Towards the Identification of Student Learning Communities using Centrality

Intissar Salhi[1], Hanaa El Fazazi[2], Mohammed Qbadou[3], Khalifa Mansouri[4]
Laboratory Signals, Distributed Systems and Artificial Intelligence ENSETM
University Hassan II, Casablanca
Morocco

***Abstract*—Emergence of universities towards "digital university" has already been present for some years. The use of digital is largely developed to ensure a good quality of education. Universities therefore use large-scale learning management systems to manage the interaction between learners and teachers. Teachers can provide online training and educational materials for students following their classes and courses, monitor their participation and evaluate their performance. Students can use interactive features such as discussion threads, videoconferences, and discussion forums. These online tools make it possible to create new social networks or connect online social interactions. This will allow us to understand the structure of this complex network and extract useful information. In this article, we report our research on the detection of student learning communities based on learner activity. We found that it is possible to group students in communities through their messages and response structures using standard community detection algorithms. Also, that their behaviours can be strongly correlated with their closest peers who belong to the same community.**

Keywords*—Student's learning communities; complex network; learner activity; community detection

I. Introduction

Learning using educational technologies has become an integral part of modern schools[1]. Scientific and technological advances are constantly improving to ensure good quality education and facilitate the engagement of students and teachers[2]. The processes for improving educational programs and teaching principles require constant adaptation to the new conditions and capacities of modern software tools[3]. Students now complete the traditional course structure with online materials. Instructors can share class materials online, have an online discussion forum, or complete questionnaires and homework submissions online. This in turn provides a wealth of new behavioural data that we can use to group students into communities using standard community detection algorithms to create qualitative and accessible software systems that will allow teachers to constantly improve their educational approaches.

The concept of community, commonly clusters or modules, is specific to online and offline social networks [4]. A community is defined in a current graph as a group of nodes that are particularly interconnected and weakly connected to the rest of the network [5]. For example, they may be individuals who interact a lot with each other and little with others. It is particularly interesting to identify these groups in order to bring out the underlying structure of the graph. It can thus be divided into natural groups of individuals (no overlap, i.e. a node belongs to a single group) that can be of any size. This identification will take into account only the structure of the graph, as well as weights of edges that provide additional information, as the level of activity of a relationship.

Communities are interesting for a variety of reasons. For example, users in a community tend to interact frequently, share interests, and trust each other to some degree. Therefore, communities are useful, for example, to guide, identify typical profiles [6], carry out targeted actions, better adjust recommendations [7], reorganize and identify central or influential actors [8], etc.

In previous work in the educational field, in paper [9], the authors have shown that students can be grouped into stable communities according to their online question and answer model. They also showed that the students' final scores were significantly correlated with those of their peers closest to the community group. In paper [10] they have also shown that learners belonging to these communities, although homogeneous in terms of performance, are not united by their incoming motivations to register for the course nor by their level of prior experience.

Until today, these results have only been found in MOOCs and the user forum, where almost all of the relevant interactions in the course occur online and where the relationship between students is the direct connection between each other's. In this paper, we have shown that interaction on forums is not the only way to establish a relationship between students in order to create communities, but that their activity can also establish it. Thus, we found that it is possible to group students in communities through their messages and their response structures using standard community detection algorithms. Furthermore, that their behaviours can be strongly correlated with their closest peers who belong to the same community.

In this article, we start with a review of the literature on community detection through social network analysis, then we explain the notions of centrality as well as the most used community detection techniques. After that, we continue with a discussion about the algorithms we used to create a student community. The resulting student community will be tested on a database and discussed. And we conclude with some recommendations for future studies.

II. Community Detection

Community detection has received a lot of attention in the research community over the last decade and several approaches have been proposed [11]. However, the majority of existing approaches mainly deal with social and biological networks, viral marketing, [12][13][14][15] etc.

In the pedagogical field, and in a set of students, the detection of subsets of vertices more densely connected than others, called student learning communities, is a problem that we find strongly as it can be beneficial to us. To control all students enrolled in an institution and even within a class and given an idea of the behaviour of each based on the descriptive characteristics of the community to which they belongs. Within a class, these communities play an important role in his organization and structure.

As a result, it is necessary to determine classes in a graph. This problem is therefore strongly related to the problem of partitioning, with the following specificity: according to the application that we want to do with these communities, classes can (or must) be disjointed or not. So we can analyse network interaction between learners to, among other things, predict, quite reliably, a list of recommendations. Just as in clustering, there are many individuals who belong to more than one community, and in this case it is reasonable to build not a partition, but a collection, that is, a system of overlapping classes. It is the same in social networks, [16] where individuals can belong to several groups.

A. Social Network Analysis

The social graph refers to the mapping of relationships within a social network [17]. Nodes are usually the interacting social actors and the links are the relationships between them. The social graph in its simplest form is modelled to form an analysable structure where all the significant links between the nodes are studied. The same goes for structural holes vertices [18], or "network closures" where there is an absence of direct links between two.

B. Presentation of a Social Network

A social network can be represented by a graph G (V, E) where V represents the set of vertices (nodes), E the set of edges and can be represented using the so-called adjacency matrix A_{ij} which indicates the connections between the nodes.

III. Notion of Centrality

In social network analysis, centrality [19] is an important concept which can be applied to all kinds of networks. The identification of the actors with the greatest centrality (leaders or influential person) makes it possible to define the structure of the network [20][21],more precisely, these actors should normally play a key role in the simulated and real behaviours.

In this part we will talk about the notion of centrality within a network.

A. Identification of Central Nodes

The importance of a vertex in a graph can be quantified simply by its neighbourhood, and it is said that a node is central if it has many neighbours [19], here we talk about the degree of centrality; it may be in terms of distance. A central node may be distant from others, so we talk about centrality closeness [22]; or more subtly, it constitutes a node of passage by the shortest way to transit from one summit to another, this is explained by centrality betweenness [23][24]. In the following parts, we will highlight these three algorithms:

1) Degree centrality: The Degree Centrality measure can help us find popular nodes in a graph. Indeed, it is the ratio between the number of outcoming links and the maximum degree possible in a network of a possible size. Thus, for a node called i and a total number of nodes n in the network:

$$D_c(i) = \frac{d_s(i)}{(n-1)} \quad (1)$$

The degree centrality reflects only a local view of the relationships between nodes in a network and does not provide information about the overall structure of the network.

2) Closeness centrality: This is the most widely used measure of centrality. The centrality of proximity of the actor i is defined as the inverse of the average degree d (i, j):

$$C_c(i) = \frac{(n-1)}{\sum_{i \neq j} d(i,j)} \quad (2)$$

If the node i has a strongest value c_c it implies that i is a central node.

The multiplication by {n-1}, where n is the number of nodes in the graph. This adjustment allows comparisons between nodes of graphs of different sizes.

3) Betweenness centrality: Intermediary is the measure of centrality of a vertex in a graph. Intermediate centrality counts the number of times a node acts as a waypoint along the shortest path between two other nodes (geodesic distance).

It is based on the counting of the geodesic distance δ_{ij} between the actors i and j, and by looking at the number δ_{ij}(m) passing through the actor m.

$$C_b(m) = \frac{2}{(n-1)(n-2)} \sum_{i \neq m} \sum_{i<j \neq m} \left(\frac{\delta_{ij}(m)}{\delta_{ij}}\right) \quad (3)$$

The larger C_b is, the higher the vertex is central since it is located at the crossroads.

The betweenness may be normalized by dividing through the number of pairs of vertices not including v, which for directed graphs is (n-1)(n-2) and for undirected graphs is (n-1)(n-2)/2.

B. Individual Relay

Problems often arise when people no longer talk to each other or interact with one another. In this case some individuals can act as relays between the two nodes[25]. These individuals belong to the shortest path between the two people.

To form groups and partition the network into disjoint sets, we must consider connections between the nodes globally. This phase is based on: the measurement of the similarity between the nodes [26], latent space model [27], approximation of the block model [28]… etc.

In this study we are interested in calculating the similarity between individuals, which is defined by the similarity of their interaction models. And we say that two nodes are structurally equivalent if they are connected to the same set of actors. The similarity in the graphs is defined in terms of neighbourhood, that means that two vertices are close (similar) if there is a strong overlap between their neighbourhoods. And it can be calculated by several approaches, namely, Jaccard coefficient [29] and cosine similarity [30].

1) Jaccard coefficient: The Jaccard coefficient is used in statistics to compare the similarity and the diversity between individuals who belong to given samples. Discovery of communities. It is the relationship between the cardinal of the intersection of the edges N connecting a node to the other nodes and the cardinal of the same edges. Let two nodes i and j, the Jaccard coefficient is as indicated below:

$$Jaccard(i,j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|} \quad (4)$$

2) Cosine similarity: The cosine similarity or cosine measure makes it possible to calculate the similarity between two nodes by determining the cosine of the angle between them. The cosine similarity between two nodes i and j is the number of common neighbours divided by the geometric mean of their degrees. This value oscillates between 0 and 1. The value 1 indicates that the two vertices have exactly the same neighbourhood, while the value 0 means that they have no neighbours in common. The cosine similarity is technically indefinite if one or both vertices have a degree of 0, but according to the convention it is said that the cosine similarity is 0, in these cases.

The cosine similarity is calculated according to formula (5) below:

$$Cosine(i,j) = \frac{|N_i \cap N_j|}{\sqrt{|N_i| . |N_j|}} \quad (5)$$

IV. Community Discovery

The discovery of communities within a network is done using several approaches, as follow:

Agglomerative Approche (hierarchical cluster analysis)

The hierarchical cluster analysis (HCA) is an iterative classification method [31].

In the case of graphs we define firstly similarity between vertices based on the adjacency matrix, and we can chain with a HCA. is the pseudocode of the HCA algorithm.

A. Divisive Approche (Girvan–Newman Algorithm)

The importance of the connection between two vertices can be materialized by the "edge betweenness"[32]. It indicates the frequency with which it is borrowed when considering the shortest path between each pair of nodes.

$$E_b(m) = \sum_i \sum_{i>j} \frac{\delta_{ij}(m)}{\delta_{ij}} \quad (6)$$

Algorithm 1: The hierarchical ascending classification

```
given a dataset (d1, d2, d3, ....dN) of size N
# compute the distance matrix
for i=1 to N:
  # as the distance matrix is symmetric about
  # the primary diagonal so we compute only lower
  # part of the primary diagonal
  for j=1 to i:
    dis_mat[i][j] = distance[di, dj]
each data point is a singleton cluster
repeat
  merge the two cluster having minimum distance
  update the distance matrix
until only a single cluster remains
```

The higher the value, the more important the connection is, because it establishes a "bridge" between groups of vertices.

The divisive approach consist to iteratively remove connections with the highest values of edge betweenness. Here is the pseudocode of the algorithm:

Algorithm 2: Girvan–Newman algorithm

1. For a given graph G = (V, E), carry out the following steps for each pair of vertices in the same component:
2. For a given pair of vertices u, v assign one unit of flow in total.
3. Find the number, k(u, v), of shortest paths from u to v.
4. Assign 1/k(u, v) units of the flow to each shortest path from u to v.
5. For each shortest (u, v)-path, record the edges in the path. After all this is finished:
6. For each edge e ∈ E count up how much flow goes through the edge e adding over all shortest paths between all pairs of vertices u, v ∈ V which use the edge e.

V. New Application: In the Educational Field

The detection of communities in an educational network aims to identify groups of learners maintaining a special relationship, so they have the same level of skills. In addition, the identification of the most influential person (Leader) may be beneficial to the teacher (learning agent) because it will facilitate his interaction with his students. In addition, the identification of the most influential person (the leader) can be beneficial for teachers (learning agents) because it will facilitate their interaction with their students. This data analytics approach will allow them to pinpoint their students, including, the behaviour, the performance and the student satisfaction in courses, it will allow them too, the prediction of the level of learners and their skills focusing only the leader, in order to enhance the learning experience by providing informed advice and optimizing learning materials and then give a list of possible recommendations.

This theme is experiencing a resurgence of interest in recent years with the development of social media (like LinkedIn, Twitter, Facebook, E-learning platform forums, etc.), multiplying the opportunities for interaction between

individuals. A community is a group of nodes with a high density of connections. This article is about to show that social network analysis techniques and community detection can be used in other areas, namely, the educational field, to control a given set of students.

We consider a particular situation where the graph is undirected, the connections between people, if they exist, are symmetrical and unweighted that is to say that the connections have the same intensity.

A. Study Case

The anonymised Students' Academic Performance Dataset [33][34] is an educational dataset which is collected from a learning management system (LMS) called Kalboard 360. Kalboard 360 is a multi-agent LMS, designed to facilitate learning through the use of leading-edge technology. Such system provides users with a synchronous access to educational resources from any device with Internet connection.

The data is collected using the experience API (xAPI), which is a learner activity tracker tool. This component is a part of the training and learning architecture (TLA) that enables to monitor learning progress and learner's actions such as reading an article and watching a training video and all activities and objects describing the learning experience.

The dataset consists of 480 student records and 16 features, between these features we choose 4 that we judged essentials to describe the behaviour of such a student. That are:

- Raised hand: define how many times the student raises his/her hand on classroom.
- Visited resources: define how many times the student visits a course.
- Viewing announcements: define how many times the student checks the new announcements.
- Discussion groups: define how many times the student participate on discussion groups.

And in order to test our program and see how it can identify communities. At first, we chose to select 20 random instances.

B. Construction of the Network

The network must be described by a symmetric Boolean adjacency matrix (values 1/0) indicating the privileged relationships (or not) maintained by the learners. To this end, we must use the database we have available to construct this adjacency matrix that will be used to build the graph and then define the leaders and determine the communities.

The adjacency matrix is calculated firstly by the cosines similarity and as a result we had a cosines similarity matrix between each instance and the others. And secondly by transforming this matrix on a Boolean one by replacing values by 0 or 1 compared on a threshold. Here the threshold is set at 0.89. Finally, we construct the graph represented in Fig. 1. Where each node in our social network represents an individual participant in the class. And the relationships between the participants are presented as arcs. We define a relationship between them by a higher degree of similarity.

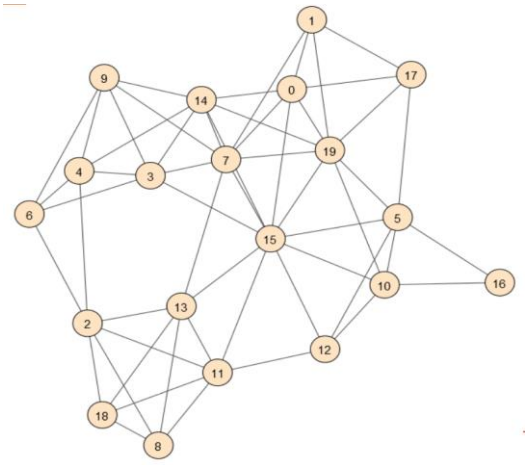

Fig. 1. The Graphical Representation of 20 Instances Chosen Randomly.

C. Identification of Central Nodes

After the calculation centrality, the program have as results the three Tables I, II and III, representing respectively the degree centrality, closeness centrality and betweenness centrality.

Reading is easier if we display the most important values in descending order by associating them with the names of the nodes.

In Table I, the vertices 15, 19, 7, 14 and 13 are highlighted. That means that these are candidate points to be central points.

Same for Table II, vertices 15, 19, 7, 14 and 13 are highlighted. Except that summit 7 is second this time.

In Table III, we observe a certain coherence with the previous results.

Obviously, individuals 15 and 7 are at the centre of relationships between group members. Then, the program choose these nodes whose highlights in the graph using appropriate colours. As shown in the following Fig. 2.

TABLE. I. TOP 5 FIRST VALUES OF THE DEGREE CENTRALITY IN DESCENDING ORDER

Node name	***degree***
15	10
19	8
7	8
14	7
13	6

TABLE. II. TOP 5 FIRST VALUES OF THE CLOSENESS CENTRALITY IN DESCENDING ORDER

Node name	***closeness***
15	0.678571
7	0.612903
19	0.558824
14	0.558824
13	0.558824

TABLE. III. TOP 5 FIRST VALUES OF THE BETWEENNESS CENTRALITY IN DESCENDING ORDER

Node name	*Betweenness*
15	49.775000
7	22.591667
13	19.250000
11	15.700000
19	15.075000

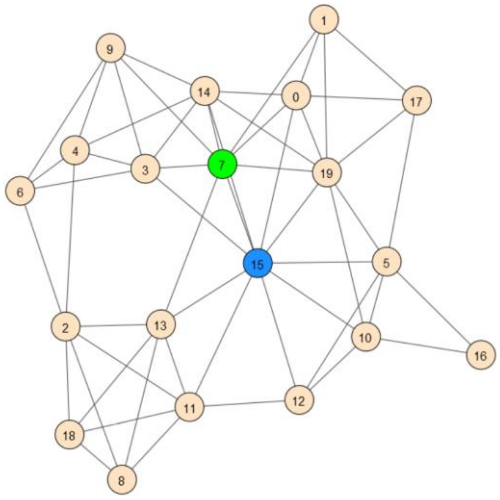

Fig. 2. The Graphical Representation with Highlighted Central Nodes.

D. Discovering Communities

We are interested in a divisive approach based on the notion of "edge betweenness". We obtain a net partition (crisp), i.e. an individual belongs to one and only one group.

In a possible partition into two groups, we note that the troublemakers' individuals 7 and 15 will actually be separated as shown in the resulting dendrogram of the hierarchical ascending classification (Fig. 3) and so this appears explicitly in the graph (Fig. 4) where both communities appear clearly.

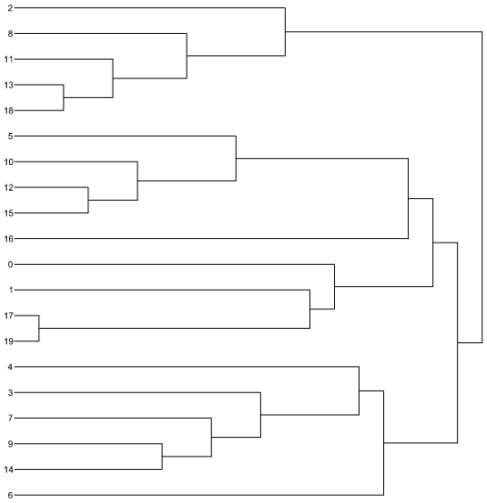

Fig. 3. The Resulting Dendrogram of the Hierarchical Ascending Classification.

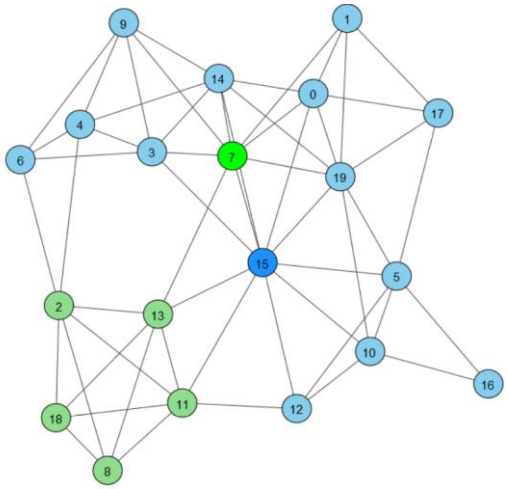

Fig. 4. The Graphical Representation with Highlighted Two Communities.

We performed the Girvan Newman clustering and the resulting clusters can be seen in Fig. 4. In this graph, nodes with dark colours represent leaders of their communities. And moreover, they are the students who can be traced to get a global idea on the other students belonging to the same community.

Community users tend to interact frequently, share interests, and trust each other to some extent. Therefore, our method will help teachers to reduce the efforts made to manage a very large number of students, to a minimum effort. In other words, to manage their students, teachers only have to observe the central players who constitute all the influential students to carry out targeted actions, organize the structure of the class, guide the learners, and adjust the recommendations, for all students belonging to the same communities.

VI. CONCLUSION

With recent technological advances, huge amounts of data are accumulating at a frantic pace in various areas of human activity, namely, the learning activity. Understanding both the universal and specific characteristics of the networks associated with this data has become a real and important task. Knowing the structure of the community makes it possible to predict certain essential characteristics of the systems under study. For example, with our approach, it is possible to discover student learning communities in the pedagogical system. We then provide a tool based on the notion of centrality and standard community detection algorithms for interpreting the local organization of a student network within a learning management system, which can be used to identify standard profiles, perform targeted actions, and better adjust recommendations. For our future work, we want to spread our study on our university students in order to integrate it in the module of recommendation of pedagogical resources of a learning management system.

REFERENCES

[1] R. Raja and P. C. Nagasubramani, "Impact of modern technology in education," J. Appl. Adv. Res., vol. 3, no. S1, p. 33, May 2018.

[2] P. Serdyukov, "Innovation in education: what works, what doesn't, and what to do about it?," J. Res. Innov. Teach. Learn., vol. 10, no. 1, pp. 4–33, Apr. 2017.

[3] W. D. Haddad and A. Draxler, "The dynamics of technologies for education," Acad. Educ. Dev., pp. 2–17, 2002.

[4] P. Bedi and C. Sharma, "Community detection in social networks," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 6, no. 3, pp. 115–135, May 2016.

[5] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Paris, "Defining and identifying communities in networks," Proc. Natl. Acad. Sci. U. S. A., vol. 101, no. 9, pp. 2658–2663, 2004.

[6] H. Cai, V. W. Zheng, F. Zhu, K. C. C. Chang, and Z. Huang, "From community detection to community profiling," Proc. VLDB Endow., vol. 10, no. 7, pp. 817–828, 2017.

[7] F. Gasparetti, A. Micarelli, and G. Sansonetti, "Community Detection and Recommender Systems," in Encyclopedia of Social Network Analysis and Mining, New York, NY: Springer New York, 2017, pp. 1–14.

[8] R. Missaoui and I. Sarr, Eds., Social Network Analysis - Community Detection and Evolution. Cham: Springer International Publishing, 2014.

[9] R. Brown et al., "Good Communities and Bad Communities: Does membership affect performance?"

[10] R. Brown et al., "Communities of performance & communities of preference." CEUR-WS, 2015.

[11] B. S. Khan and M. A. Niazi, "Network Community Detection: A Review and Visual Survey," 2017.

[12] M. Planti and M. Crampes, "Survey on Social Community Detection To cite this version : Survey on Social Community Detection," pp. 65–85, 2013.

[13] A. Dhumal and P. Kamde, "Survey on Community Detection in Online Social Networks," Int. J. Comput. Appl., vol. 121, no. 9, pp. 35–41, 2015.

[14] G. Jia et al., "Community Detection in Social and Biological Networks Using Differential Evolution," Springer, Berlin, Heidelberg, 2012, pp. 71–85.

[15] S. Combéfis, "Viral marketing and Community detection algorithms," p. 89, 2007.

[16] Q. Wang, "Overlapping community detection in dynamic networks," 2012.

[17] C. P. Diehl, G. Namata, and L. Getoor, "Relationship Identification for Social Network Discovery," AAAI Work. - Tech. Rep., vol. WS-08-04, pp. 9–14, 2008.

[18] M. Gargiulo and M. Benassi, "Trapped in Your Own Net? Network Cohesion, Structural Holes, and the Adaptation of Social Capital," Organ. Sci., vol. 11, no. 2, pp. 183–196, 2000.

[19] L. C. Freeman, "Centrality in Social Networks Conceptual Clarification," Soc. Networks, vol. 1, no. 1968, pp. 215–239, 1978.

[20] S. P. Borgatti, "Centrality and network flow," Soc. Networks, vol. 27, no. 1, pp. 55–71, Jan. 2005.

[21] S. P. Borgatti and M. G. Everett, "A Graph-theoretic perspective on centrality," Soc. Networks, vol. 28, no. 4, pp. 466–484, Oct. 2006.

[22] G. Sabidussi, "The centrality index of a graph," Psychometrika, vol. 31, no. 4, pp. 581–603, Dec. 1966.

[23] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness," Sociometry, vol. 40, no. 1, p. 35, Mar. 1977.

[24] U. Brandes, "A faster algorithm for betweenness centrality*," J. Math. Sociol., vol. 25, no. 2, pp. 163–177, Jun. 2001.

[25] Z. Lu, X. Sun, Y. Wen, G. Cao, and T. La Porta, "Algorithms and Applications for Community Detection in Weighted Networks," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 11, pp. 2916–2926, 2015.

[26] A. Rawashdeh and A. L. Ralescu, "Similarity measure for social networks-a brief survey," CEUR Workshop Proc., vol. 1353, pp. 153–159, 2015.

[27] D. K. Sewell and Y. Chen, "Latent Space Approaches to Community Detection in Dynamic Networks," Bayesian Anal., vol. 12, no. 2, pp. 351–377, Jun. 2017.

[28] E. Abbe, "Community detection and stochastic block models: Recent developments," J. Mach. Learn. Res., vol. 18, pp. 1–86, 2018.

[29] B. Bank, Jacob and Cole, "Calculating the jaccard similarity coefficient with map reduce for entity pairs in wikipedia," Wikipedia Similarity Team, 2008.

[30] S. Tariq, M. Saleem, and M. Shahbaz, "User Similarity Determination in Social Networks," Technologies, vol. 7, no. 2, p. 36, 2019.

[31] E. Cuvelier et al., "Graph Mining and Communities Detection," 2012.

[32] M. Arasteh and S. Alizadeh, "A fast divisive community detection algorithm based on edge degree betweenness centrality," Appl. Intell., vol. 49, no. 2, pp. 689–702, Feb. 2019.

[33] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," Int. J. Database Theory Appl., vol. 9, no. 8, pp. 119–136, Aug. 2016.

[34] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," in 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, AEECT 2015, 2015.