph of co-enrolled Coursera subjects, colour-coded according to community. Courses belonging to ‘Community 0’ as shown in red; ‘Community 1’ in yellow; ‘Community 2’ in green; Community 3’ in blue; and Community 4’ in purple.

### III. RESULTS

#### A. Whole network and community structure

The network graph of co-enrolled subjects is shown in Figure 3. There is a great deal of inter-connection in the graph; distinct communities are not obviously present. The community detection algorithm identified five communities, and nodes and edges are colour-coded according to the categories they were assigned to by the community detection algorithm. Note that an interactive version of Figure 3 (created using the SigmaExporter plugin for Gephi [11]) can be found online at <http://www.katyjordan.com/MOOCnetwork/>. In order to characterize the disciplinary make-up of the five communities identified within the network, the frequency of courses in different subject areas in each community are shown in Figure 4.

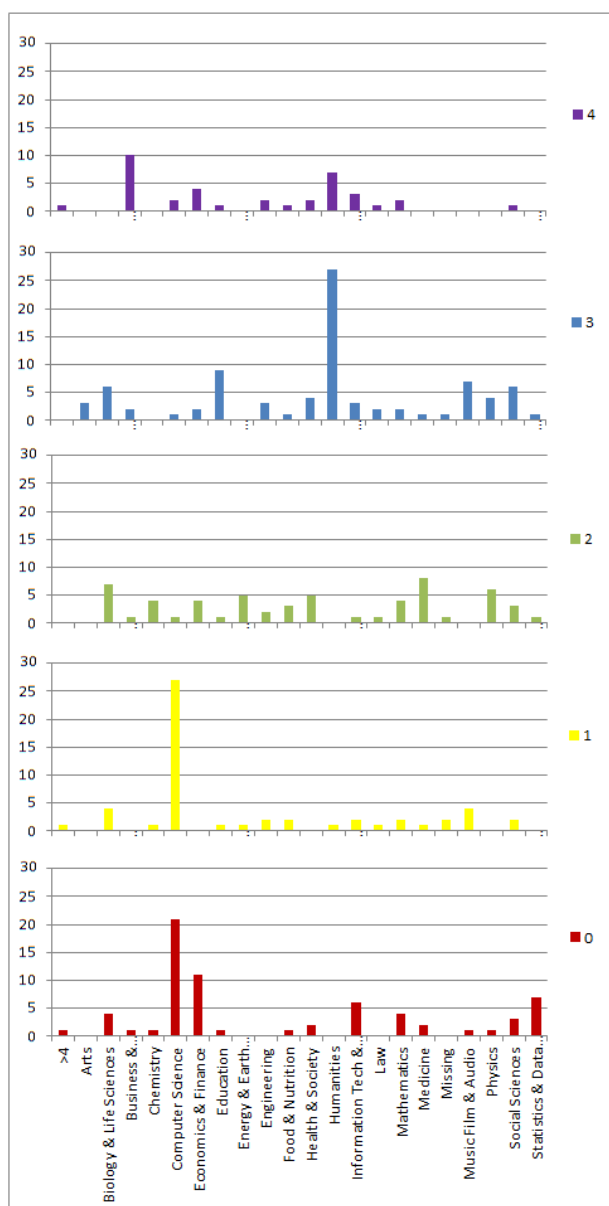


Figure 4. Bar charts showing the number of courses from each subject area represented in each community of the network. Numbers and colour-codes correspond to the communities illustrated in Figure 3.

Two of the communities (communities 0 and 1) are dominated by Computer Science courses, but differ in terms of the subjects these courses are co-studied with. Community 1 represents a more exclusively Computer Science subject community, while Community 0 is more interdisciplinary, allying Computer Science with other subjects, principally Economics and Finance, Statistics and Data Analysis, and Information Technology and Design. In contrast, Communities 3 and 4 are more strongly represented by the Humanities. In Community 3, the Humanities are allied with Social Sciences and Arts subjects, while Community 4 combines Humanities with Business-oriented subjects. Community 2 is the most interdisciplinary community, with a wide range of subject areas across the Natural and Physical Sciences represented and no single dominant area emerging. Although this gives an impression of the general focus of each community, it is also important to note that a wide range of subjects are present in every community to an extent.

#### B. Position of individual courses in the network

Basic social network analysis metrics were also used to examine which individual courses occupy notable positions within the network structure. The metrics used included weighted degree (which reflects the number of times a particular course has been studied within the sample) and betweenness centrality (a measure which reflects “the extent to which an individual node plays a ‘brokering’ or ‘bridging’ role in a network [12, p.75]). The ten courses with the greatest weighted degree are shown in Table I, and those with the greatest betweenness centrality shown in Table II.

There may be a relationship between weighted degree and time, as the majority of courses in Table I first ran in 2011 or 2012, so were relatively early established courses. It is logical that the earliest courses would have a higher weighted degree, being active for a longer period of time including a period when there were fewer courses to choose from. The courses demonstrating the greatest betweenness centrality (Table II), however, are notable for including subjects which span disciplinary areas (for example, Social Psychology and Startup Engineering) or are transferable to a range of different settings (for example, Think Again: How to Reason and Argue, and several data analysis courses).

### IV. DISCUSSION

In applying social network analysis to the MOOC courses which are co-studied by students with public profiles at Coursera, this study has identified communities of subjects which tend to be chosen together by students. In contrast to formal education, MOOC students are not restricted in their choice of courses according to a particular subjects’ syllabus. This is reflected in the network graph, which shows a good deal of overlap between the courses, and a broad range of subjects being present in all of the communities identified, to an extent. This interdisciplinarity character of the communities may be considered an example of how openness can lead to unusual behavior, in contrast to the disciplinary organization of formal Higher Education. This may pose a challenge for moves towards gaining credit for MOOCs and students who do not restrict their studies to a particular discipline. Whether this matters or not in terms of what students seek

TABLE I.  
COURSES WITH GREATEST WEIGHTED DEGREE VALUES IN THE NETWORK

Course	Institution	Date course first began	Weighted degree
Machine Learning	Stanford University	10/2011	618
Introduction to Data Science	University of Washington	01/05/2013	593
Computing for Data Analysis	Johns Hopkins University	24/09/2012	478
Startup Engineering	Stanford University	17/06/2013	431
Data Analysis	Johns Hopkins University	22/01/2013	399
Gamification	University of Pennsylvania	27/08/2012	354
Functional Programming Principles in Scala	École Polytechnique Fédérale de Lausanne	18/09/2012	349
An Introduction to Interactive Programming in Python	Rice University	15/10/2012	338
Algorithms Part I	Princeton University	12/08/2012	326
Algorithms: Design and Analysis Part I	Stanford University	12/03/2012	319

TABLE II.  
COURSES WITH GREATEST BETWEENNESS CENTRALITY VALUES IN THE NETWORK

Course	Institution	Date course first began	Betweenness centrality
Gamification	University of Pennsylvania	27/08/2012	1993.6
Machine Learning	Stanford University	10/2011	1389.0
Think Again: How to Reason and Argue	Duke University	26/11/2012	1319.5
Introduction to Data Science	University of Washington	01/05/2013	1214.8
Startup Engineering	Stanford University	17/06/2013	1149.9
Computing for Data Analysis	Johns Hopkins University	24/09/2012	913.8
An Introduction to Interactive Programming in Python	Rice University	15/10/2012	908.52
Data Analysis	Johns Hopkins University	22/01/2013	859.5
Social Psychology	Wesleyan University	12/08/2013	808.0
Model Thinking	University of Michigan	20/02/2012	685.5

to gain from participating in MOOCs is a subject for further research, and the implications in turn for formal curriculum design is an open question.

A central subject area emerged within each community, although this varied according to how broad it is in scope; for example, Computer Science dominated in Community 1 (and Community 0, allied with other subjects), while Community 2 represented the whole range of Natural and Physical Sciences. It is not clear whether this represents a shift in disciplinary boundaries, students' priorities, or reflects the types of subjects which lend themselves best to learning in a MOOC context. A social network approach such as this could provide the basis for a recommender system in order to assist students in finding their learning pathway within new emerging interdisciplinary

areas. The relationship between the extent of interdisciplinarity in a students' course choices and their likelihood of completion may be an interesting contribution to the hotly debated topic of MOOC completion rates.

This study has provided an insight into the emerging communities of subjects between individual MOOC courses, which had previously been unexplored. It is also restricted to a single MOOC platform; the ways in which students study across multiple MOOC platforms would also be an interesting area for future research. The results here only provide a snapshot of the emerging disciplinary communities; in practice, the network of subjects is dynamic. As the courses with the highest degree (reflecting those which the greatest number of students in sample signed up to) are frequently the earliest established courses at present, there is currently a skew toward these courses, which may be responsible for the dominance of Computer Science at present. As the number of courses available continues to proliferate and the number of MOOC students increases, it will be interesting to see how communities evolve over time; a wider range of communities is likely to emerge, but it remains to be seen whether these will be interdisciplinary or return to traditional disciplinary areas.

## REFERENCES

- [1] K. Jordan, "Initial trends in enrolment and completion of Massive Open Online Courses," *The International Review of Research in Open and Distance Learning*, 15(1), 133-160, 2014.
- [2] J. Daniel, "Making sense of MOOCs: Musings in a maze of myth, paradox and possibility," *Journal of Interactive Media in Education*, 2012.
- [3] S. Kolowich, "Coursera will offer certificates for sequences of MOOCs," *The Chronicle of Higher Education*, 2014, <http://chronicle.com/blogs/wiredcampus/coursera-will-offer-certificates-for-sequences-of-moocs/49581>, retrieved 22<sup>nd</sup> January 2014.
- [4] P. Hill, "Two MOOC curriculum announcements in one week", *e-Literate Blog*, 2013, <http://mfeldstein.com/two-mooc-curriculum-announcements-in-one-week/>, retrieved 22<sup>nd</sup> January 2014.
- [5] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press, 1994. <http://dx.doi.org/10.1017/CBO9780511815478>
- [6] C. Kadushin, *Understanding social networks: Theories, concepts, and findings*. Oxford: Oxford University Press, 2012.
- [7] J. Leskovec, L. Adamic and B.A. Huberman, "The dynamics of viral marketing", *ACM Transactions on the Web*, 1(1), 1-39, 2007. <http://dx.doi.org/10.1145/1232722.1232727>
- [8] Coursera website, "Terms of use", 2014. <https://www.coursera.org/about/terms>, retrieved 22<sup>nd</sup> January 2014.
- [9] M. Bastian, S. Heymann and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks", *International AAAI Conference on Weblogs and Social Media*, 2009.
- [10] V.D. Blondel, J-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment*, 2008, P1000, 2008.
- [11] S.A. Hale, "Build your own interactive network", *Interactive Visualizations blog*, 2012, <http://blogs.oii.ox.ac.uk/vis/?p=191>, retrieved 29<sup>th</sup> January 2014.
- [12] R. Ackland, *Web Social Science: Concepts, data and tools for Social Scientists in the digital age*. London: SAGE, 2013.

## AUTHORS

**Katy Jordan** is a doctoral student with the Institute of Educational Technology at the Open University UK (e-mail: [katy.jordan@open.ac.uk](mailto:katy.jordan@open.ac.uk)).

Submitted, February, 24, 2014. Published as resubmitted by the authors on May, 26, 2014.