Online-Journals.org

iJET International Journal of Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 20 No. 2 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijet.v20i02.54517

PAPER

Clustering Students Based on Online Learning Interactions Using Social Network Analysis

Amir Narimani(✉), Elena Barberà

Universitat Oberta de Catalunya (UOC), Barcelona, Spain

anarimani@uoc.edu

ABSTRACT

One of the key areas of interest within learning analytics is identifying student similarities to support collaborative applications such as score prediction, personalized recommendations, and group formation. Clustering is a prominent method for grouping students based on shared learning behaviors, enhancing peer learning, and fostering communities in online courses. This study introduces an intuitive graphical clustering approach using activity logs that track student engagement with different learning resources. These interactions are modeled as multi-dimensional vectors, and a social network of learners is constructed using cosine similarity. Social network analysis (SNA) is then applied to detect learner communities. The dataset for implementation and evaluation includes activity logs and grades from 792 students in an undergraduate study program. Results indicate that learners in the same clusters have similar interaction patterns and grade point averages (GPAs). Statistical measures, such as silhouette index and root mean square standard deviation (RMSSTD), demonstrate the method's effectiveness and benchmark its performance against K-means clustering. This approach shows significant potential for uncovering and visualizing implicit learner groups.

KEYWORDS

learning analytics (LA), learner modeling, similarity calculation, social network analysis (SNA), community detection

1 INTRODUCTION

Learning analytics (LA) has emerged as an interdisciplinary approach to tackling the challenge of analyzing massive datasets in the education sector. LA is grounded in its capacity to offer stakeholders—educators, administrators, and students—a comprehensive view of the learning process, uncovering trends and predicting outcomes to improve teaching methods and learning experiences [1]. A critical aspect of LA is the emphasis on personalization. Personalization in e-learning is often achieved by developing learner models that are extracted from various data sources, such as profile data, behavioral logs, and interaction histories. These models help in

Narimani, A., Barberà, E. (2025). Clustering Students Based on Online Learning Interactions Using Social Network Analysis. *International Journal of Emerging Technologies in Learning (iJET)*, 20(2), pp. 36–50. https://doi.org/10.3991/ijet.v20i02.54517

Article submitted 2025-01-20. Revision uploaded 2025-02-19. Final acceptance 2025-02-19.

understanding the specific needs of learners, leading to more targeted interventions and resources. One of the fundamental tasks in achieving this personalization is the classification or clustering of learners based on shared characteristics or learning behaviors [2], [3]. This process involves grouping learners by their performance, learning styles, preferences, or engagement levels, enabling educators to offer personalized learning paths that optimize the educational experience. Classification, a supervised learning approach, involves categorizing learners into predefined classes based on certain metrics. In contrast, clustering is an unsupervised method that groups learners into clusters without predefined class labels, allowing for the discovery of patterns and learner groupings that might not be initially apparent. Both techniques are crucial for identifying distinct learning behaviors, understanding group dynamics, and tailoring instructional strategies accordingly [4].

Clustering can have various collaborative applications in e-learning, significantly enhancing both teaching and learning experiences. Some of these applications include group formation for collaborative tasks, the development of recommendation systems for personalized learning pathways, peer assessment strategies that leverage student strengths, and performance prediction models that identify students at risk of underperforming [5]. Clustering algorithms draw upon a variety of data sources to categorize learners, utilizing both static learner profile data—such as age, gender, prior educational experience, and responses to questionnaires administered before the learning process—and dynamic learner actions that reflect ongoing engagement [6].

Despite the extent of information available, much of the existing literature primarily focuses on modeling learners based on static data and final grades. While such approaches offer insights into overall performance, they often overlook the details of learning behaviors exhibited throughout the educational experience. Best practices in e-learning emphasize the importance of utilizing detailed activity logs that capture a range of learning actions. Evaluating students' learning activities and participation is essential, as these factors can serve as predictive indicators of learners' performance and information literacy development [7]. Leveraging clustering algorithms to analyze the learning activities allows for a proactive approach to enhancing student learning outcomes [8].

This paper introduces a novel method for clustering learners based on their dynamic interactions with e-learning systems, focusing on learning actions rather than static profile data or final grades. Learner models are constructed from detailed activity logs, capturing interactions such as resource usage and login frequency represented as multi-dimensional vectors. Relational graphs are built to depict learner connections based on activity similarity, with clusters identified using a social network analysis (SNA) community detection algorithm. The proposed method is evaluated by analyzing behavior similarity within clusters and grade point average (GPA) as a measure of academic success. Finally, root mean square standard deviation (RMSSTD) is used to assess clustering effectiveness and compare it with the K-means algorithm. Evaluations highlight the advantages of this approach in visualizing meaningful learner groupings and envisioning their educational success. The proposed method can be customized to enhance personalized educational strategies by incorporating various metrics and types of learning activities.

2 LITERATURE REVIEW

This section outlines the foundational concepts that underpin this research: learner modeling, clustering algorithms, and SNA communities. A comprehensive

review of relevant studies in these domains will be provided, highlighting the significant contributions that have advanced our understanding of these areas. Together, these three areas create a robust framework for exploring how SNA can be utilized to cluster students effectively based on their online learning activities. This literature review discusses the interplay between these concepts and their implications for enhancing educational practices.

2.1 Learner modeling

E-learning has evolved with the advent of new learning applications such as massive open online courses (MOOCs), where personalized learning experiences can greatly enhance student engagement and achievement. Learner modeling involves creating representations of learners based on their characteristics, such as cognitive abilities, learning styles, and prior knowledge. By understanding these unique profiles, educators can tailor instructional strategies and content to better suit individual learners, fostering a more supportive and effective learning environment [9]. To construct an ideal learner model, one should identify and select the learner's characteristics that influence their learning and choose the most adapted technologies to model each characteristic with the best precision [10].

Adaptive learning systems leverage data from learners' interactions with online platforms to create dynamic profiles on a big scale that evolve over time. For example, authors utilize student academic information, social relations, and interactions to build academic social networks [11]. Researchers calculated learners' efforts by video watch time and grades history to model learners and recommend courses [12]. In another research, learner demographic profile and click history are the core features of the learner models [13]. Authors modeled student profiles upon detecting the users' learning styles and learning preferences, as well as their knowledge level and misconceptions [14].

Research has explored how grouping learners based on their profiles enables educators to deliver more precise and individualized interventions that directly address specific learning challenges [15], [16]. These approaches have been shown to enhance educational outcomes by catering to the diverse needs of students, allowing for more effective instructional strategies and personalized support that ultimately improve performance. However, the role of social factors in learner modeling, particularly in collaborative learning contexts, needs further investigation to provide a more holistic view of the learner.

2.2 Clustering algorithms

Clustering deals with the data structure partitions in an unknown area and is based on the idea of objects being more related to their peers in the same groups. At its core, clustering relies on distance or similarity metrics to ascertain relationships among data points. The applications of clustering are diverse, spanning fields such as market analysis, customer segmentation, search result clustering, recommendation systems, and pattern recognition. In the context of e-learning, clustering can be categorized into two primary types based on the focus of grouping. The first type concentrates on the relationships among learning resources, which can be classified based on attributes such as title, description, topic, instructor, language, or

institution [17], [18]. The second category focuses on learners as the primary subjects of clustering, aiming to identify individuals with similar behaviors or attributes and organize them into cohesive groups.

In their compelling research, the authors implemented a peer assessment approach based on a matching strategy. Their method relies on two essential steps that guide both the matching process and problem representation: 1) modeling each learner as an assessor, and 2) grouping assessors into categories based on their assessment competency [19]. To generate resource recommendations, the authors introduced a novel deep neural network (DNN) approach that leverages synchronous sequences and heterogeneous features. To address the cold-start problem with new learners, the approach begins by clustering learners and then integrates sequence data and heterogeneous features as embedding within the DNN model to improve recommendation accuracy [20].

2.3 Social network communities

The concepts and properties of graph theory make it highly effective for describing and visualizing clustering problems, especially in the context of social networks. A social network can be represented as a graph, where nodes depict entities, such as individuals, organizations, or any other relevant units, and edges represent meaningful relationships or interactions between these nodes. SNA is a popular and growing field that focuses on the study of social networks to uncover meaningful patterns and insights among entities. In these networks, a community typically represents a group of entities (such as individuals) that are more closely connected than the rest of the network [21].

Social network analysis community detection has a wide range of applications for studying e-learning users' interactions, particularly in addressing challenges related to collaborative learning and providing personalized resource recommendations. For instance, authors employed a community detection algorithm to recommend relevant question and answer (Q&A) forum topics in MOOCs, enhancing learners' ability to engage with the most pertinent discussions and thereby improving their learning experience [22]. Community detection has also shown promising results in enhancing collaboration within lifelong learning networks, where learners' long-term development and continuous engagement are supported through better peer connections and collaborative opportunities [23]. In another study, the authors used SNA to examine user participation, highlighting the role of active users and lurkers. The study shows that while the communities had higher weighted reciprocity compared to similar random networks, participation was concentrated among a small group of highly active users [24]. These examples show how community detection enhances e-learning by fostering peer collaboration and personalizing learning, creating more personalized and adaptive learning ecosystems.

3 PROPOSED LEARNER CLUSTERING METHOD

The primary goal of this study is to develop a practical method for clustering learners using data on their access to various learning resources. The proposed method follows a multi-step process to identify learner groups based on their activity patterns. Figure 1 highlights the key steps involved in this clustering approach.

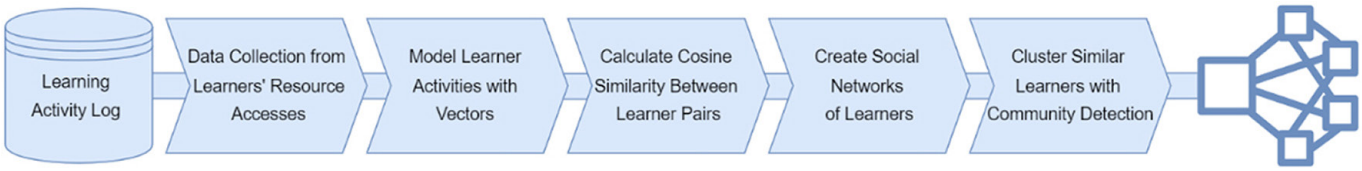


Fig. 1. The proposed learner clustering method steps

The initial step involves data collection and pre-processing. On e-learning platforms, activity logs are typically organized in tables, with each row representing a specific action performed by a user at a given time. These actions can range from watching a video, reading a chapter, and taking an exam to posting on a forum or rating learning content. Each piece of content includes an indicator identifying the resource the user interacted with. At this stage, it is necessary to group these actions by both user ID and resource ID to determine the total number of times a learner engaged with each learning resource. For every learner and action type, we define an N-dimensional vector, where N corresponds to the number of available resources within that particular action category, expressed as:

$$Learner - ActionType = \{n_1, n_2, n_3, ..., n_i | i \in [1, 2, 3, ..., N]\} \quad (1)$$

Where n_i depicts the number of times a learner accessed a resource such as watching a video, and N is the total number of resources of the same type (here, the number of videos in the subjects). Afterwards, we have $k \times l$ N-dimensional vectors since k is the number of different action types and l represents the total number of learners.

We employed the cosine function to evaluate the similarity between users' activity records. Despite its straightforward definition and implementation, cosine similarity has proven to deliver reliable and precise results in identifying user similarities [25]. This approach offers a robust method for comparing learner behaviors across different activities, enabling more accurate similarity calculations based on their interaction patterns. For each pair of learners, A and B, the calculation of vector similarity for a specific action type is performed as follows:

$$Cosine - Similarity = CS_k(A_k, B_k) = \frac{\sum_{i=1}^{n} A_{ki} B_{ki}}{\sqrt{\sum_{i=1}^{n} A_{ki}^2} \sqrt{\sum_{i=1}^{n} B_{ki}^2}} \quad (2)$$

In this formula, A_{ki} represents the number of times learner A has accessed the i-th learning material, with k serving as the indicator of action type. Since the learner-action type vectors contain only positive values, the cosine similarity yields a value between 0 and 1. A score of 0 indicates that the learners have never accessed the same resource, while a score of 1 means they have accessed the same resources the same number of times. To determine the overall similarity between learner pairs, we compute the weighted sum of individual similarity measures, as expressed by the following equation:

$$Sim(A, B) = \alpha . CS_1 + \beta . CS_2 + ... + \gamma . CS_k \quad (3)$$

Considering the importance of learners' access to various types of resources in our application, we can assign different positive values to α, β, and γ, ensuring that $\alpha + \beta + ... + \gamma = 1$. Once the total similarity between learner pairs is calculated, we can construct their network. In this social network, nodes represent the learners, and

edges between them are weighted by their similarity, provided the value is greater than 0. The larger the edge weight, the stronger the similarity between two learners, with a value of 1 indicating a perfect match.

We apply the community detection method developed and implemented through the Gephi software to form clusters of similar learners [26]. This method organizes the network structure hierarchically through iterative steps and has demonstrated strong accuracy when dealing with large networks compared to other approaches. However, as networks grow in size, community detection becomes exponentially more time-consuming. To address this, an extension of the Blondel method was introduced to limit the algorithm's time complexity [27]. The resolution parameter allows the network to be segmented at various scales within a specific time frame. This measure halts further clustering, enabling the algorithm to identify optimal intermediate partitions. In the next section, we outline the assigned value for the resolution parameter and provide case study details about the data used to implement our proposed clustering approach.

4 DATA DESCRIPTION (AN ONLINE DEGREE PROGRAM CASE STUDY)

To implement our proposed method, we utilized the dataset provided by the Universitat Oberta de Catalunya (UOC). This dataset contains anonymized information about students' profiles, their interactions with the e-learning system, and their academic performance. It covers one and a half years, from 2021 through the first semester of 2022, across all undergraduate and graduate programs. Due to the dataset's large size, consisting of 36,625,315 classroom access records and 598,059 student grades, we filtered the data by study programs to manage the scale effectively. For our method's implementation, we focused on the economics undergraduate program, selecting data from 792 unique students in their third or fourth year of study.

Our analysis concentrates on three types of student activities (Action Types), which serve as inputs for our clustering method: Activity View, Library Material Access, and Teaching Plan View. The "Activity View" data represents instances where students have viewed assigned activities on the e-learning platform. "Library Material Access" refers to the times when students accessed external learning resources through the library. Finally, the "Teaching Plan View" records access to the platform's main page, where students can view their courses and track their academic progress. Table 1 provides a detailed breakdown of the characteristics of the economics undergraduate dataset. This filtered data allows us to analyze learning actions more effectively, providing a robust foundation for our clustering approach.

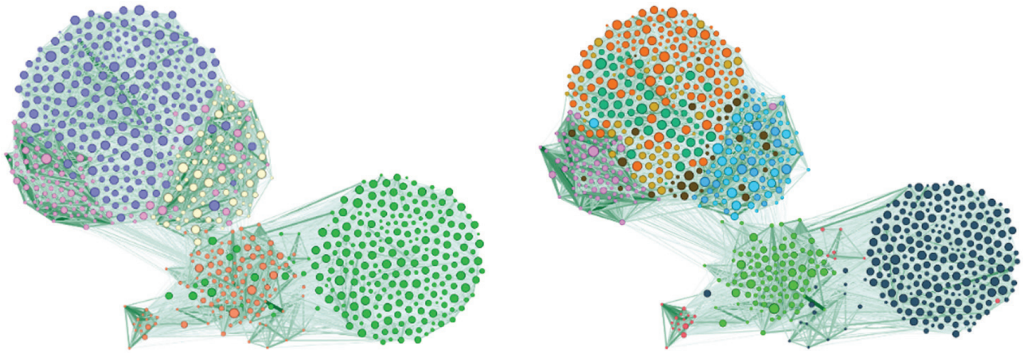
Table 1. UOC economics undergraduate studies data description

Data Table Name	Number of Records	Data Description
Student-Basic	792	Enrolled student's basic information
Activity-View	167250	Student's access to different activities
Library-Access	9683	Student's access to library materials
Teaching-Plan	12518	Times student viewed teaching plan page
Student-Grades	6596	Student's score in subjects based on a scale from 0 to 10

To begin implementing our method, we first need to organize students' actions according to the learning resources they accessed. Specifically, we have 792 students

who performed 843 different activities, accessed 729 distinct library materials, and viewed 737 teaching plans. This gives us three vectors for each student, with 843, 729, and 737 entries, respectively, capturing their learning behavior. The values in these vectors represent the number of times a student accessed the same learning activity page, with the majority of the values being 0, since students enrolled in different subjects and learning resources are very diverse. The next step involves calculating cosine similarities between every pair of students based on these vectors. For each action type, we obtain a similarity score ranging from 0 to 1, indicating the degree of similarity between two students. For the "Activity-View" vector, we computed 88,556 non-zero similarity values out of 313,236 possible pairs. For "Library-Material-Access" and "Teaching-Plan-View," we found 35,630 and 29,902 non-zero similarity values, respectively. In total, using Equation 3 from the previous section, we derived 154,088 similarity values for students in the economics undergraduate program. We assigned equal weights of $\alpha = \beta = \gamma = 1/3$ for these calculations.

As described earlier, we can construct the students' social networks, where learners are represented as nodes, and the total similarity scores serve as edge weights. To create and visualize this network, we utilized Gephi software, as introduced in Section 3. The social network provides a valuable opportunity to explore learner similarity and other dynamics relevant to various group applications. Using the community detection method, we generated learner modularity clusters using two different resolution settings. By setting the resolution parameters to 0.5 and 1.0, we formed 10 clusters and 5 clusters, respectively, for the economics undergraduate students. Figure 2 visualizes the clustered social network for the economics dataset. In this figure, the colors of the nodes represent the clusters to which learners belong, and the edges indicate the relationships between learners. The size of each node corresponds to its degree, with larger nodes representing learners who share similarities with many others, signifying that they accessed resources frequently used by other students.



a) Five clusters with a resolution parameter of 1.0

b) Ten clusters with a resolution parameter of 0.5

Fig. 2. Learner similarity network and clusters in the economics studies

The size and density of the edges in the networks represent the similarity value between two learners. Thicker and darker edges signify a higher degree of similarity between the connected nodes. For instance, a darker, thicker edge linking two smaller nodes indicates that these learners frequently accessed the same resources. Ideally, such high-similarity edges are found within the same cluster, as this would suggest that the clustering algorithm successfully grouped students with similar behaviors. It is less desirable for these strong connections to exist between nodes

in different clusters, as this would imply that learners with similar resource usage patterns were placed in separate groups, potentially weakening the cohesion of the clusters. This visual representation allows for an intuitive understanding of the clustering results, as dense, dark edges within clusters reinforce the accuracy of the grouping. Additionally, the structure of the network can help identify key learners who act as central figures in the flow of shared learning resources, shedding light on how e-learning behaviors spread within the community.

5 EVALUATIONS

This section presents the results of applying our method to the case study dataset and using different measures to evaluate the proposed method's accuracy. Evaluating clustering methods is more challenging than classification tasks, as clustering lacks labeled data. Various metrics exist for assessing clustering results that typically focus on factors such as within-cluster similarity and between-cluster differences. These metrics indicate clustering accuracy, though they may not be enough for objectively evaluating the quality of clustering algorithms, especially in specific contexts such as learner clustering. Clustering evaluation is highly context-dependent, as no universal metric can effectively assess outcomes across all datasets and applications [28].

To evaluate the efficiency of the proposed learner clustering method, we utilized several metrics, including the average Silhouette score, students' GPAs, and the RMSSTD metric. The results were compared against benchmarks established using the K-means clustering algorithm. K-means is a widely used data mining method, known for its effectiveness in e-learning tasks such as student clustering [29]. The analysis of the proposed method was conducted with resolution parameters set to 1.0 and 0.5, which resulted in the formation of five and 10 student clusters, respectively. To ensure consistency, we set the value of K to 5 and 10 in the K-means algorithm, representing the number of clusters. It is worth noting that K-means employs Euclidean distance to calculate the similarity between student access records. The detailed evaluation findings are presented in the subsequent subsections.

5.1 Evaluation by average silhouette index

The silhouette index measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates well-separated and cohesive clusters. The silhouette score for a data point i and the overall silhouette index can be calculated using Equations 4 and 5, respectively:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{4}$$

$$S = \frac{1}{N} \sum_{i=1}^{N} s(i) \tag{5}$$

In the equations, $a(i)$ represents the average distance of point i to all other points within the same cluster, while $b(i)$ denotes the average distance to all points in the other clusters. The distance is defined as the negation of the cosine similarity, calculated using Equation 3. Table 2 presents the calculated Silhouette values obtained from various examinations of the proposed method and the K-means algorithm.

Table 2. Comparison of Silhouette index for different clustering scenarios

	Proposed Method (Resolution = 1.0)	Proposed Method (Resolution = 0.5)	K-Means (K = 5)	K-Means (K = 10)
Silhouette index	0.7106	0.5544	0.6694	0.5419

5.2 Evaluation by student GPA

We used the students' GPAs (on a scale of 0 to 10) from the economics program as a reference metric to evaluate the effectiveness of the clustering approach. This metric highlights the predictive capability of the proposed algorithm in grouping similar learners, as it processes access records as input and outputs their corresponding GPAs. By calculating the average GPA within each cluster and comparing it to the overall GPA distribution of the student population, we could determine how well the clustering algorithm grouped students with similar academic performance. It is important to note that some students do not have a GPA due to either skipping exams or withdrawing from their studies, so our GPA-based analysis includes 740 students out of the original 792.

Figure 3 illustrates the average GPAs for all students, as well as the mean GPA for each cluster, using resolution parameters set at 1.0 and 0.5 for the proposed method and K-means results with K = 5 and K = 10. The overall mean GPA for all students is 7.1699, which is indicated by the green line in the figure. The difference between the blue bars, which represent the mean GPA for each cluster, and the green line highlights how each cluster's GPA mean deviates from the overall student population's GPA average. This visual comparison helps to assess and compare how well the clustering algorithms distinguish students based on academic performance.

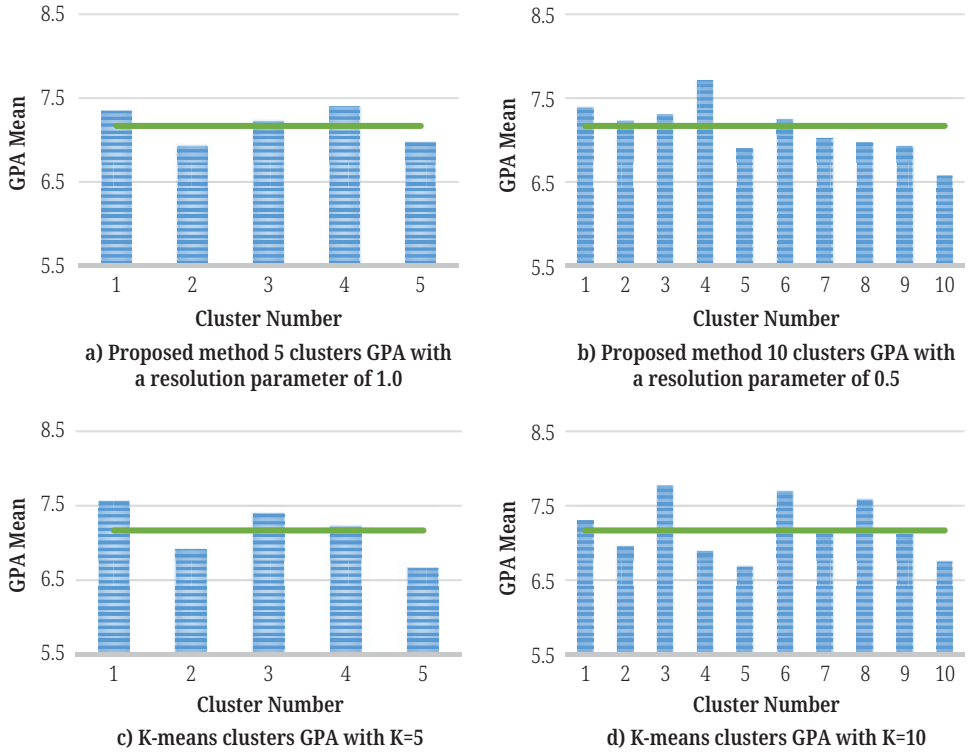


Fig. 3. GPA Means for proposed SNA-based clustering method and K-means

For instance, in Figure 3a, Cluster 1 has a mean GPA of 7.3451, indicating that students in this cluster tend to achieve higher scores compared to the overall student population.

We computed the standard deviation (STD) for GPAs of the entire student population and for each cluster separately, using both the proposed method and K-means. In Figure 4, the green line represents the STD for all students, while the blue columns depict the calculated values. A higher STD indicates that GPAs within the cluster are diverse, which is undesirable for the proposed clustering method.

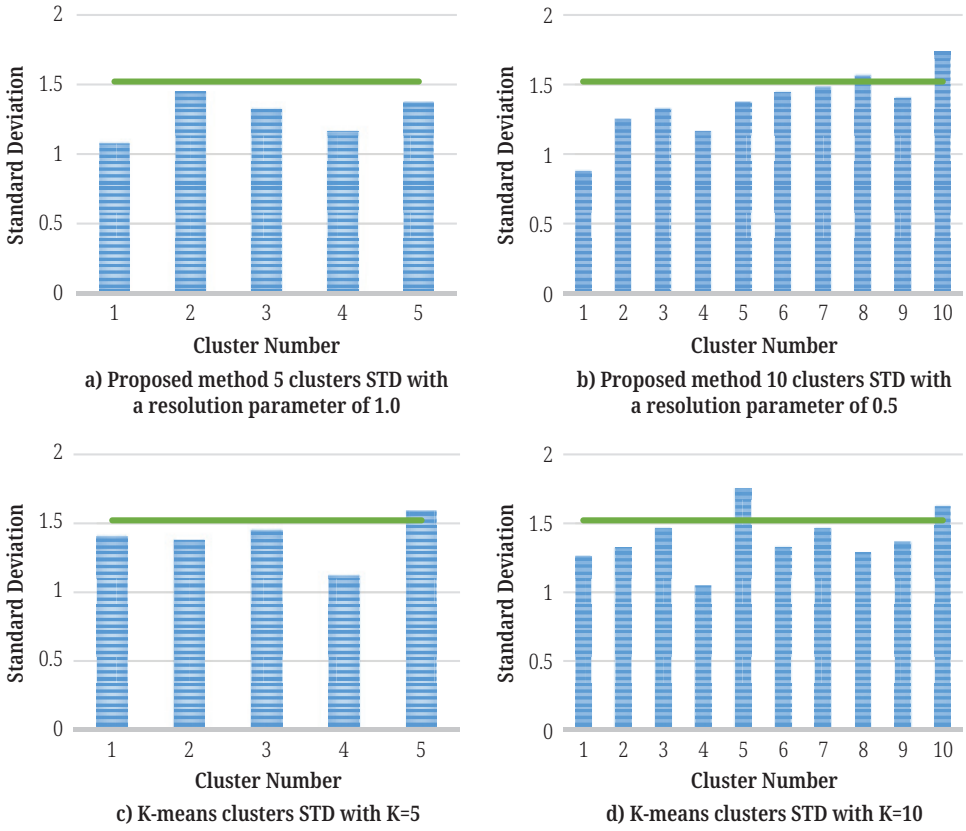


Fig. 4. GPA STD values for proposed SNA-based clustering method and K-means

For example, the smallest STD value of cluster 1 in Figure 4b indicates that students' GPAs are very close to the mean GPA of 7.3861 within the cluster (in Figure 3b). The highest STD belongs to cluster 10 with a size of 20 and a low mean GPA of 6.5766, which is equal to 1.7343. This number suggests that the small number of students in the cluster is not similar in GPAs. Possibly, their learning behavior is different than other clusters that they have grouped together by the proposed method.

5.3 Evaluation by RMSSTD

We utilized another insightful metric, the RMSSTD, to assess the quality of the clustering algorithm. A lower RMSSTD value signifies better cluster separation and improved overall clustering performance [30].

$$RMSSTD = \sqrt{\frac{\sum_{j=1\ldots p}^{i=1\ldots k}\sum_{a=1}^{n_{ij}}(x_a - \bar{x}_{ij})^2}{\sum_{j=1\ldots p}^{i=1\ldots p} n_{ij} - 1}} \quad (6)$$

Where k is the number of clusters, p is the number of independent variables in the dataset, $\bar{x}_{ij}$ is the mean of data in variable j and cluster i, and n_{ij} is the number of data in variable p and cluster k. In our evaluations, k equals 10, and the variable p is the students' GPA in different clusters.

To compare our method with K-means clustering results, we calculated the RMSSTD on different occasions. First, without considering the clusters (all population is considered as one general cluster), then calculating it with the presence of our clustering method, and then with results of K-means clustering where K = 10 and K = 5. The calculated general RMSSTD was 1.6243, and the clustered RMSSTD with the proposed SNA-based method and K-means are represented in Table 3.

Table 3. Comparison of RMSSTD for different clustering scenarios

	Proposed Method (Resolution = 1.0)	Proposed Method (Resolution = 0.5)	K-Means (K = 5)	K-Means (K = 10)
RMSSTD	1.5566	1.5031	1.5614	1.5346

The comparison shows that the proposed method grouped similar learners in the same clusters slightly better than K-means. Also, learners in the different clusters generally have less similarity based on their access to different learning resources.

6 RESEARCH CONTRIBUTION AND DISCUSSION OF RESULTS

This study introduces a novel approach to learner clustering in online education, leveraging SNA to uncover hidden patterns and groupings among students. Unlike traditional clustering methods, this approach emphasizes student interactions with online learning platforms and the types of access these students have. To our knowledge, this is the first attempt to visualize student groups through a social network representation that explicitly highlights each learner's position within the network and their similarities to peers. This foundational work provides a basis for future research aimed at enhancing collaborative online education applications. It can be further expanded by incorporating more detailed student activities, such as video watch time or participation in Q&A forums.

The evaluation of the clustering algorithm is another significant contribution, offering novel insights into addressing challenges related to clustering evaluation metrics. However, further research is required to refine clustering details, such as the optimal resolution parameter and the number and size of clusters. Our analysis reveals that these parameters critically influence evaluation outcomes. For example, based on the Silhouette index, fewer clusters generally perform better, whereas, for GPA-based and RMSSTD evaluations, higher separation resulting from a larger number of clusters yields better outcomes. This allows the educational application to identify the optimal number of clusters for the proposed approach.

Compared to the K-means algorithm, our proposed method produces clusters with greater variability in size (ranging from 20 to 220 learners) and more pronounced separation. In contrast, K-means generates more uniform cluster sizes.

Our findings reveal intriguing relationships between cluster characteristics and student performance. Cluster size notably influences GPA means. For instance, smaller clusters tend to exhibit GPA averages that deviate significantly from the overall mean. In Figure 3b, Cluster 10, comprising 20 students, shows a marked negative deviation from the overall GPA, suggesting that these students may have lower engagement with learning resources. Similarly, in Figure 4b, Clusters 1, 4, and 9 stand out as more cohesive, with lower standard deviations (STDs) in GPA and significantly different mean GPAs. These clusters, with sizes ranging from 49 to 90 students, highlight the method's ability to differentiate student activity levels. Our proposed method consistently produces clusters with lower STDs compared to K-means, demonstrating better differentiation in student engagement and GPA outcomes. Metrics such as the RMSSTD lack an established optimal scale, indicating the need for further refinement in our approach. Future work will explore enhancements to key steps in the clustering process to address these limitations.

Overall, the analysis highlights the potential of our method for clustering students based on their interactions and academic performance. Notably, dividing a cohort of 740 students into 10 groups without imposing fixed cluster sizes yields particularly effective results. As anticipated, clusters with lower GPAs tend to exhibit lower cohesion, whereas those with above-average GPAs are more cohesive. This is likely because GPAs in the 7 to 9 range are more closely distributed, while GPAs below 7 are more dispersed. By evaluating learners' GPAs, we demonstrate the method's ability to effectively group similar students and its potential as a predictive tool for identifying student clusters based on academic performance.

7 CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed a learner clustering method based on the similarity of their access to learning resources using an SNA community detection algorithm. The design of the proposed method is to calculate the similarity between learner pairs and then build the social network of the learners. We used a community detection algorithm to build similar learner clusters, and it allowed us to visualize the student similarity and the clusters with graphs. Our proposed method is highly adjustable for different contexts with the condition of learner activity data availability. However, the evaluation of clustering algorithms is the most challenging step. This challenge encourages us to implement the proposed method outcome for a user-based collaborative recommendation of learning resources in the future. It should be noted that altering the deciding factors in the proposed method can result in more suitable clustering for different contexts or applications. For future research, we will examine different learner activity types, such as contribution to forums or time spent in the learning system, to create more precise models for clustering. We will also explore different similarity calculation methods and community detection algorithms to evaluate their effectiveness in forming learner clusters. Additionally, we plan to fine-tune the parameters and formulas to enhance clustering accuracy across various datasets.

8 ACKNOWLEDGMENTS

This work was supported by the Fundació per a la Universitat Oberta de Catalunya's (UOC) grant for doctoral thesis research.

9 ETHICS STATEMENT

The authors declare that there are no conflicts of interest related to the development of this study. Generative AI tools were used exclusively for grammar checking and improving the language of the manuscript.

10 REFERENCES

[1] N. Sghir, A. Adadi, and M. Lahmer, "Recent advances in predictive learning analytics: A decade systematic review (2012–2022)," *Educ. Inf. Technol.*, vol. 28, pp. 8299–8333, 2023. https://doi.org/10.1007/s10639-022-11536-0

[2] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student engagement level in an e-learning environment: Clustering using K-means," *Am. J. Distance Educ.*, vol. 34, no. 2, pp. 137–156, 2020. https://doi.org/10.1080/08923647.2020.1696140

[3] A. J. Martin, M. Maria, and F. Sagayaraj, "Learners classification for personalized learning experience in e-learning systems," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 4, 2021. https://doi.org/10.14569/IJACSA.2021.0120485

[4] T. Le Quy, G. Friege, and E. Ntoutsi, "A review of clustering models in educational data science toward fairness-aware learning," in *Educational Data Science: Essentials, Approaches, and Tendencies, Big Data Management*, A. Peña-Ayala, Ed., Springer, Singapore, 2023, pp. 43–94. https://doi.org/10.1007/978-981-99-0026-8_2

[5] B. Albreiki, T. Habuza, N. Palakkal, and N. Zaki, "Clustering-based knowledge graphs and entity-relation representation improves the detection of at risk students," *Educ. Inf. Technol.*, vol. 29, pp. 6791–6820, 2024. https://doi.org/10.1007/s10639-023-11938-8

[6] A. Narimani and E. Barberà, "Extracting course features and learner profiling for course recommendation systems: A comprehensive literature review," *Int. Rev. Res. Open Distrib. Learn.*, vol. 25, no. 1, pp. 197–225, 2024. https://doi.org/10.19173/irrodl.v25i1.7419

[7] E. T. Khor and K. Mutthulakshmi, "A systematic review of the role of learning analytics in supporting personalized learning," *Educ. Sci. (Basel)*, vol. 14, no. 1, p. 51, 2024. https://doi.org/10.3390/educsci14010051

[8] Ü. Avcı and E. Ergün, "Online students' LMS activities and their effect on engagement, information literacy and academic performance," *Interact. Learn. Environ.*, vol. 30, no. 1, pp. 71–84, 2019. https://doi.org/10.1080/10494820.2019.1636088

[9] R. Bhalwankar and J. Treur, "Modeling learner-controlled mental model learning processes by a second-order adaptive network model," *PLoS One*, vol. 16, no. 8, pp. 1–21, 2021. https://doi.org/10.1371/journal.pone.0255503

[10] A. Abyaa, M. Khalidi Idrissi, and S. Bennani, "Learner modelling: Systematic review of the literature from the last 5 years," *Educ. Technol. Res. Dev.*, vol. 67, pp. 1105–1143, 2019. https://doi.org/10.1007/s11423-018-09644-1

[11] X. Huang *et al.*, "Course recommendation model in academic social networks based on association rules and multi-similarity," in *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2018, pp. 277–282. https://doi.org/10.1109/CSCWD.2018.8465266

[12] J. Zhang, B. Hao, B. Chen, C. Li, H. Chen, and J. Sun, "Hierarchical reinforcement learning for course recommendation in MOOCs," *Proc. Conf. AAAI Artif. Intell.*, vol. 33, no. 1, pp. 435–442, 2019. https://doi.org/10.1609/aaai.v33i01.3301435

[13] D. Agrawal and G. Deepak, "HSIL: Hybrid semantic infused learning approach for course recommendation," in *Digital Technologies and Applications. ICDTA 2022,* in Lecture Notes in Networks and Systems, S. Motahhir and B. Bossoufi, Eds., Springer, Cham, 2022, vol. 454, pp. 417–426. https://doi.org/10.1007/978-3-031-01942-5_42

[14] A. Abu-Issa, H. Butmeh, and I. Tumar, "Modeling students' preferences and knowledge for improving educational achievements," *Front. Comput. Sci.*, vol. 6, p. 1359770, 2024. https://doi.org/10.3389/fcomp.2024.1359770

[15] L. M. da Silva *et al.*, "Learning analytics and collaborative groups of learners in distance education: A systematic mapping study," *Inform. Educ.*, vol. 21, no. 1, pp. 113–146, 2022. https://doi.org/10.15388/infedu.2022.05

[16] E. Nimy and M. Mosia, "Web-based clustering application for determining and understanding student engagement levels in virtual learning environments," *E-Journal of Humanities, Arts and Social Sciences*, vol. 4, no. 12, pp. 4–19, 2023. https://doi.org/10.38159/ehass.20234122

[17] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, K. U. Sarker, and M. U. Sattar, "Predicting student performance in higher educational institutions using video learning analytics and data mining techniques," *Appl. Sci. (Basel)*, vol. 10, no. 11, p. 3894, 2020. https://doi.org/10.3390/app10113894

[18] R. S. Bhanuse and S. Mal, "Optimal e-learning course recommendation with sentiment analysis using hybrid similarity framework," *Multimed. Tools Appl.*, vol. 83, pp. 16417–16446, 2024. https://doi.org/10.1007/s11042-023-16138-7

[19] M.-A. Abrache, A. Bendou, and C. Cherkaoui, "Clustering and combinatorial optimization based approach for learner matching in the context of peer assessment," *J. Educ. Comput. Res.*, vol. 59, no. 6, pp. 1135–1168, 2021. https://doi.org/10.1177/0735633121992411

[20] F. Safarov, A. Kutlimuratov, A. B. Abdusalomov, R. Nasimov, and Y.-I. Cho, "Deep learning recommendations of E-education based on clustering and sequence," *Electronics (Basel)*, vol. 12, no. 4, p. 809, 2023. https://doi.org/10.3390/electronics12040809

[21] P. Bedi and C. Sharma, "Community detection in social networks," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 6, pp. 115–135, 2016. https://doi.org/10.1002/widm.1178

[22] A. A. Kardan, A. Narimani, and F. Ataiefard, "A hybrid approach for thread recommendation in MOOC forums," *International Journal of Computer and Systems Engineering*, vol. 11, no. 10, pp. 2250–2256, 2017. https://zenodo.org/record/1132317/files/10007978.pdf

[23] W. Rebhi, N. Ben Yahia, and N. Bellamine, "Lifelong and multirelational community detection to support social and collaborative e-learning," *Comput. Appl. Eng. Educ.*, vol. 30, pp. 1321–1337, 2022. https://doi.org/10.1002/cae.22522

[24] E. Bolger, M. Nwobi, and M. D. Caballero, "Characterizing faculty online learning community interactions using social network analysis," *arXiv preprint arXiv:2407.00193*, 2024. https://doi.org/10.48550/arXiv.2407.00193

[25] H. Khojamli and J. Razmara, "Survey of similarity functions on neighborhood-based collaborative filtering," *Expert Syst. Appl.*, vol. 185, p. 115482, 2021. https://doi.org/10.1016/j.eswa.2021.115482

[26] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. 2008, no. 10, p. P10008, 2008. https://doi.org/10.1088/1742-5468/2008/10/P10008

[27] R. Lambiotte, J.-C. Delvenne, and M. Barahona, "Laplacian dynamics and multiscale modular structure in networks," *arXiv preprint arXiv:0812.1770*, 2008. https://doi.org/10.48550/arXiv.0812.1770

[28] M. Z. Rodriguez *et al.*, "Clustering algorithms: A comparative approach," *PLoS One*, vol. 14, no. 1, pp. 1–34, 2019. https://doi.org/10.1371/journal.pone.0210236

[29] R. Vankayalapati, K. B. Ghutugade, R. Vannapuram, and B. P. S. Prasanna, "K-means algorithm for clustering of learners performance levels using machine learning techniques," *Rev. D Intell. Artif.*, vol. 35, no. 1, pp. 99–104, 2021. https://doi.org/10.18280/ria.350112

[30] P. Rujasiri and B. Chomtee, "Comparison of clustering techniques for cluster analysis," *Agric. Nat. Resour.*, vol. 43, no. 2, pp. 378–388, 2009. https://li01.tci-thaijo.org/index.php/anres/article/view/244682

11 AUTHORS

Amir Narimani is a PhD candidate in the Education and ICT (e-learning) program at the Universitat Oberta de Catalunya (UOC) in Barcelona, Spain. His research focuses on learning analytics and educational recommender systems, exploring data-driven approaches to enhance online learning experiences (E-mail: anarimani@uoc.edu).

Elena Barberà is a Professor of Psychology and Education Sciences at the Universitat Oberta de Catalunya (UOC) in Barcelona, Spain. Her research expertise includes the assessment of online learning, electronic portfolios, online teaching strategies, and knowledge construction in virtual environments (E-mail: ebarbera@uoc.edu).