Who are the top contributors in a MOOC? Relating participants' performance and contributions

Carlos Alario-Hoyos[*], Pedro J. Muñoz-Merino[*], Mar Pérez-Sanagustín[+],

Carlos Delgado Kloos[*], Hugo A. Parada G.[*]

[*]Universidad Carlos III de Madrid, Avda. Universidad, 30

E-28911 Leganés, Madrid, Spain

{calario, pedmume, cdk, hparada}@it.uc3m.es

[+]Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 486

8940000 Santiago de Chile, Chile

mdelmar.ps@gmail.com

Who are the top contributors in a MOOC? Relating participants' performance and contributions

Abstract

The role of social tools in MOOCs (Massive Open Online Courses) is essential as they connect the participants. Of all the participants in a MOOC, top contributors are the ones who more actively contribute via social tools, sometimes with posts to the emergent discussions, sometimes answering their peers' questions and concerns and sometimes even adding complementary sources of information to the course. This paper collects, analyzes, and reports empirical data from five different social tools pertaining to an actual MOOC to characterize top contributors and provide some insights aimed at facilitating their early detection. The results of this analysis show that top contributors have better final scores than the rest. In addition, there is a moderate positive correlation between participants' overall performance (measured in terms of final scores) and the number of posts submitted to the five social tools. This paper also studies the effect of participants' gender and scores as factors that can be used for the early detection of top contributors. The analysis shows that gender is not a good predictor, and that taking the scores of the first assessment activities of each type (test and peer assessment in the case study) results in a prediction that is not substantially improved by adding subsequent activities. Finally, better predictions based on scores are obtained for aggregate contributions in the five social tools than for individual contributions in each social tool.

Keywords

MOOC, social tools, learning analytics, contributions, performance

Introduction

MOOCs (Massive Open Online Courses) have caused a revolution in higher education, enabling institutions to reach millions of students worldwide who can access courses provided by elite universities, generally free of charge (Hyman, 2012). Platforms such as Coursera, edX, FutureLearn, and MiríadaX facilitate the deployment of MOOCs by both teachers and institutions. Currently, these platforms have from thousands to millions of registered users and include from tens to hundreds of courses in a broad range of knowledge areas, taught in several languages and by many different institutions across the globe (Malliga, 2013).

The large number of courses and users on these platforms enables the collection of huge amounts of low-level data related to participants' performance and behavior. These rich data open up new research opportunities in learning analytics, on a large scale, and in different knowledge areas (Sharples et al., 2013). For example, low-level data, such as the number of videos watched or the percentage of materials accessed by students (Ho et al., 2014) can be used to measure learning outcomes or to find correlations which enable the characterization and categorization of MOOC participants.

However, participants' interactions with materials (typically watching videos and solving exercises) are not the only source of users' data in MOOCs. There is also much information regarding participants' contributions that can be collected from the social tools around the MOOC (e.g., discussion forums) and that can also complement the characterization and categorization of MOOC participants. Despite a few MOOCs, such as the adaptive MOOC described by (Sonwalkar 2013), where students were assigned to different routes of enrollment and pace, the main MOOC platforms do not allow teachers to explicitly offer different learning paths, nor give personalized support to learners with problems (unlike in traditional face-to-face or non-massive online courses). This is one of the main reasons why social tools are key elements in this particular educational context, offering a space for crowd learning (Sharples et al., 2013) where participants can help each other to resolve their questions, where discussions can be arranged, and where additional materials can be shared to enrich the course (McAuley, Stewart, Siemens, and Cormier, 2010; Ho et al., 2013). Collecting and processing social interactions among learners have been recently associated with a new field called *social learning analytics* (Buckingham Shum & Ferguson, 2012), which is still in its infancy despite its importance in the MOOC context.

Recent studies in this field show that those people with a higher number of contributions (posts) in the social tools around the MOOC, from now on "*top contributors*," are actively engaged with the course (Hill, 2013). Although the role of active participants in enhancing virtual online communities has been studied for some years (Rienties, Tempelaar, Van den Bossche, Gijsalaers, and Segers, 2009), it is necessary to carry out specific research in the MOOC context. In this particular context, top contributors play a crucial role, since in some cases they perform tasks traditionally assumed by teachers. For example, platforms like edX enable MOOC teachers to assign top contributors a special role in the forums called "community TA," with permission to edit or delete messages posted by peers; this constitutes a collaborative way of maintaining the forum. Identifying who these top contributors are, how they behave, and if they master the subjects taught are research challenges whose exploration could improve the support given to MOOC participants and promote discussion around MOOCs.

This paper addresses the challenge of exploring top contributors' characteristics. For this purpose, we carry out a study with empirical data on participants' performance in different assessment activities and their use of five different social tools, obtained from a nine-week MOOC on educational technologies. In addition, we analyze the relationship between participants' overall performance and the number of posts in the social tools under analysis. Finally, we research the effect of a set of factors on the prediction of contributions in MOOCs, considering on the one hand the number of aggregate contributions in the five social tools, and on the other hand the number of individual contributions in each social tool. The ultimate goal of this work is to contribute to the categorization and early detection of top contributors, so that teachers can contact them individually to request their help as intermediaries between them and the mass of people enrolled on MOOCs.

The remainder of this paper proceeds with a review of learning analytics and social learning analytics in MOOCs. Then, the research methodology and analytical procedures are presented, including a description of the MOOC used as a case study, the set of hypotheses addressed in the paper, and the instruments and procedures applied. Next, we present the results, which include a brief overview of the course participants, the characterization of top contributors in the case study, the analysis of the correlation between participants' overall performance and contributions, and research on the effect of a set of factors on the prediction of contributions, considering aggregate and individual contributions in five social tools. Finally, the paper discusses the results and ends with conclusions and future work.

Learning analytics in MOOCs

Learning analytics is a relatively new concept that has emerged from the core of the educational data mining community, sharing with it the definition, improvement and use of data-intensive approaches for supporting basic research and practice in education (Siemens & Baker, 2012). Both learning analytics and educational data mining have been applied for years in blended and online education mediated by learning management systems and other online environments (Baker & Yacef, 2009; Perera, Kay, Koprinska, Yacef, and Zaïane, 2009; Dawson, 2010). Nevertheless, they have recently increased in popularity, drawing on the huge amounts of data that can be collected from massive environments such as MOOCs. The most widespread definition of learning analytics is that provided by the organizers of the first international conference on learning analytics and knowledge: "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Long & Siemens, 2011). However, measurement, collection, analysis and reporting apply not only to learning activities and courses that are currently under way, but can be very useful for predicting outcomes in upcoming learning activities and courses through predictive models (Romero, Ventura, Pechenizkiy, and Baker, 2010; Siemens, 2012). In this context, there are some studies which predict future variables using the available data of learning activities and courses at particular points: predicting final scores based on students' interactions with intelligent tutoring systems (Feng, Heffernan, and Koedinger, 2006; Gobert, Sao Pedro, Raziuddin, and Baker, 2013) or inferring students' behavior from survey results (Muñoz-Merino, Pardo, Muñoz-Organero, and Delgado Kloos, 2011).

Although research on learning analytics in MOOCs is fairly recent, there are already a few works which have collected data for categorizing MOOC participants based on patterns of interactions with learning materials. Kizilcec, Piech, and Schneider (2013), for instance, clustered MOOC participants according to four engagement trajectories: completing, auditing, disengaging, and sampling (i.e., learners who watch video lectures for one or two assessment periods). Hill (2013) proposes an equivalent classification, considering active participants, passive participants, drop-ins, observers and no-shows. More elaborated patterns of interactions with materials are defined by Muñoz-Merino, Ruipérez-Valiente, and Delgado Kloos (2013) for students who follow online courses based on videos and exercises, including explorers, recommendation listeners, hint avoiders, video avoiders, unreflective users and hint abusers. Other works in this area address the extension of the built-in learning analytics capabilities provided by MOOC platforms to capture and process additional low-level data related to interactions with materials (Ruipérez-Valiente, Muñoz-Merino, and Delgado-Kloos, 2013), while other studies employ the aforementioned patterns of interactions to build methods and architectures aimed at detecting students at risk of leaving MOOCs (Tabba & Medouri, 2013; Cheng, Kulkarni, and Klemmer, 2013).

Works on social learning analytics (SLA) (Buckingham Shum & Ferguson, 2012) applied to the MOOC educational context are even scarcer, and usually present preliminary studies focused on analyzing the number of interactions among students produced in a particular social tool, typically the discussion forum. Breslow et al. (2013), for example, analyzed students' posts in the discussion forum associated with the first edX MOOC; Belanger and Thornton (2013) did the same with Duke University's first MOOC; Kizilcec et al. (2013) classified participants in four aforementioned engagement trajectories (i.e., completing, auditing, disengaging and sampling) according to their level of activity in the forum; Ho et al. (2014) proposed a different classification by considering forum posts from participants who only viewed, only explored, and were certified in 16 HarvardX and MITx MOOCs; and Brinton et al. (2013) characterized forum posts in several MOOCs, proposing a model for classifying threads and ranking their relevance. Regarding SLA in other social tools around MOOCs apart from the discussion forum, the work by van Treeck and Ebner (2013) analyzed the number of interactions submitted with the Twitter hashtag associated with a MOOC during two consecutive editions of the same course, concluding that this social tool can play a relevant role in such courses. Another study conducted by Alario-Hoyos, Pérez-Sanagustín, Delgado-Kloos, Parada G., and Muñoz-Organero (2014), covering the use of five different social tools in the same MOOC, revealed the preference of participants for the discussion forum and to a lesser extent for Facebook and Twitter. Finally, Schreurs, de Laat, Teplovs, and Voogd (2014) developed an SLA tool to help teachers visualize real-time discussions in MOOCs.

All these works independently analyze and report the overall performance of MOOC participants regarding their interactions with materials, or regarding their interactions through social tools. The study by Manning and Sanders (2013) goes further and matches participants who obtained at least 10%, 60% and 90% of the final grade in 23 Coursera MOOCs with the percentage of posts these participants submitted to the discussion forum. This study concludes for instance that of the students who obtained at least 60% of the final grade, between 20% and 80% contributed through the forum (depending on the particular course), with 8 to 15% posting only once. This analysis gives some interesting insights into performance and contributions, but is still limited because it only covers one social tool, does not report the final grades of those posting more times (i.e., top contributors), and does not study factors that can help identify top contributors.

Research Methodology and Analytical Procedures

The literature review shows the increasing interest in analyzing and reporting students' performance and contributions in MOOCs. However, to the best of our knowledge, there are no works addressing the characterization of those participants with more contributions in MOOCs, nor the relationship between the number of contributions in different social tools and participants' overall performance. In addition, the effect that different factors can have on the prediction of contributions is something that needs to be researched. Analyzing how these factors are related is a first effort towards the proposal of predictive models to anticipate top contributors in MOOCs. This section presents the research methodology and analytical procedures employed to identify factors and models for characterizing and detecting top contributors. First, we present a sample MOOC that serves as a case study. Then, we define a set of hypotheses to be validated or rejected in the context of the case study. Finally, we describe the analytical instruments and procedures.

Description of case study

As a case study, we have selected a Spanish-language MOOC called “Digital Education of the Future” (DEF) with a focus on educational technologies. This MOOC was deployed on the Spanish platform MiríadaX (https://www.miriadax.net/web/educacion_digital_futuro) and ran for nine weeks in early 2013. DEF contents were mainly based on video lectures, with nine videos of about 10 minutes each per week, including a weekly interview with an expert. The teaching staff estimated the course weekly workload as three hours (27 hours in total).

The evaluation system in DEF included 13 summative assessment activities spread throughout the course. These activities were either multiple choice tests or peer assessment activities. Every week students had to complete a multiple choice test (nine in total), which represented 5% of the final grade. Every three weeks students had to complete a peer assessment activity (three in total), which represented 10% of the final grade. Peer assessment activities had two parts: first students submitted an assignment to the platform, which sent the assignment to at least three peers; then, reviewers returned comments and a grade, anonymously, following a rubric defined by the teachers. Participants had to take a final exam at the end of the course which represented 25% of the final grade. All the summative assessment activities had to be carried out within fixed dates. Other formative questionnaires were also open throughout the course, but they were not taken into account in calculation of the final scores.

The social support around DEF included five social tools: two built-in MiríadaX tools, Questions & Answers (Q&A), and a forum; and three external tools, Facebook, Twitter and MentorMob (a collaborative aggregator of learning resources in the form of a playlist). These five social tools were selected for different purposes: 1) the forum and Facebook for long discussions; 2) Twitter for short discussions; 3) Q&A for posting open questions related to the course procedures and topics; and 4) MentorMob for sharing additional materials. The inclusion of several social tools in the MOOC is a design decision that allows participants to choose those they feel more comfortable with, but requires an additional effort on the part of teachers and learners in order to identify the most relevant contributions (Alario-Hoyos et al., 2014).

The data regarding the MOOC participants, including their scores in the 13 evaluation activities, and their contributions in the two built-in tools (Q&A and forum) were extracted from analytics provided by the platform MiríadaX. The data regarding participants' contributions in Facebook, Twitter and MentorMob were extracted through an analysis of the accounts and hashtags associated with the MOOC. Unfortunately, potentially interesting high-level demographics, such as previous qualifications, age or previous online experience, as well as other low-level information, such as the number of log-ins or clicks could not be collected, either because the platform MiríadaX did not collect it or because teachers were not allowed to access that information. Teachers only had access to the names of the participants and their location, the latter being an optional field. Under the circumstances, the only demographic factor considered for the analysis was participants' gender, which could be inferred from participants' names. Gender is relevant since men or women might have a higher tendency to contribute more. Some studies show a difference in the use of Twitter between men and women (Burger, Henderson, Kim, and Zarrella, 2011), which suggests that gender might have an influence on someone's becoming a top contributor. Moreover, gender is usually included in researches in educational technology, and significant differences have been detected: e.g., in the attitude towards competition (Muñoz-Merino, Fernández Molina, Muñoz-Organero, and Delgado Kloos, 2014).

Hypotheses

We posit four hypotheses to be validated or rejected in the MOOC which serves as a case study. These hypotheses assume that: (1) a *contributor* is defined as a participant who posts at least one *contribution* in any of the social tools around the MOOC; (2) *top contributors* are the 1% of the participants who post more frequently considering aggregate contributions in all the social tools around the MOOC; (3) *performance* is the grade obtained by a participant in an assessment activity with the *overall performance* as the final grade obtained by a participant in a MOOC considering all the summative assessment activities and their weights.

- H1) Top contributors' overall performance is better than that of other contributors.
- H2) Participants' overall performance and contributions are positively related.
- H3) Gender and performance in the first assessment activities are factors which help to predict contributions.
- H4) Gender and performance in the first assessment activities are better predictors of contributions considering the number of aggregate contributions in all the social tools than considering the number of individual contributions in each social tool.

Instruments and procedures

In order to test H1, we follow several steps. First, we select top contributors as the 1% of contributors with more posts in the case study. After that, we analyze the overall performance of top contributors through their grades in the different assessment activities. Finally, we compare the grades of top contributors and the grades of the rest of the contributors, calculating the average and the standard deviation, and conducting an independent t-test on the average of grades.

In order to test H2 and see if there is positive correlation between participants' overall performance and contributions, we define the variable *contributions* representing the aggregate sum of all the posts in Q&A, Forum, Facebook, Twitter, and MentorMob for each student, and the variable *scores* representing the final score obtained by each student, normalized within the range [0, 10], with the final score calculated as the weighted sum of the 13 summative evaluation activities in DEF.

To see if gender and performance in the first assessment activities are factors which help to predict contributions and validate H3 (and eventually H4), we analyze whether contributions in social tools are influenced by gender or by the different assessment activities. The type of activity (tests and peer assessment activities) may be a relevant factor, as peer assessment activities are expected to encourage discussion further. In addition, the time when the activity takes place may be another relevant factor since the first activities occur early in the course, when participants contribute more because of their initial excitement (Alario-Hoyos et al., 2014). Given these considerations, we use a linear regression of hierarchical type because we can make reasonable hypotheses about the order of the variables (according to their importance).

In order to test H3, a linear regression model using the hierarchical method is built for the case of aggregate contributions. The dependent variable is *contributions* (i.e., the sum of the number of contributions in each social tool). The independent variables are *participants' scores* in the different assessment activities (i.e., 13 variables for the 13 assessment activities) tagged A1 to A13, where A4, A8 and A12 are peer assessment activities, A13 the final test, and the remaining ones represent weekly tests, and *participants' gender*. These predictor variables are included in order of importance, considering a balance between two arguments: peer assessment activities are expected to promote the discussion in social tools more since in this case study the answer to each of these activities is open and participants could elaborate on it; and assessment activities that happen earlier are expected to have a higher effect in predicting the total number of contributions. Finally, gender is included as the last predictor variable. Then, predictor variables that are statistically redundant and do not improve the model substantially are discarded for the next analysis. A2, A3, A5-A11, A13, and gender are discarded as their inclusion does not particularly improve the prediction of the contributions and only A4, A1 and A12 are selected.

In order to test H4, a linear regression model using the hierarchical method is built for the case of individual contributions in each social tool. Therefore, the dependent variables are the number of individual *contributions* in each social tool, and the independent variables are *participants' scores* and *participants' gender* with the following order of variables: A4, A1, A2, A3, and gender. Only activities until A4 are considered because the focus is the early detection of top contributors. Next, the variables which entail a significant improvement in prediction are selected: A4 and A1 for all cases except for Q&A, in which we selected A4 and A2.

Results

This section presents the results of the analysis carried out with the data on participants' performance and contributions extracted from DEF. These results are preceded by an overview of the course participants in order to provide contextual information to the reader. Then, top contributors are characterized by preference in terms of social tools and performance throughout the course. After that, the relationship between number of posts in social tools and final scores is analyzed. Finally, the effects of gender and performance as predictors of contributions are assessed.

General information on course participants

In total, 5,595 participants were registered in DEF at the end of the course. Of these, 456 people managed to pass the course, obtaining 50 points out of 100. Nevertheless, only 104 people from this group completed all the summative assessment activities. Some 4,791 contributions were submitted in total to the five social tools (71.4% to built-in social tools and 28.6% to external social tools), the forum being the most popular social tool, attracting almost 60% of the contributions (Alario-Hoyos et al., 2014). Nevertheless, of the 4,791 total posts, only 4,406 contributions were matched to participants registered in DEF. The reason is that some people used different user names in the course platform (MiríadaX), and in Facebook, Twitter and MentorMob. This fact precluded us from matching 152 of the total posts submitted to these three tools (11.1% of total contributions in external social tools). In addition, 233 contributions (6.8% of total contributions in built-in social tools) could not be related to any participants' performance because they deregistered the course before its end, removing any trace of interactions with the materials from MiríadaX. Finally, and for the analysis of the gender factor, it is important to say that gender could be inferred for 5,543 participants (99.1%); doubtful names were excluded from gender classification, but these have a minimal impact since in all these cases participants earned zero points, and only submitted three contributions.

Characterizing top contributors

A total of 1,031 (out of 5,595 contributors) contributed at least once in any of the five social tools around this MOOC. Of the 4,406 contributions which could be matched with participants in DEF, 2,301 (52.2%) were submitted by the 456 participants who managed to pass the course, with 298 of these participants (65.4%) submitting at least one contribution, and 64 of them (14%) submitting 10 or more contributions. Table 1 classifies participants in this MOOC into seven groups according to the number of times they posted in any of the five social tools, indicating also the total number of posts they submitted.

We define as *top contributors* the 1% who contributed most during the MOOC. Under this condition and rounding up, we obtain 11 top contributors (seven males and four females). Table 2 presents the activity of the 11 top contributors detected (TC1-TC11) in the different social tools, indicating their performance in the summative evaluation activities and their final scores. Top contributors preferred the built-in discussion forum (74.1% of top contributors' posts were sent to the forum) and Twitter to a lesser extent (20.5%). It is noteworthy that the first four top contributors (TC1 to TC4) clearly preferred the forum, using this social tool in 96.8% of their posts. Interestingly, these four top contributors were males, although this MOOC was fairly balanced regarding the number of participants by gender (males 48.2%, females 51.8%) and the number of contributions by gender (males 50.5%, females 49.5%).

The final scores indicate that 82% of top contributors in DEF completed the entire course with good grades (more than 60 points out of 100 with 50 points as the pass grade). Furthermore, top contributors in DEF obtained at least 77% of the points awarded in the assessment activities they completed. The statistical analysis reveals that top contributors had an overall performance on average of 64.15 points out of 100 (N=11, Std. Dev. 27.52). The rest of the contributors had an overall performance on average of 29.02 points (N=1020, Std. Dev. 28.78), which is much worse than that of top contributors. An independent t-test revealed a statistically significant difference of overall performance in favor of top contributors ($t=-4.03$, $df=1029$, $p<0.05$). This allows us to **validate H1**, asserting that top contributors' overall performance was better than that of the rest.

Correlation between contributions and overall performance

The results of the analysis considering the 5,595 participants and the 4,406 contributions show a significant moderate positive correlation between the variables *contributions* and *scores* ($r=0.343$, $p<0.001$). This correlation value serves to **validate H2** and indicates that there is a moderate relationship between the number of contributions posted in any of the five social tools around DEF and participants' overall performance.

Effect of gender and performance on the prediction of contributions

A regression model was built with the forced entry option considering just assessment activities A4 and A1 as predictors. Table 3 shows the results of the linear regression model with $R=0.31$ and $R^2=0.096$. This result indicates that this model, considering the first test activity plus the first peer assessment activity in the MOOC, can explain 9.6% of the variability of the contributions. In addition, a regression model was built with the forced entry option plus A12, giving $R=0.350$ and $R^2=0.122$, which means that the model, when considering these three assessment activities, can predict 12.2% of the variability of contributions. These results help to gain insights and **validate H3 regarding the factor performance in the first assessment activities** of different types. However, and although these numbers are encouraging, further research needs to be done in order to see if this prediction based on participants' scores can be improved by adding more types of assessment activities, especially at the beginning of the course (which is the appropriate time to detect top contributors), or other factors (e.g., the kind of contribution, the quality of the contribution, if the participant was already active or not in social tools before the course, etc.).

H3 is rejected regarding the factor gender. On the one hand, gender was excluded from the regression model, as its inclusion does not improve the prediction substantially. On the other hand, no significant correlations were found between gender and other factors. First, there is no correlation between the variables contributions and gender (r=-0.009 (p=0.52) >0.05). Second, partial correlation between contributions and gender taking out the effect of final scores is r=-0.027 (p=0.043), which is very low. Third, correlations between contributions and gender taking out participants' cumulative score after each summative evaluation activity are also very low.

Table 4 shows the regression models used to predict the individual contributions in each social tool considering variables A4 and A1, except for Q&A (A4 and A2). From the analysis, we extract some conclusions regarding the predictive factors. First, the prediction is not improved when each social tool is treated individually. Instead, the prediction gets worse, as the best value is for Q&A and explains 6.6% of the variability of contributions, whereas the prediction in the model which considers aggregate contributions in the five social tools is 9.6%. The results of the model support that finding: Facebook (R=0.229, R^2=0.052), Twitter (R=0.152, R^2=0.023), forum (R=0.242, R^2=0.058), MentorMob (R= 0.074, R^2=0.005) and Q&A (R=0.258,

R^2=0.066). It is important to note that although the predictive model of, e.g., MentorMob explains just 0.5% of the variability, this does not mean that removing the contributions in this tool from the total is a good idea. Instead, the presence of MentorMob in the number of aggregate contributions can be useful because it has a different purpose from that of Q&A, forum, Twitter and Facebook, so its inclusion might make a difference. These results serve to **validate H4 regarding the factor performance in the first assessment activities** of different types since this factor predicts better the variability of the contributions when considering aggregate contributions, than when considering individual contributions in the five social tools under analysis.

Discussion

The results obtained in this study are limited insofar as they were obtained from a particular MOOC in a specific domain (educational technologies and Spanish language). However, the duration of the MOOC (currently more than the average duration of courses in Coursera and edX) and the large number of participants and contributions makes it a relevant case study for analysis.

Most of the ideas presented in this work can be extended to other courses. For example, top contributors are defined as the 1% of contributors who send more posts. This definition relies only on quantitative data and considers the number of aggregate contributions from the five social tools. This definition can be adapted to other MOOCs to consider some of the following factors: (1) length (shorter MOOCs usually receive more daily contributions on average because of contributors' initial excitement and the subsequent decline of discussions) (Brinton et al., 2013); (2) workload required (the greater the workload, typically the higher the number of questions and comments raised in the social tools) (Perna et al., 2013); (3) teaching staff’s participation in social tools (a higher participation of teachers increases the discussion volume) (Brinton et al., 2013); and (4) use of social tools to calculate final scores.

It is interesting that two of the top contributors (TC5 and TC6) did not pass the course, and even got very low marks. Actually, TC5 dropped out the course after week 4, whereas TC6 completed most videos and formative activities but did not attempt to carry out most of the summative evaluation activities within the stipulated time. This behavior is not common, and although a few top contributors missed some summative evaluation activities (TC3, TC9 and TC10), most top contributors completed them all. TC6 turned out to be a socially active participant who works on the materials but does not seem to have any interest in completing the MOOC following the assessment system. The case of TC5 is also interesting because most of her contributions were posted in Twitter and during the first weeks. If she had chosen a social tool without a character limit, probably several of her contributions could have been merged in a single post and she would have not been classified as a top contributor. Further research needs to be done on other MOOCs to see whether the behavior of this participant is an exception or not, as well as on the effect of Twitter in biasing the calculation of top contributors.

The correlation value between contributions and overall performance obtained in this MOOC is lower than that reported by Macfadyen and Dawson (2012), which showed a significant strong positive correlation between students' contributions in the forum and final grades ($r=0.83$, $p<0.01$) in an LMS-supported course. The fact that a MOOC typically has a higher drop-out rate and different degrees of commitment among participants can cause this variation in the correlation values.

The analysis of the effect of assessment activities the prediction of contributions revealed that variables representing students' scores during the middle and the end of the MOOC did not improve the predictive model of the contributions so much. This is a relevant conclusion since we want to detect top contributors as soon as possible. With activities A1 and A4 (the first assessment activities of different types) we were able to make a prediction that was not considerably improved throughout the MOOC. There is one exception, i.e., A12. This peer assessment activity took place at the end of the course, when most dropouts had already happened and only the most active and resilient students remained on the course. This suggests that assessment activities at the end of a MOOC (particularly peer assessment ones) can help to improve the prediction of contributions. However, the latest activities, such as A12, may come too late from the point of view of the teacher and in terms of the final purpose of this work: detecting top contributors to support other students early in the course.

The following recommendations are distilled from the preliminary analysis of the effects of demographics and performance. First, gender should not be taken into account. Second, participants' performance throughout the course should be considered, and it is recommended that predictions for detecting top contributors should be made after the first assessment activity of each type has taken place (a test and a peer review activity in the case study). Third, the predictive models which can be obtained from aggregate contributions are better than the predictive models which only treat individual contributions in each social tool. These recommendations are expected to serve as a starting-point for other researchers who want to propose predictive models for the early detection of top contributors in MOOCs.

Conclusions and future work

Social tools play a major role in MOOCs, helping to increase the quality of the course, supporting learners with problems, and promoting discussions. This paper has presented a first characterization of top contributors in the social tools around a MOOC, revealing the existence of a moderate positive correlation between the number of contributions and overall performance, and making several recommendations useful for the early detection of top contributors: gender is not a good predictor; taking the scores of the first assessment activities of each type results in a prediction which is not substantially improved by adding subsequent activities; and better predictions are obtained for aggregate contributions than for individual contributions in each social tool. All this, using a MOOC on educational technologies as a case study for the collection, analysis, and report of empirical data.

Besides carrying out a broader analysis by employing MOOCs from different domains, future work needs to overcome some other limitations of this study. The first limitation is that the identification of top contributors is only based on the number of posts in social tools. Further research should combine the number of posts in social tools with other factors: the type of contribution and assignment of different weights to posts classified as small talk, course logistics or course-specific (Brinton et al., 2013); the quality of the contribution, promoting those posts better assessed by teachers and/or peers; the point at which the contribution is submitted, those sent on days with peak workloads or near deadlines being more important; and whether participants were already making extensive use of social tools and virtual communities before enrolling on the MOOC. In addition, top contributors should be monitored throughout the course to detect: top contributors who stop contributing after the initial excitement; and top contributors who become very active at certain times, e.g., when dealing with topics they are interested in or which are the subject of constant debate (Tobarra, Robles-Gómez, Ros, Hernández, and Caminero, 2014). Additionally (and when provided by the MOOC platform), high-level demographics of contributors, such as prior qualifications or previous online experience, as well as low-level events, such as the number of log-ins or clicks per week, should be incorporated in the analysis in order to improve the predictive model.

Alternative analyses between contributions and performance could also be conducted to follow up this research. One example would be the prediction of students' final grades in a MOOC from the number of contributions submitted in the different social tools. This analysis would require determination of a temporal threshold allowing discarding of most no-shows, observers and drop-ins (Hill, 2013), and taking into account that the initial excitement of the first days results in a large number of messages posted in social tools.

Finally, four additional lines of work include: (1) taking into consideration current literature in computer supported collaborative learning about how to identify emergent leaders (Strijbos & De Laat, 2010) to study the relationships between top contributors and leadership in MOOCs (Carte, Chidambaram, and Becker, 2006; Rienties et al*.,* 2009); (2) researching whether applying gamification strategies in the social tools around the MOOC can result in enhanced contributions, as suggested by Grünewald, Meinel, Totschnig, and Willems (2013) and also in a large number of contributions; (3) using social network analysis strategies to study the structure of the interactions between top contributors and other participants in the social tools (Rabbany, Elatia, Takaffoli, and Zaïane, 2014), in order to determine the size of the MOOC community; and (4) using the detection of top contributors as a mechanism to provide more personalized support as a step towards adaptive MOOCs (Sonwalkar, 2013).

References

Alario-Hoyos, C., Pérez-Sanagustín, M., Delgado-Kloos, C., Parada G., H.A., & Muñoz-Organero, M. (2014). Delving into participants' profiles and use of social tools in MOOCs. *IEEE Transactions on Learning Technologies* (in press).

Baker R.S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-16.

Belanger, Y., & Thornton, J. (2013). Biolelectricity: a quantitative approach: Duke University's first MOOC. Retrieved from http://dukespace.lib.duke.edu/dspace/bitstream/handle/10161/6216/Duke_Bioelectricity_MOOC_Fall2012.pdf

Breslow, L., Pritchard, D.E., DeBoer, J., Stump, G.S., Ho, A.D., & Seaton, D.T. (2013). Studying learning in the worldwide classroom: research into edX's first MOOC. *Research & Practice in* Assessment, 8, 13-25.

Brinton, C.G., Chiang, M., Jain, S., Lam, H., Liu, Z., & Wong, F.M.F. (2013). Learning about social learning in MOOCs: from statistical analysis to generative model. Retrieved from http://arxiv.org/abs/1312.2159v2.

Buckingham Shum, S., & Ferguson, R. (2012). Social learning analytics. *Educational Technology & Society, 15*(3), 3-26.

Burger, J.D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, *EMNLP'11*, 1301-1309.

Carte, T.A., Chidambaram, L., & Becker, A. (2006). Emergent leadership in self-managed virtual teams: a longitudinal study of concentrated and shared leadership behaviors. *Group Decision and Negotiation, 15*, 323-343.

Cheng, J., Kulkarni, C., & Klemmer, S. (2013). Tools for predicting drop-off in large online classes. *Proceedings of the 16th Conference on Computer Supported Cooperative Work, CSCL 2013,* 121-124. ACM.

Dawson, S. (2010). ‘Seeing’ the learning community: An exploration of the development of a resource for monitoring online student networking. *British Journal of Educational Technology, 41*(5), 736-752.

Feng, M., Heffernan, N.T., & Koedinger, K.R. (2006). Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. In *Intelligent Tutoring Systems,* Ikeda, M., Ashley, K.D., Chan, T-W. (eds.), Springer: Berlin-Heidelberg, LNCS 4053, 31-40.

Gobert, J.D., Sao Pedro, M., Raziuddin, J., & Baker, R.S. (2013). From log files to assessment metrics: measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521-563.

Grünewald, F., Meinel, C., Totschnig, M., & Willems, C. (2013). Designing MOOCs for the support of multiple learning styles. In *Scaling up Learning for Sustained Impact*, Hernández-Leo, D., Ley, T., Klamma, R. (eds.), Springer: Berlin-Heidelberg, LNCS 8095, 371-382.

Hill, P. (2013). Emerging Student Patterns in MOOCs: A (Revised) Graphical View. Retrieved from http://mfeldstein.com/emerging-student-patterns-in-moocs-a-revised-graphical-view

Ho, A.D., Reich, J., Nesterko, S., Seaton, D.T., Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: the first year of open online courses (HarvardX and MITx working paper No. 1). Retrieved from http://ssrn.com/abstract=2381263

Hyman, P. (2012). In the year of disruptive education. *Communications of the ACM*, *55*(12), 20-22.

Kizilcec, R.F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. *Proceedings of the Third International Conference on Learning Analytics and Knowledge, LAK 2013,* 170-179.

Long, P.D., & Siemens, G. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, *46*(5), 31-40.

McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). The MOOC model for digital practice. *Technical Report*, University of Prince Edward Island. Retrieved from http://www.elearnspace.org/Articles/MOOC_Final.pdf

Macfadyen, L.P., & Dawson, S. (2012). Numbers are not enough. Why e-Learning analytics failed to inform an institutional strategic plan. *Educational Technology & Society*, *15*(3), 149–163.

Malliga, P. (2013). A survey on MOOC providers for Higher Education. *International Journal of Management & Information Technology*, *7*(1), 962-967.

Manning, J., & Sanders, M. (2013). How widely used are MOOC forums? A first look. Retrieved from https://www.stanford.edu/dept/vpol/cgi-bin/wordpress/how-widely-used-are-mooc-forums-a-first-look

Muñoz-Merino, P.J., Pardo, A., Muñoz-Organero, M., & Delgado Kloos, C. (2011). Towards the prediction of user actions on exercises with hints based on survey results. In *Towards Ubiquitous Learning,* Delgado Kloos, C., Gillet, D., Crespo-García, R.M., Wild, F., Wolpers, M. (eds.), Springer: Berlin-Heidelberg, LNCS 6964, 525-530.

Muñoz-Merino, P.J., Fernández Molina, M., Muñoz-Organero, M., & Delgado Kloos, C. (2014). Motivation and emotions in competition systems for education: an empirical study. *IEEE Transactions on Education*, *57*(3), 182-187.

Muñoz-Merino, P.J., Ruipérez-Valiente, J.A., & Delgado Kloos, C. (2013). Inferring higher level learning information from low level data for the Khan Academy platform. *Proceedings of the Third International Conference on Learning Analytics and Knowledge, LAK 2013*, 112-116.

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O.R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(6), 759-772.

Perna, L., Ruby, A., Boruch, R., Wang, N., Scull, J., Evans, C., & Ahmad, S. (2013). The life cycle of a million MOOC users. *Technical Report*, University of Pennsylvania.

Rabbany, R., Elatia, S., Takaffoli, M., & Zaïane, O.R. (2014). Collaborative learning of students in online discussion forums: a social network analysis perspective. *Educational Data Mining, Studies in Computational Intelligence,* Peña-Ayala, A. (ed.), Springer International Publishing, vol. 524, 441-466.

Rienties, B., Tempelaar, D., Van den Bossche, P., Gijsalaers, W., & Segers, M. (2009). The role of academic motivation in computer-supported collaborative learning. *Computers in Human Behavior*, *25*(6), 1195-1206.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R.S. (2010). *Handbook of educational data mining*, Chapman and Hall/CRC Press, Taylor & Francis. Data Mining and Knowledge Discovery Series.

Ruipérez-Valiente, J.A., Muñoz-Merino, P.J., & Delgado-Kloos, C. (2013). An architecture for extending the learning analytics support in the Khan Academy framework. *Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality, TEEM 2013*, 277-284.

Schreurs, B., de Laat, M., Teplovs, C., & Voogd, S. (2014). Social learning analytics applied in a MOOC-environment. *eLearning Papers*, *From the field*, 36(4), 45-48.

Sharples, M., McAndrew, P., Weller, M., Ferguson, R., FitzGerald, E., Hirst, T., & Gaved, M. (2013). Innovating Pedagogy 2013: Exploring new forms of teaching, learning and assessment, to guide educators and policy makers. *Technical Report*, The Open University.

Siemens, G. (2012). Learning analytics: envisioning a research discipline and a domain of practice. *Proceedings of the Second International Conference on Learning Analytics and Knowledge, LAK 2012*, 4-8.

Siemens, G., & Baker, R.S. (2012). Learning analytics and educational data mining: towards communication and collaboration. *Proceedings of the Second International Conference on Learning Analytics and Knowledge, LAK 2012*, 252-254.

Sonwalkar, N. (2013). The first adaptive MOOC: a case study on pedagogy framework and scalable cloud architecture—Part I. *MOOCs Forum*, *1*(P), 22-29.

Strijbos, J.-W., & De Laat, M.F. (2010). Developing the role concept for computer-supported collaborative learning: An explorative synthesis. *Computers in Human Behavior*, *26*(4), 495-505.

Tabba, Y., & Medouri, A. (2013). LASyM: a learning analytics system for MOOCs. *International Journal of Advanced Computer Science and Applications*, *4*(5), 113-119, 2013.

Tobarra, L., Robles-Gómez, A., Ros, S., Hernández, R., & Caminero, A. C. (2014). Analyzing the students' behavior and relevant topics in virtual learning communities. *Computers in Human Behavior*, *31*, 659-669.

Van Treeck, T., & Ebner, M. (2013). How useful Is Twitter for learning in massive communities? An analysis of two MOOCs. *Twitter & Society*, Weller, K., Bruns, A., Burgess, J., Mahrt, M., Puschmann, C. (Eds.), Peter Lang, 411-424.

Table 1 Number of people posting and number of posts submitted in any of the five social tools.

Number of times posting	Number of people (%)	Number of posts (%)
1	438 (42.5%)	438 (9.9%)
2-5	404 (39.2%)	1154 (26.2%)
6-10	101 (9.8%)	743 (16.9%)
11-20	61 (5.9%)	868 (19.7%)
21-40	16 (1.6%)	465 (10.6%)
41-100	10 (1%)	572 (13%)
>100	1 (0.1%)	166 (3.8%)
TOTAL	1031 (100%)	4406 (100%)

Table 2 Characterization of top contributors in DEF, indicating gender, number of aggregate contributions, number of contributions per social tool, final score, number of summative evaluation activities completed, and percentage of points obtained in the summative evaluation activities. Means and Std. Dev. are provided for contributors and for the rest of contributors.

Id	Gender	Number of contributions	Q&A	Forum	Facebook	Twitter	MentorMob	Final score	Number of activities completed	% of points obtained in the activities completed
TC1	Male	166	0	155	5	6	0	77.05	13	77.05
TC2	Male	95	0	95	0	0	0	77.95	13	77.95
TC3	Male	81	3	78	0	0	0	63.7	10	79.62
TC4	Male	64	0	64	0	0	0	81.95	13	81.95
TC5	Female	51	0	2	1	48	0	8.75	2	87.5
TC6	Male	49	0	38	2	9	0	12.5	3	83.33
TC7	Female	49	0	9	5	35	0	87.35	13	87.35
TC8	Female	47	1	29	17	0	0	77.65	13	77.65
TC9	Male	46	2	40	0	4	0	68.85	11	86.06
TC10	Female	46	0	25	0	21	0	65.35	11	81.69
TC11	Male	44	0	12	2	28	2	84.6	13	84.6
TC Mean (N=11)		67.09	0.55	49.73	2.91	13.73	0.18	64.15	10.45	82.25
TC Std. Dev. (N=11)		36.75	1.04	45.57	5.05	16.78	0.6	27.52	4.08	3.86
Rest of contributors Mean (N=1020*)		3.60	0.52	2.09	0.52	0.42	0.04	29.02	5.33	73.07
Rest of contributors Std. Dev. (N=1020*)		4.80	1.06	3.56	1.38	2.09	0.28	28.78	4.70	18.93

*Except for the last column in which N = 772 (participants who completed no activities were excluded from the analysis)

Table 3 Multiple regression models between the number of aggregate contributions in the five social tools (dependent variable) and participants' scores in the different assessment activities until A4 (step 1), and considering all the activities (step 2).

N=5,543; Dependent variable: number of aggregate contributions in the five social tools	Unstd. Coef.	Std. Coef.
Step 1: considering A1 and A4, R^2=0.096		
Constant	0.237	
A4	0.041	0.218
A1	0.020	0.158
Step 2: considering A1, A4 and A12, R^2=0.122		
Constant	0.189	
A4	0.025	0.132
A1	0.017	0.135
A12	0.038	0.189

Table 4 Multiple regression models between the number of individual contributions in each social tool (dependent variables) and participants' scores in the different assessment activities until A4.

N=5,543; **Dependent variable: number of individual contributions in each social tool**	**Unstd. Coef.**	**Std. Coef.**
Regression model for Facebook, R^2=0.052		
Constant	0.033	
A4	0.005	0.150
A1	0.003	0.129
Regression model for Twitter, R^2=0.023		
Constant	0.013	
A4	0.005	0.087
A1	0.004	0.099
Regression model for Q&A, R^2=0.066		
Constant	0.042	
A4	0.004	0.173
A2	0.002	0.135
Regression model for Forum, R^2=0.058		
Constant	0.132	
A4	0.026	0.173
A1	0.012	0.120
Regression model for MentorMob, R^2=0.005		
Constant	0.003	
A4	0.000	0.042
A1	0.000	0.048