Future Generation Computer Systems 125 (2021) 590–603

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Data-driven detection and characterization of communities of accounts collaborating in MOOCs

José A. Ruipérez-Valiente [a,*], Daniel Jaramillo-Morillo [b], Srećko Joksimović [c], Vitomir Kovanović [c], Pedro J. Muñoz-Merino [d], Dragan Gašević [e]

[a] Department of Software Engineering and Artificial Intelligence, Complutense University of Madrid, Spain
[b] Departamento de Telemática, Universidad del Cauca, Popayán, Colombia
[c] Education Futures, University of South Australia, Australia
[d] Department of Telematics Engineering, Universidad Carlos III de Madrid, Spain
[e] Faculty of Information Technology, Monash University, Australia

ARTICLE INFO

Article history:
Received 2 December 2020
Received in revised form 7 June 2021
Accepted 4 July 2021
Available online 13 July 2021

Keywords:
Learning analytics
Educational data mining
Collaborative learning
Massive open online courses
Artificial intelligence

ABSTRACT

Collaboration is considered as one of the main drivers of learning and it has been broadly studied across numerous contexts, including Massive Open Online Courses (MOOCs). The research on MOOCs has risen exponentially during the last years and there have been a number of works focused on studying collaboration. However, these previous studies have been restricted to the analysis of collaboration based on the forum and social interactions, without taking into account other possibilities such as the synchronicity in the interactions with the platform. Therefore, in this work we performed a case study with the goal of implementing a data-driven approach to detect and characterize collaboration in MOOCs. We applied an algorithm to detect synchronicity links based on their submission times to quizzes as an indicator of collaboration, and applied it to data from two large Coursera MOOCs. We found three different profiles of user accounts, that were grouped in couples and larger communities exhibiting different types of associations between user accounts. The characterization of these user accounts suggested that some of them might represent genuine online learning collaborative associations, but that in other cases dishonest behaviors such as free-riding or multiple account cheating might be present. These findings call for additional research on the study of the kind of collaborations that can emerge in online settings.

1. Introduction

Massive Open Online Courses (MOOCs) are online courses that cater to large numbers of students, are designed for open participation and can be accessed by anyone via the Internet [1,2]. MOOCs have become a promising worldwide educational medium which have attracted much attention from different stakeholders, and many institutions have chosen to incorporate them into their educational programs, including for academic credit [3,4]. The entrance of MOOCs in the higher education sector has also facilitated the collection of large amounts of data from students distributed around the globe, which in turn has helped thrive data analytics in education. The analysis of educational data can help improve the quality and effectiveness of teaching and learning, and transform the current infrastructure into a modernized data-driven higher education. Many studies have focused on analyzing and characterizing student behaviors in these courses and thus, generated inputs that can help improve the learning process in digitally-mediated educational environments [5,6].

On the other hand, MOOCs support the social constructivism theory of learning that enables group interaction, mutual work, discussion, and collaborative knowledge formation. In this way, collaboration is considered as one of the main drivers of learning [7], and many learning theories promote the benefits of collaborative learning, both in face-to-face and online courses. Then, it is no surprise that there have been numerous researchers that have studied collaboration in MOOCs through the use of communication tools such as forums, or collaborative projects [8,9]. Teachers encourage student participation in the course through the technology and often use third-party tools and plugins to provide additional collaboration functionalities to students, such as social networks, messaging, or video conferencing tools [10]. Collaboration can also emerge through different student activities

* Corresponding author.
E-mail addresses: jruipere@ucm.es (J.A. Ruipérez-Valiente), dajaramillo@unicauca.edu.co (D. Jaramillo-Morillo), Srecko.Joksimovic@unisa.edu.au (S. Joksimović), vitomir.kovanovic@unisa.edu.au (V. Kovanović), pedmume@it.uc3m.es (P.J. Muñoz-Merino), dragan.gasevic@monash.edu (D. Gašević).

https://doi.org/10.1016/j.future.2021.07.003

such as commenting, responding, updating and sharing through discussion forums, increasing student participation [10,11]. In this way, learning can also arise from the connections between students in a spontaneous way and not only from the interaction with content. However, few authors have studied students' collaborations beyond what is visible online, for example, through the analysis of the interaction with the courseware, such as the course navigation, content visualization, or the submission of the scheduled exams. Therefore, we find that this approach that tries to reveal traces of collaboration that happen in the background, is currently missing in the literature.

Studies on collaboration typically take place in controlled environments or online classrooms where there are small numbers of students. However, MOOCs are a unique playground for examining how students collaborate at a larger scale [12]. Apart from having large amounts of data to analyze, MOOC students have very heterogeneous profiles, beliefs, and reasons to participate in the courses [13]. Previous work studied how students behave in an online course and much of the work highlights the benefits of collaboration in learning environments. However, it has also been found that not all collaborative student behaviors are good. Numerous unethical behaviors have been found, such as helping friends to pass their exams, or even using different accounts to obtain feedback through multiple attempts to questions [6]. For example, Hellas et al. [14], Lan et al. [15], and Waters et al. [16] identified potential unethical collaborations through the analysis of similarities in the scheduling of the activities taken, and the start and end times of take-home exams. Therefore, it is important to better understand *how* students are actually collaborating in MOOCs.

Besides, we found several previous studies on collaborative learning in MOOC environments focused on analyzing tools for course collaboration and the behavior of students in the discussion forums [10,17–19]. However, we did not find any work that performed a data-driven detection and characterization of collaborations based on students' interaction data. This refers to 'invisible collaborations' that cannot be detected by simply looking at online social interaction in forums or similar collaborative tools. In this paper, we present a novel data-driven approach to characterize students' collaborations in MOOCs. This work builds on top of an algorithm to detect collaborations that we developed in previous work [20], and that operationalizes collaboration as the synchronization of students when they submitted their quizzes to the MOOC platform. With respect to our previous study [20], this work takes place within the same context of Coursera MOOCs, using the same data set, and considering similar variables (Sections 3.1, 3.2, and 3.3). Then, we re-use the algorithm to detect collaborators from previous work [20] in the same context where it was previously applied and using the same parameters that we previously validated (Section 3.4). The new methodological contribution comes afterwards, by proposing a data-driven characterization of those accounts that were detected as collaborators (Section 3.5), which is completely novel in the literature. We present insights about the types of accounts and characterize the different emerging associations, while also connecting these findings with the current literature and theory. The methodology presented in this paper provides new ways to use digital trace data to understand e-learning collaboration and potentially provide feedback to instructors and students. Furthermore, because collaborations are not unique to MOOC courses, the depicted methodology can be re-used in new research applied to other online learning contexts. Specifically, we have the following Research Questions (RQs):

RQ1 What are the types of students' accounts based on their interaction with the MOOC platform?

RQ2 What are the behavioral characteristics of the detected associations of accounts?

The remainder of the paper is organized as follows. Section 2 reviews the related work in the area of student behavioral modeling, collaboration in MOOCs, and academic dishonesty. Section 3 presents the methodology applied to conduct this research, while Section 4 describes the results regarding behavioral characterization of the different accounts and associations. Section 5 discusses results comparing with the literature and, finally, Section 6 concludes the paper.

2. Background

In this background, we focus on presenting an overview of the three research directions that are more closely related to our work. First, in Section 2.1 we review studies that have applied techniques from educational data mining and learning analytics to model student behavior. Then, in Section 2.2 we focus on the studies that have analyzed collaboration behavior in MOOCs. Finally, in Section 2.3 we examine studies that tackled academic dishonesty behaviors in MOOCs.

2.1. Analysis of student behavior in MOOCs

There is a high diversity in the kind of work published within the context of student modeling in MOOCs. Much of it has been focused on modeling students' motivations to participate in these courses and their preferences [21–23]. A number of studies have specifically focused on students' motivation with gamification features, for example, to analyze their perceptions toward earning badges in a gamified MOOC [24] or to propose metrics to infer which students are earning badges intentionally [25]. These studies aim to better understand the motivations of MOOC learners in order to adapt the materials and better cater to learners' needs and interests.

Another predominant purpose of modeling students' behavior has been to predict learners' attrition in MOOCs. For example, Halawa et al. [26] presented a dropout predictor based on the interaction activity of students with the MOOC platform that can provide a trustworthy dropout risk factor. Ramesh et al. [22] also presented a framework for modeling and understanding student engagement in online courses based on trace data, using a probabilistic model to connect student behavior with course completion. These studies have sought the possibility to implement systems that can help improve MOOC completion.

Moreover, another key research line in MOOCs has been the investigation of which behaviors affect learning outcomes. For example, Al-Shabandar et al. [27] conducted two experiments to analyze which behavioral features were related to engagement levels and positive learning outcomes. In addition, Ruipérez-Valiente et al. [28] conducted a study on a Khan Academy instance building a prediction model of learning gains that included different activity indicators and behavioral data. They found a number of behaviors positively correlated with learning gains (e.g., students who follow recommendations made by course instructors), while others were negatively correlated (e.g., unreflective behaviors). Results from these kind of studies can help understand instructors and researchers which behaviors can have a positive or negative impact on learning outcomes, and thus enable the possibility of promoting or discouraging certain behaviors.

A large body of clustering studies in MOOCs have applied these techniques to find different behavioral profiles of students based on how they interacted with the activities [25,29–31]; there are nuances between these studies, for example [25] aimed to infer profiles of engagement with respect to the gamification features

of Khan Academy, Chen et al. [31] focused on extracting self-regulated learning strategies patterns, and both [29] and [30] focused on extracting different subpopulations of learners based on how they engaged with the activities. Other studies have applied clustering for alternative purposes; for example, Li and Li [32] used clustering approaches to provide personalized recommendations of MOOCs to users based on their characteristics or [33] applied it to study different profiles of participation in MOOC discussion forums. Moreover, clustering has also been used within MOOC studies for group formation purposes. For example, Lynda et al. [34] used it to group learners with similar profiles for the peer-review process and Sanz-Martínez et al. [35] used it to group alike learners for collaborative learning activities. As we see, the majority of the studies have used clustering either to find profiles of students in MOOCs, for recommendation purposes, or for group formation in order to develop some sort of activity between peers. However, to the best of our knowledge, clustering has not been applied within MOOCs for the purpose of characterizing collaborations.

The studies mentioned in this subsection have demonstrated diverse purposes to perform behavioral modeling of MOOC learners. However, even though student collaboration is one of the outstanding opportunities in MOOCs, few papers reported results regarding behavioral modeling that is performed to detect or characterize collaboration in MOOCs; our research study is focused in this direction.

2.2. Collaboration in MOOCs

Although numerous studies have focused on analyzing how students behave in MOOC environments, only few of them have delved into students' collaborations. In this direction, Claros et al. [36] presented several reflections about monitoring and assessment processes from two collaborative learning systems: The first one was defined with the aim of engaging students in a social process around the composition of interactive multimedia learning objects, while the second one sought to help the instructors in the design of collaborative learning scenarios with a set of services embedded into Moodle. By experimenting with these two collaborative learning approaches, the authors provided recommendations on how to apply these approaches to MOOCs in order to reduce instructors' workload. However, they did not analyze the collaborations and interactions between students that took place in the courses.

On the other hand, the majority of MOOC platforms offer limited technical functionality for collaborative work. After examining the collaboration support across Coursera, edX, Udacity, and MiriadaX MOOC platforms, Staubitz et al. [17] encouraged future work to improve features to support collaborative learning in MOOCs. Based on the analysis, the authors implemented a set of tools that can support collaboration on the OpenHPI MOOC platform. This set of tools consisted of a general virtual space for collaborative online learning, which supports study groups, topic-centered learning, and teams in both public and private working groups. For online communication, a combination of synchronous and asynchronous tools was added, such as a lab collaboration space that provides learning groups with the opportunity to share artifacts. Staubitz and Meinel [37] continued this line of work by examining the practical implications of some forms of collaborative learning that were implemented in the OpenHPI platform. The most important conclusion of their study was that the number of participants contributing to the forum increased considerably when instructors participated in the collaborative process. Their results also confirmed that forum participation in MOOCs actually works better with a large number of participants, as both students and instructors are more active because there are more interactions in the forums.

Several studies have looked into the effects that collaboration may have on different learning outcomes in MOOCs. In this sense, Brooks et al. [38] investigated whether participating in a MOOC with friends or colleagues can improve both course completion and student social interaction during the course. In this study, they sent surveys to students to analyze those who enrolled with friends, and the results suggested that enrolling in a MOOC with peers correlated positively with course completion rate, level of achievement, and use of the discussion forum. They demonstrated that there was a positive effect on student academic achievement and an increased online interaction when students enrolled with friends or colleagues. Li et al. [39] investigated the benefits of collaborations in MOOCs through an inverted classroom case study. Their results suggested that students in MOOCs prefer to study in groups, and that social facilitation within study groups can make learning difficult concepts a more enjoyable experience. The students reported a high overall satisfaction with this study group learning approach and the research revealed that students liked to be in sync with the group while watching the MOOC videos and completing the assessments. However, neither of these two studies analyzed the actual behaviors that these students performed in the MOOC platforms while collaborating together.

Collaboration in MOOC discussion forums has also been a common topic in the literature [18,40]. For example, Cohen et al. [18] used learning analytics methods to retrieve and analyze data of students' interaction with the course forums. The authors showed that 20% of the students were collaborating in the forums throughout the course and they were responsible for 50% of the total posts. Similarly, Ezen-Can et al. [40] presented a study of MOOC discussion forums with the aim of automatically extracting the structure of discussions posts to understand how students collaborate with each other.

Most studies on collaboration in MOOCs explored how students interacted through a collaboration tool or what benefits are gained from these collaborations. However, our approach is very different from these studies, as we use a data-driven algorithm to detect and characterize students' accounts that are collaborating when there is no specific encouragement to collaborate or additional tools to do so. We seek to know how students collaborate and whether these collaborations are learning-oriented or geared towards effortlessly obtaining a certificate; no approaches like this one have been reported in the literature thus far.

2.3. Academic dishonesty in MOOCs

While collaboration has been depicted as a great opportunity to improve online learning [41], there is also a delicate line between healthy collaborations and academic dishonesty. Previous work has been exploring this issue, for example [16] presented a framework for detecting collaboration between students in online or take-home tests, which depending on the course rules could be labeled as academic dishonesty. The authors developed a method to detect collaborations by making use of the SPARFA (SPARse Factor Analysis) framework. With this, Lan et al. [15] proposed two Bayesian hypothesis tests to detect collaboration in educational data sets. The first test examines the number of matches between couples of students given by SPARFA and uses this information to infer the probability of collaboration. The second test examines the sequence of joint responses by couples of students using a specific model of collaboration and assesses the probability that such patterns will emerge independently. However, this method has not been tested in MOOCs.

Academic dishonesty in MOOCs has received much attention in the literature, where several authors have proposed algorithms for the detection of CAMEO (Copying Answers using Multiple

Existences Online) behaviors [6,42–44]. CAMEO is one of the reported methods of cheating in MOOCs, where harvester (fake) accounts are used to get correct answers using the automatic feedback of the system, which are then used by a master account to achieve the grade that allows the student to get a certificate. Bao [42], Northcutt et al. [43], and Alexandron et al. [6] presented algorithms for identifying student submissions that were performed applying this CAMEO method; the algorithms were based on several heuristics that make use (among other things) of the IP addresses of the students and the timestamps of the submissions. Moreover, Ruipérez-Valiente et al. [44] presented a supervised machine learning algorithm that detected CAMEO without using IP addresses by using a previously labeled sample of CAMEO submissions. This algorithm used as input several features about the submissions, students, and the design of the problem to predict the likelihood of a submission being completed using CAMEO.

Following the line of data-driven detection of academic dishonesty, Ruipérez-Valiente et al. [20] proposed an algorithm that detects collaboration links between students in online learning environments, which is the one that we use in this study. Specifically, the study presented a method developed to detect links between students based on the students' temporal closeness or synchronization when submitting their quizzes [20]. The study found that the detected students needed significantly less activity with the courseware to get a certificate of completion. However, the authors concluded the paper indicating that more work was needed in the future to characterize students' behaviors based on the interaction data with the platform to determine whether students were involved into any behaviors that can be characterized as dishonest, which is our goal in this study.

Overall, we have detected a consistent gap in the literature that warrants a need to propose a data-driven method to characterize collaborations in MOOCs. This can be particularly important to differentiate between fruitful collaborations and dishonest behaviors that can lead to free-riding [45,46]. In this manuscript, we address this gap by implementing the aforementioned method to detect collaborations [20], and then we perform a novel data-driven characterization of the different associations that we have detected.

3. Methodology

3.1. Context of the study

The data used in the study comes from two MOOCs offered on Coursera platform by a large research university in the United Kingdom. First, *Introduction to Philosophy* (`PHIL`), which presents the main areas of research in contemporary philosophy, and *Fundamentals of Music Theory* (`MUSIC`), which introduces students to the theory of music providing basic skills to read and write on Western music notation.

From an instructional design perspective, both courses implemented auto-graded quizzes ever week, lasting seven and five weeks respectively. Both courses had one graded quiz per week, with around 6–12 (`PHIL`) and 10–14 (`MUSIC`) questions per quiz. Since our algorithm relies on finding synchronous submissions to quizzes, the fact that both MOOCs have weekly quizzes and large numbers of students, were our primary reasons to select them. The passing grade of `PHIL` was 50 points and the one for `MUSIC` 65 points, over a total of 100 possible points in both cases. The students did not receive any specific instructions to encourage collaboration, and therefore we assume that students either knew each other beforehand or met while taking the course.

3.2. Data collection

We used Coursera's raw student interaction data, which included actions and clicks performed by the student while interacting with the MOOCs. Coursera provides raw SQL exports, clickstream logs, and demographic data for session-based courses. The SQL exports of the course can be imported into a relational database and queried via traditional SQL statements.

A total of 53,831 and 89,896 students enrolled in `PHIL` and `MUSIC` MOOC respectively. Since the focus of the study is to detect collaboration across the course, we filtered out those students that did not persist through it. We operationalized this by selecting a sub-sample of only those students that submitted all the quizzes in a course. The final amount of students that passed this criteria and are included in the study are 2359 (4.38% from total) and 5159 (5.73% from total) students from the `PHIL` and `MUSIC` courses, respectively.

3.3. Considered variables

We implemented scripts to perform feature engineering based on the raw data provided by Coursera. We decided to implement metrics related to different dimensions: the academic engagement (grades and submissions) and behavioral engagement with the platform (general activity levels, interaction with videos and discussion forums). The rationale to select these dimensions was based on having different aspects to characterize the collaborations. The initial selection of features was based on the experience of the co-authors in MOOC research. For the academic engagement we implemented the following features:

- `FinalGrade`: The final numeric course grade (between 0 and 100).
- `GotCertificate`: Boolean variable indicating whether a given student obtained a certificate in a given course or not.
- `SubmissionCount`: The total number of submissions to graded assignments that a particular student attempted.
- `SubmissionUnique`: The number of submissions to different graded assignments that a particular student attempted.
- `SubmissionAverage`: The average number of submissions per graded assignment attempted.

Then, for the behavioral engagement, we implemented the following features for the general activity levels, videos, and discussion forums:

- `ActiveDaysCount`: The total number of days that a particular student was active in the course.
- `ActiveWeeksCount`: The total number of weeks that a particular student was active in the course.
- `DistinctVideoCount`: The total number of unique lecture videos accessed or downloaded by a given student.
- `VideoSeekCount`: The total number of video seek events generated by a given student.
- `VideoPauseCount`: The total number of pause events generated by a given student.
- `DistinctThreadCount`: The total number of unique discussion topics accessed by a given student.
- `DistinctThreadsPosted`: The total number of threads of discussion posted in the forum.
- `DistinctCommentsPosted`: The total number of comments posted in threads of discussion.

Fig. 1 shows a boxplot visualization with the distribution of all these features per course and divided for those that acquired a certificate or not. Moreover, we also computed the variables `SubmissionTimes` for the detection algorithm, and `Order` for the community characterization. These variables are defined as follows:

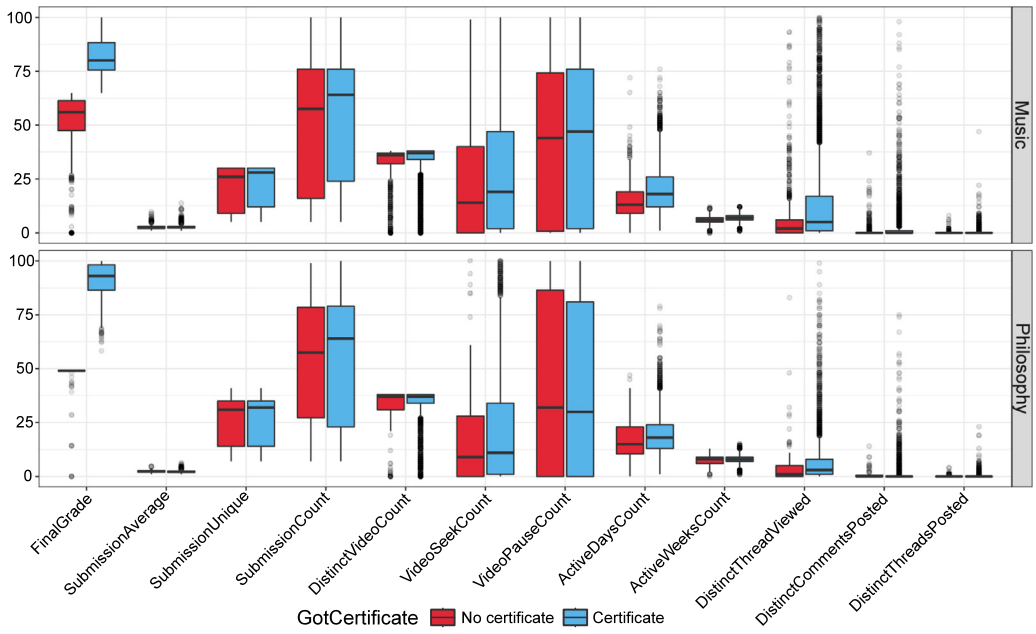

Fig. 1. Boxplot visualization of the continuous variables considered for this study separated by `GotCertificate` and computed for each course separately.

- `SubmissionTimes`: The list of timestamps of all submissions to course graded problems by a given student.
- `Order`: For a pair of collaborator accounts, this variable ranges from -1 to 1 indicating the order in which the submissions were done. A value of 1 signals that the first account always submitted the quizzes before the second account and analogously, a value of -1 indicates that the first account always submitted the quizzes after the second account. The values in between indicate relative difference between these two extremes.

3.4. Overview of the detection of collaborators

3.4.1. Definition of collaboration in this study

As we have seen in the related work, collaboration and collaborative learning have been defined and operationalized in many different ways. In this study, we focused on the previously reported notion of temporal synchronicity as a state in which the activities of a collaborating group are synchronized across time, that is, when group members are working on the same activity at the same time, we have that a collaboration is emerging [47]. A systematic literature concluded that the temporal analysis in collaborative learning can help increase scholar understanding in terms of theory and potential methodologies [48], which presents a strong alignment with our work. In our case scenario, we detected this synchronicity via students' timestamps when they submitted their quizzes. The rationale is that the statistical likelihood of two or more accounts submitting their quizzes at almost the same time every week is very low, specially given that these courses do not have due dates. For example, given that `MUSIC` course had seven quizzes, the probability of finding by chance a community of four students that always submitted their quizzes around the same time window of five minutes is extremely low given that the tests did not have due dates. We refer to this as 'invisible collaborations' that cannot be detected by simply looking at online social interaction in forums or other social tools. These are the underlying conceptual foundations of the algorithm that we detail next and the rationale why we selected it.

3.4.2. Algorithm

The algorithm that we implemented is based on the previous work by Ruipérez-Valiente et al. [20] and consists on identifying user accounts on the MOOC platform that always submit their assignments very close in time. The algorithm provides a systematic approach to detect synchronicity between students, which can be an indicator of collaboration, and can be easily applied to any online environment where students have to complete certain learning activities.

The algorithm is based on the comparison of the timestamps of all quiz submissions done by a student with respect to the rest of the students of the course and calculating how close they are in time, thus obtaining a distance matrix DS. The algorithm uses a dissimilarity matrix $DS \in R^{NxN}$ as follows:

$$DS = \begin{pmatrix} ds_{1,1} & ds_{1,2} & ds_{1,3} & \cdots & ds_{1,N} \\ ds_{2,1} & ds_{2,2} & ds_{2,3} & \cdots & ds_{2,N} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ ds_{N,1} & ds_{N,2} & ds_{N,3} & \cdots & ds_{N,N} \end{pmatrix} \quad (1)$$

Each entry $ds_{i,j}$ is a real number representing the dissimilarity between students i and j based on the differences in their assignment submission times and where N is the number of students in the course. Each element of matrix DS is calculated by a chosen dissimilarity function $diss(\vec{sp}_i, \vec{sp}_j) \in R$ which operates on vectors of student submission timestamps, with $\vec{sp}_i$ defined as:

$$\vec{sp}_i = [sp_{i,1}\ sp_{i,2}\ \cdots\ sp_{i,M}], i \in \{1 \cdots N\} \quad (2)$$

where $sp_{i,1}$ would be the timestamp of the submission to quiz 1 of the student i computed based on the variable `SubmissionTimes` and M is the number of quizzes in that course. After the DS matrix with the distances between all course participants is computed, we establish a threshold to classify a couple of students a collaborators. Then, we extract from matrix DS all unique entries $d_{i,j}$ where the value of the cell is below said threshold.

3.4.3. Collaborators detected

In this study, the dissimilarity measure is the mean absolute deviation (`MAD`), since it provides a comprehensive value to understand how closely two students submit their exams. The `MAD`

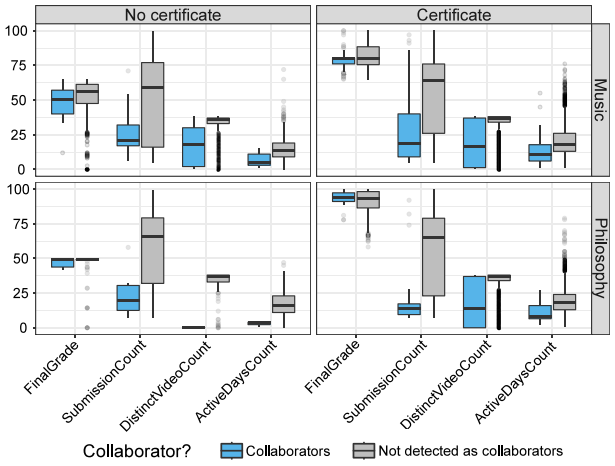

Fig. 2. Boxplot visualization that shows differences in the selected indicators for those accounts detected as collaborators and the rest, separated by `GotCertificate` and computed for each course separately.

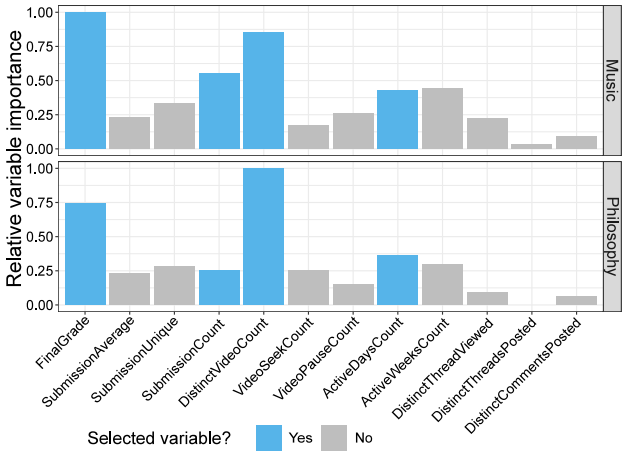

Fig. 3. Bar plot of the relative variable importance after running the clustering method with all the continuous considered variables. Blue denotes that the variable was selected for the final clustering analysis.

measure is defined as follows:

$$diss_{MAD}(\vec{sp}_i, \vec{sp}_j) = \frac{1}{M}\sum_{k=1}^{M}|sp_{i,k} - sp_{j,k}| \quad (3)$$

We used a `MAD` threshold of 30 min, which is based on experimenting with different thresholds and dissimilarity measures in our previous study [20]. Based on this procedure, we detected the following collaborators:

- `MUSIC`: 30 couples, two three-member communities, one four-member community, three five-member community, and one 14-member community. Overall, 99 different student accounts.
- `PHIL`: 11 couples and one four-member community. Overall, 26 different student accounts.

Fig. 2 shows a comparison of the selected indicators between those accounts detected as collaborators and those that are not detected. The differences between the two types of accounts are statistically significant, therefore confirming that we are detecting a different subpopulation of accounts.

3.5. Overview of the community characterization

3.5.1. Clustering method and metrics

We used the IBM SPSS Statistics Two-Step clustering method [49]. As part of the options of the algorithm, we selected the Euclidean distance as distance measure, we let the algorithm decide the optimum number of clusters automatically (range 2–15), and we used as clustering criterion the Bayesian Information Criterion (BIC). We pre-scaled the input variables by computing the z-scores of each variable (i.e. $z = \frac{x-\mu}{\sigma}$ where μ is the mean and σ the standard deviation of x). The algorithm automatically performs the following two steps:

- First, it identifies the appropriate number of clusters through agglomerative hierarchical clustering. In order to select the appropriate number of clusters, it will maximize the silhouette coefficient value, as described by Kaufman and Rousseeuw [50].
- Second, it applies k-means with the identified optimal number of clusters and Euclidean distance as dissimilarity metric to assign each one of the students to a cluster.

We used the relative variable importance as provided by IBM SPSS Statistics Two-Step to evaluate the importance of each predictor, where for certain variable i we have that:

$$VI_i = \frac{-log_{10}(sig_i)}{\max_{j\in\Omega}(-log_{10}(sig_j))} \quad (4)$$

where Ω denotes the set of features introduced to the clustering algorithm, and sig_i is the significance or p-value computed from applying a t-test or ANOVA when appropriate [49].

3.5.2. Selected variables

To avoid over-fitting of the relatively small data set, we decided to perform a feature selection to optimize the modeling. We made an initial run of the IBM SPSS Statistics Two-Step [49] with all of the continuous considered variables in Section 3.3; the algorithm was run separately for each one of the MOOCs. Then, we plotted their relative variable importance as shown in Fig. 3.

We decided to keep the variable with the highest importance for each one of the dimensions that we indicated before; based on what we see in Fig. 3 and our own judgment as experts in this area, we selected `FinalGrade`, `SubmissionCount`, `ActiveDaysCount`, and `DistinctVideoCount`. We did not select any of the variables related to forum activity because all of them have low importance, and as we see in Fig. 1, the majority of learners did not interact with the forum.

3.5.3. Characterization of the couples and communities

Once we have detected those accounts that are collaborators, our first RQ is to characterize these accounts. To solve this problem, clustering techniques are normally applied when we do not have a clear idea of the underlying groups in a population, and subjects are then clustered on the basis of some inherent similarity among them [51]. Therefore, we apply the clustering methodology in Section 3.5.1 to find different types of student accounts based on their engagement with the learning platform. This clustering process is applied separately to `PHIL` and `MUSIC` collaborators. The silhouette coefficient value for `PHIL` is 0.7 and for `MUSIC` 0.6, which can be considered as good values [50], and thus we conclude that the final clusters are valid. This kind of clustering approaches to find different profiles of students in MOOCs have been used in previous studies successfully [25,29–31].

Then, we represent the student collaborations on a network graphic, where the nodes represent students, the edges link two

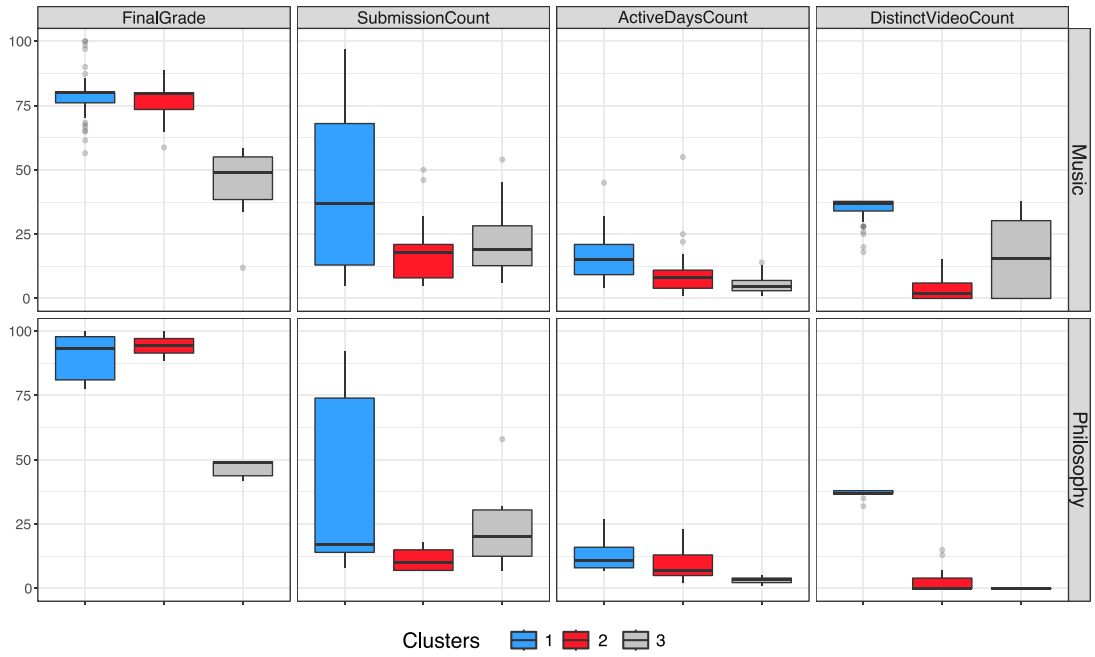

Fig. 4. Clustering results showing a boxplot visualization of the input variables separated by cluster and course.

students detected as collaborators, and the color of the node codifies the cluster assignment. This way, we represent collaborators in communities depending on how many accounts they were collaborating with. Finally, we analyze the indicators and clusters of the detected associations, connecting them with previously reported literature in order to perform a theory-driven validation of our findings.

4. Results

4.1. RQ1. Types of accounts based on the clustering analysis

We applied the clustering methodology as described in Section 3.5 to classify student accounts based on their interactions with the MOOC platform. Fig. 4 shows a boxplot with the clustering results where each input indicator is separated by cluster (on the *x*-axis) and by course (top row for `MUSIC` and bottom row for `PHIL`). The highest relative variable importance for clustering lied in the variables `FinalGrade` and `DistinctVideoCount`. `SubmissionCount` had the lowest importance. As shown in the plot, the variance of `SubmissionCount` was the highest of all, and thus it was not the one defining the clusters. The three clusters obtained are described below:

- Cluster 1: This group is composed of 34.6% of the `PHIL` course accounts and 41.41% of the `MUSIC` course accounts. The accounts that belong to this cluster had a high `FinalGrade` and the highest median values for the `ActiveDaysCount` and `DistinctVideoCount` variables. Additionally, the variable `SubmissionCount` had a very high variance, thus there were different types of accounts regarding the amount of submissions. Overall, since this cluster had the highest values for the two activity variables (`ActiveDaysCount` and `DistinctVideoCount`), and also a high value of `FinalGrade` variable, these accounts put effort and invested time on the course achieving high grades and obtaining certificates of completion.
- Cluster 2: A total of 42.3% accounts of the `PHIL` course belonged to this cluster and 42.42% of the accounts of `MUSIC` course. This cluster contains accounts that also had a high value of the `FinalGrade` variable. However, there were important differences in comparison to cluster 1 regarding the rest of the variables. Most importantly, we found that in terms of `DistinctVideoCount`, accounts in cluster 1 had a very high use of videos (most of the videos were seen by the users of the accounts in this cluster), whereas in cluster 2 this was quite the opposite case, where the users of most accounts watched very few videos. Additionally, the value of `SubmissionCount` and `ActiveDaysCount` variables were also lower than in cluster 1. Therefore, the users of the accounts in this cluster achieved high grades and obtained certificates, and they were able to accomplish this by watching very few videos, being active fewer days and with fewer submissions than the users of the accounts in cluster 1. Therefore, our hypothesis is that either the students running these accounts already had prior knowledge regarding the topic of the course and they just solved the required activities to get the certificates, or they might have been performing some illicit actions as part of the collaboration that had facilitated their way into obtaining a certificate without much effort.
- Cluster 3: This group is composed of 23.1% of the `PHIL` course accounts and 16.16% of the `MUSIC` course accounts. The last cluster of user accounts is clearly distinguishable from the other two clusters by its `FinalGrade`, which was much lower than in the other two with the median value of 50%. This means that most accounts in this cluster did not achieve a certificate of completion. The value of `ActiveDaysCount` was also the lowest one of all clusters, with very few days active. It is also interesting to see that the median value of `SubmissionCount` was higher than those of the other clusters in `PHIL` and higher than that of cluster 2 in the case of `MUSIC`. Therefore, although these accounts did not receive certificates and were active only very few days, they did make many submissions, in fact, this cluster has the highest median value of submissions in `PHIL` course. Finally, for the `DistinctVideoCount` variable, in the case of `PHIL`, the median value was 0 and none of those accounts watched any videos; in the case of `MUSIC` the variable had a high variance and the median value was higher than that

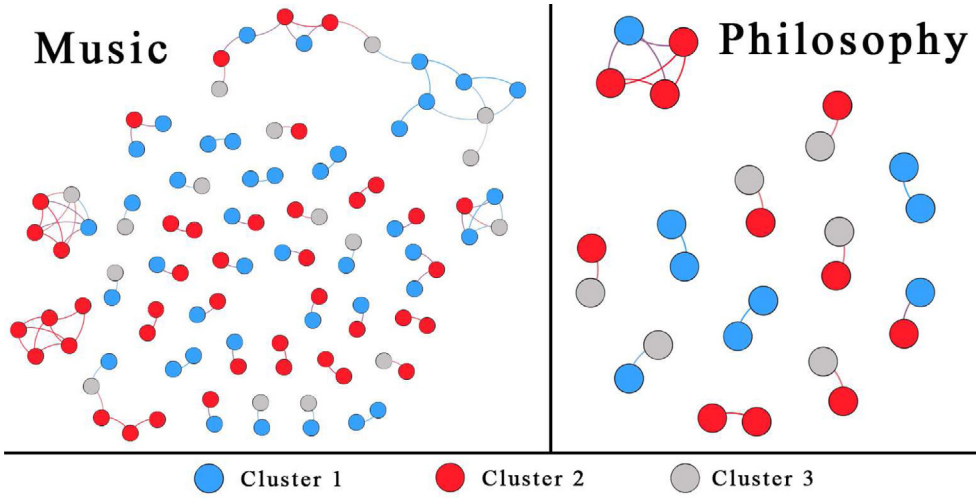

Fig. 5. Network graph of the couples and bigger communities detected by the algorithm and colored based on their cluster assignment. Each node represents an account, and the edge between two of them indicates the collaborating relationship.

Table 1
Examples of couples for each of the cluster associations found.

Association	Cluster	MAD	Order	Final Grade	Sub. Count	Act. Days Count	Dist. Video Count
Fruitful collaboration	1 1	2.65	+0.14	100 100	92 12	26 16	35 32
Free-riding	1 2	17.07	+1	81 98.6	74 16	7 19	37 1
Illicit collaboration	2 2	2.64	+0.71	97.1 91.4	7 18	5 5	0 1
CAMEO helper	1 3	1.21	−1	94 49	28 58	11 5	38 0
CAMEO premeditated	2 3	1.27	−1	96.4 48.5	7 32	14 4	0 0

for cluster 2. Our hypothesis is that this cluster of accounts represents the *harvesting accounts* that have been reported in previous research about CAMEO [6,43]; these accounts were created for the mere purpose of harvesting correct solutions by using exhaustive search (i.e., each quiz item has several attempts available and students receive feedback on the correctness after the submission). The correct solutions can be used later in the main account that would receive a certificate. This hypothesis is plausible since the accounts in cluster 3 did not achieve a certificate, were not very active in the course but still made many attempts to the quizzes.

Finally, Fig. 5 shows networks of the couples and bigger communities that were detected by the algorithm. In these networks, the circle (node) represents each one of the accounts, and the line (edge) linking the accounts indicates that those two accounts were detected as collaborators. Additionally, the color of each circle represents the cluster assignment. For example, on the top-left network of the PHIL course, we see four accounts collaborating together, three from cluster 3 and one from cluster 1, and all of them are connected with each other. This way, we are able to see the different cluster associations in the couples and communities. Next Sections 4.2.1 and 4.2.2 report the findings for the couples and communities detected, respectively, based on their cluster assignment and associations between accounts.

4.2. RQ2. Behavioral characteristics of the detected associations of accounts

4.2.1. Couples of accounts

This subsection describes the associations between the couples of accounts regarding their cluster assignment. Table 1 exemplifies each cluster association with the variables of one the detected couples per association:

- Association 1 "Fruitful collaboration" (cluster 1 and cluster 1 — PHIL 3/11 and MUSIC 5/30): This association represents two students from cluster 1 working together. As we reported in the previous subsection, the users of the accounts from cluster 1 put considerable amounts of effort on the platform to achieve certificates, with high values of ActiveDaysCount and DistinctVideoCount. Therefore, this association might represent two students that were taking the course seriously, and were collaborating reciprocally with each other in order to achieve better grades. In the example of this association in Table 1, the two accounts obtained the highest possible grade.
- Association 2 "Free-riding collaboration" (cluster 1 and cluster 2 — PHIL 1/11 and MUSIC 11/30): This association represents one student of cluster 1 and one of cluster 2, which might be a genuine association between two real students; however, this relationship is not equitable. According to the chosen clustering variables, cluster 1 has a higher platform interaction than cluster 2, but in both clusters, high grades are achieved. In this association, the student of cluster 1 would put effort in their work on the platform, whereas

student of cluster 2 did not make much effort but still would get a certificate with the help of the student of cluster 1. The value of the `Order` variable for the "Free-riding" collaboration was close to 1. That means that the account of cluster 1 almost always submitted the assignments before the peer, and we can see exactly that in the example of Table 1.

- Association 3 "Illicit collaboration" (cluster 2 and cluster 2 — `PHIL` 1/11 and `MUSIC` 5/30): In this association both accounts belong to cluster 2, therefore this case represents two accounts that did not demonstrate much effort in the course in terms of videos watched or active days, but still were able to receive certificates of accomplishment.
- Association 4 "CAMEO helper" (cluster 1 and cluster 3 — `PHIL` 1/11, `MUSIC` 6/30): This association represents one account from cluster 1 and one from cluster 3. In this case, we have one account that achieved a certificate investing a significant effort, and the second one that could potentially be a *harvesting account* based on previous literature [43,44], since it did not achieve a certificate, watched only few videos, and made many submission attempts.
- Association 5 "CAMEO premeditated" (cluster 2 and cluster 3 — `PHIL` 5/11, `MUSIC` 3/30): This association represents one account from cluster 2 that was able to achieve a certificate with little effort and one from cluster 3 that could potentially be a *harvesting account* [6]. In both "CAMEO helper" and "CAMEO premeditated", the `Order` variable tended to be close to −1, meaning that the account from cluster 3 almost always submitted the quiz first to get the correct responses. We can see this in both examples of Table 1.
- Association 6 (cluster 3 and cluster 3 — `PHIL` 0/11, `MUSIC` 0/30): We found no associations of two accounts of cluster 3. This makes sense as we generally label accounts from cluster 3 as *harvesting accounts* and hence it would not have a lot of sense to find two of them coupled (unless the student dropped the course).

4.2.2. Communities of more than two accounts

In the case of the communities of accounts, it was harder to present an overall view, since the size and associations between the different members of the community varied from one case to another. Therefore, it was difficult to provide a systematic general approach to describe all communities. Instead, we delve into the specifics of two community examples. The extracted indicators for each member of the selected communities can be seen in Table 2:

- Community 1: The first community in Table 2 belongs to `PHIL` and is composed by three accounts from cluster 2 and one account from cluster 1. The account of cluster 1 watched all the videos in the course, whereas the rest of accounts watched fewer videos. They had similar values for `FinalGrade`, `ActiveDaysCount` and `SubmissionCount`. Additionally, we can support our hypothesis with Fig. 6, where each quiz is represented on the x-axis and the time difference between the submissions of the accounts for that quiz on the y-axis. The plot shows that for Community 1, the submissions of all accounts for each quiz were always done within a 5 min timeframe (except for the submission of account 1 to Quiz 1). They always met one day each week (either a Monday or a Tuesday) and solved together the weekly quiz.
- Community 2: The second community represented in Table 2 is more complex than community 1 and belongs to `MUSIC`. There is one account from cluster 1, three from cluster 2 and one from cluster 3. With the exception of the account from cluster 1 that watched 20 videos, the rest of the accounts watched none or very few of them. For this community we found that, all 25 submissions made by the 5 accounts, were done in a interval of time of only 68 min during the same day.

Table 2
Description of the extracted indicators for each member of the two selected communities of accounts.

Community	Cluster	Final Grade	Sub. Count	Act. Days Count	Dist. Video Count
1	2	92.14	14	8	15
	1	91.79	14	8	38
	2	91.55	16	7	7
	2	88.57	14	12	13
2	1	56.47	71	8	20
	2	69.86	21	6	2
	2	79	5	10	0
	3	38.55	19	1	0
	2	80	27	25	4

5. Discussion

The section is divided in two parts, first, Section 5.1 discusses the results of the different types of associations that have been found and Section 5.2 the potential implications.

5.1. Different types of associations

In the current study, we detected different collaboration behaviors among accounts, and we hypothesized that some of them could be strongly related to academic dishonesty in MOOCs, while others might be beneficial for students. We first applied the algorithm described in Section 3.4 to detect collaborators, and then implemented the clustering approach in Section 3.4 to characterize the collaborations. We remark that the main underlying idea for this characterization was that couples or communities of students detected by our method had always submitted their assignments very close in time to each other; therefore, this time closeness represents a suspicious and possibly an illicit behavior. One important finding was that despite the fact that we applied the cluster analysis to both `MUSIC` and `PHIL` courses separately, we obtained the same cluster types for courses of different topics, suggesting that this finding could generalize beyond the data set used in the current study. However, the transferability of the clustering results to other courses should be analyzed more deeply since just two courses were used in this study. In addition, we can observe some differences on the values of variables in both courses for some clusters, e.g., regarding the grade values. This could be due to the fact that the course difficulty in `MUSIC` and `PHIL` are different. Therefore, the course characteristics such as the difficulty of the topic should be taken into account when changing the context.

The clustering method detected three different clusters with different characteristics. The accounts in cluster 1 received a certificate by investing a great effort, they watched most videos, and were active many days. The accounts in cluster 2 made a small effort, they almost did not watch videos, were active a moderate amount of days, and made few submissions. Still, they managed to get high scores and received certificates. Finally, the accounts in cluster 3 were active few days and did not watch any videos, but they still made many quiz submissions and did not receive certificates. Thus, we were able to identify three different types of students' collaborations in the form of couples (associations 1, 2, and 3 presented in Table 1). We did not consider associations

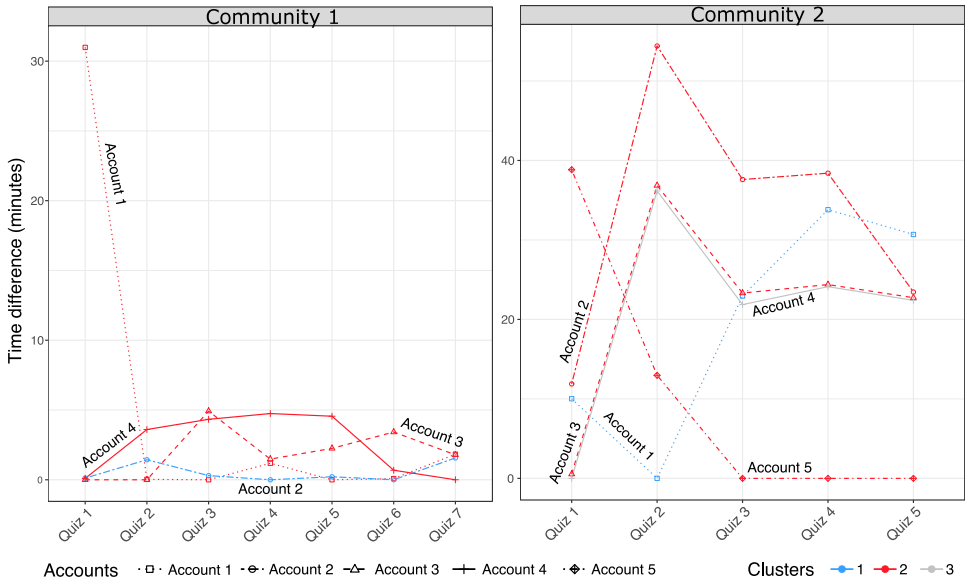

Fig. 6. Time difference between the submissions of each one of the accounts of the community for each quiz. For example, in the case Community 2, the Accounts 3 and 4 submitted first the Quiz 1, and Accounts 1 and 2 submitted 10 and 12 min later respectively.

4 and 5 as real collaborations; instead, we hypothesized these to be CAMEO [6,42,43], and hence both accounts in associations of type 4 and 5 were likely run by the same student. The discussion delves into these findings now.

Associations 1 "Fruitful collaboration" are composed of accounts that potentially worked together and had high degrees of commitment according to the variables `ActiveDaysAccount` and `DifferentVideoCount`. These associations can potentially represent two students who made an effort in the course by watching videos, they tried to learn and understand the contents, and met to submit their assignments together, potentially solving together in a collaborative way the quizzes in a sort of equitable relationship. The motivation here can be the ambition to improve the grades, and we might argue that this relationship does not represent a severe problem for the learning process of these students. The two accounts would work together to achieve high grades and in the same way, they showed an effort on the platform. This is the only type of association that demonstrated a behavior that has the more positive characteristics of collaborative learning [52].

Associations 2 "Free-riding" might represent an inequitable collaboration. In this case, there was a less balanced interaction where students from cluster 1 potentially have a passive attitude and pass the answers to the students from cluster 2 (potentially a friend or acquaintance), so that this latter account could obtain a certificate without investing much effort in the course, practicing the behavior known as free-riding [45,46]. Indeed, the literature has reported that one typical behavior toward cheating is that, one copies from the other ('active'), and the other allows the others to copy ('passive') [53]. This definition resembles quite well this situation where the student of cluster 1 would usually submit an assignment before the student in cluster 2 as the variable `Order` has values close to 1. Additionally, letting others to copy from you is regarded as less severe than actually copying from others [54]. In the case of this specific association, the impact on the learning process of the students from cluster 2 is obviously more severe.

In a similar way, in associations 3 "Illicit collaboration", the accounts had very limited interaction with the platform but both obtained a certificate. Thus, we can assume that these associations performed some kind of strategy that allowed them to get answers to exam questions without studying the contents of the course. In this case, students might have been applying "gaming the system" strategies, where a learner attempts to succeed in an educational environment by exploiting properties of the system's help and feedback rather than by attempting to learn the material in order to accomplish a passing grade without investing the necessary effort [55]. This kind of behavior can be severe for the learning process, since in several studies authors found gaming the system behaviors to be associated with poor learning outcomes [56,57]. This can also affect the future beliefs and attitudes of these students, as they might come to think that they are able to accomplish goals without putting much effort.

Finally, we have associations 4 "CAMEO helper" (cluster 1 and cluster 3) and associations 5 "CAMEO premeditated" (cluster 2 and cluster 3). As cluster 3 had low level of interaction with the content on the platform, many submission attempts and low scores without receiving certificates, the hypothesis that we described was that these were harvesting accounts as described extensively in the CAMEO literature [6,42,43]. Since accounts from cluster 3 were present in both associations, most probably these were CAMEO associations. Therefore, we can conclude that in these two association types, both accounts were managed by the same student. First, the association between a cluster 1 account and a cluster 3 account might represent a slightly less severe situation, because the cluster 1 account invested an effort to study on the platform and might be using the harvesting account to secure and achieve a passing grade without struggle. This association is closer to the idea of applying CAMEO as a "helper-mode" that was reported by Alexandron et al. [6]. The second scenario is an association of group 2 and group 3 accounts, which could represent a more severe situation, since the student is managing to receive a certificate without investing any effort and seems to be closer to the "premeditated-mode" that was reported by Alexandron et al. [6].

We also detected a number of communities with more than two accounts collaborating together. However, it is difficult to systematically characterize associations in each community because there were numerous accounts. Hence, we described two examples in Table 2 from the set of communities that we found. While in community 1 there were some associations that could

represent genuine collaborative behaviors, the accounts in community 2 exhibited elements of explicit dishonest behavior. Therefore, this is an analysis that needs to be performed for each community separately.

More work will be needed to assess if students knew each other prior to starting the course or if they met online in study groups [58], and then decided to engage into an 'unethical collaboration.'

5.2. Implications

While this work started with the aim of detecting and characterizing collaborations that may arise in MOOCs, we have found numerous behaviors that can be considered dishonest where students exhibit a deliberate behavior with no intention of learning the course contents. This study can be a good complement to previous work that focused on CAMEO [6,42,43]. This kind of dishonest collaborations probably have high prevalence due to the certification provided by MOOC-based online programs. For example, the literature has shown that students performed CAMEO more frequently on those questions that had higher weight towards the final grade of a MOOC [6].

Our work has focused on characterizing a number of collaborations following a data-driven approach. However, we believe that there are some limitations in the findings reported in this paper. We did not have a clear threshold value in our detection methodology, and there might be other types of collaborations that have not been captured by the algorithmic approach used in this study; therefore, having a different threshold of the accounts that we categorized as collaborators would impact the precision and recall of the algorithm. Additionally, although the collaborators discovered provided solid evidence since the differences are statistically significant, we do not have a ground truth that can help us refine the algorithm and evaluate its real quality. In fact, there could be other potential explanations to the results that we reported. Furthermore, the context may be a strong determinant for the existence of different collaborations and behaviors [59]. The subject matter and design of the course [60], the platform where it took place [61], and the audience to whom the course is addressed [62], could have an important influence that could lead to collaborations with different behaviors than these presented here. Therefore, a wider study with different contexts would be necessary in order to generalize the findings that we report.

In the future we plan to add new data sources in our analysis in order to improve the insights and characterization. For example, forum interactions could bring information about how students interact in the forum and we could contrast this information with these results. Additionally, we could detect healthy interactions from a text mining analysis among a group of students that belong to cluster 1, giving more insights about a healthy and fruitful collaboration. Moreover, mixed-methods studies that involve collection of qualitative data such as interviews and focus groups could help validate some of the inferences made in this paper. While students who were involved in dishonest behaviors might be reluctant to disclose details of their behavior, qualitative studies at least could be beneficial to corroborate findings about healthy collaboration links identified in this study.

The aim of this work was to shed some new light on the understanding of students' collaborations online, behaviors, motivations, and needs. This can also help to better understand the role that online collaboration can have in learning outcomes and provide the teacher with tools that allow them to somehow improve the design and development of MOOCs to promote more collaboration. However, the high prevalence of collaborations where students were clearly passing a course thanks to some kind of 'free-riding' limits the positive conclusions that we can draw from the study. Previous work [6] found that some course design aspects, such as randomization, could greatly help to deter academic dishonesty. Another easy possibility to control CAMEO, would be to link each student account to a physical person, instead of allowing the registration of multiple accounts [5]. There are numerous design options that can help minimize these issues and future work should invest time on creating helpful guidelines for online course designers and practitioners.

The findings offer a significant novel contribution to the literature, as for the first time, this study confirms that it is possible to characterize collaborations emerging in MOOC environments without having previous knowledge about the existence of such collaborations. All previous work has been centered on studying collaborations that were self-reported, controlled or visible through collaborative tools such as forums. Our study has also shown that the majority of the students have used this anonymous environment to collaborate in dishonest ways, a practice that has facilitated their way into a completion certificate. The results that we have reported could have implications on the design of dashboards for teachers to help them understand the types of collaborations that are taking place. These dashboards can provide opportunities to teachers to assess the quality of the collaborations, intervene, and provide feedback as appropriate in each case scenario. These findings also open the possibility of new research built on top of the methodology proposed in the paper. While we have applied it to MOOCs, the methodology could easily be adapted to other online learning environments taking into account the specific contextual characteristics, thus, opening new research horizons on collaborative learning.

6. Conclusions

Nowadays, we frequently find that the design of MOOCs is no longer focused on having students collaborate together to construct knowledge. Still, many social or collaborative tools such as forums or peer-review activities are maintained. In addition, teachers encourage student participation in the course through technology platforms and resources or tools available on the platforms [10,63]. Moreover, collaborations in MOOCs might emerge spontaneously because people can meet in the forums or on virtual working groups, or because friends decide to take a course together. However, while collaborations in MOOCs are generally considered positive for the learning process, this work has revealed that not all students' collaborations can be considered as good or beneficial. This phenomenon is not new, and in traditional classroom courses, researchers and practitioners have frequently reported inequitable or dishonest collaborations [14–16]. This study has extended the state of the art by implementing a data-driven characterization of different collaboration types in MOOCs.

Collaboration can be an important factor in students' outcomes in any type of course, since learning can arise from the spontaneous connections between students and in many of the works we found, the advantages of collaboration are highlighted. However, the majority of the associations that we detected have shown a low interest in learning the courseware and explicit dishonest behaviors. Therefore, we argue that there is still the need to more profoundly study the types of collaborations that can emerge in MOOCs and other types of online courses, to really understand which of those can be positive for the learning outcomes of students.

CRediT authorship contribution statement

José A. Ruipérez-Valiente: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project administration. **Daniel Jaramillo-Morillo:** Formal analysis, Writing – original draft, Writing – review & editing. **Srećko Joksimović:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision. **Vitomir Kovanović:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision. **Pedro J. Muñoz-Merino:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision. **Dragan Gašević:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Authors want to acknowledge support from PROF-XXI project, Spain (609767-EPP-1-ES-EPPKA2-CBHE-JP), the European Commission and the Spanish Ministry of Economy and Competitiveness through the Juan de la Cierva Formación program (FJCI-2017-34926).

References

[1] S. Downes, Connectivism and Connective Knowledge: Essays on Meaning and Learning Networks, National Research Council Canada, 2012, pp. 1–616.
[2] T. Liyanagunawardena, S. Williams, A. Adams, The impact and reach of MOOCs:a developing countries' perspective, ELearning Pap. 33 (2013) 38–46.
[3] M. Perez-Sanagustin, J. Maldonado, N. Morales, Estado del arte de adopcion de MOOCs en la Educacion Superior en America Latina y Europa, Techreport WPD1.1, MOOC-Maker Constr. Manag. Capacit. MOOCs High. Educcation, 2016.
[4] C. Impey, Higher education online and the developing world, J. Educ. Human Develop. 9 (2) (2020).
[5] D. Jaramillo-Morillo, J. Ruipérez-Valiente, M.F. Sarasty, G. Ramírez-Gonzalez, Identifying and characterizing students suspected of academic dishonesty in SPOCs for credit through learning analytics, Int. J. Educ. Technol. Higher Educ. 17 (1) (2020) 45.
[6] G. Alexandron, J.A. Ruipérez-Valiente, Z. Chen, P.J. Muñoz Merino, D.E. Pritchard, Copying@Scale: Using harvesting accounts for collecting correct answers in a MOOC, Comput. Educ. 108 (2017) 96–114.
[7] J. van der Linden, G. Erkens, H. Schmidt, P. Renshaw, Collaborative learning, in: New Learning, Springer, 2000, pp. 37–54.
[8] A.-M. Nortvig, R.B. Christiansen, Institutional collaboration on MOOCs in education—A literature review, Int. Rev. Res. Open Distrib. Learn. 18 (6) (2017).
[9] L. Bacon, L. MacKinnon, The challenges of creating successful collaborative working and learning activities in online engineering courses, in: Proceedings of the 14th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Engineering Innovations for Global Sustainability", Latin American and Caribbean Consortium of Engineering Institutions, 2016.
[10] J. Chauhan, An insight to collaboration in MOOC, Int. J. Adv. Eng. Res. Develop. 4 (7) (2017).
[11] J. Chauhan, S. Taneja, A. Goel, Enhancing MOOC with augmented reality, adaptive learning and gamification, in: 2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE), 2015, pp. 348–353.
[12] J. Reich, Rebooting MOOC research, Science 347 (6217) (2015) 34–35, Publisher: American Association for the Advancement of Science Section: Education Forum.
[13] V. Kovanović, S. Joksimović, O. Poquet, T. Hennis, P. de Vries, M. Hatala, S. Dawson, G. Siemens, D. Gašević, Examining communities of inquiry in massive open online courses: The role of study strategies, Internet Higher Educ. 40 (2019) 20–43.
[14] A. Hellas, J. Leinonen, P. Ihantola, Plagiarism in take-home exams: Help-seeking, collaboration, and systematic cheating, in: Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education - ITiCSE '17, ACM Press, 2017, pp. 238–243.
[15] A.S. Lan, A.E. Waters, C. Studer, R.G. Baraniuk, Sparse factor analysis for learning and content analytics, J. Mach. Learn. Res. (2013) 1959–2008, arXiv:1303.5685.
[16] A.E. Waters, C. Studer, R.G. Baraniuk, Bayesian pairwise collaboration detection in educational datasets, in: 2013 IEEE Global Conference on Signal and Information Processing, 2013, pp. 989–992, ISSN: null.
[17] T. Staubitz, T. Pfeiffer, J. Renz, C. Willems, C. Meinel, Collaborative learning in a MOOC environment, ICERI2015 Proc. (2015) 8237–8246.
[18] A. Cohen, U. Shimony, R. Nachmias, T. Soffer, Active learners' characterization in MOOC forums and their generated knowledge, Br. J. Educ. Technol. 50 (1) (2019) 177–198.
[19] A.C.A. Holanda, P.A. Tedesco, E.H.T. Oliveira, T.C.S. Gomes, MOOCOLAB - a customized collaboration framework in massive open online courses, in: V. Kumar, C. Troussas (Eds.), Intelligent Tutoring Systems, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 125–131.
[20] J.A. Ruipérez-Valiente, S. Joksimović, V. Kovanović, D. Gašević, P.J. Muñoz Merino, C. Delgado Kloos, A data-driven method for the detection of close submitters in online learning environments, in: Proceedings of the 26th International Conference on World Wide Web Companion, in: WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2017, pp. 361–368.
[21] S. Zheng, M.B. Rosson, P.C. Shih, J.M. Carroll, Understanding student motivation, behaviors and perceptions in MOOCs, in: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, in: CSCW '15, ACM, New York, NY, USA, 2015, pp. 1882–1895.
[22] A. Ramesh, D. Goldwasser, B. Huang, H. Daumé III, L. Getoor, Learning latent engagement patterns of students in online courses, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, in: AAAI'14, AAAI Press, Québec City, Québec, Canada, 2014, pp. 1272–1278.
[23] C. Alario-Hoyos, I. Estévez-Ayres, M. Pérez-Sanagustín, C.D. Kloos, C. Fernández-Panadero, Understanding learners' motivation and learning strategies in MOOCs, The International Review of Research in Open and Distributed Learning 18 (3) (2017).
[24] A. Ortega-Arranz, E. Er, A. Martínez-Monés, M.L. Bote-Lorenzo, J.I. Asensio-Pérez, J.A. Muñoz Cristóbal, Understanding student behavior and perceptions toward earning badges in a gamified MOOC, Univ. Access Inform. Soc. 18 (3) (2019) 533–549.
[25] J.A. Ruipérez-Valiente, P.J. Muñoz-Merino, C. Delgado Kloos, Detecting and clustering students by their gamification behavior with badges: A case study in engineering education, Int. J. Eng. Educ. 33 (2-B) (2017) 816–830.
[26] S. Halawa, D. Greene, J. Mitchell, Dropout prediction in MOOCs using learner activity features, Proc. Second Eur. MOOC Stakeholder Summit 37 (1) (2014) 58–65.
[27] R. Al-Shabandar, A.J. Hussain, P. Liatsis, R. Keight, Analyzing learners behavior in MOOCs: An examination of performance and motivation using a data-driven approach, IEEE Access 6 (2018) 73669–73685, Conference Name: IEEE Access.
[28] J.A. Ruipérez-Valiente, P.J. Muñoz Merino, C.D. Kloos, Improving the prediction of learning outcomes in educational platforms including higher level interaction indicators, Expert Syst. 35 (6) (2018) e12298.
[29] R. Ferguson, D. Clow, Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs), in: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, 2015, pp. 51–58.
[30] M. Khalil, M. Ebner, Clustering patterns of engagement in massive open online courses (MOOCs): the use of learning analytics to reveal student categories, J. Comput. Higher Educ. 29 (1) (2017) 114–132.
[31] B. Chen, Y. Fan, G. Zhang, Q. Wang, Examining motivations and self-regulated learning strategies of returning MOOCs learners, in: Proceedings of the Seventh International Learning Analytics & Knowledge Conference, 2017, pp. 542–543.
[32] Y. Li, H. Li, MOOC-FRS: A new fusion recommender system for MOOCs, in: 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), IEEE, 2017, pp. 1481–1488.
[33] H. Tang, W. Xing, B. Pei, Exploring the temporal dimension of forum participation in MOOCs, Distance Educ. 39 (3) (2018) 353–372.
[34] H. Lynda, B.-D. Farida, B. Tassadit, L. Samia, Peer assessment in MOOCs based on learners' profiles clustering, in: 2017 8th International Conference on Information Technology (ICIT), IEEE, 2017, pp. 532–536.

[35] L. Sanz-Martínez, A. Martínez-Monés, M.L. Bote-Lorenzo, J.A. Munoz-Cristóbal, Y. Dimitriadis, Automatic group formation in a MOOC based on students' activity criteria, in: European Conference on Technology Enhanced Learning, Springer, 2017, pp. 179–193.
[36] I. Claros, A. Garmendía, L. Echeverría, R. Cobos, Towards a collaborative pedagogical model in MOOCs, in: 2014 IEEE Global Engineering Education Conference (EDUCON), IEEE, 2014.
[37] T. Staubitz, C. Meinel, Collaborative learning in MOOCs approaches and experiments, in: 2018 IEEE Frontiers in Education Conference (FIE), (ISSN: 2377-634X) 2018, pp. 1–9.
[38] C. Brooks, C. Stalburg, T. Dillahunt, L. Robert, Learn with friends: The effects of student face-to-face collaborations on massive open online course activities, in: Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15, ACM Press, 2015, pp. 241–244.
[39] N. Li, H. Verma, A. Skevi, G. Zufferey, J. Blom, P. Dillenbourg, Watching MOOCs together: investigating co-located MOOC study groups, Distance Educ. 35 (2) (2014) 217–233.
[40] A. Ezen-Can, K.E. Boyer, S. Kellogg, S. Booth, Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach, in: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, in: LAK '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 146–150.
[41] C. Haythornthwaite, Facilitating collaboration in online learning, J. Asynchronous Learn. Netw. 10 (1) (2006) 7–24.
[42] Y. Bao, Detecting Multiple-Accounts Cheating in MOOCs (Ph.D. thesis), TU Delft, 2017, URL: http://resolver.tudelft.nl/uuid:64ee5526-8c9e-4013-9019-c63a63413ca2.
[43] C.G. Northcutt, A.D. Ho, I.L. Chuang, Detecting and preventing "multiple-account" cheating in massive open online courses, Comput. Educ. 100 (2016) 71–80.
[44] J.A. Ruipérez-Valiente, P.J. Munoz-Merino, G. Alexandron, D.E. Pritchard, Using machine learning to detect 'multiple-account' cheating and analyze the influence of student and problem features, IEEE Trans. Learn. Technol. 12 (1) (2017) 112–122.
[45] R. Swaray, An evaluation of a group project designed to reduce free-riding and promote active learning, Assess. Eval. Higher Educ. 37 (3) (2012) 285–292.
[46] O. Viberg, A. Mavroudi, Y. Fernaeus, C. Bogdan, J. Laaksolahti, Reducing free riding: CLASS – a system for collaborative learning assessment, in: E. Popescu, A. Belén Gil, L. Lancia, L. Simona Sica, A. Mavroudi (Eds.), Methodologies and Intelligent Systems for Technology Enhanced Learning, 9th International Conference, Workshops, Springer International Publishing, 2020, pp. 132–138.
[47] V. Popov, A.v. Leeuwen, S.C.A. Buis, Are you with me or not? Temporal synchronicity and transactivity during CSCL, J. Comput. Assisted Learn. 33 (5) (2017) 424–442.
[48] J. Lämsä, R. Hämäläinen, P. Koskinen, J. Viiri, E. Lampi, What do we do when we analyse the temporal aspects of computer-supported collaborative learning? A systematic literature review, Educ. Res. Rev. (2021) 100387.
[49] SPSS Statistics IBM, Twostep cluster analysis, 2021, Online; accessed 21 May 2021, https://www.ibm.com/docs/en/spss-statistics/27.0.0?topic=features-twostep-cluster-analysis.
[50] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Vol. 344, John Wiley & Sons, 2009.
[51] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O.P. Patel, A. Tiwari, E.M. Joo, W. Ding, C.-T. Lin, A review of clustering techniques and developments, Neurocomputing 267 (2017) 664–681.
[52] M. Laal, M. Laal, Collaborative learning: what is it?, Procedia - Social and Behavioral Sciences 31 (2012) 491–495.
[53] J. Eisenberg, To cheat or not to cheat: effects of moral perspective and situational variables on students' attitudes, J. Moral Educ. 33 (2) (2004) 163–178.
[54] J. Yardley, M.D.R. Ph.D, S.C. Bates, J. Nelson, True confessions?: Alumni's retrospective reports on undergraduate cheating behaviors, Ethics Behav. 19 (1) (2009) 1–14.
[55] R. Baker, J. Walonoski, N. Heffernan, I. Roll, A. Corbett, K. Koedinger, Why students engage in "gaming the system" behavior in interactive learning environments, J. Interactive Learn. Res. 19 (2) (2008) 185–224.
[56] M. Cocea, A. Hershkovitz, R.S.J.d. Baker, The impact of off-task and gaming behaviors on learning: Immediate or aggregate? in: Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems that Care: from Knowledge Representation to Affective Modelling, IOS Press, NLD, 2009, pp. 507–514.
[57] R.S.J.d. Baker, A.T. Corbett, I. Roll, K.R. Koedinger, Developing a generalizable detector of when students game the system, User Model. User-Adapted Interact. 18 (3) (2008) 287–314.
[58] C. Lampe, D.Y. Wohn, J. Vitak, N.B. Ellison, R. Wash, Student use of facebook for organizing collaborative classroom activities, Int. J. Comput. Supported Collabor. Learn. 6 (3) (2011) 329–347.
[59] S. Joksimović, O. Poquet, V. Kovanović, N. Dowell, C. Mills, D. Gašević, S. Dawson, A.C. Graesser, C. Brooks, How do we model learning at scale? A systematic review of research on MOOCs, Rev. Educ. Res. 88 (1) (2018) 43–86.
[60] D. Gašević, S. Dawson, T. Rogers, D. Gasevic, Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success, Internet Higher Educ. 28 (2016) 68–84.
[61] B. Thoms, E. Eryilmaz, How media choice affects learner interactions in distance learning classes, Comput. Educ. 75 (2014) 112–126.
[62] S. Joksimović, A. Manataki, D. Gašević, S. Dawson, V. Kovanović, I.F. De Kereki, Translating network position into performance: importance of centrality in different network configurations, in: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, 2016, pp. 314–323.
[63] M. Zapata-Ros, El diseño instruccional de los MOOC y el de los nuevos cursos abiertos personalizados, Rev. Educ. Dist. (RED) (45) (2015).

José A. Ruipérez-Valiente completed his B.Eng. and M.Eng. in Telecommunications at Universidad Católica de San Antonio de Murcia (UCAM) and Universidad Carlos III of Madrid (UC3M) respectively, graduating in both cases with the best academic transcript of the class. Afterwards, he completed his M.Sc. and Ph.D. in Telematics at UC3M while conducting research at Institute IMDEA Networks in the area of learning analytics and educational data mining. He completed two postdoctoral periods, one at MIT and a second one at the University of Murcia with the prestigious Spanish fellowship Juan de la Cierva. He is currently an Associate Professor of Software Engineering and Artificial Intelligence at Complutense University of Madrid.

Daniel Jaramillo-Morillo completed his B.Eng in Electronic and Telecommunications and his M.Eng in Telematic Engineering at the Universidad del Cauca in 2017. He was a Young Researcher with a scholarship from Colciencias (Colombia) in 2017 and is currently a Ph.D Student in Telematic Engineering. He is researcher and administrator of a learning platform at the Universidad del Cauca.

Srecko Joksimovic completed in 2017 a Ph.D. in Learning Analysis and Information Technology from the University of Edinburgh and Simon Fraser University respectively. He is a Senior Lecturer in Data Science at the Education Futures, University of South Australia. His research is centered around augmenting abilities of individuals to solve complex problems in collaborative settings. Srecko is particularly interested in evaluating the influence of contextual, social, cognitive, and affective factors on groups and individuals as they solve complex real-world problems.

Vitomir Kovanovic is Research Fellow at the School of Education, University of South Australia and a Data Scientist at the Teaching Innovation Unit, University of South Australia. His research focuses on the development of novel learning analytics systems using learners' trace data records collected by learning management systems with the goal of understanding and improving student learning. He obtained his Ph.D. in Informatics, at the University of Edinburgh, United Kingdom in 2017.

Dr. Pedro J. Muñoz-Merino is Associate Professor at Universidad Carlos III de Madrid. His main areas of expertise are on data analysis, educational data mining, learning analytics and adaptive systems. He teaches on data science topics at his university and at other institutions such as INAP (National Institute of Public Administration). Pedro is a Telecommunications and Telematics Engineer from the Universidad Politecnica de Valencia and a Ph.D. in Telematics Engineering from the Universidad Carlos III de Madrid.

Dragan Gašević is Professor of Learning Analytics in the Faculty of Information Technology and Director of the Centre for Learning Analytics at Monash University. Before the current post, he was Professor and Chair in Learning Analytics and Informatics in the Moray House School of Education and the School of Informatics and Co-Director of Centre for Research in Digital Education at the University of Edinburgh. He is B.S. in Computer engineering and informatics at Military Technical Academy, M.S. in Software systems and electrical engineering and Ph.D. in Information systems at University of Belgrade.