Visualizations of Online Course Interactions for Social Network Learning Analytics

http://dx.doi.org/10.3991/ijet.v11i07.5889

Ángel Hernández-García[1], Inés González-González[2], Ana I. Jiménez Zarco[3], Julián Chaparro-Peláez[1]
[1] Universidad Politécnica de Madrid, Madrid, Spain
[2] Universidad de Navarra, Pamplona, Spain
[3] Universitat Oberta de Catalunya, Barcelona, Spain

***Abstract*—Social network learning analytics aims to extract useful information to improve the learning process, but the variety of learning management systems makes this task burdensome and difficult to manage. This study shows how Gephi, a general-purpose, open-source social network analysis application, can be used by instructors and institutions to extract and visualize relevant information that is commonly hidden or difficult to observe for course coordinators and teachers. The empirical case study uses data from one cross-curricular course with 656 students at the Open University of Catalonia (UOC) and showcases the use of Gephi as a social network learning analytics tool. The study further discusses the potential of social network learning analytics to improve online instruction by visualization of educational data.**

***Index Terms*—Gephi, learning analytics, learning data visualization, online learning, social network analysis.**

I. INTRODUCTION

Virtual classrooms add a whole new set of challenges for instructors. One of them is the difficulty for teachers to keep track of students' progress and activity in the course due to the impossibility of direct observation of the different interactions occurring in the classroom. This problem has several implications: first, without the existence of physical cues, it is not easy for teachers to determine the level of engagement and understanding of individual students and student groups [1]; second, as learning strategies shift toward self-directed learning, teachers may need to change their role from deliverers of instructional content and knowledge, to facilitators or guides [2]; and third, instructors lack the information to observe the social dynamics of the class which, if available, would allow them to act on the most influencing or the disconnected students, depending on their needs. Furthermore, although current online learning systems store large amounts of interaction-related data in their databases, these data are generally separated from course information and contents, and are therefore generally not available to course instructors due to the burdensome task of filtering and presenting only the relevant and useful information for the learning process to the different agents [3].

Learning analytics emerges as a solution to face this set of challenges. Learning analytics refers to "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" [4, p.34]. Course-level learning analytics may be classified depending on whether they focus on identification and categorization [5], post-mortem analysis to improve course design upon analysis of finished course data, or modeling of alert systems' to detect and correct abnormal learning behaviors, such as at-risk students [6]. Ultimately, learning analytics generally relies in visualization techniques. However, despite some efforts in this field –e.g., the *Social Networks Adapting Pedagogical Practice (SNAPP)* tool for Moodle, Blackboard or Sakai, and *Meerkat-ED* and *Forum Graph* for Moodle, to visualize student interactions– visualization of learning data and student interactions is not standardized and therefore it requires tailoring the output of data to specific requirements. This study shows the potential of Gephi –an open source and general purpose social network analysis application– for the analysis of educational data, including general guides on how to further explore its possibilities from a social network learning analytics approach.

In order to do so, this paper is structured as follows: section II discusses the need for analysis of class interactions in online distance learning and how social learning analytics –and, particularly, visualizations of social network learning analytics– can help interpreting learning data; section III presents the different functionalities of Gephi, as well as some guidelines on how to use it for social network learning analytics purposes; in order to illustrate its use, section IV introduces a case study with data from the Open University of Catalonia (UOC), while sections V and VI exhibit results of the visual and social network analyses, respectively; the final section draws the conclusions from the study and depicts ongoing avenues of research in this field.

II. ONLINE CLASS INTERACTIONS AND SOCIAL NETWORK LEARNING ANALYTICS

When compared to traditional in-class lectures, online distance learning is characterized by the lack of physical contact between instructors and students –and among students– within a class, as well as how instructional content is delivered –with students assuming the responsibility of learning. To overcome this constraint and facilitate communication in the classroom, learning systems have built-in synchronous and asynchronous capabilities; chat rooms are an example of the former, while the latter usually are implemented as message boards and interpersonal messaging systems.

Communication tools allow teachers to observe only one side –the most visible one– of the dynamics of the course: the messages exchanged by active students; from there, teachers can assess whether or not the different concepts and lessons are being understood, as well as the

construction of discourse and knowledge, and then decide on when to intervene to help and support the students in the learning process. However, much of the passive activity which occurs in the system may pass unnoticed to instructors. As a consequence, it is extremely difficult for teachers to detect the involvement of students who are not actively engaging in conversations. Moreover, the students sharing a course may have very different learning styles, and sometimes the lack of active engagement does not mean a lack of involvement or that learning is not happening [7]. For instance, some students may expand their learning by searching for external resources on the topic, and they may decide not to share them with the rest of the class; other students may act like learning witnesses or "invisible students" [8] and rely on content shared by others. Therefore, it is sometimes very hard for instructors, and especially in courses with a large number of students, to determine whether participation and learning are actually happening, be it for individual students or the class as a whole. Furthermore, even when they have that information, teachers may not know if the social dynamics –the study of the relationship between individual interactions and group level behaviors [9]– occurring in the classroom are the most appropriate.

Ref. [10] identifies three main levels of learning analytics: a) identification of suitable indicators for analysis; b) identification, understanding and explanation of learning behaviors; and 3) mechanisms for adaptive learning. While they can be related, understanding and explanation of students' learning behaviors has raised most interest among scholars and practitioners, given its immediacy –interpretation of results is almost straightforward– and its value for theory building. In order to understand and explain online learning processes, researchers analyze students' activity in online learning platforms, assuming that "data speak for themselves" [11]. However, other situational information –e.g., the social nature of the co-construction of knowledge in networks of practice [12]– is most commonly neglected.

Ignoring this situational information may be a bad idea in online learning, especially when the instructional method relies heavily on collaborative and teamwork-based online learning because it makes it difficult to detect dysfunctional groups or lack of students' engagement. Social network learning analytics may provide this supplementary information to make informed decisions to improve the learning process (e.g., [13]).

Social learning analytics refers to a distinctive subset of learning analytics that is socially situated [14]. Ref. [15] defines five levels of social analytics, differentiating between inherently social analytics and socialized analytics. According to Buckingham-Shum and Ferguson, inherently social analytics may be divided into Social Network Analytics –derived from the analysis of interpersonal relationships– and Discourse Analytics –focused on language-based constructed knowledge. This study focuses on the former, and aims to offer ways to visualize and analyze the different interactions between students and teachers, and among students, in an online course. In formal online learning contexts, the interactions, participation, social exchanges and discourse-based knowledge building processes happen essentially in course forums. Therefore, it is only natural that this study uses information from message boards to describe, explain and understand the social dynamics of online courses.

In general, social network learning analytics scholars have centered their attention in identifying relevant actors in the classroom [16], be it influential students, at-risk students or "broker" students –students who connect different groups. An increasing number of scholars –e.g. [2, 17-22]– have already validated the usefulness of social network learning analytics for instructors based on the use of social network metrics.

Conceptually, social network analysis (SNA) considers each actor as a node of the whole network, while the different relationships between them are conceptualized as lines connecting the nodes and known as edges or ties; edges can in turn be undirected –that is, the edges are not oriented and the edge (a, b) is equal to the edge (b, a), with *a* and *b* being network nodes– or directed –when the edges are oriented. Edges can also have different weights depending, for example, on the strength or the number of interactions between two nodes.

SNA offers individual centrality measures which are useful for social network learning analytics, such as degree centrality, betweenness centrality [23], hubs and authorities [20], among others.

Degree centrality represents the number of nodes that a certain node is connected to, and it is generally associated with how influential a node is within the network, with higher degree values corresponding to more influential nodes. Degree centrality has been found to be a predictor of sense of community [2] and to be related to academic performance [20].

Betweenness centrality refers to which nodes are connecting groups of nodes, acting as a "bridge" between them. In social network learning analytics, this is key to identify which actors –students and teachers– may spread more effectively and quickly any information and knowledge across the whole network. Ref. [24] associates betweenness to learning performance, although its relation is weaker than that of degree and closeness centrality, and [17] find that tutors and instructors show higher betweenness values, a result also found in [20].

Apart from individual centrality measures, SNA can also provide global network values, such as the average degree of the network –average number of incoming, outgoing, or global links of a node in the network–, network density –the number of total edges present in the network relative to the number of edges in a full-connected network– or network diameter –the largest number of nodes that must be traversed in order to travel from one node to another.

As stated earlier, the benefits of social network learning analytics do not rely exclusively in the calculation of these network parameters, which may be difficult to understand by instructors, but also in that it allows creating a much more easy-to-understand graphical (visual) representation of the network. The following section explains how Gephi may help calculating SNA parameters values from educational data and showcases Gephi's data visualization capabilities, so that course instructors may extract relevant information about the learning process and take decisions based on this information.

III. GEPHI, AN OPEN SOURCE TOOL FOR SOCIAL NETWORK ANALYSIS

Gephi [25] is an open-source software under the GPL3 (GNU General Public License) for "interactive visualiza-

tion and exploration for all kinds of networks and complex systems, dynamic and hierarchical graphs."[1] Gephi allows visualization of social networks, as well as calculation of SNA parameters, and nodes' and edges' partitioning, ranking and filtering [26].

Gephi has additional features of interest for social network learning analytics purposes: first, it has superior capabilities for data analysis when compared to other tools such as SNAPP, Meerkat-ED or Forum Graph, and offers more functionalities and attribute support than other SNA applications [26-27]; second, it allows data import/export in many different formats; and third, it is easily expandable and customizable by developing NetBeans-based plugins, which makes it a reasonable choice in case additional development is required.

IV. CASE STUDY: A FINANCIAL COURSE AT THE OPEN UNIVERSITY OF CATALUNYA (UOC)

A. Course context and characteristics

In order to illustrate the use of Gephi as a social network learning analytics tool, this study uses data from the semester-long course "Introduction to financial information" from the Business Administration, Marketing, Work Relationships, and Tourism Degrees at the Open University of Catalonia (UOC). The course took place between September 2013 and January 2014, and the main reasons for this choice were the structure of the course and the high heterogeneity of students –the course is a core course in the Business Administration degree, but an elective course in the other three degrees. This is interesting because there is a positive association between student-student interactions in core and elective courses, larger in the case of core courses [28].

The course is structured as follows: regarding the actors involved in the course, there is one coordinating professor –who takes the responsibility for course design and coordination, student tracking, course content updating and management of consultant teachers–, ten consultant teachers –who are assigned one classroom and are responsible for the actual teaching– and students –each student is randomly assigned to a class, with a maximum of 70 students per classroom.

With regards to the functioning of virtual campus, there are three different areas: 1) planning: course plan, schedule, key dates and other relevant information, such as assignments and answers; 2) communication, where interaction among students and teachers take place; and 3) assessment, a mailbox for students to send the assignments for continuous assessment. There are five assessment activities –following a fixed schedule– and students must deliver and pass at least four of them to pass the course.

The second area, communication, is the most relevant for this study and is divided into three sections: announcements' board, where only teachers may post to provide information about course issues, such as answers to questions of general interest, and requirements for assignment submission and assessment, deadlines, etc.; discussion board, where each consultant teacher may propose course-related case studies for the students to discuss; and general forum, an open message board for student questions and insights on course topics –typical posts here include questions, comments, shared links of interest, suggestions for improvement of course materials, etc.

It is worth noting that the UOC is characterized for an atypical student profile, in terms of age and professional experience, with a mean age of 32.4 years old in 2012 and approximately 90 percent of UOC students also actively participating in the labor market.

B. Data collection

Practically all the interaction between students and teachers –and among students– in the "Introduction to financial information" course at UOC occurs in the communication area. Therefore, data for this study was extracted from the learning system's activity log, from September 2013 –starting date– to January 2014 –end date. A total of 114.756 records were retrieved from 656 students –distributed along 10 classrooms–, 10 consultant teachers and 1 coordinating professor. Unfortunately, at the time of data collection, the legacy system log was not designed with a social network learning analytics in mind, and thus some further processing in MS Excel was required before proceeding to data import in Gephi. Additionally, and in order to better explain the results of the SNA, information about each student's final continuous assessment grade was retrieved and incorporated as node attributes.

C. Gephi in action

In order to show the full capabilities of Gephi for social network learning analytics, and following [22], three different datasets were generated from the original data extracted from the system log: 1) in Dataset1, each node represents an actor –student or instructor–, and each directed edge from node (user) *a* to node *b* has a weight that represents the number of times that node *a* replies to a post created by node *b*; 2) Dataset2 is similar to Dataset1, but with edges representing the number of times that node *a* reas a message posted by node *b*; both datasets include node attributes for further filtering capabilities, such as the classroom that each actor –node– was assigned to, as well as his or her final grade and role –student or teacher–, total number of reads, number of messages opening a discussion, and number of replies; and 3) Dataset3, where each node represents a forum post, and directed edges –with a weight of one– show the relation between messages –i.e., which message was written as a reply to another message; this third dataset includes additional information for each node (message) as attributes, such as number of times that a message has been read, the classroom where the message was posted, the type of message board, and the day when it was posted.

The main objectives behind this segmentation of the original dataset is three-fold, with regards to the expected results of their analysis and visualization:

- Dataset1 provides meaningful information about the visible learning interactions taking place in the classroom –that is, who is actively participating or not in the course– and shows the social graph of active interactions; as a direct consequence of this, calculation of SNA parameters of this dataset facilitates the observation of the role played by students –e.g. influencers, knowledge brokers, hub or authorities– and instructors.

[1] http://gephi.org/about

- Dataset2, on the other hand, provides meaningful information about the invisible learning interactions which occur in the classroom, and allows identification of "learning witnesses" and passive learners, as well as most read actors.
- Dataset3 offers relevant information about the popularity of topics and discussions (which might be indicative of potential value for student engagement, or of topics that require further explanation), and their distribution along time.

V. DATA ANALYSIS AND RESULTS

First, and in order to provide further insight from the social network learning analysis of the three datasets, the distribution of final grades for the course was calculated. For reference, figure 1 shows how many students passed, failed or abandoned (did not finish, DNF) the course in each classroom.

From figure 1, success rate is slightly over 50%, with three courses notably underperforming (6, 8 and 9) and two classrooms having a very good overall performance (5 and 7). It is also worth noting the unique results in classroom 3, with just one student failing to pass the continuous assessment, but an attrition rate of nearly 40 percent.

Once the dataset is loaded, the resulting social network is shown in Gephi's Graph window in the *Overview* view. It is then possible to interact with the graph and calculate the different social network parameters (see figure 2).

The different actions that can be performed at this point are the following[2].

Calculation of network parameters, such as node degree centrality, average network degree, network diameter, network density, cluster analysis, modularity analysis for detection of communities, or betweenness and closeness centrality. In order to do so, the corresponding *Run* button in the window *Statistics* must be clicked. After the calculation is finished, the results and distribution diagrams pop up, and the results are incorporated to the *Data laboratory* view, facilitating the export of results.

Advanced filtering: when the window *Filters* is selected, the different parameters may be used to filter the network. When a new network parameter is calculated, it is added to the filter options. Gephi allows to use many types of filters, such as discrete, range and partition filters, which can be added on top of each other, giving users flexibility to achieve a high level of refining. User-defined filters may be saved for later reusability.

Graph manipulation, by using the icons in the *Graph* window. As seen in figure 2, at this stage the social graph is most likely to be somewhat confusing for analysis, and it may be very difficult to extract any meaningful conclusion from its observation. However, it is possible to select and move individual nodes and groups of nodes, get relevant information about each node, zoom in and out, among other actions[3]. It is recommended to use the different preset layouts and the *Ranking* window to display the social graph.

[2] A quick tutorial may be found at https://gephi.org/tutorials/gephi-tutorial-quick_start.pdf

[3] See https://gephi.org/tutorials/gephi-tutorial-visualization.pdf for the complete set actions that can be performed in the *Graph* window.

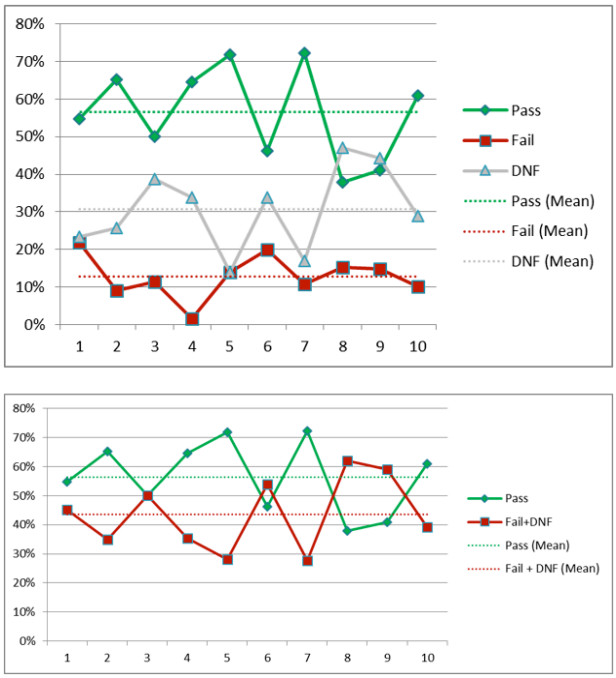

Figure 1. Success, failure and abandon rates by classroom

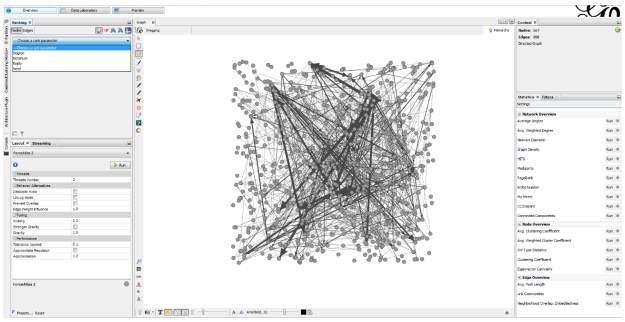

Figure 2. Overview view after import of Dataset1

The use of the *Ranking* window is pretty straightforward, but extremely useful to display relevant information. The different options allow using one parameter of classification –including social network metrics once they have been calculated– and assign relative sizes and colors to the nodes depending on parameter values. Nevertheless, and in order to improve the analysis, some prior layout transformations are recommended.

Layout transformations can be made using the options in the *Layout* window. This study will focus on three of them: ForceAtlas 2, Fruchterman-Reingold and Radial Axis, which will be detailed in the next subsection.

A. Visualization of learning data in Gephi

As seen in figure 2, once data is loaded Gephi displays the resulting network, but it does so randomly, and it becomes a hard task to extract meaningful information from it. This section shows how to use different types of visualizations to help interpreting learning data from a social network learning analytics perspective.

The first of them, ForceAtlas2 [29] creates a network based on forces of attraction and repulsion. ForceAtlas2 aims to provide a generic and intuitive way to spatialize networks with good performances for networks of fewer than 100000 nodes, and keeping a continuous and dynam-

ic layout, resulting in a better user experience. In this kind of layout, the nodes repulse each other based on their degree –the number of their incoming and outgoing edges– while edges attract connected nodes. As a result, high-degree nodes tend to separate from other high-degree nodes and to attract low-degree nodes to which they are connected. This visualization uses some parameters to control for the global dispersion of the network, such as gravity –which controls how disperse separate disconnected components are–, repulsion forces between nodes, or avoiding hubs and overlapping from occurring. Due to its visual interpretation of modularity, the Lin-Log version of this layout is extremely appropriate to assess interaction between actors, such as those from active interactions in message boards –in this study, those from Dataset1– as well as to analyze "invisible" behaviors, such as reading messages, as well as to discover useful content and assess passive class interactions –e.g. Dataset2.

As an example, figure 3 depicts the network from Dataset1, where *Ranking* has been adjusted to show node size representing weighed in-degree –total number of replies received– and node color represents number of new messages –with red nodes having sent a lesser number of messages. Additionally, a third layer of information may be added for analysis by displaying some of the nodes' attributes or metrics –for example, class number or grade–; this third layer has been omitted in figure 1 for clarity purposes.

The second layout is the classic Fruchterman-Reingold layout [30], which is considered a force-directed algorithm that emphasizes placing connected nodes next to each other, but not too close. The algorithm also tries to establish an even distribution of the nodes in the graph, minimizing edge crossings; the result is a visualization such as that in figure 4, where Dataset2 has been used, with node size representing the average in-degree –number of total reads for a given user– and node color representing the total number of messages sent by a user. As expected, this graph is denser than the one for Dataset1, and it is easy to identify the 10 different classrooms. This visualization, as it will be shown later on, is useful to analyze student and teacher active behaviors on a per-classroom basis instead of global course analysis because it provides high quality visualizations for smaller network sizes [29].

Finally, the Radial Axis layout is a type of circular layout which facilitates secondary hierarchical classification of nodes depending on the selected criteria. Radial Axis layout is installed in Gephi as a free plug-in named Circular Layout[4] and its main underlying principle is that nodes are evenly distributed around a circle based on one criterion, and radially distributed based on a second selection criterion. Similar kinds of visualization have been previously been used in learning analytics contexts [31] as a means to display temporal information about the interactions taking place in online classrooms. Therefore, it may be extremely adequate to represent temporal data, and this implies that the imported data must include this information; however, the analysis of temporal data is not straightforward, and therefore including an attribute indicating the time passed in seconds, minutes, hours or days relative to the beginning of the course should be considered as an alternative –in practice, this would require to subtract the timestamp of the starting date to all the records in the learning system log prior to proceed to data import.

[4] https://marketplace.gephi.org/plugin/circular-layout

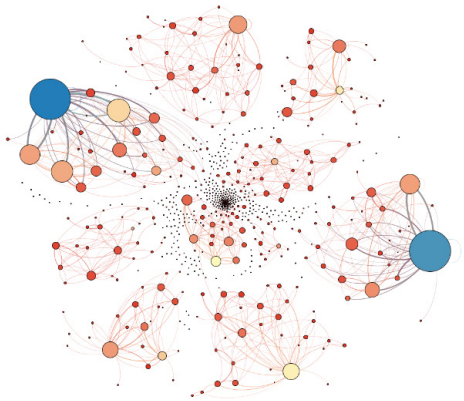

Figure 3. ForceAtlas2 visualization of Dataset1.

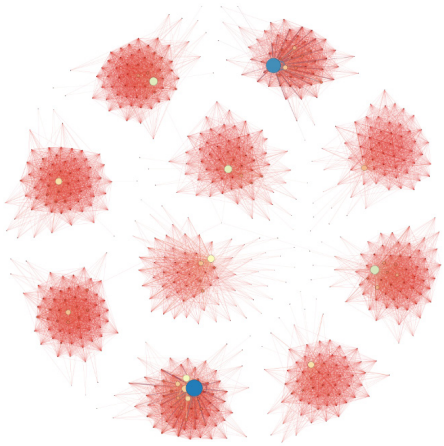

Figure 4. Fruchterman-Reingold visualization of Dataset2.

VI. VISUALIZING THE DATASETS: SOME EXAMPLES

Given the broad set of possibilities derived from the different filtering options available in Gephi, this section uses some of them as an example of how Gephi can be used as a tool to extract meaningful information from the learning systems' log data and facilitate identification of relevant actors and learning behaviors.

A. *Dataset1: Send/reply message networks*

As explained earlier, Dataset1 represents a network of the teachers and students (nodes) and their relations (who communicated with whom in the message boards; these are directed weighed edges, with weight representing the number of times *user a* replied *to user b*). Additional node attributes used for analysis are each student's final grade, total number of messages sent and total number of replies. This facilitates filtering and ranking operations, and it offers additional information about each node when selected. The analysis and visualization of this kind of networks with ForceAtlas2 gives an idea of the interactions occurring in the online classroom. Depending on the selection criteria, it may also offer insights about the relation between student participation and performance (figure 5), group cohesion (figure 6), and student activity and identification of relevant actors who may be acting as knowledge hubs, authorities or "brokers" (figures 7a-c).

Figure 5. ForceAtlas2 visualization of Dataset1. Node colors represent final grade and node size represents out-degree.

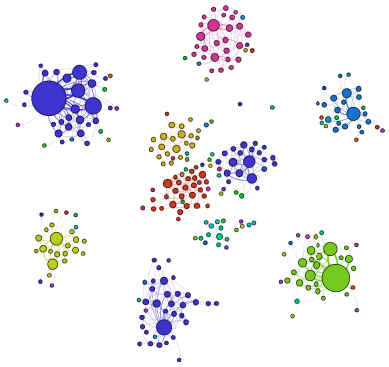

Figure 6. ForceAtlas2 visualization of Dataset1. Node colors represent strongly connected components and node size represents weighed degree.

This visualization (figures 7a-c) provides information about which students are not participating at all in the course dynamics –disconnected students, i.e. unlinked nodes surrounding the sub-networks– as well as the role of teachers –white nodes– in each classroom (the biggest white nodes act as knowledge hubs, authorities or information brokers, respectively). Inspection of figure 5, with bigger nodes (higher participation) having blue colors (higher grades) confirms prior evidence that student's active participation is positively related to academic performance (e.g. [32-34]), and allows the coordinating professor to use this information as proxy of the relative level of engagement in each classroom, so that instructions may be given by the coordinating professor to consultant teachers in order to foster participation in their classroom. Furthermore, it helps identifying relationships between students, an information that can be useful for group configurations based on relatedness, in case group assignments are planned. In figure 6, the majority of connected classmates in a given classroom are identified by the same color but there are some students with different colors; this visualization uses the concept of strongly-connected component as selection criteria, and facilitates the identification of students who are weakly-connected to their peers –i.e., they may not be fully integrated in the classroom– and that may require teacher intervention. In turn, figures 7a-c provide useful information about relevant actors; the bigger nodes in each network represent which students act as conversation hubs, authorities or brokers, respectively. Figure 7b also suggests that higher authority values are associated with higher final grades.

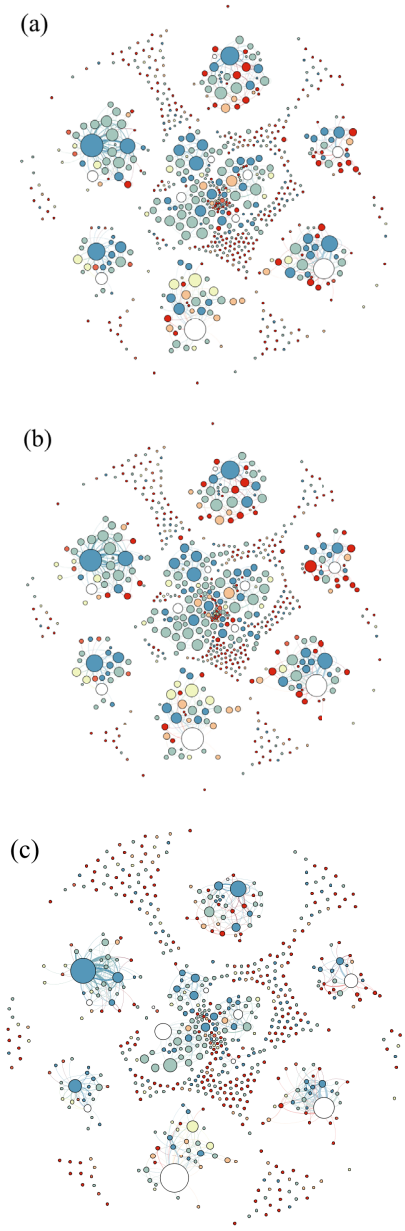

Figure 7. ForceAtlas2 visualization of Dataset1. Node colors represent final grade and node size represents (a) hub; (b) authority; and (c) betweenness centrality.

It must be noted that these visualizations of the whole network may be most helpful for the coordinating professor. However, since the level of activity in each classroom may initially differ largely from one classroom to another –due to many factors, such as teaching styles– outliers in one group could have a great impact in the visualization of the whole network –for example, regarding relative node sizes when using the *Ranking* function. Therefore, further inspection should be made on a per-classroom basis, using the filtering function and recalculating SNA parameters so that the rankings may also be recalculated [26]. These differences will be shown in the next section with Dataset2.

The Fruchterman-Reingold layout may also be helpful to analyze data from Dataset1. Nevertheless, as depicted in figure 8, the information provided by this visualization may be more useful when the network is filtered by classroom.

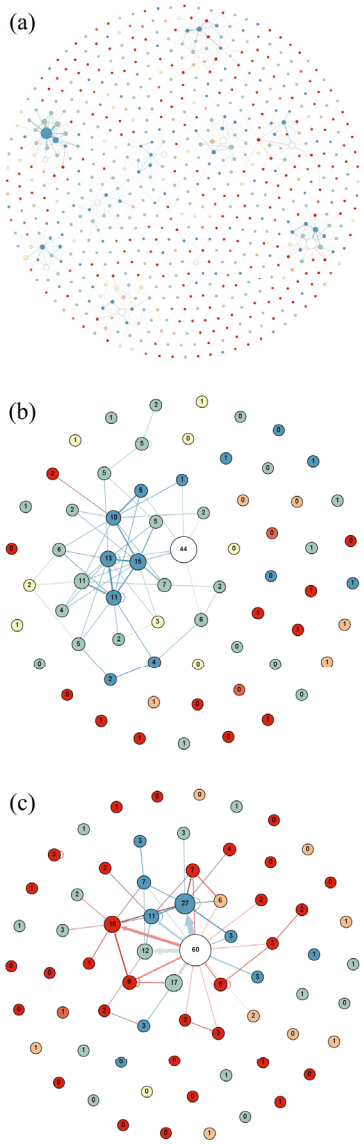

Figure 8. Top-down: Fruchterman-Reingold visualization of (a) Dataset1; (b) classroom 7; and (c) classroom 8. Node sizes and number labels represent number of new messages sent, node colors represent final grade, and edge thickness represent number of replies.

As seen in figures 8b and 8c, very different behavioral patterns may be observed, especially those related to the role of teachers (student-centered learning on classroom 7, and teacher-centered learning on classroom 8). Although a thorough analysis of these data goes beyond the scope of this study, the visual analysis of each classroom showed that classrooms with lower performance had instructors who tended to intervene too much or too little.

B. Dataset2: Read messages network

The analysis of this kind of networks is analogous to that of Dataset1. Nevertheless, as even lesser active students tend to read the messages posted in the message board –at least at the beginning of the course–, the resulting networks tend to be much denser and more difficult to interpret (see figure 4), and therefore the use of Fruchterman-Reingold is not recommended. However, a ForceAtlas2 visualization would be able to provide information about students who have no active or passive engagement whatsoever with the course. In this case, it is advised to

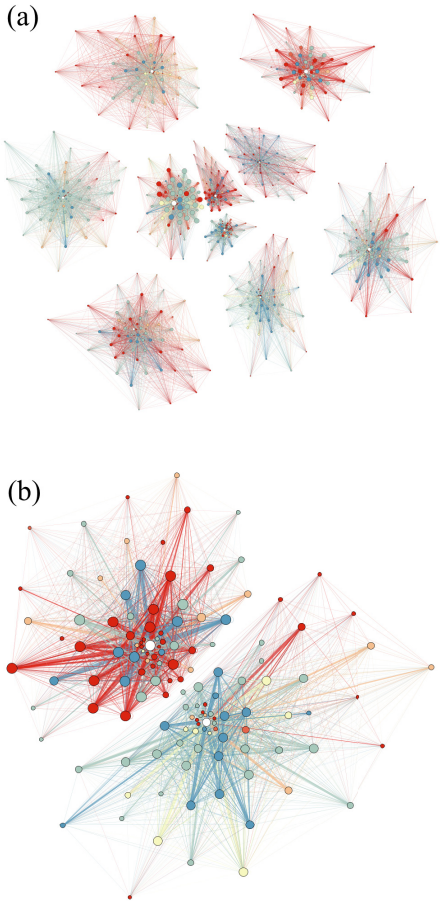

Figure 9. Top-down: ForceAtlas2 visualization of Dataset2 for (a) the whole network, and (b) classrooms 7 and 8. Node sizes represent number of new messages sent, node colors represent final grade and edge thickness represent the total number of read messages.

perform a filtering by classroom in order to avoid the effect of outliers (see figure 9). Observation of the classroom graph allows to observe "lurking" behaviors, and also disconnected students.

Whereas ForceAtlas2 visualizations of Dataset1 provide information about participation, ForceAtlas2 visualizations of Dataset2 provide information on four different but important aspects of learning in online courses.

First, it allows instructors to easily identify –by filtering the resultant network– students who are completely disconnected from the course, and therefore the visualization offers valuable and actionable information for interventions aiming to prevent students from dropping the course. Second, this analysis complements the results from the ForceAtlas2 visualization of Dataset1, by helping instructors to identify lurkers –students who are following the course but not actively participating in it– and relating their activity to academic performance. The coordinating professor and consultant teachers may use this information to make instructional decisions, depending on the desired course dynamics. Third, it provides insight on which users are contributing the most valuable and interesting content –bigger nodes, with thicker and higher number of edges incoming. Finally, the visualization allows detection of students who may be active readers but are not performing well. Teachers may use this information in order to make sure that the concepts are clear and that no misunderstandings or misconceptions are happening.

C. *Dataset3: Posted messages network*

In Dataset3, each node represents a message posted to any message board. Therefore, a ForceAtlas2 layout would be helpful, but since threads tend to increase along time, its use might not be recommended for large courses (see figure 10, from which no useful information can be directly extracted). Instead, if temporal data is included, a Radial Axis layout is highly recommended (see figure 11). The use of Radial Axis visualization has, however, one caveat: the radius used for the visualization tends to increase as the number of threads –conversations– and messages per thread increases, and therefore the usefulness of this layout is determined by the total number of threads and thread messages. In the example used in figure 11, the network was filtered by classroom, and a Connected-Components test was run using the calculation from the *Statistic* window; then, the Radial Axis layout was performed, grouping the nodes by (weakly) connected ComponentID[5] and order in the axis by their ID –so that nodes farther from the center represented messages posted later in time; then, node size was determined by read count of each message –attribute from the preprocessing stage– and colored by date –from red to blue– since course start to control for temporal evolution.

From figure 11, it is easy for instructors to see which topics have not been replied yet and which are, or have been, more popular, engaging or controversial during the course. This visualization then becomes a valuable source of information either to alter course dynamics when convenient or for future course planning –posting topics of relevance in advance, for example.

Radial axis layout, as in figure 11, concentrates a large amount of information in a single figure. For instance, node color shows whether posting activity is uniformly distributed along the course or concentrated in specific periods; axis length indicates thread popularity, which –in the case of longer elements– might refer to interesting topics for discussion; isolated nodes indicate unanswered threads, if posted by students ("S" nodes); on the other hand, isolated "T" nodes –threads initiated by teachers with no replies– generally refer to announcements; node size may be an indicator of message relevance, as bigger nodes have higher number of reads; finally, edges offer a simple way to observe the flow of discussion in a thread, and thus they may help detecting when a discussion is going off-topic, messages that promote debate or issues that require further clarification. Interestingly enough, figure 9 shows how in both courses student-initiated threads generate more discussion than threads started by the teachers.

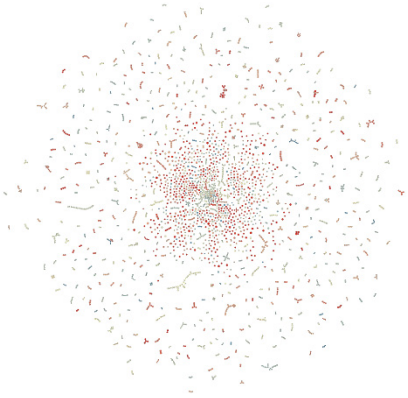

Figure 10. ForcedAtlas2 layout of Dataset3 (whole network).

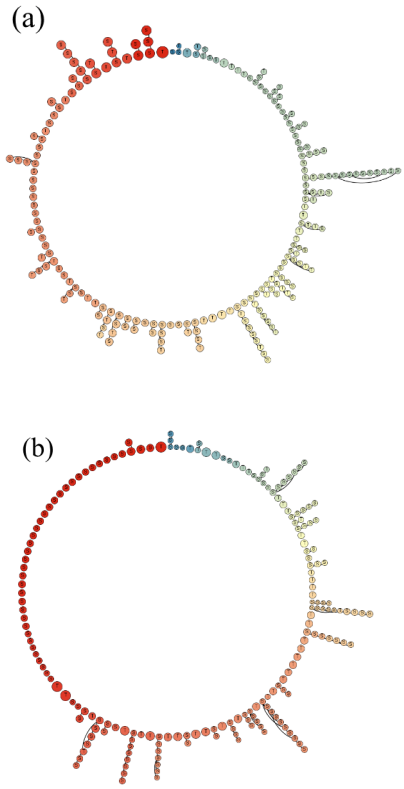

Figure 11. Radial Axis layout for Dataset3 for (a) classroom 1; and (b) classroom 7. Node size represents read count, node color represents day since starting date and label identifies type of poster (student or teacher). In (a), the distribution of messages is more or less uniform during the course. In (b), the higher posting activity occurs in the first third of the course.

VII. SNA METRICS

As an intermediate step in the visualization process, SNA parameters were also calculated for each dataset, both for the complete network and on a per-classroom basis, in order to provide further insights on the use of social network learning analytics.

Regarding centrality measures, we calculated their correlation – both parametric and non-parametric – with final grade (only the 456 students who passed the continuous assessment were included in this analysis). A significant but low correlation (between 0.18 and 0.24) was found, with weighed out-degree (the number of times that a student read or replied to a message) having the highest correlation with academic performance. These results suggest that instructors and course coordinators should choose to rank by weighed out-degree in the initial phases of analysis and visualizations in Gephi. Interestingly enough, a significant positive relation between closeness centrality and student outcomes was found in Dataset1, but this correlation was significant and negative in Dataset2.

[5] Interestingly, the Radial Axis layout seem to be limited to group a maximum of 127 nodes for secondary axis; therefore when the number of threads is higher, from the 128th thread, all nodes are connected along the main axis. This limitation does not seem to be addressed yet in Gephi documentation.

From the global network parameters of the whole network and each classroom's network, no significant differences were found between better and worse performing classrooms, but when instructors were excluded from the analysis, the courses with higher drops in degree-related and path-related parameters corresponded to those with higher drop-out rates. Furthermore, the analysis of SNA metrics suggested that higher network diameters were more characteristic of better performing learning networks and should probably be used as a first means to rapidly compare performance across courses.

Finally, a one-way ANOVA of centrality measures showed significant differences between groups, but the post-hoc test with Bonferroni and Tukey-b corrections did not provide useful information to group classrooms with similar patterns. Nonetheless, the analysis showed that classrooms 1 and 4 differed the most with the other classrooms, a result that would require further investigation.

VIII. CONCLUSION

Despite the fact that current online learning systems store a great amount of information about students' and teachers' activity in their databases, online distance education still faces some important challenges for instructors, especially when it comes to student tracking. In the past years, a large body of research in the emerging field of learning analytics has focused on solving this problem by designing systems able to extract, summarize and visualize that information so that action could be taken based on the processed data.

Social network learning analytics focuses on interaction data among the different actors involved on the learning process, and relies heavily on data visualization as a means to display this information. However, efforts have focused on the use of tools tailored to the specifics of each learning system, tools that fall short of the features and functionality required for advanced analysis [27]. This study shows how Gephi, an open source SNA application, may be used as a tool for social network learning analytics, using data from one online course at the UOC as an example.

The main objective of this study was to facilitate and encourage researchers and practitioners around the world to further explore the possibilities of this kind of approach to the study of online learning interaction, by demonstrating how to use existing tools for social network learning analytics. For that reason, it was considered that an in-depth analysis of the course data and visualizations in this article would fall beyond the scope of the study and would only create confusion in the reader –yet, some results have been given and explained along the text; more avid readers may find further examples in [26].

Furthermore, it must be warned that although this paper is oriented towards the general public, the large number of options available in Gephi may become a barrier for many non-expert users. Therefore, further research on social network learning analytics is needed in order for experts to be able to fine-tune Gephi and offer easy-to-use and customized settings of the program options. This customization should be accompanied by training programs for teachers willing to delve into the possibilities of Gephi on how to better understand and manipulate the network visualizations and analysis results.

It must also be pointed out that extended use of general purpose SNA tools is at this moment still hindered by the need to process the information from the learning system's format to the SNA tool's format [35]. Nevertheless, the diversity of formats supported by Gephi and the relative simplicity of the data structure required for analysis make it relatively easy to develop specific plug-ins to export data from learning systems logs (e.g., *GraphFES* for Moodle [25-26]), and therefore this study also serves as a call to the programming community to develop this kind of interfaces.

Additionally, this study focuses on social network learning analytics based on interactions, but the uses of Gephi could extend to other types of learning analytics, such as visualizations of discourse analytics; for example, using SPARQL queries if we include semantic information as attributes –although this approach would most likely require additional preprocessing of data.

Finally, it has to be emphasized that this study might be seen as just a "tip of the iceberg": social network learning analytics is a relative new field and, as such, we are just beginning to be aware of its possibilities and scope of application in the educational landscape, both from a learning data visualization perspective and from the statistical analysis of SNA metrics. Therefore, it is the authors' hope that this study may inspire other scholars to promote the research and use of these approaches in education, in order to improve learning processes in online education.

REFERENCES

[1] C. Reffay, and T. Chanier, "How social network analysis can help to measure cohesion in collaborative distance-learning. Designing for change in networked learning environments: *Computer-Supported Collaborative Learning*, Vol. 2, 2003, pp. 343–352.

[2] S. Dawson, "A study of the relationship between student social networks and sense of community". *Educational Technology & Society*, Vol. 11(3), 2008, pp. 224–238.

[3] L. Macfadyen and S. Dawson, "Numbers are not enough. Why e-learning analytics failed to inform an institutional strategic plan". *Educational Technology & Society*, Vol. 15(3), 2012, pp. 149–163.

[4] P. Long and G. Siemens, "Penetrating the Fog: Analytics in Learning and Education", *EDUCAUSE Review*, Vo. 46(5), 2011, pp. 31–40.

[5] Á. F. Agudo-Peregrina, S. Iglesias-Pradas, M. Á. Conde-González and Á. Hernández-García, "Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning", *Computers in Human Behavior*, Vol. 31, 2013, pp. 542–550. http://dx.doi.org/10.1016/j.chb.2013.05.031

[6] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an ''early warning system" for educators: A proof of concept". *Computers & Education*, Vol. 54(2), 2010, pp. 588–599. http://dx.doi.org/10.1016/j.compedu.2009.09.008

[7] A. F. Wise and S. N. Hausknecht, "Learning Analytics for Online Discussions: A Pedagogical Model for Intervention with Embedded and Extracted Analytics". *Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13)*, 2013, pp. 48–56. http://dx.doi.org/10.1145/2460296.2460308

[8] M. F. Beaudoin, "Learning or lurking? Tracking the "invisible" online student". *The Internet and Higher Education*, Vol. 5(2), 2002, pp. 147–155. http://dx.doi.org/10.1016/S1096-7516(02)00086-6

[9] S. N. Durlauf and H. Peyton Young, *Social Dynamics*. Cambridge, MA: MIT Press, 2001.

[10] Á. Hernández-García and M. Á. Conde, "Dealing with complexity: educational data and tools for learning analytics". *Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM '14)*. New

York:ACM, 2014, pp. 263–268. http://dx.doi.org/10.1145/2669711.2669909

[11] F. Mosteller, S. E. Fienberg and R. E. R. Rourke, *Beginning Statistics with Data Analysis*, 1983.

[12] M. De Laat and F. Prinsen, "Social learning analytics: Navigating the changing settings of Higher Education". *Research & Practice in Assessment*, Vol. 9, 2014, pp. 51–60.

[13] J. Oshima, R. Oshima and Y. Matsuzawa, "Knowledge building discourse explorer: a social network analysis application for knowledge building discourse". *Educational Technology Research and Development*, Vol. 60(5), 2012, pp. 903–921. http://dx.doi.org/10.1007/s11423-012-9265-2

[14] D. Vu, P. Pattison and G. Robins, "Relational event models for social learning in MOOCs". *Social Networks*, Vol. 43, 2015, pp. 121–135. http://dx.doi.org/10.1016/j.socnet.2015.05.001

[15] S. Buckingham-Shum and R. Ferguson. "Social learning analytics". *Journal of Educational Technology & Society*, Vol. 15(3), 2012, pp. 3–26.

[16] D. Cambridge and K. Perez-Lopez, "First steps towards a social learning analytics for online communities of practice for educators". *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, 2012, pp. 69–72. http://dx.doi.org/10.1145/2330601.2330622

[17] K. Nurmela, E. Lehtinen and T. Palonen, "Evaluating CSCL log files by social network analysis". P*roceedings of the 1999 conference on Computer support for collaborative learning (CSCL '99)*, Article No. 54, 1999. http://dx.doi.org/10.3115/1150240.1150294

[18] B. Sundararajan, "Emergence of the most knowledgeable other (MKO): Social network analysis of chat and bulletin board conversations in a CSCL system". *Electronic Journal of E-Learning*, Vol. 8(2), 2010, pp. 191–208.

[19] R. Rabbany, S. ElAtia, M. Takaffoli and O. R. Zaïane, "Collaborative learning of students in online discussion forums: A social network analysis perspective". *Educational Data Mining: Applications and Trends*. Berlin, Heidelberg, 2013, pp. 1–30.

[20] D. García-Saiz, C. Palazuelos and M. Zorrilla, "Data mining and social network analysis in the educational field: An application for non-expert users". *Educational Data Mining: Applications and Trends*. Berlin, Heidelberg: Springer Berlin/Heidelberg, 2013.

[21] L. Tobarra, A. Robles-Gómez, S. Ros, R. Hernández and A. C. Caminero, "Analyzing the students' behavior and relevant topics in virtual learning communities". *Computers in Human Behavior*, Vol. 31, 2014, pp. 659–669. http://dx.doi.org/10.1016/j.chb.2013.10.001

[22] Á. Hernández-García, I. González-González, A. I. Jiménez-Zarco and J. Chaparro-Peláez, "Applying social learning analytics to message boards in online distance learning: A case study", *Computers in Human Behavior*, Vol. 47, 2015, pp. 68–80. http://dx.doi.org/10.1016/j.chb.2014.10.038

[23] L. C. Freeman, "Centrality in social networks: Conceptual clarification". *Social Networks*, Vol. 1(3), 1979, pp. 215–239. http://dx.doi.org/10.1016/0378-8733(78)90021-7

[24] H. Cho, G. Gay, B. Davidson and A. Ingraffea, "Social networks, communication styles, and learning performance in a CSCL community". *Computers & Education*, Vol. 49(2), 2007, pp. 309–329. http://dx.doi.org/10.1016/j.compedu.2005.07.003

[25] M. Bastian, S. Heymann and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks". *Proceedings of the Third International ICWSM Conference*, 2009, pp. 361–362.

[26] Á. Hernández-García and I. Suárez-Navas, "GraphFES: A web service and application for Moodle message board social graph extraction". *Big Data and Learning Analytics in Higher Education. Current Theory and Practice*. Springer International Publishing, in press.

[27] J. Chaparro-Peláez, E. Acquila-Natale, S. Iglesias-Pradas and I. Suárez-Navas, "A web services-based application for LMS data extraction and processing for social network analysis". *New Information and Communication Technologies for Knowledge Management in Organizations. Lecture Notes in Business Information Processing*, Vol. 222, Springer International Publishing, 2015, pp. 110-121.

[28] S. Joksimović, D. Gašević, T. M. Loughin, V. Kovanović and M. Hatala, "Learning at distance: Effects of interaction traces on academic achievement". *Computers & Education*, Vol. 87, 2015, pp. 204–217. http://dx.doi.org/10.1016/j.compedu.2015.07.002

[29] M. Jacomy, S. Heymann, T. Venturini and M. Bastian, "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software". *PLOS One*, Vol. 9(6), e98679, 2012. http://dx.doi.org/10.1371/journal.pone.0098679

[30] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement". *Software: Practice and Experience*, Vol. 21(11), 1991, pp. 1129–1164. http://dx.doi.org/10.1002/spe.4380211102

[31] D. A. Gómez-Aguilar, R. Therón and F. J. García-Peñalvo, "Semantic spiral timelines used as support for e-learning". *Journal of Universal Computer Science*, Vol. 15(7), 2009, pp. 1526–1545.

[32] J. Alstete and N. Beutell, "Performance indicators in online distance learning courses: a study of management education". *Quality Assurance in Education*, Vol. 12 (1), 2004, pp. 6–14. http://dx.doi.org/10.1108/09684880410517397

[33] J. Davies and M. G. Graff, "Performance in e-learning: online participation and student grades". *British Journal of Educational Technology*, Vol. 36 (4), 2005, pp. 657–663. http://dx.doi.org/10.1111/j.1467-8535.2005.00542.x

[34] S. Hrastinski, "What is online learner participation? A literature review". *Computers & Education*, Vol. 51 (4), 2008, pp. 1755–1765 http://dx.doi.org/10.1016/j.compedu.2008.05.005

[35] D. Amo Filvà, F. J. García-Peñalvo, M. Alier Forment, "Social network analysis approaches for social learning support". *Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM '14)*. New York:ACM, 2014, pp. 269–274.

AUTHORS

Á. Hernández-García is with Departamento de Ingeniería de Organización Administración de Empresas y Estadística, Universidad Politécnica de Madrid, Madrid, Spain (e-mail: angel.hernandez@upm.es).

I. González-González is with Universidad de Navarra, Pamplona, Spain (e-mail: ines.gonzalez@unavarra.es).

A.I. Jiménez-Zarco is with Universitat Oberta de Catalunya, Barcelona, Spain (e-mail: ajimenezz@uoc.edu).

J. Chaparro-Peláez is with Departamento de Ingeniería de Organización Administración de Empresas y Estadística, Universidad Politécnica de Madrid, Madrid, Spain (e-mail: julian.chaparro@upm.es).

An earlier version of this study titled *Usare Gephi per visualizzare la partecipazione nei corsi online: Un approccio di Social Learning Analytics* was published in TD Tecnologie Didattiche, Vol. 22(3), pp. 148–156, under a CC-BY-NC license. This article is an extended and modified version of a paper presented at the Fifth International Workshop on Adaptive Learning via Interactive, Collaborative and Emotional approaches (ALICE 2015) in conjunction with the Seventh International Conference on Intelligent Networking and Collaborative Systems (INCOS 2015), held in Taipei, Taiwan, September 2-4, 2015. Submitted 26 May 2016. Published as resubmitted by the authors 29 June 2016.