Challenges to Research in MOOCs

Helene Fournier
Research Officer
Institute for Information Technology
National Research Council of Canada
Moncton, NB E1A 7R1 CANADA
helene.fournier@nrc-cnrc.gc.ca

Rita Kop
Dean
Faculty of Education
Yorkville University
Fredericton, NB E3C 2R9 CANADA
rkop@yorkvilleu.ca

Guillaume Durand
Research Council Officer
Institute for Information Technology
National Research Council of Canada
Moncton, NB E1A 7R1 CANADA
guillaume.durand@nrc-cnrc.gc.ca

Abstract

Over the past five years, the emergence of interactive social media has influenced the development of learning environments. Learning management systems have come to maturity, but because they are controlled by educational institutions and are subsequently used to support institutional learning, have been seen by learning technologists as not capturing the spirit and possibilities that new media have to offer for learning. Academics and researchers are currently investigating a different learning environment, more open and networked, while the underpinning learning theory is moving from social constructivism towards connectivism. Research in open learning environments is only in its infancy and researchers have only started to become interested in massive open online courses (MOOCs) as a topic of investigation. Recent research and development efforts have focused on generating technologies that might facilitate learning within a self-directed information and communication stream. In this paper, the authors report on an exploratory case study of PLENK, a connectivist-style MOOC, and highlight some of the challenges in the research and analysis process, especially as significant amounts of both quantitative and qualitative data were involved. Important findings related to activity levels and important dimensions of self-directed learning in an open learning environment are presented.

Keywords: massive open online course (MOOC), connectivist massive open online course (cMOOC), connectivism, activity level, learner motivation

Introduction, Background, and Research Design

The New Media Consortium's annual *Horizon Report* forecasts the expected trends in technology and education and learning. Several points are highlighted in the 2011 edition of the report (Johnson, Smith, Willis, Levine, & Haywood, 2011), such as the challenges of ubiquitous information: "The abundance of resources and relationships made easily accessible via the Internet is increasingly challenging us to revisit our roles as educators in sense-making, coaching, and credentialing" (p. 1), and the expectation of mobility that the technology affords: "People expect to be able to work, learn, and study whenever and wherever they want, that the world of work is increasingly collaborative, giving rise to reflection about the

way student projects are structured, that the technologies we use are increasingly cloud-based, that the perceived value of innovation and creativity is increasing" (p. 7).

Moreover, the European Union reports on specific trends for education and learning, such as the rise of informal learning, personalization and collaboration. The structure of the learning environment, the place and presence of learners and educators within institutional boundaries, the nature of knowing and learning are all challenged by the fast pace of technological change (Weller, 2010). Emergent technologies offer different models and structures to support learning. They disrupt the notion that learning should be controlled by educators and educational institutions, simply because information and "knowledgeable others" are readily available at the press of a button for all who are interested in expanding their horizons.

Tenets of emergent theories of knowledge and learning, such as connectivism, argue that online social networks can help interpret and validate information (Downes, 2007; Siemens, 2006). They promote a learning organization whereby there is not a body of knowledge to be transferred from educator to learner, and where learning does not take place in a single environment. Rather, it is distributed across the Web and people's engagement with it constitutes learning (Bell, 2011).

Since 2009, the National Research Council of Canada (NRC) has been engaged in the research and development of a pedagogical platform that could support networked learning in all its facets outside formal education. Such an environment, referred to as a personal learning environment (PLE), would combine (intelligent) information streams with editing and publishing tools, and also provide scaffolding, communication, and support structures for learners. One component of the PLE research has focused on information gathering and investigating educational issues in massive open online courses (MOOCs). Research has been conducted in the context of two connectivist MOOCs (cMOOCs), namely Critical Literacies (CritLit) and Personal Learning Environments, Networks, and Knowledge (PLENK). cMOOCs are based on a number of principles stemming from connectivist pedagogy, including aggregation, remixing, repurposing, and feeding forward ("Massive Open Online Course," 2013). Surveys were used to gather information in the CritLit MOOC and findings were published on important factors to consider in the design and development of a pedagogical platform. Results from the survey pointed to important elements that would make learners think critically about resources accessed (Fournier & Kop, 2011). Educational research was pursued in the context of a second MOOC, namely PLENK. This MOOC was scheduled over 10 weeks from September to November 2010, and was offered as a joint venture between the NRC and the Technology Enhanced Knowledge Research Institute (TEKRI) at Athabasca University. In the remainder of this paper, the focus is on findings from a case study of PLENK as an exploration of learning and activity levels in an open learning environment.

Learner Motivation and Self-Directed Learning

In an open online learning environment, the control of learning no longer rests with an educational institution but with the learners themselves. It is argued that there are a number of factors influencing the success of learning in such an environment (Bouchard, 2009). Bouchard clustered the factors into four dimensions: one dealing with psychological issues, one with pedagogical issues, and two with contextual matters. The first dimension, *conative*, relates to psychological issues such as drive, motivation, initiative and confidence. The *algorithmic* dimension relates to pedagogical issues, for instance the sequencing, pacing and goal setting in learning, and the evaluation of progress and final evaluation. These are clearly tasks that in the past were carried out by the educator, but in an autonomous learning environment are issues that learners have to resolve themselves. The dimension that Bouchard called the *semiotics of learning* is an environmental factor related to the delivery model of resources. This dimension of learning has drastically changed in recent years and moved from the use of resources such as books and paper to electronic texts and multimedia, which might be stored in searchable databases and linked through hyperlinks. It could also include contributions in blogs, wikis, and synchronous and asynchronous communication. Information is obtained through social networks and learners will need to be able to evaluate and navigate this new information landscape. The *importance of economy* was recognized as a fourth dimension of learning, which includes the perceived and actual value of the learning, the choice to learn for personal gain (such as for future employment), and the possible cost of other study options.

Bouchard (2009) saw these factors as related to self-directed learning, and the authors' previous research showed that these factors might influence each other (Kop & Fournier, 2011). When examining the concept of motivation, for instance, the non-psychological factors influence the psychological ones,

such as the level of stimulus to participate in the learning. Research by Hartnett, St. George, and Dron (2011) confirms that student motivation is "not a one-dimensional trait, but is complex, multifaceted, and influenced by both person and context" (p. 31).

As was highlighted by Bouchard (2009), the onus is on the learners to organize their learning activities, to match them both to the best possible technologies to use and to the most suitable persons to communicate with on their social network in order to reach their learning goal. Analysis of PLENK data, according to Bouchard's four dimensions of learning, will be presented in the qualitative data collection and analyses section.

New Ethics and Privacy Issues in Networked Environments

Every researcher has to consider the ethical implications of the chosen methods of obtaining data for a study as well as the use of the data. Sometimes obtaining data is a matter of accessing statistics or documents. When human subjects are involved in the research, careful consideration of the level of informed consent by participants is also required. Miller and Bell (2002) argued that gaining informed consent is problematic if it is not clear what the participant is consenting to and where "participation begins and ends" (p. 53). Several ethical issues were raised in the literature, of which misuse of data and privacy issues were the most important. Boyd (2010) and van Wel and Royakkers (2004) caution that data could pose a threat to subjects when either misused or used for different purposes than for which it was supplied. Researchers should at least anonymize data in order to respect privacy issues (Boyd, 2010; Rogers, McEwen, & Pond, 2010; van Wel & Royakkers, 2004). It has also been suggested by network researchers that people should have the choice to opt in or opt out of the use of their data. If someone is not aware that the data is being collected or how it will be used, that person has no real opportunity to consent or withhold consent for its collection and use. This "invisible data gathering" is common on the Web (van Wel & Royakkers, 2004, p. 133) and highlights some new decisions related to ethics that researchers will have to make. Researchers have a responsibility to carefully consider the context of their research, and also the process that takes place between observing, collecting and analyzing "big data" – data that is left by traces of activities that might not at all be related to the visible participation of learners.

In this study big data was captured out on open networks. The research team set out the boundaries of the research on the consent form that participants were asked to read at the start of the course. They were informed that data collection would include learning-related activities in the course environment and also learning activities that happened outside the course, where the course hashtag "#PLENK2010" was being used.

Data on PLENK 2010 were collected according to the following principles: using quantitative as well as qualitative measures, asking for informed consent, and using the course hashtag to identify course-related data outside the course environment included in the research and specified in the consent request.

Results

Participation Levels

When the PLENK course started, 846 participants had registered, and this number steadily increased to 1641 at the end of the course, as shown in Figure 1. Twice-weekly meeting sessions were hosted on Elluminate Live! (now known under the name Blackboard Collaborate, an online learning and collaboration platform): once a week with an invited speaker and once as a discussion session amongst the group and facilitator(s). Actual presence at these synchronous sessions decreased over the weeks from 97 people in the second week, when attendance was the highest, to 40 in the final week with a similar downward trend in accessing recordings for the sessions. Global participation and multiple time zones influenced who could be present during the live sessions and who accessed the archived recordings after the fact. A high number of blog posts were generated related to the course (900) and an even higher number of Twitter contributions (3,104).

The "#PLENK2010" identifier facilitated the easy aggregation of blog posts, social bookmarking links (such as Delicious), and Twitter messages produced by participants (which highlighted a wide number of resources and links back to participant's blogs and discussion forums, thus connecting different areas of the course). Although the number of course registrations was high, an examination of contributions across weeks (i.e., Moodle discussions, blogs, Twitter posts marked with the "#PLENK2010" hashtag,

and participation in synchronous web-conferencing sessions through Elluminate Live!) suggested that on average, about 40 to 60 individuals actively contributed to the course on a regular basis by producing blog posts and discussion posts, while the remaining participants' visible participation rate was much lower. Figure 1 helps to illustrate the number of times people used particular tools, but these statistics do not provide insight into the factors that contributed to their use. Survey data were thus explored in greater depth for additional information on the factors contributing to participation in the PLENK MOOC.

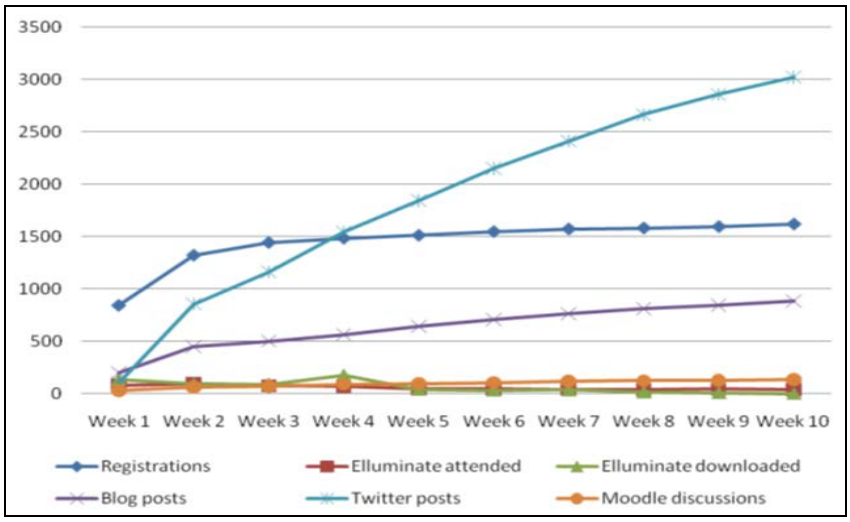

Figure 1. *PLENK participation rates*

Survey Results

Three surveys were carried out at the end of the course in order to capture and explore various factors affecting learning experiences in PLENK, namely: the end of course survey (*n* = 63), an active producers' survey (*n* = 32) that was filled out by participants after an invitation was posted in the course blog for people who had produced more than two digital artifacts, and a lurkers' survey (*n* = 74) that was filled out after a similar call for people who had limited their participation in the course to producing less than two digital artifacts and whose behavior was characterized as "consuming" rather than "participating."

Active participants in the PLENK MOOC reported on what urged them to produce something in the course – for example, a blog post, Moodle discussion post, video, word cloud, or concept map. Figure 2 highlights some of the important factors which prompted greater participation in the course, including: discussions posted by someone else (64%), blog post from someone else (54%), or someone connecting difference concepts that made me want to produce something (50%). Participants (39%) also offered additional remarks on factors that urged them to participate in an open-ended comment sections, including: *"the need for self-reflection," "inspiration by the connections I was making," "recommendations,"* as well as the *"example of others."* In a survey addressing lurkers in the PLENK MOOC, participants selected the following responses as factors that contributed to their silence (or non-active participation), including: "time restrictions, job, family, and other commitments impacted on my ability to particpate actively in the course" (76.8%), "being more of a listener and reflector when it comes to subject matter so sitting back was comfortable for me" (35.4%), and "feeling that lurking is a legitimate learning strategy and it is the way I want to continue to participate in MOOCs" (31.7%).

Among the most important factors in helping participants to learn were "receiving feedback from a knowledgeable person" (68%) and "reading material related to the learning activity" (64%). Participants also reflected on how the PLENK MOOC affected their learning.

Figure 3 presents findings from the end of course survey, which was used to collect data highlighting factors that had the most impact on learning. Factors such as "freedom to do and read as I felt like" (81%) and "how the course was organized" (76%) were important for the majority of participants.

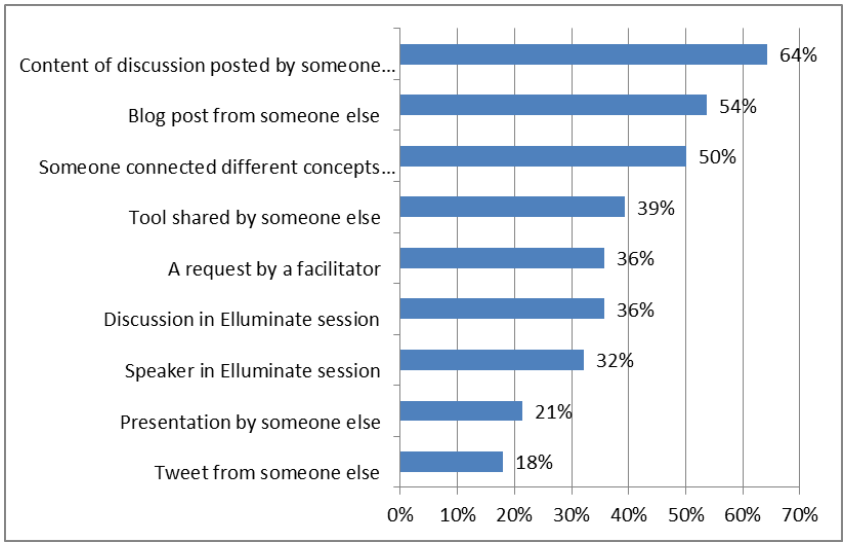

Figure 2. *Factors that urged participation among active PLENK MOOC participants*

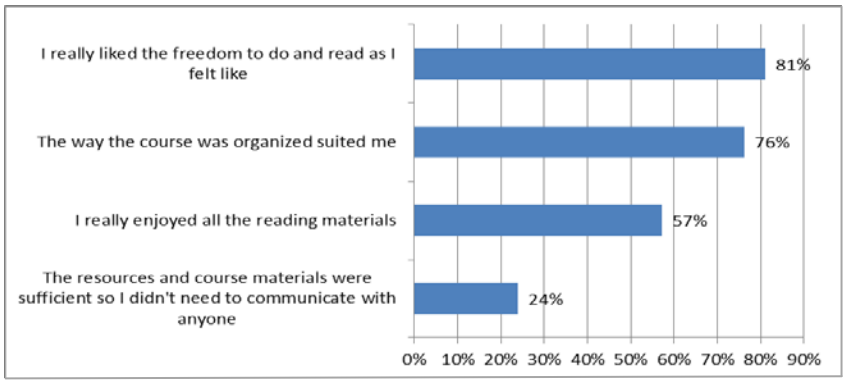

Figure 3. *Factors that had the most impact on learning for PLENK MOOC participants*

Also, 26% of survey respondants offered additional open-ended responses with regards to the factors that had the most impact on their learning. A sampling of comments include: *"More feedback on my input* would *have been useful," "time zones wiped out any proper chance to participate in the live sessions & this lack was very important to me & the way the course developed for me," "volunteer helpers to comment on blog posts and discussions would have been motivating,"* and *"some organization to develop sub-groups based on various backgrounds/interests would also have been helpful."*

Virtual Ethnography and Qualitative Research Results

Qualitative methods in the form of virtual ethnography were used to further explore rich data from Moodle discussion and blogs. An ethnographer was working on the course, collecting qualitative data through observation of activities and engagement. She also interviewed and surveyed a number of participants during the final week and held a focus group with lurkers after the course to gain a deeper understanding of particular issues related to the active participation of learners. As part of the educational research on open online learning environments, an examination of the processes taking place and the perspectives and understandings of the people in the setting – what Denzin (1989) describes as the "details, context, emotion, and the webs of social relationships that join persons to one another" (p. 83) – was conducted. Hine (2005) stresses that on the Web, the technology itself and the artifacts it produces should be taken into consideration in the online ethnography as these are part of the research setting and might influence the human interactions researched. As vast amounts of discursive data were generated in this form of networked learning in an open environment, computational tools such as NVivo were used for analyses and interpretation of the qualitative research data. It was fairly easy to capture vast amounts of qualitative data through the aggregation tools, software that was used

to collect and distribute course related digital artifacts and other information for/from participants, such as the gRSShopper aggregator (Downes, 2008).

Discourse Analysis of Moodle Forum Discussions and Blogs

The analysis of the discursive data generated by PLENK 2010 participants was challenging because of the volume of data. The aim was to use this data to gain in depth insights into the learners' experiences, such as their motivation, which made it necessary to use qualitative analysis measures. NVivo software was chosen, which is promoted as a technological application capable not only of being able to analyze the data (Bazeley, 2007), but also to possibly enhance validity and transparency. However, there are also researchers who highlight the restrictive nature of the search functions of computerized tools, such as NVivo (Welsh, 2002), that use quantitative measures, namely counting particular words, rather than interpreting them as a human researcher might do.

It could be argued that a combination is possible, and, that in the present case, with a high volume of data, there was no other choice than to utilize a computer program to aid in organizing the data and increase rigor by coding all data systematically (in particular themes) in such an environment. Once the coding was finished, connections were made between nodes in order to interrogate the data. Coding was accomplished by two researchers to allow for comparison of the coding for similarity. Coding was initially completed in a free-flowing manner. After all coding of the Moodle discussion forum messages, blog posts and comments, and Twitter messages had been carried out, they were organized in nodes and sub-nodes according to Bouchard's (2009) four dimensions of self-directed learning: conative, algorithmic, semiotic, and economic. This framework was well suited to the analysis of dimensions related to learning a MOOC. The method used to categorize and code the data – that is, in Moodle forum discussions and blogs – originates from grounded theory literature (Glaser & Strauss, 1967; Strauss, 1986), content analysis (Babbie, 1979), and qualitative data analysis (Bryman, 1988).

Table 1 presents examples of content coded under conative factors, which include psychological issues related to self-directed learning in Bouchard's (2009) model. Semiotic factors related to delivery models of resources in self-directed learning, including factors such as social networks, search activity, and artifacts, were collapsed into "social interaction."

Table 1. *Conative factors related to self-directed learning*

Factors related to psychological issues	Affective issues (sub-themes: frustration, interest, suffering), behavior (sub-themes: efficiency, focus, cognitive flexibility), competency, confidence, critical literacies, learning styles, motivation, past experience, among others		
Coding: Themes and sub-themes	Content	Contributor	Coded Segment (excerpts)
Motivation	Moodle forum discussion	Facilitator	*"I put the 'strong motivation' in the top of my list of personal requirements to build and use successfully a PLE/Ns."*
Past experience	Blog	Participant	*"I have a long history of online teaching (facilitating actually) and now I want to find online places where I enjoy communicating and learn more."*
Reflective	Blog	Participant	*"I'm not convinced either way at the moment, but there is no doubt that the PLENK course was the best professional development I've experienced in many years."*

Algorithmic factors, such as teaching, peer support, practice, and assessment, were collapsed under "pedagogy and learner support," while sequencing, pacing, navigation, and requirements amongst other sub-themes were collapsed under the broader theme of "structure and organization." Content was coded under economic factors when statements were made about the perceived and actual value of the learning, the choice to learn, and reasons for learning (i.e., personal gain, such as future employment), and statements about cost and the value of other alternatives.

A fifth general coding category evolved out of the initial data analysis, which allowed for coding of factors which went beyond the scope of Bouchard's (2009) four dimensions of self-directed learning, including

broad sub-themes such as agency, learning, education, knowledge, factors related to the learning environment and design, and issues related to theory and research.

Overall, discussions around general factors were predominant in Moodle forum discussions (total coded segments = 1,332) and included themes such as the learning environment and design, agency, learning, theory and research, as well as education (see Figure 4).

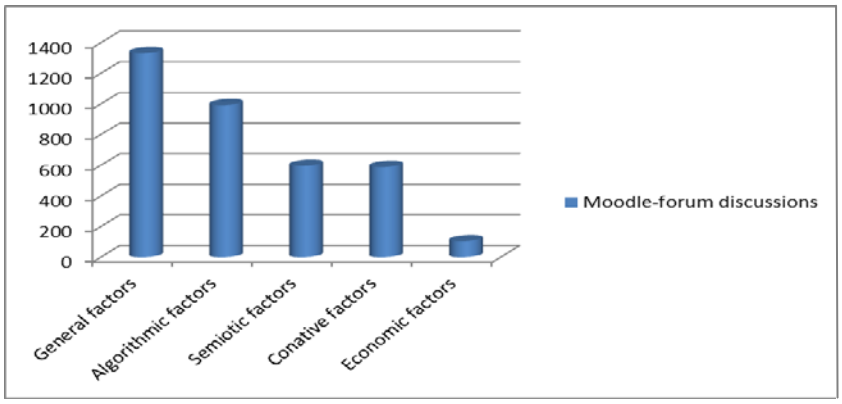

Figure 4. *Overall counts for coded segments in Moodle forum discussions*

Discussions around algorithmic factors (total coded segments = 991) in the context of Moodle forum discussions included sub-themes such as tools and resources, pedagogy and learner support, as well as structure and organization. Semiotic factors accounted for 595 coded segments, followed closely by conative factors at 588. Lastly, economic factors accounted for only 103 of the total coded segments in Moodle forum discussions (Figure 4).

The data generated by 1,641 registered PLENK 2010 participants amount to thousands of potential data points to explore, which is well beyond the scope of the current study. Instead, a representative sample of nine participants (see Table 2), focused on Moodle forum discussions and blog contributions, was selected for closer examination and exploration.

Table 2. *Focus group participants*

Participant	Age	Computer Skills	Country	English Proficiency	Gender	MOOC Experience	Profession
1	37-42	High	Canada	High	Female	No	Post-secondary
2	43-48	High	U.S.	High	Female	Yes	Writer/ Editor
3	49-54	High	U.K.	High	Male	No	University Lecturer
4	49-54	High	U.S.	High	Male	No	Consultant/ Trainer
5	Over 55	Moderate	Australia	High	Female	Yes	Teacher/ Mentor
6	Over 55	Moderate	Australia	High	Female	Missing data	IT Teacher/ Librarian
7	Over 55	Moderate	Israel	Moderate	Female	Missing data	Tutor
8	Over 55	Moderate	Finland	Moderate	Female	Yes	Retired
9	Over 55	High	U.S.	High	Female	Yes	Librarian

This sample of nine participants was selected on the basis of completeness of the data and profile information, as well as for its representation of different age groups, gender, computer skills, and

previous MOOC experience. In addition, data from course facilitators' participation were coded and analyzed.

A sub-sample of nine participants and four facilitators were compared on their contributions to Moodle forum discussions and blog posts. Figure 5 reveals numbers for categories of coded content in order of importance: algorithmic factors were highest in participant blogs (978), semiotic factors were highest in facilitator blogs (454), general factors (i.e., theory and research, learning environment and design, agency) in facilitator blogs (357), and conative factors in participant blogs (340).

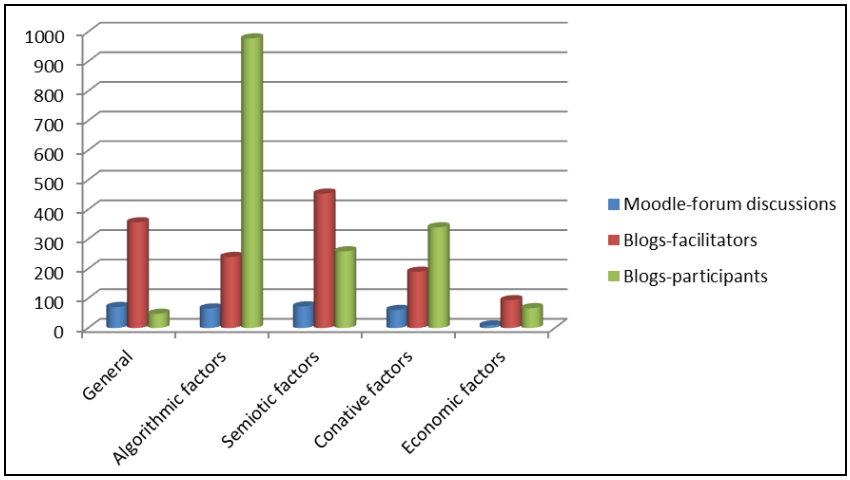

Figure 5. *Moodle forum discussions and blog activities for nine participants and facilitators*

Participants' blogs highlighted the importance of algorithmic factors such as pedagogy and learner support, and assessment. Facilitator blogs focused on semiotic factors such as the production of resources, communication and language, tags (keywords for search engines), and information and data. Facilitator blogs also included more discussion around general factors, including conversations around theory and research (111 coded segments), while participant blogs included more content related to conative factors such as motivation (162), past experience (20), and reflection (17). Since the control of learning in an open online learning environment rests in large part on the learners themselves and on psychological factors such as drive, motivation, initiative and confidence, these conative factors were examined in greater depth and are reported on next.

In addition, NVivo data analyses and exploration revealed strong positive correlations between conative factors such as past experience and intentions, motivation and intentions, reflection and intentions, as well as past experience and motivation across content in blogs and Moodle forum discussion. Table 3 presents some of the stronger positive correlations ranging from .5 to .9.

Table 3. *Strong positive correlations between motivation and related factors*

Strong Positive Correlation	Pearson Correlation Coefficient
Participant blogs: past experience and intentions	0.853
Participant blogs: motivation and intentions	0.844
Moodle discussion forums: reflection and intentions	0.771
Participant blogs: past experience and motivation	0.764
Participant blogs: past experience and Moodle discussion forums: reflection	0.717
Participant blogs: motivation and Moodle discussion forums: reflection	0.710
Participant blogs: social environnement and Participant blogs: intentions	0.548
Moodle discussion forums: reflection and Moodle discussion forums: learning styles	0.521

Further analyses of motivation through the NVivo group query function revealed that discussions hinging on motivation were widespread across participant and facilitator blogs and Moodle forum discussions, and in some cases were sustained over weeks. These findings prompted further exploration of participation data from the Moodle environment more specifically, as this was where the hub of activity was centered throughout the course and where important amounts of data records existed. The next section will explore results from educational data mining (EDM) efforts within the PLENK MOOC.

Educational Data Mining of the Moodle Environment

In order to explore differences in participation levels ever further, analyses of activity levels were conducted using Moodle logs. Moodle is an open-source learning management system allowing educators to manage learning by creating a shared space where participants could post material, participate in a forum, and post comments in a course wiki. Prior to enrollment in the online course, users had the opportunity to register by filling out a form. Each registered participant was asked to provide demographic information such as name, age range, gender, level of computer skills, language skills, MOOC experience, mother tongue, country, and e-mail address. The main objective of this exploratory study was to discover some links, beyond the course survey data, between the activity level of the learners, their motivation to learn, and the demographics, in order to exploit the obtained results in future sessions of the MOOC. Moodle logs all actions of registered users in a dedicated database table. At the time of the study, the table included 232,000 lines of MOOC-related activity (e.g., log identification, date and time of log record, user identification, course, type of module, and module identification) for 1,641 registered participants.

Mining of the Moodle data involved an iterative three-step process described by Romero and Ventura (2007) and Romero, Ventura, and García (2008), namely: data collection, data pre-processing, and data analysis (see also Mazza, 2010; Sheard, 2010). In our EDM study of the MOOC, the analysis process involved numerous iterations between data pre-processing and analysis. A first analysis using a high level activity quantifier was used; then a second analysis was conducted with different activity indicators. For both studies, 672 completed profiles created by users were collected. The loss of more than half of the created profiles in the study sample can be explained by the free registration process. Many users did not go through the registration process and/or did not fill in the information needed to complete their profiles and/or filled the forms with non-consistent values. The profile list was first verified, then a sample of 330 complete profiles was extracted. Table 4 offers some selected descriptive statistics of the sample.

Relationships between the different demographic variables are very weak, and a Pearson and Spearman correlation study yielded no significant linear relationships between the variables. This was confirmed by a "Mutual Information" study that highlighted at best a very weak relationship between country and mother tongue (I = 1.932, *bins* = 10).

Moodle logs contain users' activity tracks. An indicator called "Activity count" counted all the different lines of logs for each participant of the MOOC. Considering the 330 profiles of the initial sample, the indicator took values comprised within range of 4 to 4,021, with the mean being 174.48, the median being 56, and the standard deviation 428.57. The standard deviation in this case was large, and it indicated that the data points were far from the mean. In other words, the activity level of the users was highly variable from one user to another.

Figure 6 illustrates this trend by showing the different user activity counts and the average counted actions by age ranges.

Younger individuals, especially in the 25-30 age bracket, seem to have participated on average less than the older participants with a standard deviation by range that is lower. A preliminary statistical analysis was run, and no linear correlation between age and activity account was discovered.

No significant Pearson and Spearman correlations between demographic variables and the activity accounts were found. This result was unexpected since it would be natural to have a more significant relation between computer skills and the level of activity. After investigating further, it was discovered that most of the participants considered themselves to be at ease (maybe wrongly) with computers. The few participants (2) having difficulties had a lower count than the median value, but their participation was not significantly lower in comparison with other computer skills categories.

Figure 7 presents the relationship between the counted actions and the computer skills levels.

Table 4. *Descriptive statistics of the sample*

	Frequency	Percent
Age		
18-24	9	2.73
25-30	15	4.55
31-36	43	13.03
37-42	57	17.27
43-48	63	19.09
49-54	63	19.09
Over 55	80	24.24
Gender		
Male	171	51.82
Female	159	48.18
Computer skills		
Low	2	0.6
Moderate	85	25.76
High	243	73.64
English skills		
Low	8	2.42
Moderate	60	18.18
High	261	79.39
MOOC experience		
Yes	126	8.18
No	204	61.82
Mother tongue		
English	189	57.27
Other	141	42.73

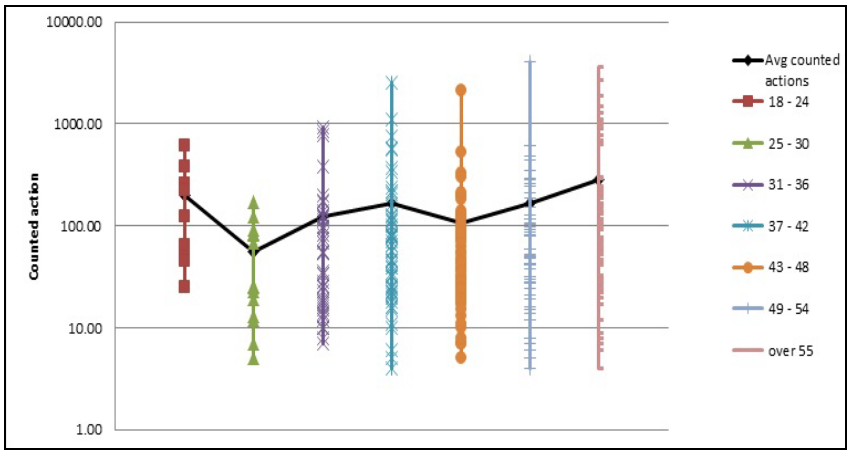

Figure 6. *Variability of activity by ages*

As mentioned in Romero et al. (2008), the clustering method can be interesting to identify groups of students. Analyses were undertaken to see if there were particular levels of participation for specific demographic groups. The discretization of the demographics variable in the MOOC included age, gender, computer and English skills, and MOOC experience and language, as presented in Table 5. The K-Means clustering algorithm was used, and the best results obtained by considering three clusters (k = 3, *max runs* = 10, *max optimization steps* = 100) were reported.

In order to use the K-Means algorithm, demographics data were discretized as follows. Counted action was discretized by size in four ranges or levels: [0, 25.5] (93 items), [25.5, 67] (88 items), [67, 188.5] (89 items), [188.5, +∞](60 items). Levels were identified by numbers going from 0 to 3 respecting range orders.

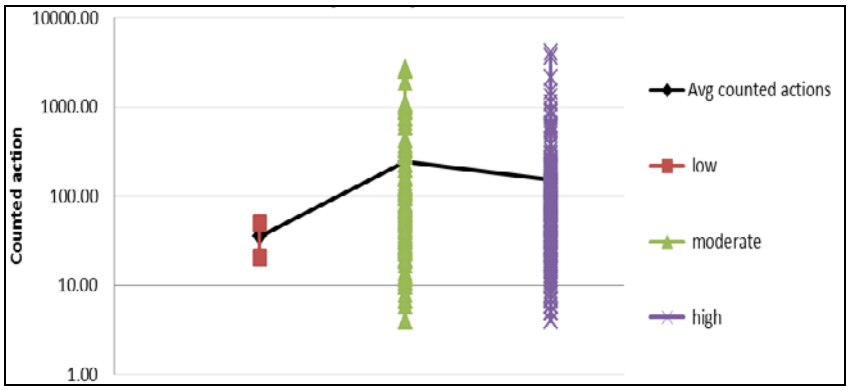

Figure 7. *Variability of activity by computer skills*

Table 5. *Discretization of the demographics variable*

	Discrete Value Assigned
Age	
18-24	1
25-30	2
31-36	3
37-42	4
43-48	5
49-54	6
Over 55	7
Gender	
Male	1
Female	0
Computer skills	
Low	0
Moderate	1
High	2
English skills	
Low	0
Moderate	1
High	2
MOOC experience	
Yes	1
No	0
Mother tongue	
English	1
Other	0

Table 6 shows Cluster 1 is very different from the two others and seems to be associated with moderately active participants (Counted Action Level (CAL) = 1.298). Contrary to other clusters, the Cluster 1 profiles seem to count mainly people in their thirties (3.193) being non English language natives (1.726). Cluster 2 and Cluster 3 were similar in terms of demographic results but were very different in term of activity (CAL = 0.535, CAL = 2.435). It is difficult with the information provided to distinguish between low performers and high performers. Even by changing the discretization of the measured activity (Number of Counted Action level) or by modifying the size of the clusters, high performers could not be isolated from low performers in terms of their measured activity and their demographics. The Moodle log table was particularly rich in information since it logged actions and was informative in terms of the type of actions and where the actions occurred (Table 7). The course Moodle activity logs allowed for sharper differentiation between activities like active participation from passive participation in the forum posts characterized by "surfing" activity in the course environment compared to the writing of a post.

Table 6. *Tabloid table of the three clusters*

	Cluster 1 (*n* = 124)	**Cluster 2** (*n* = 114)	**Cluster 3** (*n* = 92)
Counted Action Level	1.298	0.535	2.435
Mother tongue	0.403	0.693	0.652
Age	3.193	5.991	6.196
Gender	0.468	0.579	0.511
Computer skills	1.774	1.737	1.663
English	1.726	1.851	1.728

Table 7. *Actions log used from the Moodle log table*

- cnt_wiki
  - cnt_wiki_edit
  - cnt_wiki_viewcnt_wiki_link
- cnt_forum
  - cnt_forum_add
  - cnt_forum_update
  - cnt_forum_view
  - cnt_forum_subscribe
- cnt_upload
- cnt_blog
  - cnt_blog_view
  - cnt_blog_add
  - cnt_blog_update
  - cnt_blog_delete
- cnt_course
  - cnt_course_view
  - cnt_course_addcnt_ course_update
- cnt_twitter

As a result the former activity count indicator was included in several more precise instances counting their related number of log lines for each user. This analysis helped in identifying some of the tools that have been widely used during the MOOC experience. For each of those tools, the K-Means clustering algorithm was executed and several analyses of action logs and demographics attempted. Age was not used in the clustering since it systematically led to clusters with common activity counts. Table 8 shows a typical structure for clusters. Two types of participants are clearly identified by either their non-active or active participation in the forum but none of the demographics allowed for a characterization based on specific profile information. The highest level of activity was in the Moodle forum.

Table 8. *Example of active participant in the forum clusters*

Attribute	**Cluster 0** (*n* = 188)	**Cluster 1** (*n* = 142)
Cnt_forum_add	0.271	2.423
Mother tongue	0.596	0.542
Gender	0.553	0.472
Computer skills	1.761	1.690
English	1.809	1.718
MOOC experience	0.340	0.437

The use of more detailed activity information did not improve the results significantly. The quantitative demographic information collected from the user profiles could not explain some of the user participation behaviors. Although some relationships, clusters, and patterns were revealed in the Moodle forum data, additional efforts and tools are needed to better understand how activity level data is conductive to the learning process and how participation and learning could be assessed with greater ease.

Discussion

The authors' exploration of learning in a cMOOC has highlighted some of the challenges in organizing and working with large, incomplete, and dispersed data sets. A mixed-methods approach was found to be the most appropriate means to analyze and interpret the data. The volume of data made it impossible to analyze the data systematically in the traditional forms of manual coding and arranging the data in

themes. It was highlighted that the use of NVivo to analyze a large dataset and to investigate issues related to motivation had its challenges. The program did not do what was expected; that is, take the labor out of the coding exercise by doing it automatically or by using the text search query function. Additional recoding of the high quantity of data was thus required. As highlighted in the results presented in the previous section, it was possible to extract motivational factors from the data and their correlation with other factors, but more in depth analyses of types of learners and their activity levels would help to inform the research on informal learning, personalization, and collaboration. The findings of the present study point to the importance of pedagogical factors, the delivery of resources, and the psychological factors as highlighted by Bouchard's (2009) model, which are all challenged by the fast pace of technological change (Weller, 2010).

The findings obtained underscored how information and "knowledgeable others" were readily available at the click of a button, but the scope of the analyses performed did not provide valuable insights into the role of key participants and the impact, if any, of the artifacts they produced in encouraging and motivating lurkers in the course. Instead, the analyses focused on learner motivation and self-directed learning and related psychological factors (i.e., conative factors). More in depth analyses should be carried out on semiotic factors, especially given the high number of artifacts produced, the significant use of tags to promote and share resources, and their possible impact on psychological factors such as motivation, interest and reflection.

EDM offered limited results for two main reasons. First, there were inconsistencies in the datasets. The dataset contained low-level information, mainly demographic information filled by learners at the time of registration. Though the datasets were cleaned during the pre-processing, there may still have been some inconsistencies left. Another source of inconsistency comes from the Moodle authentication system that allows "guest" users to participate in almost any activity using the same shared guest profile. By analyzing login activities in the platform, it was discovered that a large number of users had participated in the platform using their own account and other times a guest account.

A second reason for limited results comes from the complexity and richness of the dataset. The dataset contained poor quantitative information, and the qualitative study clearly shows that qualitative data leads to more significant results. Last but not least, pre-processing and methods used in knowledge discovery are always questionable, especially when they do not allow for firm conclusions. The datasets were challenging but cannot exclude that other machine learning algorithms would have brought better results considering the same data.

Conclusion

The combined methods of quantitative and qualitative data analysis as reported in this paper have facilitated in-depth exploration of massive amounts of MOOC and blog data. The application of Bouchard's (2009) dimensional model of learning has helped to confirm what has been highlighted in the literature as important new formats and strategies for learning and teaching, and their relevance and effectiveness in creating high quality learning experiences. Bouchard justly concludes that "only through the careful application of multi-dimensional models can progress be made towards creating environments that truly support the emergence and development of self-directed learning" (p. 21).

EDM and activity level analyses also lead us to conclude that in certain cases, human interpretations of qualitative data yielded more consistent and meaningful datasets than knowledge discovery and data mining alone. The current qualitative analyses point to important triggers for motivation, engagement and participation. The ability to dig even deeper into the various nuances of learner motivation through a mix of complementary approaches to analytics could possibly reveal the factors which help or hinder participation directly – factors which are especially important in environments which promote self-directed and self-regulated learning. The analysis of motivation in an online learning environment involves more than simple usage tracking, and its cognitive aspects are much more difficult to deduce by means of log files.

The authors' previous PLE research and evaluation efforts (Fournier & Kop, 2011; Kop & Fournier, 2011) have provided important baseline data about user experiences with existing tools, applications, systems and desirable features for creating new and improved personal learning environments. Their case studies pointed to the importance of human factors such as motivation, incentives, support (organizational, social network, either online or in the community) in creating high-quality learning experiences. Bouchard's (2009) framework provided an important basis for exploring learner autonomy

and self-directed learning in an open networked learning environment but also highlighted some of the challenges of cMOOCs.

According to Koutropoulos et al. (2012), MOOCs need to evolve in areas such as data capture and learner and learning analytics. Their categories include: determining who is merely "window shopping" in the initial periods of a MOOC, who is a lurker, who is an active participant, and when and why participants drop out completely. These MOOC researchers also state that in order to better understand the learners and MOOC participation, rich and informative data needs to be available for analysis and systems that facilitate this collection need to be built.

There is a need for powerful tools to allow analyses of important factors contributing to the quality of the learning experience "at a glance" to be useful to either the facilitator or those participants who are in need of targeted remediation at crucial points in the course – before their participation starts to decline or participants drop out altogether. Automated methods of EDM could readily add temporal information and clarity to the levels of learning activity of participants, with possible suggestions as to the nature of support, feedback, and tools they desire and require. Intelligent data analysis would overcome the limitations and challenges in using qualitative software analysis tools which are labor intensive and time consuming. However, publications about the usefulness of the tools just described are not yet available and further research will help to clarify if they might help learners on their self-directed learning journey.

References

Babbie, E. (1979). *The practice of social research* (3rd ed.). Belmont, CA: Wadsworth.

Bazeley, P. (2007). *Qualitative data analysis with NVivo*. London, UK: Sage.

Bell, F. (2011). Connectivism: Its place in theory-informed research and innovation in technology-enabled learning. *The International Review of Research in Open and Distance Learning*, *12*(3), 98-118. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/902/1664

Bouchard. P. (2009). Pedagogy without a teacher: What are the limits? *International Journal of Self-Directed Learning*, *6*(2), 13-22.

Boyd, D. (2010, April). *Privacy and publicity in the context of big data*. Keynote speech delivered at the 19th International World Wide Web Conference, Raleigh, NC. Retrieved from http://www.danah.org/papers/talks/2010/WWW2010.html

Bryman, A. (1988). *Quantity and quality in social research*. London, UK: Unwin Hyman. doi:10.4324/9780203410028

Denzin, N. K. (1989). *The research act: A theoretical introduction to sociological methods* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Downes, S. (2007, February 3). What connectivism is [Web log post]. Retrieved from http://halfanhour.blogspot.ca/2007/02/what-connectivism-is.html

Downes, S. (2008). gRSShopper [Web log post]. Retrieved from http://www.downes.ca/cgi-bin/page.cgi?post=44682

Fournier, H., & Kop, R. (2011). Factors affecting the design and development of a personal learning environment: Research on super-users. *International Journal of Virtual and Personal Learning Environments, 2*(4), 12-22. doi:10.4018/jvple.2011100102

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. New York, NY: Aldine.

Hartnett, M., St. George, A., & Dron, J. (2011). Examining motivation in online distance learning environments: Complex, multifaceted, and situation-dependent. *The International Review of Research in Open and Distance Learning, 12*(6), 20-38. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/1030/1954

Hine, C. (2005). Internet research and the sociology of cyber-social-scientific knowledge. *The Information Society*, *21*(4), 239-248. doi:10.1080/01972240591007553

Johnson, L., Smith, R., Willis, H., Levine, A., & Haywood, K., (2011). *The Horizon Report: 2011 edition*. Austin, Texas: The New Media Consortium. Retrieved from http://net.educause.edu/ir/library/pdf/hr2011.pdf

Kop, R., & Fournier, H. (2011). New dimensions to self-directed learning in an open networked learning environment. *International Journal of Self-Directed Learning*, *7*(2), 1-18. Retrieved from www.sdlglobal.com/IJSDL/IJSDL7.2-2010.pdf#page=6

Koutropoulos, A., Gallagher, M. S., Abajian, S. C., de Waard, I., Hogue, R. J., Keskin, N. O., & Rodriguez, C. O. (2012). Emotive vocabulary in MOOCs: Context & participant retention. *European Journal of Open, Distance and E-Learning, 2012*(1). Retrieved from http://www.eurodl.org/?p=archives&year=2012&halfyear=1&article=507

Massive open online course. (2013). In *Wikipedia*. Retrieved July 26, 2013, from http://en.wikipedia.org/wiki/MOOC

Mazza, R. (2010). Visualization in educational environments. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. d. Baker (Eds.), *Handbook of educational data mining* (pp. 9-26). Boca Raton, FL: CRC Press. doi:10.1201/b10274-4

Miller, T., & Bell, L. (2002). Consenting to what? Issues of access, gate-keeping and "informed" consent. In M. Mauthner, M. Birch, J. Jessop, & T. Miller (Eds.), *Ethics in qualitative research* (pp. 53-69). London, UK: Sage.

Rogers, P. C., McEwen, M. R., & Pond, S. (2010). The use of web analytics in the design and evaluation of distance education. In G. Veletsianos (Ed.), *Using emerging technologies in distance education* (pp. 231-248). Edmondton, Canada: Athabasca University Press. Retrieved from http://www.aupress.ca/books/120177/ebook/12_Veletsianos_2010-Emerging_Technologies_in_Distance_Education.pdf

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135-146. doi:10.1016/j.eswa.2006.04.005

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384. doi:10.1016/j.compedu.2007.05.016

Sheard, J. (2010). Basics of statistical analysis of interactions data from web-based learning environments. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. d. Baker (Eds.), *Handbook of educational data mining* (pp. 27-42). Boca Raton, FL: CRC Press. doi:10.1201/b10274-5

Siemens, G. (2006, November 12). Connectivism: Learning theory or pastime of the self-amused? [Web log post]. Retrieved from http://www.elearnspace.org/Articles/connectivism_self-amused.htm

Strauss, A. L. (1986). *Qualitative data analysis for social scientists*. Cambridge, UK: Cambridge University Press.

van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology*, *6*(2), 129-140. doi:10.1023/B:ETIN.0000047476.05912.3d

Weller, M. (2010). The centralisation dilemma in educational IT. *International Journal of Virtual and Personal Learning Environments*, *1*(1), 1-9. doi:10.4018/jvple.2010091701

Welsh, W. (2002). Dealing with data: Using NVivo in the qualitative data analysis process, forum. *Forum: Qualitative Social Research*, *3*(2). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/865/1880