Linking Engagement Profiles to Academic Performance Through SNA and Cluster Analysis on Discussion Forum Data

Pamella L. S. de Oliveira[1], Gabriel C. da Silva[1], **Raphael A.** Dourado [2,3] and Rodrigo L. Rodrigues [1]

[1] *Universidade Federal Rural de Pernambuco, Rua dom Manuel de Medeiros, s/n, Recife, 52171-900, Brasil*
[2] *Universidade Federal de Pernambuco, Av. Prof. Moraes Rego, 1325, Recife, 50670-901, Brasil*
[3] *Instituto Federal da Paraíba, Campus Soledade, R. das Trincheiras, 275, João Pessoa, 58011-000, Brasil*

Abstract

Given the increasing popularity of online education, it is important to understand how social interaction takes place in the discussion forums commonly used in the platforms that support these courses – the Virtual Learning Environments (VLEs). Since interaction and engagement are two intertwined phenomena in online learning, and the later influences academic performance, it is possible to identify students at risk of dropping out by analyzing their interaction patterns in the discussion forums. In this work, we applied Social Network Analysis (SNA) and cluster analysis to forum data to identify groups of students with different engagement profiles. We identified four profiles and, when analyzing the particularities of each one and relating them to students' grades, we found a connection between engagement profiles and academic performance. Therefore, the characteristics of each engagement profile identified in our work can help teachers and managers in defining strategies to increase student interaction in the virtual environment and thus prevent dropout.

Keywords

Distance education, social network analysis, cluster analysis, engagement, academic performance

1. Introduction

Despite the growing interest and acceptance of distance education, the lack of student-student and student-teacher social interaction is still an open problem in this domain, which compromises communication and causes a feeling of isolation among students [1].

Interaction, dialogue and collaboration are factors that determine the nature of learning, since the quality of distance learning courses is directly related to these factors [2]. In addition, interaction fosters student engagement [3]; He [4], for instance, found positive correlations between the number of questions students send to the instructor and their final grades. In sum, the literature shows a relationship between engagement and learning and its impact on academic performance, acquisition of knowledge, and motivation.

Social Network Analysis (SNA) has been used in previous studies to measure student interaction and engagement. In [5], the authors show how the results from SNA analysis motivated teachers to look

LALA'21: IV LATIN AMERICAN CONFERENCE ON LEARNING ANALYTICS - 2021, October 19–21, 2021, Arequipa, Perú

EMAIL:pamella.lesilvao@gmail.com (A. 1); gabcandidods@gmail.com (A. 2); rasd2@cin.ufpe.br (A. 3); rodrigomuribec@gmail.com (A. 4)
ORCID: 0000-0001-8251-0992 (A. 1);0000-0002-7470-8613 (A. 2); 0000-0001-6445-6590 (A. 3); 0000-0002-3598-5204 (A. 4)

for new ways to monitor their work and of their students', thus improving collaboration and offering personalized help to students.

Although previous studies have used SNA to understand peer interaction, there is still a gap in the literature when it comes to show how interaction between students can impact engagement and, as a consequence, influence academic performance. Therefore, the goal of this work is to answer the following questions: 1) Is it possible to find different engagement profiles on discussion forums using SNA metrics?; and 2) What are the relationships between engagement profiles and students academic performance? The remainder of this paper is divided into four sections: background (section 2), methodology (section 3), results and discussions (section 4), and conclusion (section 5).

2. Background: Engagement and Social Networks Analysis

Current literature offers three categories of definitions for educational engagement: (1) the cognitive, which is related to the idea of effort; (2) the behavioral, which is usually measured from quantitative aspects in relation to the actions of students; and (3) the emotional, which is associated with the feeling of belonging to a group [6].

In addition, collaboration is one of the social aspects that can be used to measure student engagement in educational activities, since it can be estimated by the improvement in the volume and quality of student involvement, satisfaction, engagement and learning [3].

One of the research areas that seeks to understand and measure collaboration is Social Network Analysis (SNA). According to Saqr et al. [5], SNA is a distinct type of analysis that can be used to map the relationships and interactions between agents within groups in participatory environments. This technique is widely used in the literature to measure the level of interaction between students and student-teacher. In [2], for instance, the authors used SNA to find interaction patterns in discussion forums and, in this way, help instructors in the longstanding issue of following students' learning progress effectively.

Many studies use the SNA to analyze social interaction and show how this type of analysis can support teachers and administrators. In [7], the authors use SNA metrics to build a system for tracking interactions in forums. In the study of [8], SNA techniques were used in conjunction with the Random Forest prediction algorithm to predict and improve student performance, thus improving motivation and providing individualized feedback, either by a teacher in a small course or by an automated system in a massive course.

In a previous work [6], we show that grouping students by engagement profile can help teachers and management staff reflect and adjust the course to prevent retention and dropouts. In the study of [9], the authors used clustering techniques to analyze student participation in forums; they concluded that such analysis is useful to identify groups with distinct behavioral characteristics, often imperceptible to managers and teachers.

3. Methodology

The methodology we used followed an iterative process consisting of eight phases: understanding the problem, choosing the database, understanding the database, extracting SNA metrics, summarizing and viewing SNA results, applying clustering algorithms, validating clustering algorithms, and summarizing the resulting groups. The following sections describe some of these phases in detail.

3.1. Choosing and understanding the database

After collecting information from the students, in order to understand the two questions of this research, it was necessary to search for articles in the literature that dealt with social network analysis and engagement. Therefore, readings were made in systematic reviews that addressed engagement and SNA.

We used data from a Biology program offered online by *Universidade de Pernambuco* (a public state university in Brazil) through the Moodle LMS platform. The collected data included grades of webquest activities, graded posts in discussion forums, and face-to-face tests plus the interaction logs of chat and discussion forums. We included in the analysis only the students that participated in the four graded forums required by each course in the program, which totalled 616 students. The whole Biology program is composed of fifteen courses and lasts eight semesters.

3.2. Extracting SNA metrics and summarizing the results

We extracted the SNA metrics for student engagement (measured by their interaction in the discussion boards) using RStudio (https://www.rstudio.com/) and the *igraph* network analysis package (https://igraph.org/). In this way, we obtained the values for the indegree, outdegree, degree, closeness and betweenness metrics. To plot the graphs, we used Gephi (https://gephi.org/).

Table 1
Application of SNA metrics in the educational context.

Metrics	Importance in the educational context
Indegree	Indicates how many connections the student received, thus being useful to measure their popularity.
Outdegree	Indicates how many connections the student has made, showing each student's contribution to the others.
Degree	Indicates whether the interaction is centralized in a small group of students or evenly distributed in the network.
Closeness	Indicates the distance between students in the network. Therefore, those with higher values are considered isolated and have received little information and influence from the network.
Betweenness	Shows students who are responsible for distributing information among students, and thereby connecting multiple groups.

Table 1 describes the importance of each metric in the educational context. These metrics allow the identification of isolated students, the popularity of each student, those who influence the network, and those who are central to the dissemination of information in the network.

3.3. Application and validation of clustering algorithms

We used cluster analysis to search for engagement profiles in discussion boards. This technique is useful to classify the data in different groups or categories initially unknown based on automatically identified patterns found by manipulating the characteristics of the data [10].

Initially, we used the Hopkins statistic to find the best subset in our data. Then, we used two approaches to find the ideal number of clusters: the Elbow and hierarchical methods. Finally, we validated the clusters using internal and external validations.

4. Results and Discussions

In this section, we present the results from the SNA and cluster analysis. Subsections 4.1 to 4.4 detail the results of each step in our analysis process and how each of them builds on the results of the previous one. Subsection 4.5 discusses the relationships we found between engagement profiles and academic performance. Finally, Subsection 4.6 characterizes the four engagement profiles identified in our work.

4.1. Extraction of SNA metrics

Table 2 shows the average, median, standard deviation, minimum, and maximum statistics for the SNA metrics calculated for each student in our dataset. The low average values for the indegree, outdegree, degree, and betweenness metrics indicate that most students do not interact much in the environment. On the other hand, the high maximum values for these same metrics reveal that some few isolated students show high interaction levels with other students, which is confirmed by the low values for the “closeness” metric.

Table 2
Average values for SNA metrics

Statistic/Metric	Indegree	Outdegree	Degree	Closeness	Betweenness
Average	1.062	1.000	3.012	0.471	0.677
Median	0.750	1.000	2.000	0.500	0.000
Standard Deviation	1.501	2.757	2.867	0.399	3.417
Minimum	0.000	0.000	0.250	0.000	0.000
Maximum	10.000	19.250	19.250	1.100	41.750

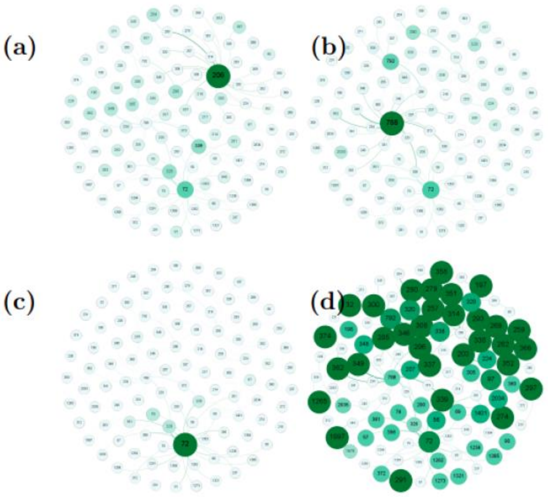

Figure 1: Graph representation of SNA metrics

Figure 1 shows the graphical representation for four SNA metrics: (a) indegree, (b) outdegree, (c) closeness e (d) betweenness, generated with Gephi. Each node represents a unique student, each edge represents the interaction between two students, and the size and color of the nodes represents the level of interaction for the student --- the greener and bigger the node, the higher that student interaction was. These graphs reinforce the uniformity in students' interaction levels shown in Table 2: while some students interact a lot (the greener nodes), most of the others show low interaction levels.

Therefore, given that interaction in the discussion boards does not occur uniformly, teachers can benefit from knowing the different engagement profiles to personalize their practice accordingly.

4.2. Clustering of SNA metrics

After extracting and visualizing the SNA metrics, we further analyzed them using clustering techniques to look for different engagement profiles and understand how they relate to students' academic performance. First, we used the Hopkins statistic to find the best set or metrics for clustering. This statistic evaluates the grouping tendency of a dataset; according to Kassambara [11], the closer the Hopkins statistic result is to zero, the greater is the possibility of finding significant clusters in the dataset. We tested five combinations and the one that yielded the best results (Hopkins' statistic closest to zero) was the mean of the SNA metrics "indegree", "outdegree", "degree", and "closeness" calculated for each of the four forums available in the course. Therefore, we adopted this combination.

Then, we used two methods to find the ideal number of clusters: Elbow and Hierarchical. The Elbow method tests the data variance in relation to the number of clusters; it finds the optimal number of clusters (cut-off point) when an increase in this number does not result in a significant gain value [6]. The hierarchical method, on the other hand, works by successively grouping or dividing elements, in which elements are aggregated or disaggregated in order to build a hierarchy of clusters. The result of the hierarchical grouping is represented through a cluster tree, also called Dendrogram [10]. In both methods, the results indicated that the optimal number of clusters for our dataset was four.

4.3. Cluster validation measures

To evaluate the quality of our clusters, we ran two types of validation: internal and external. For internal validation, we used the Dunn index and the Silhouette Coefficient methods. The Dunn index is useful for identifying compact and well-separated clusters, which is the case when the distance between the clusters is large and their diameter is small. Thus, large values for the Dunn index indicate the presence of compact and well-separated clusters [12]. The Silhouette Coefficient returns a value between -1 and 1; a value close to 1 indicates a good clustering. The tests were performed using the K-Means algorithm with values varying from two to six for K. The respective results for K considering the Dunn index are: 1.19, 0.63, 0.56, 0.52, 0.64. As for the silhouette coefficient, they were: 0.49, 0.45, 0.49, 0.50, 0.52.

The value six for K presented the best result for the silhouette coefficient and the second best for the Dunn index. However, the results from the hierarchical and Elbow method indicated an optimal number of four clusters. Therefore, it was necessary to run an external validation to decide between using 4 or 6 clusters.

For external validation, we used the Rand index, which provides a measure to assess the similarity between two randomly fitted partitions. This index ranges from -1 (no agreement) to 1 (perfect agreement) [11]. We then used it to verify the difference of the grouping with values of four and six for

K. The test returned a value of 0.94, indicating that there was a perfect agreement between these groups, which allowed the choice of either one.

4.4. Graphical representation of the clusters

Finally, as the Rand index indicated an agreement between the values of 4 and 6 for the number of clusters, we analyzed the balance of the clusters to make a decision. In the 4-cluster scenario, the size of each cluster was, respectively,: 282, 89, 13, and 232 students. In the 6-cluster scenario, the sizes were: 225, 8, 271, 22, 85, and 5. Therefore, the 4-cluster scenario resulted in more balanced clusters, which led us to choose this option to proceed with the study.

The choice of the K-Means algorithm was because it is considered the most used non-hierarchical algorithm and when compared to the hierarchical method, this method is faster [10]. The graph in Figure 2 shows the cluster formed by the K-Means algorithm. It is possible to see in the graph that the four groups have small intersections between them, indicating that some observations are very close to groups different from yours. However, most observations are distant from other groups, indicating that they actually belong to the group in which they are found.

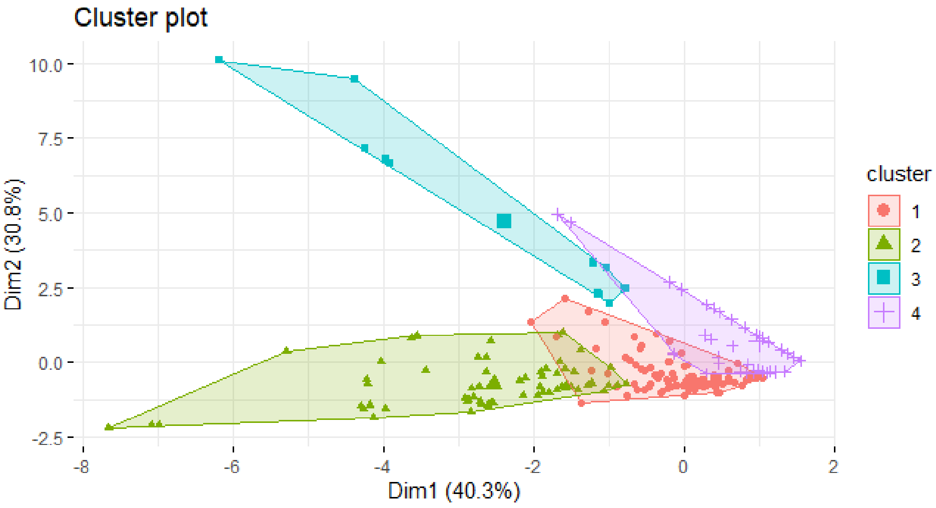

Figure 2: Graph generated by the K-Means algorithm, where K=4.

4.5. Linking SNA metrics and academic performance

In this section, we analyze the clusters behavior in relation to SNA metrics and students' average grades in the webquest activities, exams, and graded forum posts. These results are shown in Figures 3 and 4 through boxplots, which allow the identification of central tendencies, variability, and outliers. Analyzing how interaction takes place within the discussion forums and identifying the different group profiles is essential so that teachers can adopt methodologies based on the profile of each group.

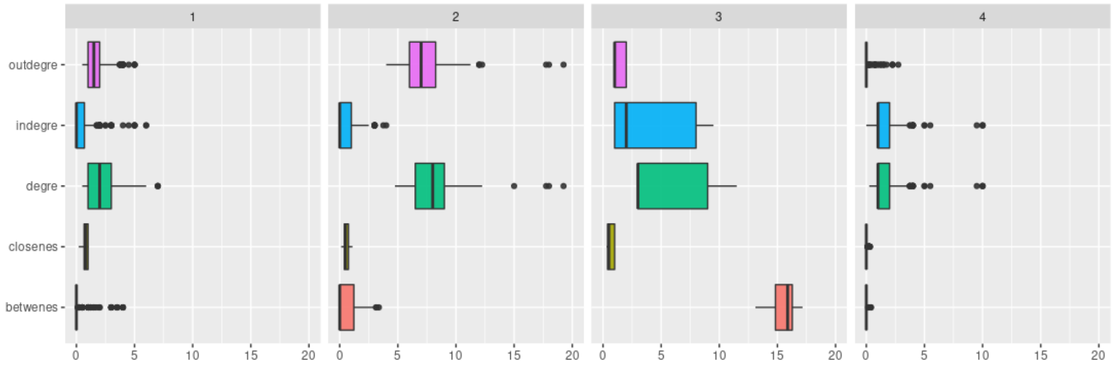

Figure 3: Clusters behavior in relation to SNA metrics.

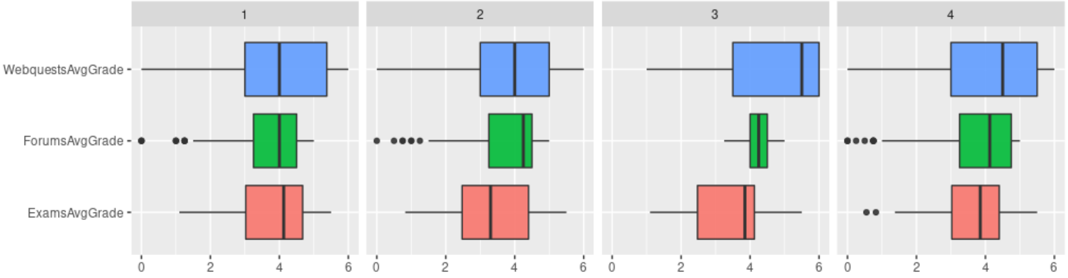

Figure 4: Clusters behavior in relation to average grades, by type of activity.

By observing the metrics for each cluster in Figures 3 and 4, we can spot several distinct characteristics. In cluster 1, both the SNA metrics and grades' average values coincide with the median, indicating that students belonging to this group have an irregular interaction pattern, which leads to also irregular grades. Cluster 2 shows good median results for the "outdegree" and "degree" metrics as well as for the average grades; however, the median of "indegree" and *ExamsAvgGrade* are not as good, which suggests that students in this group interact consistently in the forums but other students do not interact back with them, thus explaining the low values for the *ExamsAvgGrade* metric. Cluster 3 stood out among the four groups, since it is the only one that does not have any outlier and shows the highest median value for the "betweenness" and *WebquestsAvgGrade*; this means that students in this group interact with different groups, building bridges between them, which resulted in the good grades shown for this group in Figure 4. Finally, cluster 4 showed the worst results, with median values of zero for three out of the five interaction metrics and many outliers in the grades metrics (Figure 4); therefore, students in this group show the lowest interaction levels and high variance in their grades.

Therefore, the results presented above suggest that there is a relationship between engagement, interaction, and academic performance. This is in line with previous studies that suggest a correlation between student engagement and successful learning outcomes, as discussed in [13].

4.6. Characterization of engagement profiles

Table 3 describes the four engagement profiles identified in this work for students' interaction in forums plus the particular characteristics of each group. In general, it can be concluded that group

1, the “sporadic” ones, are students who interact irregularly in the forums. Group 2, “socializer”, represents the most active students in the forums, the ones who interact regularly. Group 3, “integrator”, are the students who interact with different student profiles, building bridges in the class. Finally, group 4, “isolated”, are the ones with lower participation level.

Table 3

Engagement profiles in discussion forums

Group	Name	Characteristics
1	Sporadic	It is a group made up of people who use the forum feature with irregular periodicity. Students in this group are likely to be students who only try to interact when they have questions. In this way, they obtain average grades that are a reflection of the irregularity in the interaction.
2	Socializer	It's a very active group in the forums. Students belonging to this group are very communicative and are the ones who most seek to interact with other students, even when their counterparts do not respond to these interaction attempts. It is possible that these students interact mostly in an attempt to have their questions about the course subjects answered by classmates.
3	Integrator	Students belonging to this group have the ability to interact with colleagues from groups with different interaction profiles, building bridges between the groups. As a result, these students are able to gain knowledge in different ways and contribute to increase collaboration in the forums. This behavior also leads to higher grades when comparing other groups.
4	Isolated	This is the group with the lowest level of engagement in forum discussions. Students in this group are the ones who rarely seek to interact with others. However, some students in this group are contacted by other students in search of interaction.

In sum, these results can change the perception of educators about how interaction happens in forums, which may influence the way they plan and monitor such learning activities in search of improving students’ interaction and engagement. Also, analyzing the particularities of the different interaction profiles can help in adopting personalized pedagogical interventions for each group.

5. Conclusion

This study aimed to examine the interaction in discussion forums using social network analysis to find the values of the indegree, outdegree, degree, closeness and betweenness metrics. This allowed us to confirm that the interaction in the forums does not happen in a uniform way. We then used cluster analysis on the SNA metrics to identify the different student profiles, which revealed four groups: *sporadic, solicializer, integrator, and isolated*.

We also looked for relationships between engagement profiles and academic performance. For this purpose, we analyzed the behavior of each group and the grades they obtained in the course activities, which revealed a relationship between the SNA metrics and grades.Students in the 'Integrator' group are seen as bridges between groups and obtain the best grades. The 'Socializers' group represents the most communicative students, although there is a high variation in their grades. The ‘Sporadics’ group shows an irregular participation in the forums and average grades. Finally, the group of ‘Isolated’ students is the one with the lowest interaction levels and also low grades.

Overall, this study was able to show that 1) it is possible to find different engagement profiles through SNA metrics, and 2) there is a relationship between engagement and academic performance. Therefore, knowledge of these different engagement profiles can help educators in making decisions to avoid student failure and dropout. As future work, we intend to apply the same analysis techniques used in this work to other datasets (such as other courses or programs) and also test other EDM techniques to discover new behaviors or predict student behaviors.

References

[1] H. Karal, V. Nabiyev, A. K. Erümit, S. Arslan, A. Çebi, Students’ opinions on artificial intelligence based distance education system (artimat), Procedia - Social and Behavioral Sciences 136 (2014) 549–553. URL: https://www.sciencedirect.com/science/article/pii/S1877042814038555. doi:https://doi.org/10.1016/j.sbspro.2014.05.374, GLOBAL CONFERENCE on LINGUISTICS and FOREIGN LANGUAGE TEACHING (LINELT-2013).

[2] E. Gottardo, R. V. Noronha, Social networks applied to distance education courses: Analysis of interaction in discussion forums, Association for Computing Machinery, New York, NY, USA, 2012. URL: https://doi.org/10.1145/2382636.2382710. doi:10.1145/2382636.2382710.

[3] I. Messias, L. Morgado, M. Barbas, Students’ engagement in distance learning: Creating a scenario with lms and social network aggregation, 2015, pp. 44–49. doi:10.1109/SIIE. 2015.7451646.

[4] P. He, Evaluating students online discussion performance by using social network analysis, in: 2012 Ninth International Conference on Information Technology - New Generations, 2012, pp. 854–855. doi:10.1109/ITNG.2012.72.

[5] M. Saqr, U. Fors, M. Tedre, Como o estudo da aprendizagem colaborativa online pode orientar professores e prever o desempenho dos alunos em um curso de medicina, BMC Med Educ 18 24 (2018). doi:10.1186/s12909-018-1126-1.

[6] P. Oliveira, R. Rodrigues, J. Ramos, J. Silva, Uma análise de algoritmos de clusterização para descoberta de perfis de engajamento, in: Anais do XXXI Simpósio Brasileiro de Informática na Educação, SBC, Porto Alegre, RS, Brasil, 2020, pp. 1012–1021. URL: https://sol.sbc.org.br/index.php/sbie/article/view/12857. doi:10.5753/cbie.sbie.2020.1012.

[7] A. Silva, Figueira, Depicting online interactions in learning communities, in: Proceedings of the 2012 IEEE Global Engineering Education Conference (EDUCON), 2012, pp. 1–8. doi:10.1109/EDUCON.2012.6201094.

[8] S. G. Cabeza, A. L. Arredondo, P. Massaferro, N. Rubido, A. M. Hirt, Redes profesionales en procesos educativos en línea, in: Proceedings of the 2nd Latin American Conference on Learning Analytics, 2019.

[9] R. Rodrigues, J. Ramos, J. Sedraz, A. Gomes, Discovery engagement patterns moocs through cluster analysis, IEEE Latin America Transactions 14 (2016). doi:10.1109/TLA.2016.7785943.

[10] J. Ramos, R. Rodrigues, J. Sedraz, A. Gomes, R. Silva, A comparative study between clustering methods in educational data mining, IEEE Latin America Transactions 14 (2016) 3755. doi:10.1109/TLA.2016.7786360.

[11] A. Kassambara, Practical Guide To Principal Component Methods in R: Unsupervised Machine Learning, STHDA, 2017.

[12] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, Journal of Intelligent Information Systems 17 (2001). doi:10.1023/A:1012801612483.

[13] M. O. Maia, J. Figueiredo, D. Serey, Online student engagement: A case study in teaching of programming, Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE 2019) (2019). doi:10.5753/cbie.sbie.2019.51