Education and Information Technologies (2022) 27:3529–3565
https://doi.org/10.1007/s10639-021-10742-6

Revealing latent traits in the social behavior of distance learning students

Rozita Tsoni[1] · **Christos T. Panagiotakopoulos**[2] · **Vassilios S. Verykios**[1]

Received: 25 May 2021 / Accepted: 2 September 2021 / Published online: 29 September 2021

Abstract

This paper proposes a multilayered methodology for analyzing distance learning students' data to gain insight into the learning progress of the student subjects both in an individual basis and as members of a learning community during the course taking process. The communication aspect is of high importance in educational research. Additionally, it is difficult to assess as it involves multiple relationships and different levels of interaction. Social network analysis (SNA) allows the visualization of this complexity and provides quantified measures for evaluation. Thus, initially, SNA techniques were applied to create one-mode, undirected networks and capture important metrics originating from students' interactions in the fora of the courses offered in the context of distance learning programs. Principal component analysis and clustering were used next to reveal latent students' traits and common patterns in their social interactions with other students and their learning behavior. We selected two different courses to test this methodology and to highlight convergent and divergent features between them. Three major factors that explain over 70% of the variance were identified and four groups of students were found, characterized by common elements in students' learning profile. The results highlight the importance of academic performance, social behavior and online participation as the main criteria for clustering that could be helpful for tutors in distance learning to closely monitor the learning process and promptly interevent when needed.

✉ Rozita Tsoni
rozita.tsoni@ac.eap.gr

Christos T. Panagiotakopoulos
cpanag@upatras.gr

Vassilios S. Verykios
verykios@eap.gr

[1] School of Science and Technology, Hellenic Open University, Patras, Greece

[2] Department of Primary Education, School of Humanities and Social Sciences, University of Patras, Patras, Greece

Keywords Distance learning · Learning Analytics · Social Network Analysis · Principal Components Analysis · Clustering · Discussion Forum

1 Introduction

Distance Learning emerged as a field almost a century ago (Valentine, 2002). However, lately, new powerful tools and massive changes in the educational demand have brought Distance Learning into the spotlight of our personal and social lives. Two milestones cultivated the hype of online learning in the past decade: the rise of MOOCs and the Covid-19 pandemic. The revolution brought about by the massive participation in MOOCs and the proliferation of Open Educational Resources originated in conventional universities that mainly provided face-to-face courses (MIT, Stanford, Yale, etc.) New educational models that fueled participation were adopted but at the same time, the dropout rates were skyrocketed (Laurillard et al., 2018). The massive public acceptance and adoption of online courses created significant challenges and urged researchers of the educational field to find answers to important questions concerning teaching and learning quality, learners' behavior, policies, educational design, and many other relevant issues.

In this already marginal situation, the crisis of the Covid19 pandemic came along, forcing the need for emergency measures to be taken hastily. To the ears of the inexperienced and novice in Distance Education methodology, every effort allocated towards the preservation of the educational process offering from distance was simply considered a "distance education program". Nevertheless, in most cases, what was actually happening can be best described as *"emergency remote education"* (Bozkurt, et al., 2020). To avoid an escalated risk from the degeneration of *distance education* to *emergency remote education,* a reflection of the basic values that established the Distance Education field is needed.

The impact of Learning Analytics and Knowledge (LAK) and Educational Data Mining (EDM) in the Distance Education field is crucial because they can offer solid proof based on the plethora of data created by the online educational activity. There is a large number of analytical methods that can be applied to this data such as social network analysis, sentiment analysis, influence analysis, discourse analysis, classification, clustering, Bayesian modeling, relationship mining, and discovery with models (Siemens & Baker, 2012), and even larger number of tools that offer an implementation of such methods. Driven by this notion that *"learning analytics is about learning. As such, the computational aspects of learning analytics must be well integrated within the existing educational research"* (Gašević et al., 2015), we adopt a three-layered learning analytics methodology that is immersed into our educational research of investigating traits of students that facilitate learning in a distance education scenario. Hence, we propose a sequence of methodological steps aiming to lead to a better understanding of a distant students' community. Several data sources were combined to mine primitive variables concerning students' academic performance and social interaction.

As it is known, the physical distance of the persons involved in the educational process deprives important information and impoverishes their interaction (Bouhnik & Marcus, 2006). Thus, it is necessary to replace those missing properties so that the educational activity could regain its social character. Michael G. Moore (2007) addresses the problem of geographical separation in the transactional distance theory, using the concepts of structure (that concerns the educational methodology) and dialogue (between learners, tutors and teaching material) to explain how learners' autonomy is affected in a distance learning course. At the same time, Conectivism, a new educational theory, fine-tuned with the internet era, emerged by Siemens and Downes (2008). Connectivism expands the social aspect of learning in the digital age stressing that knowledge lies within networks, where nodes of information are linked in a meaningful way. Thus, SNA techniques enriched our data with network related metrics and Principal Components Analysis (PCA) analysis allowed us to create a new latent space of lower dimensionality. Consequently, we were able to cluster students' community, dentify common learning patterns and create actionable information.

2 Research problem statement

The problem of observing and understanding students' behavior in order to improve their educational experience and their learning outcomes has been in the spotlight of educational research since the beginning of typical education. In online learning environments, the lack of direct communication is compensated by the vast amount of available data generated in the educational context. Therefore, we are aiming to propose a multilayered methodology that would make good use of these data to produce actionable conclusions for the improvement of the educational process as it is conceived by the students. Two main aspects of learning are taken into consideration in our proposed methodology: the academic performance indices and the social interaction indices. The objective of our research is not focused on the prediction of the final achievement. Instead, we attempt to explore features that go beyond Key Performance Indicators (KPI) and uncover deeper aspects of learning. We aim to identify groups of typical students' profiles, so that tutors could easier make targeted interventions. Thus, data from a full academic year, along with students' grades, were used to create an integrated image about their learning behavior and evolution as members of a learning community. Hopefully, a more sophisticated taxonomy than the dipole "good student-weak student" will occur. The identification of learning behavioral patterns could lead further research about performance prediction and the creation of early warning systems. In addition, it is of high importance to quantitatively imprint the contribution of each variable concerning different aspects of the learning process deriving by students' digital traces. This result would allow the creation students' profiles with increased accuracy. We summarize the goals of our study in the following research questions:

RQ1: *Which methodological steps along with specialized tools and algorithms can combine simple metrics and derived variables to reveal aspects of students'*

participation and performance, difficult to monitor aspects of students' participation and performance in distance learning settings? This methodology should be threefold, in the sense that it should include: a) a process that would capture learners' social interaction, b) a process that would group variables and would explain their relations and c) a clustering method for the students' community based on the intergraded features. Every layer of analysis would provide new variables based on the previous layer in order to create a new space where latent features can best describe students' interactions of high complexity.

RQ2: *Within a number of available primitive and derived variables, what is the contribution of each one of them, how are they correlated and which principal components can capture the majority of the variance?* In particular, we attempt to explore the contribution of observable features (like the number of views or the number of posts), together with academic performance features (like grades) and more advanced attributes concerning students' interaction. By definition, some of these variables are interrelated. Thus, it is necessary to explore their relations. In addition, in the search of a simple, yet elaborated way of describing the learning behavior in an online setting, it is necessary to indicate factors that would bring together the important attributes and would transform our data, preventing biased and skewed results and reducing the multidimensionality of the problem.

RQ3: *Which obvious or latent features signify the evolution of learning between different learning communities and how do they differentiate the structure of the learning community?* Each methodology which is tested in the context of a specific course, produces results which are strongly related with a number of factors concerning the scientific field of the course, the educational design, the teaching and assessment methods and many others. However, the learning process is much alive, constantly transforming the learning community. Moreover, the learning process itself changes within the learning community as the learners and their tutors adapt to the demands of a society that constantly builds up ties and balances the behavior of different personalities and learning profiles. Therefore, it is meaningful to investigate different learning communities so that will be able to identify those aspects that capture their evolution as a whole but also individually.

To address the research questions, a knowledge discovery process is proposed based on students' data from their activity in an online learning environment. SNA is used, not as a final step, but as a pre-processing phase to produce indices for PCA. The eigenvectors of the new space produce third-level variables that are used for clustering. Consequently, the groups that occur, sort students based on hidden characteristics revealing deeper aspects of their learning behavior.

3 Related work

Increasing amounts of important data across the board of student academic life and performance are becoming openly available, providing more resources with a high potential for improving learning (Bates, 2019). Educational organizations have already started to reap the benefits from this potential and researchers show

increased interest. A significant factor for successful tutoring and student guidance in a Distance Learning Course goes hand in hand with the direct access of the educational stakeholders to accurate and updated information. Therefore, a Learning Analytics (LA) approach that can provide for this actionable information that is missing from the process itself, is a necessity in Distance Education. Often there is a mismatch between the ideal dynamic of a group of students and their actual potential (Hernández-García & Suárez-Navas, 2017). The importance and the benefits that LA offers to education are highly recognized (Sergis & Sampson, 2017) as it can shift the educational research in deeper understanding (Viberg et al., 2018) and provide the means to tailor learning experience according to individual learners needs (Tsoni et al., 2019a, b, c). The knowledge originated by LA in order to be actionable has to be meaningful. However, not all meaningful knowledge is actionable. As Gašević et al. (2015) pointed out, some significant predictors of academic performance (i.e., the number of logins) cannot be used in a practical sense to improve academic performance if they don't come along with suggestions for teaching and learning improvement.

Students' profiling, predictive models, personalization, and adaptive learning are some of the areas that LA and EDM are focusing on (Siemens, 2013). In several studies students' data, including grades, log files, and forum participation data, were utilized to build predictive models, referring to the students' final achievement (Bayer et al., 2012; Crossley et al., 2017; Gkontzis et al., 2018; Romero et al., 2013a, b). Chiu & Hew (2018) demonstrated that the number of views in a forum, had greater predictive power than the number of posts a student makes. The need that the forum activity has to be investigated in a deeper level than by simply looking to posts' counts was also proven by Sun et al. (2018) who studied the impact of students' grouping tactics on the forum interaction. The importance of selecting advanced, tailored tools that can provide analysis that captures the complexity of group or community formation in distance learning was stressed by Hernández-García, and Suárez-Navas, (2017). Methodological issues are treated thoroughly by the LA community. LA approaches can be applied in a wide range of educational problems (Klašnja-Milicevic & Ivanovic, 2018) however, the power of Artificial Intelligence has changed the traditional hypothesis-testing research approach.

Educational research is a sensitive domain and it is important to maintain the balance between theory-driven, humanistic approaches and advanced algorithms with high predictive power. Sharma et al., (2019) proposed a methodology that combined theory-driven research questions with data modeling for feature extraction and prediction in educational settings. A synergy of pedagogical criteria and machine learning techniques of analysis lead Gkontzis et al., (2020) to accurate performance prediction by dividing the academic year of students into six periods.

Regarding social interaction, SNA was used by numerous scientists to study the complex relations between the members of learning communities. Several studies were focusing on tools and applications (Batagelj & Mrvar, 1998; Borgatti et al., 2002; Bastian et al., 2009; Chen et al., 2010; Csardi, 2013) while others aimed to identify network features that can improve the educational research (Sternitzke et al., 2008; Yusof, & Rahman, 2009; Traxler et al., 2018). Two successive studies (Lotsari et al., 2014; Kagklis et al., 2015) leveraged Network

Analysis of forum activity to create students' profile based on their online participation. Network representation through time revealed the evolution of students' community and along with polarity analysis can provide insights into the social aspect of their learning behavior (Tsoni et al., 2019a, b, c). The idea of a detailed analysis, including the content of the discussion fora, was one of the main findings of a literature review conducted by Cela et al., (2015). Additionally, social network metrics derived from the forum activity were used as indicators of the structure of communities of practice and communities of inquiry (Jan & Vlachopoulos, 2019). Amano et al. (2019) proposed a collaborative forum-based learning design, including pre-learning and post-learning activities based on SNA findings.

Finally, De-Marcos et al. (2016) conducted PCA of network metrics and additionally examined the correlation between those metrics and students' academic achievement. They found a moderate correlation between most centrality measures and learning achievement. However, they pointed out that this result is affected by the educational design that included graded forum related assignments. PCA was also used to identify certain activities within the use of Web 2.0 that students had selected, as indicators of students' success (Giovannella et al., 2013).

Although SNA, PCA, and Clustering have benefited the educational research, in the vast majority of the relevant literature, they are used separately. SNA is used as a final modeling method to reveal information about the community of learners. PCA and Clustering are modeling primitive variables drawn from LMSs or questionnaires. The sequential combination of the abovementioned techniques could address the challenge to intergrade data that represent various aspects

Table 1 A taxonomy of variables describing students' explicit features

	Academic Performance	Online Participation
Primitive	• WA_1 • WA_2 • WA_3 • WA_4 • WA_5 • WA_6 • Final Grade	• Views
Derived	• Pass/Fail • Exams	• Forum Participant • Degree • Weighted Degree • Harmonic Closeness Centrality • Betweenness Centrality • Eigenvector Centrality • Eccentricity • Authority • PageRank

of learning such as academic achievement, online participation, and social interaction.

4 Feature engineering

Two main types of variables were chosen as the most relevant to our research questions: variables concerning the academic performance and variables concerning their online participation (Table 1). Furthermore, there is a distinction between primitive and derived variables. Primitive variables were mined directly from the online environment that support learners' activity, while derived variables were produced by a process that followed the initial data mining.

4.1 Academic performance features

To assess the academic performance of the students, the grades of their mandatory written assignments were used. In the first-year course students have to hand in up to six written assignments and an average grade of 50% is prerequired for their participation in the final exams (variables WA_1 to WA_6). The same applies in the second-year course but there are 5 written assignments instead. The variable "exams" is a binary variable that represents whether a student has the right to participate in the final exams, that is, whether she/he have gained a minimum average grade of 50% in the written assignments.

4.2 Online participation features

Concerning online participation, the variable "views" captures the total number of forum views of each participant for the whole academic year. We have to point out that the binary variable "forum participant" isn't based on the number of views as all of the students in both courses have visited the forum just to read other posts and also denote a rather passive participation. Instead, the degree (derived by the SNA) of each user was used (that has to be non-zero for a student to be characterized as a forum participant) in order to capture active participation. That means that a forum participant has at least one forum post during the academic year.

Attributes derived from the online participation of students were based on network metrics which are strongly affected by the type of network that was used to imprint the students' interaction in their forum. One-mode undirected networks were created for both courses including all students and tutors. Each node represents a person and the edges represent ties between nodes/participants that have posted in the same thread. Thus, it is obvious that students who did not post on the forum, appear to have zero values in all network-related metrics.

In order to describe the metrics and the algorithms related to the network of students' communication, we created a small, sample network with the same properties. That is, an undirected, one-mode network of five nodes and nine edges (Fig. 1). The

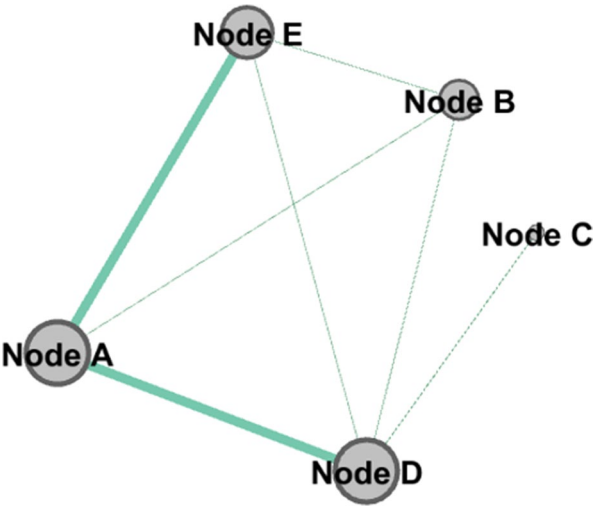


Fig. 1 A sample undirected one-mode network of five nodes

Table 2 The list of links between nodes in the sample network

Edges' list	
Node A	Node B
Node A	Node D
Node A	Node E
Node B	Node D
Node B	Node E
Node C	Node D
Node D	Node E
Node A	Node D
Node A	Node E

Table 3 Basic metrics of the sample network's nodes

Id	Node A	Node B	Node C	Node D	Node E
Degree	3	3	1	4	3
Weighted Degree	5	3	1	5	4
Closeness Centrality	0.8	0.8	0.571429	1	0.8
Harmonic Closeness Centrality	0.875	0.875	0.625	1	0.875
Eccentricity	2	2	2	1	2
Betweenness Centrality	0	0	0	3	0
Eigenvector Centrality	0.919775	0.919775	0.324944	1	0.919775
Authority	0.482044	0.482044	0.169648	0.523566	0.482044
Hub	0.482045	0.482045	0.169651	0.523561	0.482045
PageRank	0.208537	0.208537	0.090383	0.284005	0.208537

small number of nodes allows creating a network of an appropriate magnitude so that the relations between nodes to be obvious and, at the same time, to be adequate so that complex metrics to be meaningful. A list of edges between the nodes is presented in Table 2 and the metrics that emerged for the sample network are listed in Table 3. The theoretical underpinning of the SNA which is of high importance for our methodology based on which we attempt to build new knowledge, is presented below.

The *degree* is the simplest metric of the network showing the number of nodes adjacent to a given node. Node's A degree is equal to three because it is linked to three nodes: B, D and E. In the forum participation network, the degree of a participant reveals the number of other persons who posted in a common thread with her/him. Additionally, the *weighted degree* adds up the frequency of each interaction. Two participants who have the same degree interacted with the same number of peers but a higher weighted degree indicates a more active participant because she/he made more posts. In the sample network the weighted degree of Node A is equal to five because it is connected five times with other nodes (there is one link with node B, two links with node D and two links with node E).

Closeness centrality (x), (1) was proposed by Latora and Marchiori (2001) that can be used in both connected or unconnected graphs. Closeness centrality is defined as:

$$C(x) = \frac{N-1}{\sum_{x \neq y} \mathrm{d(x,y)}} \tag{1}$$

where x,y are two vertices in the network G, d(x,y) denotes their distance, that is the number of edges in a shortest path connecting them, and N is the number of nodes in the graph. Thus, a node with high closeness centrality is a central node in the network. In other words, if the sum of the distances is large, then the closeness is small and vice versa (Metcalf & Casey, 2016). In the sample network, the node D is directly connected with all the other nodes, thus, it has the highest closeness centrality. On the contrary, node C is directly connected only with node D, so it has the lowest closeness centrality. *Harmonic Closeness Centrality* $H(x)$, can be computed as follows:

$$H(x) = \sum_{y \neq x} \frac{1}{d(y,x)} \tag{2}$$

In case that there is no path between x, y the value of $\frac{1}{d(y,x)}$ equals zero. Harmonic closeness centrality was chosen over closeness centrality in our study due to the disconnected nodes that were included in the networks that represent the non-active participants. High harmonic closeness centrality of the nodes in the students' communication network indicates participants, in the core of the network, that are directly or closely related with a large number of others.

Eccentricity $e_{G(v)}$, is a distance measure that is considered to be much simpler than closeness centrality (Sereni et al., 2018). In a given network G the eccentricity $e_{G(v)}$ of a node v is the maximum distance between node v and u over all the nodes of the network. From the definition below (Eq. 3) it is obvious that a node with high eccentricity is a distant node.

$$e_{G(v)} = max\{dist_G(v,u) : u \in V\} \in N \cup \{\infty\} \tag{3}$$

In the sample network all nodes' eccentricity is equal to two because it takes maximum two steps for each one of them to reach any other node. An exemption is node D that is directly connected with all other nodes; therefore, its eccentricity equals to one. Certain nodes might have an important role in connecting two, otherwise separate, groups of the network called bridges. In a learning community, this is a very important feature that indicates a person who facilitates communication between distant members. This feature is captured by the metric of betweenness centrality. *Betweenness centrality* $C(x)$, (4) is loosely defined as the number of times that a node is part of the shortest path between two other nodes. Thus:

$$BC(x) = \sum \frac{\sigma_{(v,w)}(x)}{\sigma_{(v,w)}} \tag{4}$$

Node D lies in the shortest path between C and A, C and B and C and E, thus its betweenness centrality equals to three whereas, all other nodes' is zero because they don't participate in other shortest paths. *Eigenvector centrality* $EC(x_i)$, (5) captures the influence a node has. It is proportional to the sum of centralities of the nodes who are straightly linked to it. Thus, a participant with significant neighbors gains significance itself. Given matrix A, an eigenvector for this matrix is a vector x that satisfies the equation $Ax = \lambda x$ for some constant λ. This would give the equation:

$$EC(x_i) = \frac{1}{\lambda} \sum_{j \in M(i)} x_j \tag{5}$$

where $j \in M(i)$ means that the sum is overall j such that the nodes i, j are connected. Each node gets a centrality score affected by the value of the nodes to whom is connected with. Therefore, all nodes in the sample network are benefited from its connection with the highly influential node D that adds value on them. Eigenvector centrality is more meaningful in larger networks where nodes have different neighbors.

Advanced metrics of higher complexity are derived by elevated algorithms aiming to illustrate a node's value in a network in accordance with the quality of its neighbors and the strength of their ties. In this respect, there are two widely used algorithms: HITS that produces the metrics "Authority" and "Hub" and the PageRank algorithm that was initially designed as a measure of influence. Both are based on the Principle of Repeated Improvement that is an iterative process where an initial value is assigned to a node and then a re-weighting process begins re-assigning new values according to each node's connections until the convergence criteria are met. In a students' communication network, this process allows to efficiently imprint the augmented influence of a person in the community as she/he establishes her/his relations with other participants considering their level of influence also. In the next subsection, the main characteristics of these algorithms are briefly discussed.

4.3 HITS and pagerank algorithms

Hyperlink-Induced Topic Search (HITS), is a link analysis algorithm for webpages ranking (Kleinberg, 1999). HITS algorithm is also known as "Hubs and Authorities". Initially, a hub and an authority value are assigned in each node according to its incoming and outgoing edges. An iterative process begins correcting these values until a default point of convergence is met.

A high value of hub means that the node points to high authorities i.e., nodes with valuable information, represented as nodes with high in-degree in a directed network. Respectively, a node with high authority is being pointed by good hubs in a mutually reinforcing relationship. A good hub adds value to an authority and subsequently, the authority becomes better, adding more value to the hub in a recurrent process that, after several iterations, converges to a final result. The degree of convergence *e (epsilon),* that determines the ending point of the iterative algorithm is the maximum divergence between two sequential results. Thus, at first, each node p is assigned with a hub value and an authority value equal to one ($x^{<p>} = 1$ and $y^{<p>} = 1$). Then, these values are recalculated according to the incoming and the outgoing links of the nodes respectively. The operation that updates the weights for hubs and authorities ($x^{<p>}$:authority update rule, and $y^{<p>}$:hub update rule), respectively are:

$$x^{<p>} \leftarrow \sum_{q:(q,p)\epsilon E} y^{<p>} \tag{6}$$

and

$$y^{<p>} \leftarrow \sum_{q:(q,p)\epsilon E} x^{<p>} \tag{7}$$

where V is a collection of nodes represented as a directed graph G = (V, E) and a directed edge (p, q) indicates the presence of a link from p to q. In a one-mode undirected network the incoming links (represented by the in-degree) and the outgoing links (represented by the out-degree) are the same because of the reciprocity of each edge. Therefore, hub and authority result to same value. In the sample network the highest hub and authority is assigned to Node D and the lowest to the node C. Nodes A, B and E have the same value of hub and authority because they are connected to the same nodes creating triangles (that means that each one of them is directly linked with all the direct connections of its neighbors).

HITS algorithm is used in a wide range of fields like biology (Lei et al., 2019; Szczurek & Horeni, 2018), traffic modeling (Tran & Draeger, 2021), location-based services (Farahat et al., 2006), education and social sciences (Capocci et al., 2005; Hu et al., 2017; Zhou et al., 2018).

Another similar advanced algorithm is PageRank. PageRank is a widespread scoring function that measures objectively the subjective notion of importance initially introduced as a Google feature for Webpages ranking (Brin & Page, 1998). It is also based on the idea of assigning value to a node depending on its connections following the principle of repeated improvement. The PageRank is defined

for directed graphs; however, some studies use it in undirected graphs (Abbassi & Mirrokni, 2007; Andersen et al., 2006; Iván & Grolmusz, 2011; Perra & Fortunato, 2008; Wang et al., 2007) in several different fields (Brown, 2017; Jiang et al., 2017; Kandiah & Shepelyansky, 2012; Lazova & Basnarkov, 2015; Mooney et al., 2012; Mukai, 2013).

As a measure of influence, it is related to the number of connections a node has. Furthermore, Lotfi et al., (2019) showed that there is a PageRank vector ordering that is proportional with the variance of degree sequence in an undirected graph. The widespread use of PageRank in the past years brought certain modifications in the algorithm including fairness-aware link analysis (Tsioutsiouliklis et al., 2020) to avoid bias against a protected group defined by the value of a sensitive attribute.

In the sample network PageRank provides analogous information as the HITS algorithm. This is due to the small size of the network where all nodes have the same neighbors. In larger networks HITS and PageRank algorithms capture different details in their structure as they exhibit many different features (Grover & Wason, 2012). One of them is that HITS is applied to the local neighborhood of a node whereas PageRank is applied to the entire network. There is a significant interrelation between metrics of the network determining largely the analysis process followed. A more thorough presentation of the abovementioned metrics could point out their differences, their interrelations and their special features, however, this is out of the scope of this study. Unlike most of the educational research incorporating SNA in their methodology and presenting its metrics as their final results, in this study SNA is rather a pre-processing step, providing the derived variables for the next steps of the analysis.

5 A Three-layered learning analytics approach for understanding students learning behavior

The research problem that we are confronting concerns the improvement of understanding students' behavior by making the most of the data accumulated by their online learning activity. There is enough theoretical evidence to drive the selection of data about their academic performance, their online presence, and their position in the learning network as it is expressed in the course's forum community. To achieve our research goal three main types of modeling are used:

a) Social Network Analysis, to capture the social aspect of learning
b) Factor Analysis, to group primitive and derived variables and reveal their significance
c) Clustering, to identify common features in students' learning behavior.

Our model follows a Learning Analytics cycle (Fig. 2). LA cycles were actually evolved from older learning theories and enclose adequate theoretical grounding (Clow, 2012).

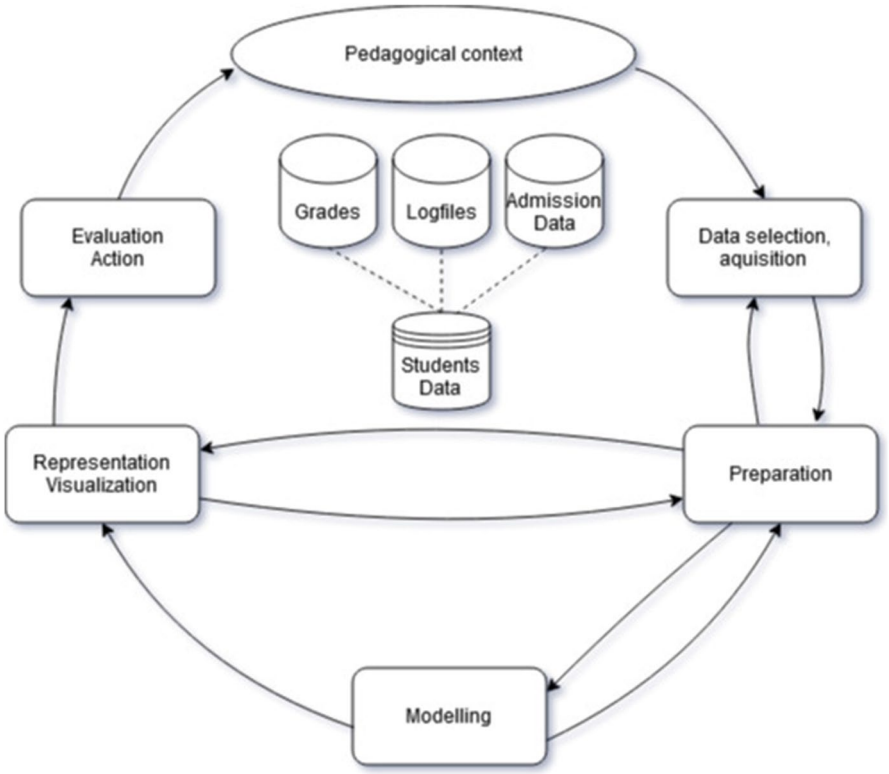


Fig. 2 The LA cycle

Starting out from Pask's *conversational theory* (1976) and Kolb's *learning cycle* (1984) and moving on to more recent views like the *reflection-in-action* and the *reflection-on-action* of Schön (1991) and finally, to Laurillard's *conversation of action* and *conversation of conception* (2013), several similar iterative processes of generating data, producing metrics and visualization and proving feedback, aim to improve the education. The LA cycle narrates briefly the idea of a methodology for discovering meaningful information concerning the learning process bases on observable and countable features.

5.1 Social network analysis techniques

Social Network Analysis is an interdisciplinary field that collaborates Network Analysis, Social Science, and Graph theory. Some common tasks are the identification of most central nodes, the detection of inner communities and cliques, and the investigation of the link strengthens (strong and weak ties). There are two main SNA approaches. The *socio-centric approach* faces the network as a whole aiming to discover structural patterns. The *egocentric approach,* on the other hand, focuses

on individuals aiming to discover types of relations and contributions in the community. A mixed approach is often used to meet the challenges of high complexity problems.

There are several software packages (e.g., Ucinet, Gephi, Pajek, NetMiner, and R packages) offering networks' creation, visualization, and manipulation of networks, quantitative statistical analysis, and community detection. Our data size and our research goals lead us to the selection of the software Gephi (Bastian et al., 2009). It provides visualization, appearance and layout formatting, statistical analysis and metrics exportation, and a large number of add-ons for advanced features.

Students' forum data were imported to Gephi for analysis. A two-node network (participant-to-thread) was initially created and the first step in Gephi modeling was its projection into a unimodal network (participant-to-participant). A unimodal network serves best our research goals due to the focus on actors' behavior regardless of the topic of the discussion. The second step was the appearance and layout customization using color and size partition for certain nodes' features and the Force Atlas2 algorithm (Jacomy et al., 2014) to achieve a clear visualization of participants interaction. The third step was to run statistics to get centrality measures and more complex algorithms' outputs (HITS and PageRank). For the iterative algorithms the degree of convergence epsilon used was $e = 0.0001$. Finally, the metrics' output was exported and merged with the dataset to prepare the next level of analysis. Evidently, the purpose of Gephi analysis was twofold: to produce information directly used to understand participants' place and role in the learning community and to derive metrics for further analysis.

5.2 Dimensionality reduction techniques

Studying human behavior has always been a complex and compound task. Even though we investigate certain interactions in a closed and relatively controlled environment, without including cultural or personal data, the high complexity remains, creating a large volume feature space, difficult to interpret. Dimensionality reduction is a statistical process providing fewer parameters and a simpler structure maintaining the majority of variation. Exploratory factor analysis can reveal can provide insight into the learning process are reveal traits that weren't obvious in the preliminary multidimensional space. There are several methods of dimensionality reduction. We chose to operate PCA. PCA is an orthogonal transformation that provides principal components as a linear combination of the initial variables weighted by their variance. This allows us:

a) to simplify our study coping with fewer, unrelated components,
b) to easily visualize the new lower-dimensional feature space,
c) to use the main components that capture the majority of variance for further analysis (see cluster analysis below),
d) to examine the impact of the initial variables and interpret the results.

The IBM SPSS syntax editor was used to perform the analysis mainly due to the flexibility and the repeatability that can provide. SPSS syntax editor is a programming language that uses commands rather than the graphical interface of SPSS and also the total sequence of commands can be stored and repeated precisely. The new dataset contains the network metrics along with the initial students' data is imported to SPSS. Firstly, simple statistics, distributions, and visualization were conducted. Sixteen variables were standardized and used for PCA. Additionally, parallel analysis determined the number of factors to be kept. Finally, we visualized the latent variables that were produced in the new space.

5.3 Cluster analysis

Until this point, the sequence of actions created a new space with latent variables that explain most of the variance of our data. However, a grouping method is needed to indicate students' common features that cannot be directly observed using the initial variables. Thus, exploratory cluster analysis was conducted using the SPSS syntax editor. The process consisting of three steps using two algorithms is presented below:

a) Hierarchical Cluster Analysis (between-group linkage method using the Euclidean distance) to determine the optimal number of clusters.
b) K-means clustering to identify the clusters of students with common features.
c) Computation of mean squared error of cluster distances to examine the cohesion of the clusters.

In the next section, the experimental testing of the proposed methodology is presented.

6 Experimental design

The Hellenic Open University is the only university in Greece that provides exclusively distance education. Additionally, it is the only university with open admission (no written exams required). Students living in remote areas can also attend because there is no obligatory face-to-face participation apart from some practical laboratory courses with short duration in the field of science that take place in summer vacation. The curriculum includes group sessions for advisory and support. The sessions are optional and students can choose to participate either in a face-to-face group or in an online group. Students have to hand on up to six written assignments (WA). A minimum average grade of five out of ten is a prerequisite in WA to establish the right to participate in the final exams where a passing grade of five out of ten leads to the successful completion of the course.

Two datasets were used to test the proposed methodology and conduct a comparative analysis between learning communities with different characteristics. In that sense, two different courses were selected. The courses are offered in a

Table 4 Basic information about the courses

	Course A	Course B
Number of Students	175	126
Numbers Tutors	7	6
Number of Posts	1310	451
Active Forum Participants[1]	56	71
Participation Rate[2]	31,4%	56,3%
Average Number of Posts per Participant	23,4	6,4

[1] At least one post

[2] Based on active forum participants

postgraduate program of the School of Science and Technology at the Hellenic Open University. The Learning Management System that supports the courses is based in a Moodle platform. A main feature of the program is that it is targeting at a wide range of attendants leading to a significant heterogeneity in the community of learners. The first course (hereafter Course A) is a compulsory first-year course that signifies students' first contact with the program, and for many of them, their first experience in Distance Learning. The second course (hereafter Course B) is a second-year optional course. This choice is driven by *RQ3* that refers to the development of the students within the learning process in a Distance Learning course. Basic information about Courses A and B is presented below in Table 4.

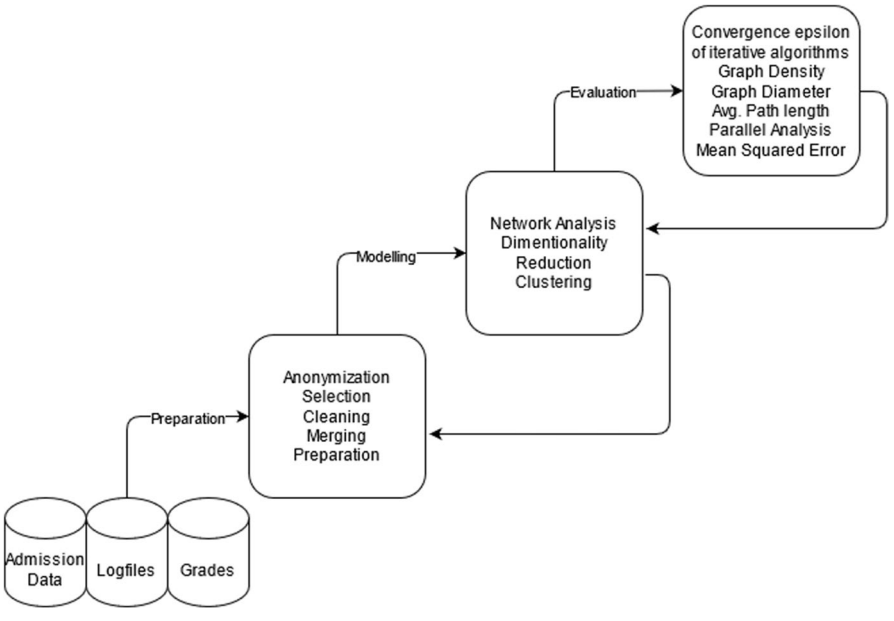


Fig. 3 The Knowledge discovery process

The proposed methodology was tested in four main steps. The first step was the data retrieval from several sources. The second step included all the pre-processing actions (anonymization, data selection and exclusion, data cleaning and merging, and preparation for the modeling tools). The next step was the three-layered modeling process described in Sect. 5. Results emerging from a layer of modeling were used as derived variables for the next modeling action. Finally, evaluation metrics were used to ensure the validity of the results (Fig. 3).

7 Experimental results and comparison

The results are presented in the following three sub-sections. The first sub-section is about Course A, the second concerns Course B, and the third sub-section covers its comparison. Simple results about explicit features are initially presented, followed by latent features derived from PCA and Clustering.

7.1 A data-based description of the first-year course's community

Course A, as is abovementioned, signifies the first chance for the students to experience the program's actual workload and its demands. For many of them is also their first experience in an online learning setting. Consequently, the learning outcome that is imprinted in their final grade is rather low with an average final grade 4,26 (std = 3,29). The high standard deviation reflects the heterogeneity of the participants' performance that is also consistent with the admission data that show a wide range of different backgrounds of students (Kagklis et al., 2017). Their online participation styles also vary. The average number of forum views is 429,38 (std = 545,62) and their degree in the forum network range from 0 to 23 (average degree = 1,93, std = 4,61).

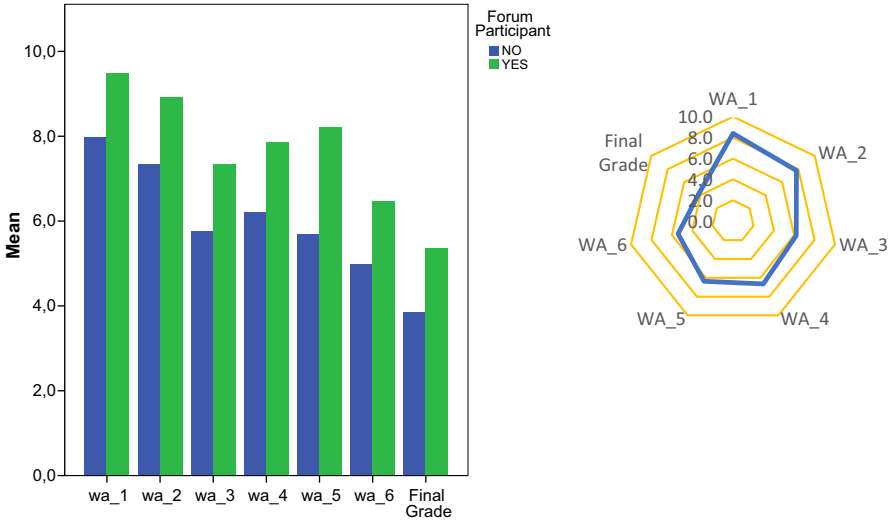


Fig. 4 Average grades in Course A

Table 5 Network metrics for the ten more active participants in Course A based on their Pagerank score

Id	Grade	Views	Degree	WD[1]	HCC[2]	BC[3]	Authority	PR[4]	EC[5]
tutor1A	N/A	1652	39	6337	0,86	608,68	0,33	0,07	1,00
std145A	0,0	1006	22	694	0,70	106,33	0,25	0,03	0,75
std51A	8,0	1771	23	957	0,71	79,98	0,29	0,03	0,86
std86A	9,7	1083	22	711	0,70	103,91	0,27	0,03	0,80
tutor2A	N/A	1510	22	325	0,70	113,24	0,26	0,03	0,78
std18A	8,0	5139	21	9141	0,69	68,76	0,26	0,03	0,78
std46A	7,7	791	20	742	0,68	51,42	0,25	0,03	0,76
std96A	5,6	751	17	130	0,65	75,64	0,21	0,03	0,63
std15A	5,3	918	15	567	0,63	34,78	0,19	0,02	0,57
std66A	5,0	1333	14	155	0,62	17,53	0,20	0,02	0,59

[1] Weighted Degree

[2] Harmonic Closeness Centrality

[3] Betweenness Centrality

[4] PageRank

[5] Eigenvector Centrality

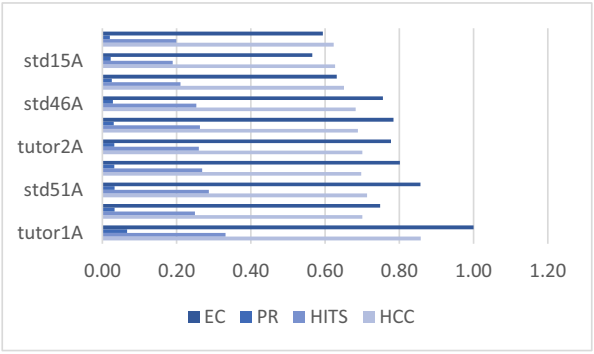


Fig. 5 Network metrics for the ten more active forum participants in Course A

There is a descending pattern in students' grades, that is in accordance with the gradually increasing difficulty of the course (Fig. 4), while there is an almost steady precedence of students who actively participate in the forum.

Different learning styles and behavioral patterns are revealed from the study of the ten most active participants of the forum community (Table 5). The most active participant is tutor1A. Tutor2A is also very active, confirming tutors' mediative role in the collaboration network. All of the highly active students completed successfully the course except std145A. Std145A is an interesting case due to his/her high participation in the forum community and his/her high grades in the written assignments (all grades above 9/10) but he/she didn't show up in the final exams leading to the course failure.

Additionally, the comparison of the metrics concerning std51A and std18A shows a very different learning attitude. Std51A has far fewer views and weighted degree than std18A. However, they have a similar degree and std51A features higher values of betweenness centrality, authority, and eigenvector centrality, indicating a more focused behavior. Their final grade is exactly the same showing that the differences in their actions have not any implication in their final performance, however, std18A seems to have followed a more time-consuming learning path (Fig. 5).

The visualization of the community network reveals two discrete areas (Fig. 6). Area 1 includes forum viewers that do not actively participate in the discussions and Area 2 includes forum active participants. Green nodes represent tutors and orange nodes represent students. There is a very active and strongly connected center and several weak linked, peripheral nodes. The majority of the tutors have a central strategic position in the network confirming their leading role. Moreover, the structure of the network indicates a novice community with low levels of autonomy and relatively weak ties between fellow students.

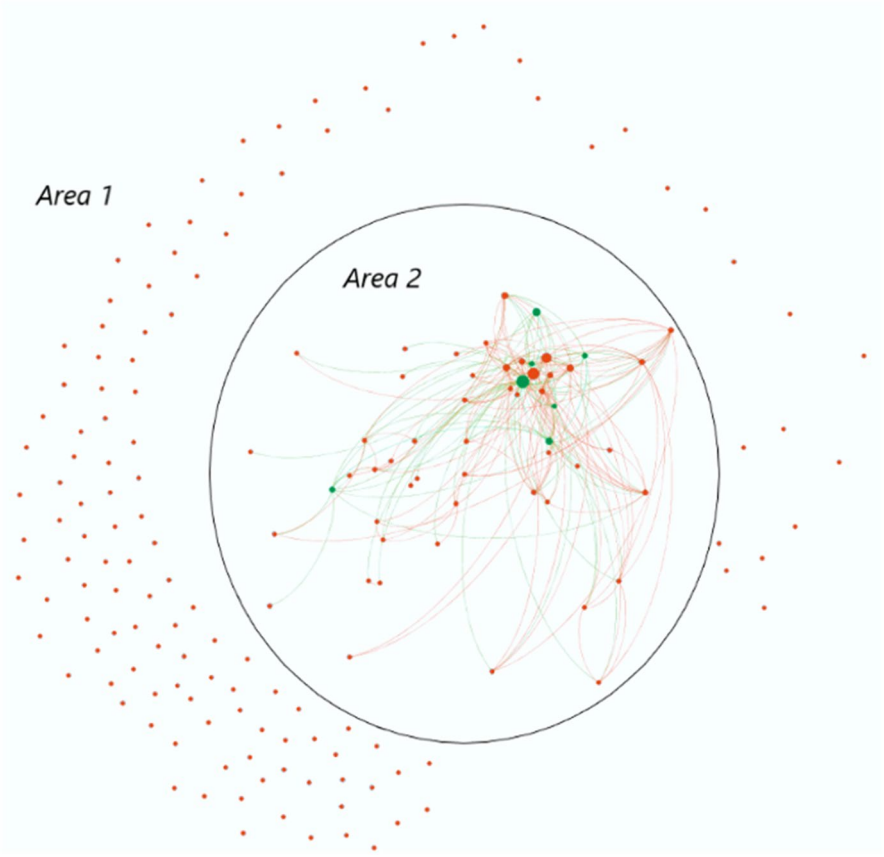
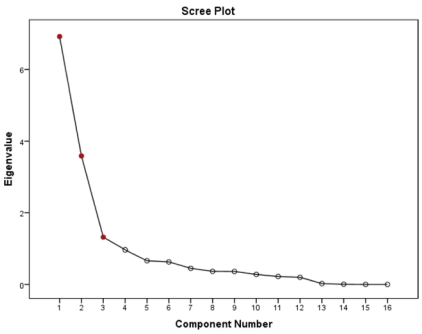


Fig. 6 The network of students' forum interaction in Course A

Table 6 Suitability testing for PCA (Course A dataset)

KMO and Bartlett's Test		
Kaiser–Meyer–Olkin Measure of Sampling Adequacy		,798
Bartlett's Test of Sphericity	Approx. Chi-Square	5658,903
	Df	120
	Sig	,000



Initial Eigenvalues			
Component	Total	% of Variance	Cumulative %
1	6,920	43,252	43,252
2	3,588	22,426	65,678
3	1,315	8,220	73,898

Fig. 7 Principal components for Course A dataset (left: Scree Plot, right: Initial eigenvalues and variance explained)

Previous to the factor analysis, suitability testing is needed to ensure that our data are correlated. The Kaiser–Meyer–Olkin Measure of Sampling Adequacy indicated a sufficient value KMO (Table 6). Additionally, Bartlett's Test of Sphericity confirms the suitability of our data.

Three principal components emerged (eigenvalues > 1) that explain 73,9% of the variance (Fig. 7). Additionally, parallel analysis was conducted resulting in random eigenvectors with values that were all below one. Thus, we can keep all three components of the PCA analysis.

Varimax rotation technique was used to adjust the coordinates of the main components and provide more explainable results that represent how data correlate with each principal component in a more apparent way (Table 7). The first component compiles the majority of the network metrics, explaining 43,25% of the variance. The second factor that explains 22,42% of the variance, sums up the academic performance since it is strongly correlated exclusively with students' grades. The third factor, which explains 8,22% of the variance, concerns students' online activity due to its strong correlation with the number of views and the degree that shows the number of peers a person has interacted with within the forum community.

The 3D scatterplot of the eigenvectors that emerged in the latent space shows two different groups of students. The distinctive feature is their active participation in the forum community (Fig. 8).

Table 7 Rotated component matrix

	Component		
	1	2	3
WA_1	,105	**,722**	-,063
WA_2	,075	**,787**	-,004
WA_3	,085	**,796**	,082
WA_4	,074	**,878**	,089
WA_5	,157	**,834**	,074
WA_6	,081	**,723**	,207
Final Grade	,050	**,825**	,100
Views	,288	,385	**,767**
Degree	**,908**	,142	,319
Weighted Degree	,266	,008	**,854**
Eccentricity	**,807**	,173	-,239
Harmonic Closeness Centrality	**,942**	,178	-,063
Betweenness Centrality	,398	-,056	,105
Eigenvector Centrality	**,915**	,140	,298
Authority	**,913**	,139	,300
PageRanks	**,913**	,142	,310

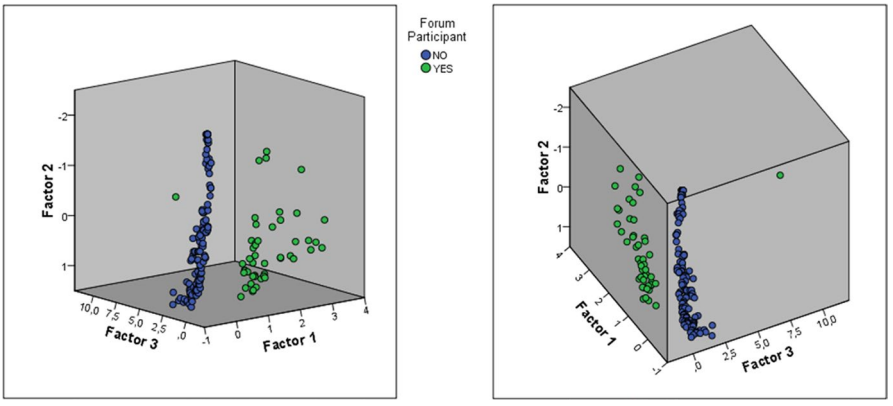


Fig. 8 The 3d Scatterplot for the three main components in the latent space partitioned by forum participation for Course A

However, forum participants (green dots) appear a wider diffusion. Cluster analysis will eventually reveal groups of students with common latent characteristics. Initially, Hierarchical Cluster Analysis provided us the optimal number of clusters. Subsequently, k-Means clustering for 4 clusters was conducted and produced the following results (Table 8).

Table 8 k-Means clustering for Course A

Cluster ID	Number of students
1	44
2	89
3	1
4	41
Mean squared error = 0,49	

The first cluster contains 44 students. They're all active forum participants, with a high number of views and high grades. The students in the second cluster are not active forum participants with an average number of views and good grades. Cluster 3 has captured an outlier. The outlier is std18A that his/her profile differs a lot from his/her peers with a number of views over ten times above the average and a value of weighted degree almost 150% above the weighted degree of the second most active participant of the forum (Table 5). The fourth cluster gathers all the students with low academic achievement regardless of their social behavior. Therefore, it includes both active and non-active forum participants (Fig. 9).

7.2 A data-based description of second-year course's community

Course B is offered in the second academic year. It is an optional course with high demand. The average final grade is 7,1 (std = 1,7) and students appear to be active with an average number of views equal to 346,63 (std = 230,67) and an average degree equal to 11,12 (std = 15,34). Similar to Course A, students who actively participate in the forum community tend to have higher grades in the written assignments and the final examination (Fig. 10).

Two out of the ten most active participants are tutors (Table 9). Tutor2B is the most active participant who facilitates interaction (he/she has the highest eigenvector

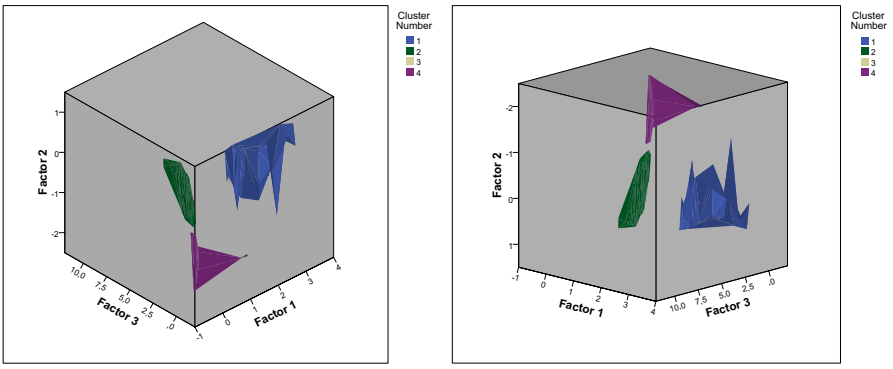


Fig. 9 The students' groups in the latent space (Course A)

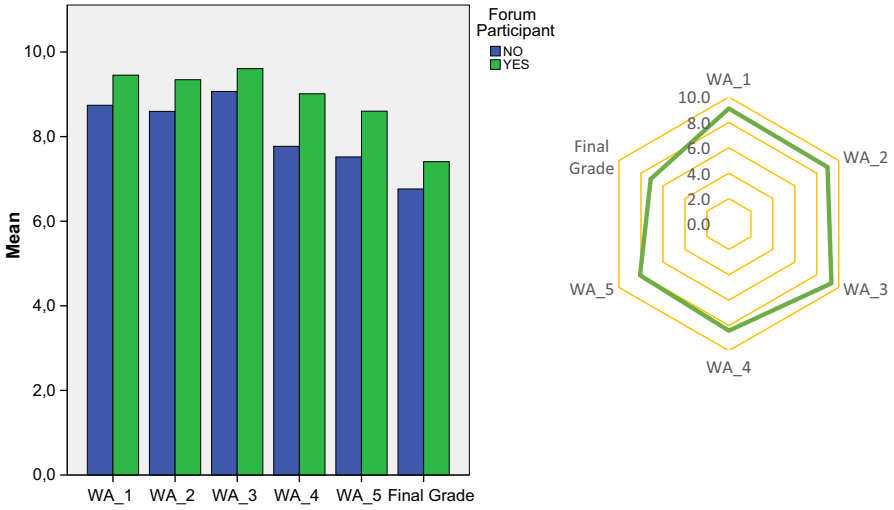


Fig. 10 Average Grades in Course B

Table 9 Network metrics for the ten more active forum participants in Course B based on their Pagerank score

Id	Grade	Views	Degree	WD	HCC	BC	HITS	PR	EC
tutor2B	N/A	407	56	269	0,90	271,36	0,20	0,03	1,00
std13B	9,0	888	51	173	0,86	229,21	0,19	0,03	0,92
std14B	9,5	747	53	232	0,88	169,45	0,20	0,03	0,99
std43B	7,5	157	45	94	0,82	112,43	0,19	0,02	0,94
std31B	7,5	509	44	80	0,81	89,36	0,19	0,02	0,93
std59B	8,0	340	43	99	0,80	46,01	0,19	0,02	0,93
tutor4B	N/A	2587	29	126	0,71	228,16	0,08	0,02	0,43
std28B	8,0	567	43	105	0,80	47,20	0,19	0,02	0,92
std77B	6,5	782	42	138	0,79	39,47	0,19	0,02	0,92
st112	6,5	595	40	96	0,78	20,74	0,18	0,02	0,90

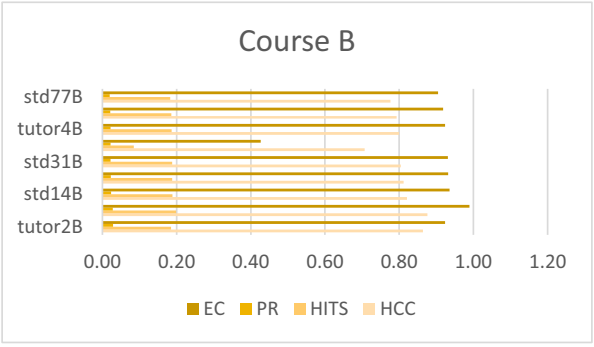


Fig. 11 Network metrics for the ten more active forum participants in Course B

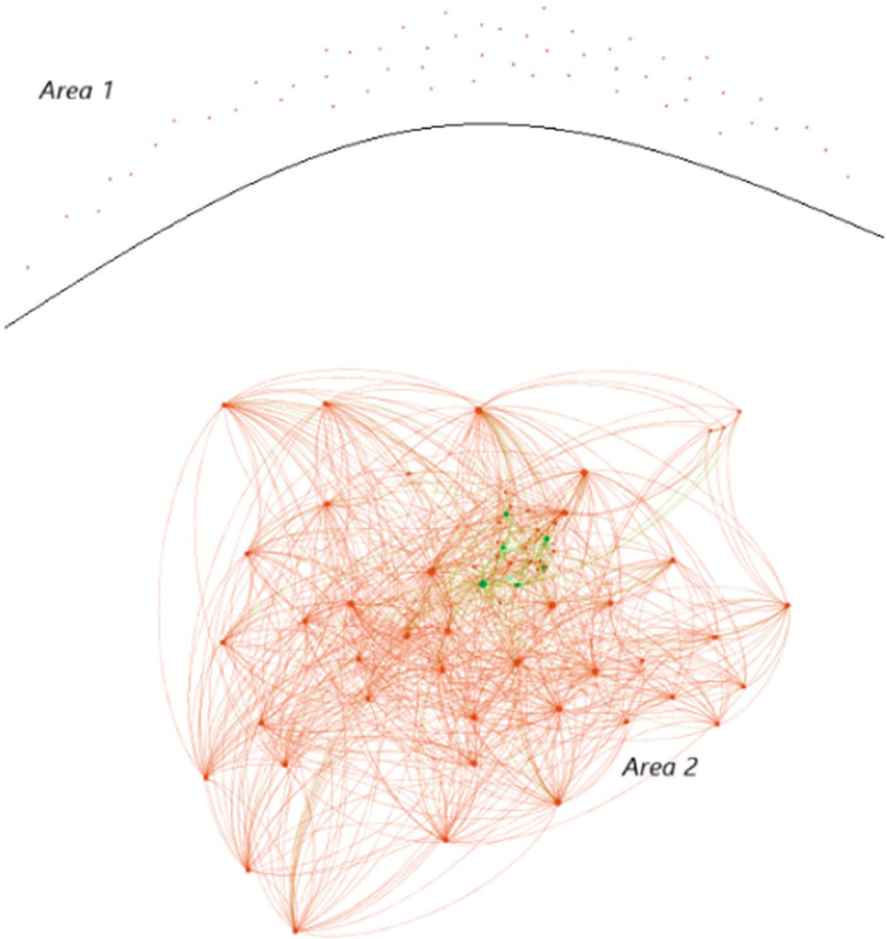
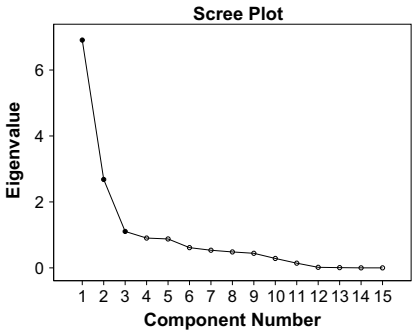


Fig. 12 The network of students' forum interaction in Course B

centrality communication) and mediates communication (he/she has the highest betweenness centrality). On the other hand, tutor4B adopts a less intervening type of assisting learning with much more views, relatively high betweenness centrality but far less posting (Table 9).

Table 10 Suitability testing for PCA (Course B dataset)

KMO and Bartlett's Test		
Kaiser–Meyer–Olkin Measure of Sampling Adequacy		,756
Bartlett's Test of Sphericity	Approx. Chi-Square	4047,525
	Df	105
	Sig	,000



	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	6,908	46,055	46,055
2	2,681	17,875	63,930
3	1,105	7,364	71,294

Fig. 13 Principal Components for Course B dataset (left: Scree Plot, right: Initial eigenvalues and variance explained)

Table 11 Rotated component matrix

	Component		
	1	2	3
WA_1	,119	**,731**	,000
WA_2	,095	**,798**	-,026
WA_3	,066	**,722**	,060
WA_4	,169	**,633**	-,371
WA_5	,112	**,628**	,189
Final Grade	,082	**,712**	-,025
Views	,022	,068	**,940**
Degree	**,981**	,113	-,026
Weighted Degree	**,895**	,107	,050
Eccentricity	**,710**	,243	-,094
Harmonic Closeness Centrality	**,896**	,213	-,062
Betweenness Centrality	**,667**	-,006	,171
Eigenvector Centrality	**,966**	,113	-,035
Authority	**,965**	,113	-,035
PageRanks	**,983**	,119	-,022

All active students develop a common pattern of behavior where a more central position in the network comes with a higher grade (Fig. 11).

The forum network in Course B is also parted into two areas (Fig. 12). Area 1 contains the non-active participants who only visit the forum just to read the posts and area 2 that shows the interaction of active students. Green dots represent tutors and orange dots represent students. A large number of edges indicate that there is a lot of interconnection in the network with a small number of visible peripheral members.

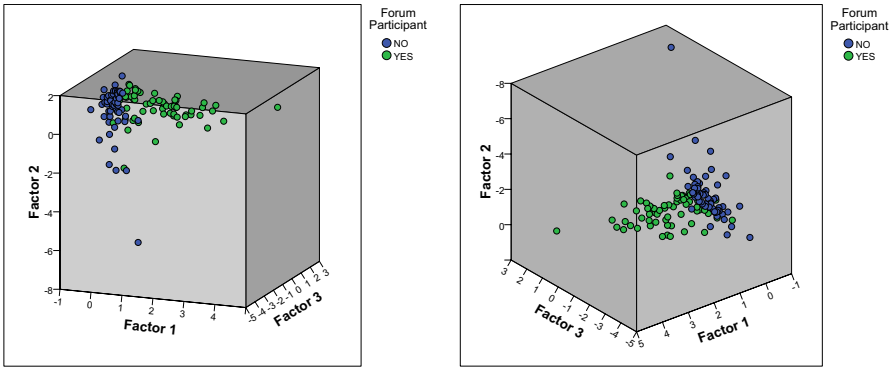


Fig. 14 The 3d Scatterplot for the three main components in the latent space partitioned by forum participation for Course B

Table 12 k-Means clustering for Course B

Cluster ID	Number of students
1	8
2	36
3	49
4	33
Mean squared error = 1,14	

Dimensionality reduction was also conducted in the dataset of Course B since the suitably testing confirmed that our data are fitting for Principal Component Analysis (Table 10).

Three main components met the criterion of having an eigenvalue above 1 (Fig. 13). Additionally, the parallel analysis produced all random eigenvalues below one so three factors were kept.

The rotated component matrix (Table 11) presents some highly explainable results. The first factor, which explains 46% of the variance (Fig. 13), gathers all network metrics. The second factor concerns academic performance since it is strongly correlated with students' grades and the third factor is about online attendance, mainly related to the number of views.

In the case of Course B, there is no large distinction between active forum participants and non-active forum participants. The two groups are separate but close together, indicating that active forum participation might be more important distinctive factor in Course A (Fig. 14).

Clustering was conducted following the same process as in the Course A dataset. Hierarchical clustering also indicated an optimal number of 4 clusters and k-Means clustering produced the results shown in Table 12.

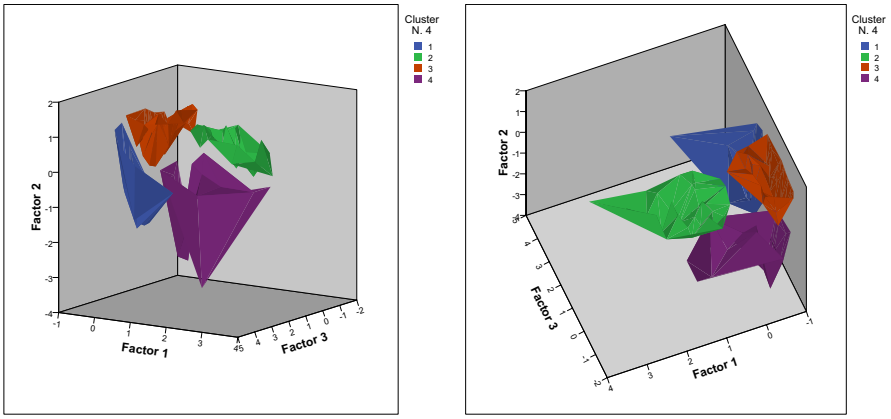


Fig. 15 The students' groups in the latent space (Course B)

The first cluster represents a group of students who participated in the course's forum and had a relatively high number of views yet, they had poor grades indicating that they were facing severe difficulties. The second cluster, on the other hand, contains highly active students with very good grades and important social interaction. The students in the third cluster had good academic performance but mediocre social interaction. The fourth group captures low-activity students with low grades (Fig. 15).

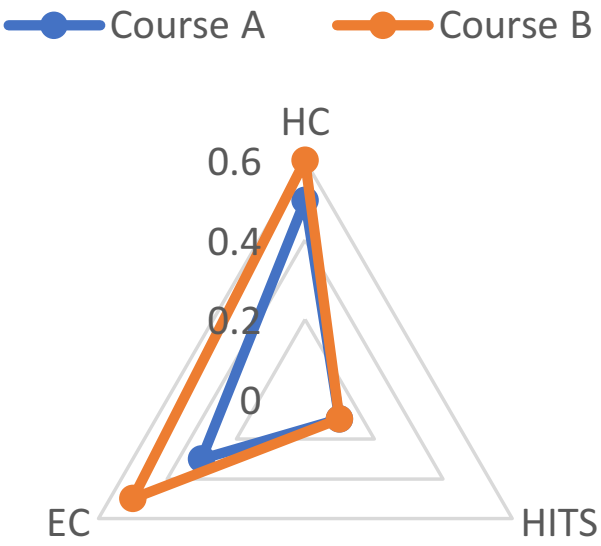


Fig. 16 Comparison of three network metrics (HC: Harmonic Closeness Centrality, EC: Eigenvector Centrality, HITS: Hub and Authority)

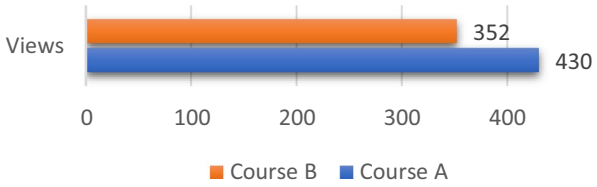


Fig. 17 Average number of participants' views in Courses A and B

7.3 A comparison between students' communities

Students in the two courses differ in their social behavior and their academic performance. In both courses, written assignments start with high grades in a descending pattern. However, Course A has lower completion rate. Participation and social behavior are reflected in views and forum network metrics. Students from course A have a higher average number of views per participant while network's metrics that represent the mediative role and their position in the network (i.e., eigenvector centrality and harmonic closeness centrality) present lower values (Figs. 16, 17). This is an indication of a society seeking to be connected where participants are interested in interacting however there is not a solid structure of actively connected participants.

The evaluation of the structure of the network in each course is imprinted in metrics that concern the total network (Table 13). Both networks (Fig. 18) have the same diameter. The four degrees of separation indicate that the most distant

Table 13 Evaluation of Graphs' structure

	Course A	Course B
Graph Diameter	4	4
Graph Density	0,014	0,091
Avg, Path Length	2,02	1,8
Modularity	0,12	0,22

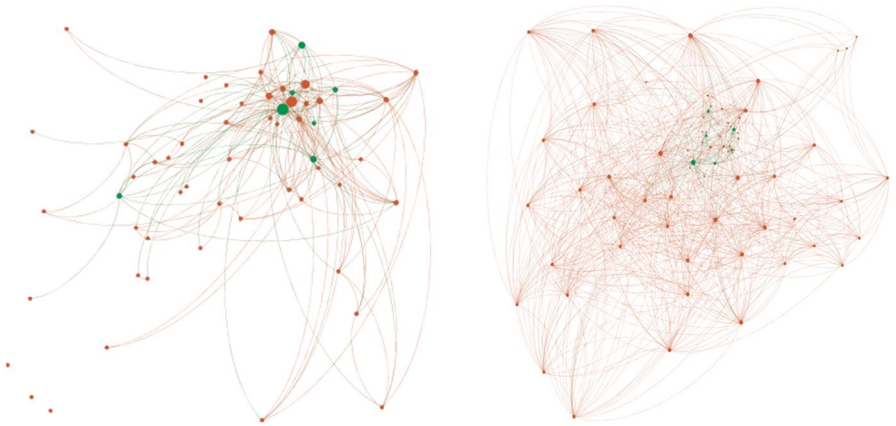


Fig. 18 The diversity in forum participation networks' structure

participants are four links away. The average path length is smaller in Course B indicating a closer relationship between participants. There is a significant difference between the density of the networks. The small size of both networks prevents us from securely assuming a small world effect. Participants in Course B are connected with stronger ties in a more condensed structure but with a higher tendency to form inner communities demonstrated by its higher modularity.

The comparison of the visualized networks is clearly showing the difference in the interaction level. Green nodes that denote tutors are in the center of both networks. Yet, in Course A this center is the core of interaction that connects several peripheral nodes but in Course B even distant nodes are well connected with a larger number of peers as well.

Another type of visual representation that provides a valid comparison of the social interaction is the distribution of some typical network metrics. As it was mentioned above, weighted degree in the one-mode network of forum participants represents the number of participants a person communicated with through a post in a common thread, weighted by the number of messages he/she has posted. Thus, for an active participant weighted degree rises in a non-proportional way. The distribution of weighted degree in two courses (Fig. 19) shows that in Course A there are some very highly active participants while the rest of them have significantly lower participation. On the other hand, in Course B the participation of active students does not differ so much from a regular participant. It is a community where everyone communicates in a balanced manner rather than few persons monopolize the discussion that seems to happen in Course A.

A participant with high betweenness centrality joins parts of the network that would otherwise have been apart. This mediative role usually is assigned to tutors, especially at the beginning of the semester. In Course A all tutors embrace that task, but also there is a student with very high betweenness centrality. He/she is the same student who also stands out in Fig. 19. He/she participates in almost any thread and often dominates the discussion with persistence. Apart from this finding, the main

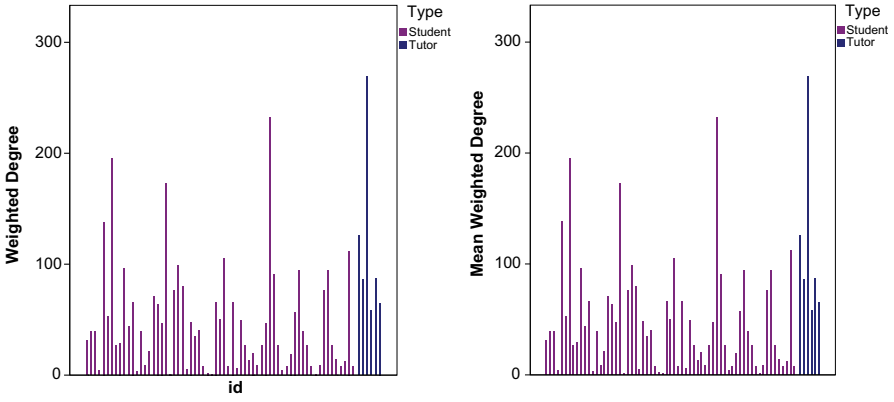


Fig. 19 The distribution of weighted degree. The left graph refers to course A and the right graph to Course B. Purple bars denote students and blue bares denote tutors (note that the grand difference in y-axis scale between two courses)

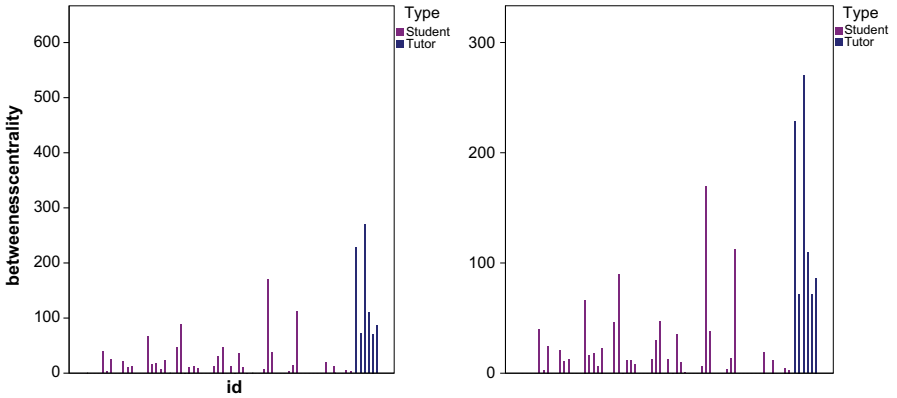


Fig. 20 Betweenness Centrality distribution for courses A and B. The left graph refers to Course A and the right graph to Course B. Purple bars denote students and blue bares denote tutors

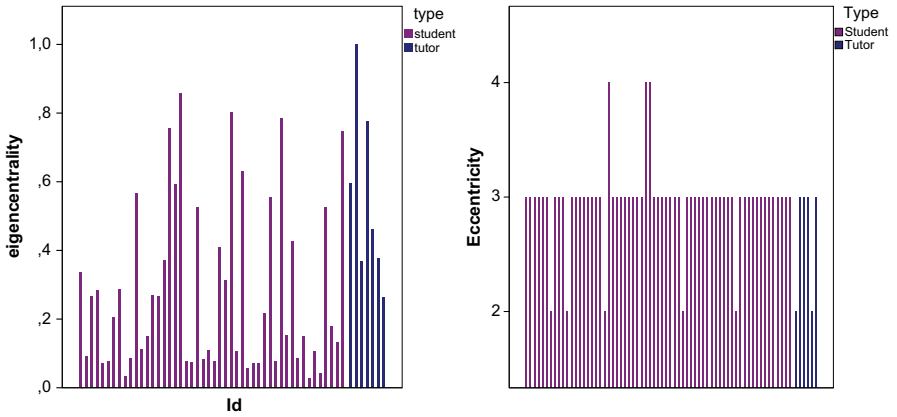


Fig. 21 The distribution of Eccentricity. The left graph refers to Course A and the right graph to Course B. Purple bars denote students and blue bares denote tutors

difference between the communities of the two courses is that in Course B more participants act meditatively and facilitate communication and collaboration, while in Course A there are fewer students in this role.

Distant nodes in the collaboration network with high eccentricity are indicating that these students did not establish strong ties with their peers and stayed in the periphery of the community. In the second-year course, there are only three remote students, far fewer than the first-year's remote students.

The authority distribution tells a similar story. In Course A there are a lot of authoritative participants having a leading role in the community. At the same time, there is a large number of low authority participants.

All the distributions presented in this section (Figs. 19, 20, 21 and 22) designate that there are two types of communities: a "leaders-followers" community in Course

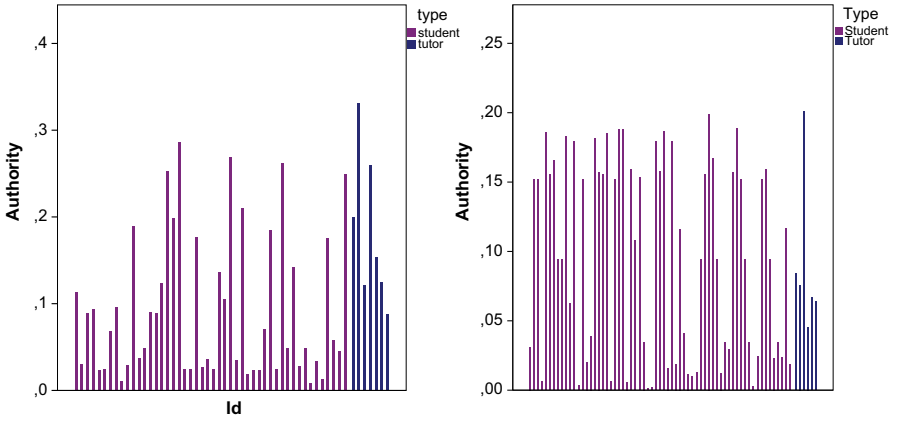


Fig. 22 Authority distribution for both courses. The left graph refers to Course A and the right graph to course B. Purple bars denote students and blue bares denote tutors

A, and a collaborative/democratized community in Course B. This is a strong indication of a maturation process that concerns the online students who gradually adapt to the distance learning environment, create bonds with their tutors and their peers that help them to cross the barriers of physical distance and communicate effectively and enhance their learning autonomy and, at the same time, becoming members of a collaborative community.

8 Discussion and pedagogical reflections

Borje Holmberg who strongly supported empathy and individualization in Distance Learning, even before technological tools revolutionize communication, in his book "Theory and Practice of Distance education" (2005) indicated three types of guided didactic dialogue: a) commentary on the learning process b) personal support of the learner's reflection and c) referents for evaluating learning competence. All three types of conversation can be supported by the discussion forum where, additional, those who do not actively participate can be benefited too. The first research question (*RQ1*) was aiming to propose a process that takes into account both academic and social aspects of learning. The proposed methodology was thoroughly described in Sect. 6. The methodology generally follows the LA cycle that has its roots in grounded learning theories and has several specific steps that serve the purpose of this research. The results provide a rich description of the learning and the social behavior of students using a combination of relevant indices.

Regarding the identification of common features between the communities of the courses, it was found that in the heart of the interaction tutors and highly graded students were placed (Tables 5 and 9). Active participants with a high academic profile can act as the *More Knowledgeable Other* (Vygotsky, 1978) enhancing the learning process. Additionally, the combination of grades and network metrics revealed outliers in both courses. The first extreme case is a student in Course A, a "super-poster"

who heavily contributed to the forum community. However, a mass quantitative contribution does not necessarily imply a qualitative contribution (Huang et al., 2014). Network metrics indicate that this is the case with the certain student as he/she is engaged in a large number of threads with a reply rate far greater than the average, insisting on long dialogues often only with his/her tutors. The second outlier is a student in Course B whose academic and social performance was steadily high. However, he/she didn't sit the final exam. It is obviously a case where personal barriers led the student to fail the course. Similar cases stress the importance of the communication and phycological support that goes beyond tutoring and provide second chances for students in need. The combination of metrics also reveals small differences between students that denote different learning styles and behaviors like the example presented in Sect. 7.1. One of the main findings of this study regarding the contribution of certain variables to the explanation of students' learning behavior comes from the dimensionality reduction process. The PCA pointed three unrelated principal components that explain over 70% of the variance in both courses. The first factor is the *network position factor* that gathers all network metrics, proving the importance of the social aspect of learning as it is express in the discussion forum community. The second factor is the *academic performance factor* relying on students' grades, and the *online presence factor* is mainly related to the number of views. This result describes all students, regardless of the course they participated in even though the communities that were analyzed had significant differences in their structure and their participants' behavior.

Several findings indicated the maturation of the learners' community as they pass from a first-year compulsory course to a second-year optional course (RQ3). Already by comparing some of the primitive variables some simple, yet important differences occur. Students performed better in Course B. Moreover, in Course B a greater percentage of students participated actively in the forum community. On the other hand, students in Course A were visiting more often the forum just to read other participants' posts, indicating passive participation. The online participation becomes more substantial in Course B, where derived network metrics indicate that more students undertook central roles in the interaction community. The higher quantity and quality of second-year students' participation in the forum implies an increasing acceptance of the facilitating function of the forum community in the learning process. The network visualization illustrates complex relations that reveal the structure of the community as a whole, as well as individual features derived from each node's position. The differences in the networks' structure (Fig. 18) reveals that Course A was a novice, tutor-centered community that contained many peripheral participants. On the contrary, the graph created by Course's B participants interaction was a more "small-word"-like network, where students tend to act more autonomously, building relationships with peers in a more coherent community. These results are also confirmed by the distribution of the network metrics presented in Sect. 7.3. The final step of data modeling included a cluster analysis. Two out of the four clusters that emerged in each course had common features: the first one contained high-graded students who actively participated in the forum community, and the second contained less sociable students with good academic performance. However, in Course A the same cluster contains all poorly

performed students, regardless of their social behavior, whereas, in Course B there were two separate clusters of low-graded students: one with those who participated actively in the forum community and one with the non-participants (who also had slightly worse grades). The cluster of the low-graded and socially active students is a typical group of students where intervention has to be made. They were probably interacting with the learning community seeking support, but unsurpassed barriers prevented them from succeeding. Finally, it is important to note that the clusters in Course A are much more coherent (Mean Squared Error = 0,49) than clusters in Course B (Mean Squared Error = 1,14) indicating more homogeneous groups of participants.

Students in the learning community were classified by Wegner (1999) into four groups depending on the level of their contribution:

i. Full participation
ii. Legitimate peripherality
iii. Marginality
iv. Full non-participation

Even though this classification was rather a theoretical approach in the Community of Practice educational theory, LA confirms the existence of these groups. The experimental testing of the proposed methodology produced clusters that, in some cases, match the abovementioned classification. Cluster 1 of the Course A and cluster 2 of the Course B could match the *full participation group (i)*. Clusters of *Full non-participation (iv)* were identified in both courses (cluster 4 in both courses) while clusters 3 and 2 of the Course A and Course B respectively, could be included in the *Legitimate peripherality (ii)* group. However, cluster 1 of the Course B that contains active participants with low academic performance cannot be assigned to any of the four groups of Wegner's classification. This fact provides a strong indication that other parameters apart from forum activity should be intergraded when studying learning communities.

As it concerns the classification of learners, in a relevant study about the use of learning tools in an LMS (including forum use) Lust et al. (2013) identified four disparate groups that match Wegner's pre-LMS classification:

i. no-users,
ii. intensive active learners
iii. selective users
iv. intensive superficial

Although there are numerous studies in the field of learning analytics incorporating discretely SNA, PCA and clustering methods, authors were not able to find any relevant studies where clustering is based on the combination of the level of social interaction with actual learning outcome metrics in a latent space created by principal components. Due the multidimensionality of the research problem it has to be

stressed that the groups found in this study are probably subject-related sensitive. Moreover, the level of education might also differentiate the results.

9 Conclusion

This paper proposes an exploratory, multilayered methodology aiming to investigate students' learning behavior in an online environment. The experimental design and evaluation brought into light some interesting results for students' profiling as well as for their communities' structure. Dawson et al. (2019) highlighted the need for an integrative, complex, and holistic view to understand the dynamics outside of a specific course that influence learning performance as learning experience does not limit to a single course. The selection of two successive courses allowed us to compare the learning habits of the participants and get a sense of their evolvement in the learning community. There are some common features between courses but also some important distinctions signifying the change that comes along in students learning path. However, the three principal components that similarly emerged for the two student communities, highlight the important factors that can describe learning behavior based on observable attitudes of their digital traces. The repeatability of the results is an issue of many related aspects. In our future work, we plan to automate the proposed LA process, flexibly, so that the experiments could be repeated in different settings, providing results with higher generalizability.

References

Abbassi, Z., & Mirrokni, V. S. (2007). A recommender system based on local random walks and spectral methods. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (pp. 102–108).

Amano, K., Tsuzuku, S., Suzuki, K., & Hiraoka, N. (2019). Learning Together for Mastery by Using a Discussion Forum. *In 2019 International Symposium on Educational Technology (ISET)* (pp. 165–169). IEEE.

Andersen, R., Chung, F., & Lang, K. (2006). Local graph partitioning using pagerank vectors. In 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06) (pp. 475–486). IEEE.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open-source software for exploring and manipulating networks. *Icwsm, 8*(2009), 361–362.

Batagelj, V., & Mrvar, A. (1998). Pajek-program for large network analysis. *Connections, 21*(2), 47–57.

Bates, A. W. (2019). *Teaching in a Digital Age* – (2nd ed.). Tony Bates Associates Ltd.

Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T., & Popelinsky, L. (2012). Predicting Drop-Out from Social Behaviour of Students. *International Educational Data Mining Society.*

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). Ucinet for Windows: Software for social network analysis. *Harvard, MA: analytic technologies*, 6.

Bouhnik, D., & Marcus, T. (2006). Interaction in distance-learning courses. *Journal of the American Society for Information Science and Technology, 57*(3), 299–305.

Bozkurt, A., Jung, I., Xiao, J., Vladimirschi, V., Schuwer, R., Egorov, G., … & Rodes, V. (2020). A global outlook to the interruption of education due to COVID-19 Pandemic: Navigating in a time of uncertainty and crisis. *Asian Journal of Distance Education, 15*(1), 1–126.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Brown, S. (2017). A PageRank model for player performance assessment in basketball, soccer and hockey. arXiv preprint arXiv:1704.00583

Capocci, A., Servedio, V. D., Caldarelli, G., & Colaiori, F. (2005). Detecting communities in large networks. *Physica A: Statistical Mechanics and Its Applications, 352*(2–4), 669–676.

Cela, K. L., Sicilia, M. Á., & Sánchez, S. (2015). Social network analysis in e-learning environments: A preliminary systematic review. *Educational Psychology Review, 27*(1), 219–246.

Chen, J., Fagnan, J., Goebel, R., Rabbany, R., Sangi, F., Takaffoli, M., … & Zaiane, O. (2010). Meerkat: Community mining with dynamic social networks. *In 2010 IEEE International Conference on Data Mining Workshops* (pp. 1377–1380). IEEE.

Chiu, T. K., & Hew, T. K. (2018). Factors influencing peer learning and performance in MOOC asynchronous online discussion forum. *Australasian Journal of Educational Technology,* 34(4).

Clow, D. (2012). The learning analytics cycle: closing the loop effectively. *In Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 134–138).

Crossley, S., Dascalu, M., McNamara, D. S., Baker, R., & Trausan-Matu, S. (2017). Predicting success in massive open online courses (MOOCs) using cohesion network analysis. Philadelphia, PA: International Society of the Learning Sciences.

Csardi, M. G. (2013). Package 'igraph'. Last accessed, 3(09), 2013

Dawson, S., Joksimovic, S., Poquet, O., & Siemens, G. (2019). Increasing the impact of learning analytics. *In Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 446–455).

De-Marcos, L., García-López, E., García-Cabot, A., Medina-Merodio, J. A., Domínguez, A., Martínez-Herráiz, J. J., & Diez-Folledo, T. (2016). Social network analysis of a gamified e-learning course: Small-world phenomenon and network metrics as predictors of academic performance. *Computers in Human Behavior, 60*, 312–321.

Farahat, A., LoFaro, T., Miller, J. C., Rae, G., & Ward, L. A. (2006). Authority rankings from HITS, PageRank, and SALSA: Existence, uniqueness, and effect of initialization. *SIAM Journal on Scientific Computing, 27*(4), 1181–1201.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71.

Giovannella, C., Scaccia, F., & Popescu, E. (2013). A PCA study of student performance indicators in a Web 2.0-based learning environment. *In 2013 IEEE 13th International Conference on Advanced Learning Technologies* (pp. 33–35). IEEE.

Gkontzis, A. F., Kotsiantis, S., Tsoni, R., & Verykios, V. S. (2018). An effective LA approach to predict student achievement. *In Proceedings of the 22nd Pan-Hellenic Conference on Informatics* (pp. 76–81).

Gkontzis, A. F., Kotsiantis, S., Kalles, D., Panagiotakopoulos, C. T., & Verykios, V. S. (2020). Polarity, emotions and online activity of students and tutors as features in predicting grades. *Intelligent Decision Technologies*, 1–28.

Grover, N., & Wason, R. (2012). Comparative analysis of pagerank and hits algorithms. *International Journal of Engineering Research & Technology (IJERT), 1*(8), 1–15.

Hernández-García, Á., & Suárez-Navas, I. (2017). GraphFES: A web service and application for Moodle message board social graph extraction. In *Big data and learning analytics in higher education* (pp. 167–194). Springer, Cham.

Holmberg, B. (2005). *Theory and practice of distance education* (p. 51). Routledge.

Hu, J., Liang, J., & Dong, S. (2017). ibgp: A bipartite graph propagation approach for mobile advertising fraud detection. Mobile Information Systems, 2017.

Huang, J., Dasgupta, A., Ghosh, A., Manning, J., & Sanders, M. (2014). Superposter behavior in MOOC forums. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 117–126)

Iván, G., & Grolmusz, V. (2011). When the Web meets the cell: Using personalized PageRank for analyzing protein interaction networks. *Bioinformatics, 27*(3), 405–407.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS One, 9*(6), e98679.

Jan, S. K., & Vlachopoulos, P. (2019). Social network analysis: A framework for identifying communities in higher education online learning. *Technology, Knowledge and Learning, 24*(4), 621–639.

Jiang, B., Kloster, K., Gleich, D. F., & Gribskov, M. (2017). AptRank: An adaptive PageRank model for protein function prediction on bi-relational graphs. *Bioinformatics, 33*(12), 1829–1836.

Kagklis, V., Karatrantou, A., Tantoula, M., Panagiotakopoulos, C. T., & Verykios, V. S. (2015). A learning analytics methodology for detecting sentiment in student fora: A case study in Distance Education. *European Journal of Open, Distance and E-Learning, 18*(2), 74–94.

Kagklis, V., Lionarakis, A., Marketos, G., Panagiotakopoulos, G. T., Stavropoulos, E. C., & Verykios, V. S. (2017). Student admission data analytics for open and distance education in Greece. *Open Education: The Journal for Open and Distance Education and Educational Technology., 13*(2), 6–16.

Kandiah, V., & Shepelyansky, D. L. (2012). PageRank model of opinion formation on social networks. *Physica a: Statistical Mechanics and Its Applications, 391*(22), 5779–5793.

Klašnja-Milicevic, A., & Ivanovic, M. (2018). Learning Analytics-New Flavor and Benefits for Educational Environments. *Informatics in Education, 17*(2), 285–300.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM), 46*(5), 604–632.

Laurillard, D. (2013). *Rethinking university teaching: A conversational framework for the effective use of learning technologies*. Routledge.

Laurillard, D., Kennedy, E., & Wang, T. (2018). *How could digital learning at scale address the issue of equity in education? Learning at scale for the global south*. Foundation for Information Technology Education and Development.

Latora, V., & Marchiori, M. (2001). Efficient behavior of small-world networks. Physical review letters, 87(19), 198701.

Lazova, V., & Basnarkov, L. (2015). PageRank approach to ranking national football teams. arXiv preprint arXiv:1503.01331

Lei, X., Wang, S., & Wu, F. (2019). Identification of essential proteins based on improved HITS algorithm. *Genes, 10*(2), 177.

Lotfi, A., Ghorbani, M., & Mesgarani, H. (2019). A Study of PageRank in Undirected Graphs. *Mathematics Interdisciplinary Research, 4*(2), 157–169.

Lotsari, E., Verykios, V. S., Panagiotakopoulos, C., & Kalles, D. (2014). A learning analytics methodology for student profiling. In *Hellenic Conference on Artificial Intelligence* (pp. 300–312). Springer, Cham.

Lust, G., Elen, J., & Clarebout, G. (2013). Students' tool-use within a web enhanced course: Explanatory mechanisms of students' tool-use pattern. *Computers in Human Behavior*, 29(5).

Metcalf, L., & Casey, W. (2016). Cybersecurity and applied mathematics. Syngress.

Mooney, B. L., Corrales, L. R., & Clark, A. E. (2012). MoleculaRnetworks: An integrated graph theoretic and data mining tool to explore solvent organization in molecular simulation. *Journal of Computational Chemistry, 33*(8), 853–860.

Moore, M. G. (2007). The Theory of Transactional Distance. In M. G. Moore (Ed.), *(2007) The Handbook of Distance Education* (2nd ed., pp. 89–108). Lawrence Erlbaum Associates.

Mukai, N. (2013). PageRank-based traffic simulation using taxi probe data. *Procedia Computer Science, 22*, 1156–1163.

Pask, G. (1976). Styles and strategies of learning. *British Journal of Educational Psychology, 46*(2), 128–148.

Perra, N., & Fortunato, S. (2008). Spectral centrality measures in complex networks. *Physical Review E, 78*(3), 036107.

Romero, C., López, M., Luna, J., & Ventura, S. (2013b). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education, 68*, 458–472.

Romero, C., Espejo, G. Zafra, A., Romero, J., & Ventura, S., (2013). Web usage mining for predicting marks of students that use Moodle courses. *Computer Applications in Engineering Education*.

Schön, D. A. (Ed.). (1991). *The Reflective Turn: Case studies in and on educational practice*. Teachers College Press.

Sereni, J. S., Krnc, M., Škrekovski, R., & Yilma, Z. B. (2018). Eccentricity of networks with structural constraints. *Discussiones mathematicae*, 1–22.

Sergis, S., & Sampson, D. G. (2017). Teaching and learning analytics to support teacher inquiry: A systematic literature review. *Learning analytics: Fundaments, applications, and trends* (pp. 25–63). Springer.

Sharma, K., Papamitsiou, Z., & Giannakos, M. (2019). Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach. *British Journal of Educational Technology, 50*(6), 3004–3031.

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist, 57*(10), 1380–1400.

Siemens, G., & Downes, S. (2008). Connectivism & connective knowledge. Universidad de Manitoba.
Siemens, G., & Baker, R. S. D. (2012). Learning analytics and educational data mining: towards communication and collaboration. *In Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252–254).
Sternitzke, C., Bartkowski, A., & Schramm, R. (2008). Visualizing patent statistics by means of social network analysis tools. *World Patent Information, 30*(2), 115–131.
Sun, B., Wang, M., & Guo, W. (2018). The influence of grouping/non-grouping strategies upon student interaction in online forum: A social network analysis. In *2018 International Symposium on Educational Technology (ISET)* (pp. 173–177). IEEE.
Szczurek, P., & Horeni, M. (2018). Using Link Analysis Algorithms to Study the Role of Neurons in the Worm Connectome. In 2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA) (pp. 651–657). IEEE
Tran, M., & Draeger, C. (2021). A data-driven complex network approach for planning sustainable and inclusive urban mobility hubs and services. *Environment and Planning B: Urban Analytics and City Science*, 2399808320987093.
Traxler, A., Gavrin, A., & Lindell, R. (2018). Networks identify productive forum discussions. *Physical Review Physics Education Research, 14*(2), 020107.
Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2020). Fairness-Aware Link Analysis. arXiv preprint arXiv:2005.14431
Tsoni, R., & Verykios, V. S. (2019). Looking for the "More Knowledgeable Other" through Learning Analytics. *In proceeding of 10th International Conference in Open and Distance Learning, 10*(3A), 239–251.
Tsoni, R., Paxinou, E., Stavropoulos, E. C., Panagiotakopoulos C., & Verykios, V. (2019). Looking under the hood of students' collaboration networks in distance learning. *The Envisioning Report for Empowering Universities*, 39–41.
Tsoni, R., Samaras, C., Paxinou, E., Panagiotakopoulos, C., & Verykios, V. S. (2019). From Analytics to Cognition: Expanding the Reach of Data in Learning. In *Proceedings of CSEDU*.
Tsoni, R., Stavropoulos, E. C., & Verykios, V. Leveraging Learning Analytics with the Power of Words. (2019). *The Envisioning Report for Empowering Universities*. pp. 24–27
Valentine, D. (2002). Distance learning: Promises, problems, and possibilities. *Online Journal of Distance Learning Administration*, *5*(3).
Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior, 89*, 98–110.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
Wang, J., Liu, J., & Wang, C. (2007). Keyword extraction based on pagerank. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 857–864). Springer.
Yusof, N., & Rahman, A. A. (2009). Students' interactions in online asynchronous discussion forum: A Social Network Analysis. *In 2009 International Conference on Education Technology and Computer* (pp. 25–29). IEEE.
Zhou, J., Zeng, A., Fan, Y., & Di, Z. (2018). Identifying important scholars via directed scientific collaboration networks. *Scientometrics, 114*(3), 1327–1343.