Decoding Student Relationships Based on Academic Performance using Social Network Analysis in Online Education Paradigm

ASHISH LEDALLA, B18EE008
NISHANT SARRAFF, B18CSE064

Abstract: With COVID-19, many academic institutions are rapidly changing the basic way they do their work. The COVID-19 pandemic forced academic institutions worldwide to incorporate the online education paradigm to execute academic activities without any hindrance. This makes it difficult for students in terms of interaction and thus affects the quality of education. Because of online mode, students have become more isolated and less socially active which is also affecting their mentality. And since, there is no traditional classroom setting, teachers are not able to understand the changes in students, their abilities, their pace, their behavior, and grasp on the courses. From the perspective of the faculty, the student community is like a black box and creates a problem for the proper evaluation of students. And this is the reason that we need some better understanding about the students. For example, observing the behavior of a particular student or a group of students, finding similar-minded students or groups of students, etc. In this work, we try to analyze this situation using the tools and methods of SNA. The SNA studies the interactions and connections among entities in a network based on some relationship. We employ correlation-based similarity measures using the academic performance of students to generate a network of students.

Keywords: *Correlation, Online Education, Academic Performance, Social Network*

1 INTRODUCTION

The COVID-19 has resulted in the shutting down of all the schools and institutions all over the world. Many schools and institutions are rapidly changing their basic way of educating the students. As a result, education has changed dramatically, with the rapid rise of online learning, whereby teaching is undertaken remotely and on online platforms. Research suggests that since online learning has been to increase retention of information and take less time, many institutions decided to start educating students only through online mode. According to several studies, students who learn online retain 25-60% more content than those who learn in a classroom retain only 8-10%. This is mostly due to students' ability to learn more quickly online; e-learning takes 40-60% less time than traditional classroom learning since students can learn at their own pace, going back and re-reading, skipping, or accelerating through ideas as they see fit. But the issue with online education is that it is weakening the relationship among students and teachers.

The sudden shift in the way of teaching and learning has resulted in a lack of motivation among students and the faculty. Students do not get a chance to participate in academic activities in a full-fledged manner. This results in poor education quality. On the other hand, faculty members lack experience in online teaching and thus it makes it difficult for them to assess the students and provide better learning experiences for the students. The traditional way of offline face-to-face education methods is difficult to replicate in a remote education system where everyone is isolated. Moreover, some institutions are starting to complete online programs to facilitate accessibility and flexibility for the students and the importance of online learning will only increase in the future. Thus it is very crucial to come up with methods and systems that can demonstrate how students are coping up with the shift.

For this project, we will try to analyze this situation using the tools and methods of SNA. The SNA studies the interactions and connections among entities in a network based on some relationship. To analyze the behaviour of a student or a group of students, and decode the student's communities for a particular course, we will be using a correlation-based similarity measure, which is one of the SNA measures, to form a network. Such type of analysis network is also known as Correlated-based Network Analysis (CNA). Correlation-based network analysis is based on mathematically defined (dis)similarity measures that correlate different components to each other and the resulting correlation coefficients reflect the magnitude of the co-linear relationship of the components. Furthermore, we will be using the centrality measures to understand the properties of a particular node or a group of nodes. Some measures include common neighbor, number of degrees, betweeness, etc.

Social network analysis is proving to be useful in characterizing interaction among the actors forming the network. For this specific purpose, we are going to use the academic performance of students to form a relationship between students. To characterize this relationship, we employ statistical similarity measures such as Spearman or Pearson correlation coefficients. Pairwise correlation coefficients of all the students' academic performance will let us generate a correlation-based network which can then be worked upon using the tools of SNA and graph theory. Thus, the actors in the network will be the students and the connections between them will be defined by the correlation coefficient reflecting the correlation in the performance of students in different quizzes and assignments. We intend to utilize this network to better understand the student communities based on the network topology. This will let us establish methodologies to improve the remote education system. Applications can include identifying like-minded students, trying to prevent possible cheating in online exams, improving the evaluation schemes, temporal developments in the network etc.

2 PROBLEM STATEMENT

2.1 Relationship between the Survey network and the correlation network

The network obtained using the survey data is based on the student relationships in academic interactions. The data collected through the survey is used in building the network of students. Each student represents a node in the network, and the edges between the nodes represent the academic interaction as reported by the students.

The correlation network is built using the academic performances of the students. Pairwise Pearson correlation is taken, and the correlation coefficient is taken as the weight quantifying the relationships in the network. To compare these, we first consider degree distribution in the survey network and the strength distribution, which is equivalent to degree distribution in weighted networks. Then we also assess each student's degree vs the strength of each student in the correlation network.

2.2 Experimenting ICM on the Survey Network

Utilizing the independent Cascade Model for information diffusion, we experimentally try to understand to what extent the academic information (such as class notes) can circulate within the network formed using the survey network. This is compared with the hard thresholded correlation network.

2.3 Temporal changes in the Correlation Network formation

Correlation-based network analysis is based on mathematically defined (dis)similarity measures that correlate different components to each other. In our project, a correlation network is obtained using the marks distribution of each student within a semester. And using this network, we can observe how many students think similarly for a given exam. And since students' minds change over time, we have compared the correlation network formations over time to observe students' behavior using different SNA tools, like community detection and similarity measurement between clusters, structural analysis, etc. In this way, it may help the professor understand the students' communities, understand how many students' thought processes are the same, and understand how they can cope with the course.

2.4 Identifying the unique minded students through Correlation-Network Analysis

We used the correlation network over the whole input data to identify the unique students whose thought processes differ from other students. This will result in a network to figure out the unique students whose thinking process is different from others, who didn't think the same way and didn't get similar marks.

3 RELATED WORK

Correlation-based network analysis (CNA) has grown in popularity as a data-mining approach for visualising and evaluating biological correlations in big data sets over the last decade. Molecular elements (e.g., metabolites or genes) and their correlation coefficient (strength and sign) are represented by vertices and edges in this sort of network. Correlation analysis inferred edges represent a coordinated behaviour between vertices across the data collection (treatments, genotypes, conditions, and time). The type of correlation must be chosen based on the data's parametrical distribution. Data must be checked for normalcy in large population studies using known techniques, such as the Shapiro-Wilk test. Pearson correlation should be used on normally distributed data, while Spearman's rank correlation should be used on data that does not follow the normal distribution assumption. CNA has been successfully applied to a variety of biological systems, revealing metabolic markers related to plant growth and biomass in Arabidopsis thaliana recombinant inbred lines (RIL) and introgression lines (IL), the role of gene Col5a2 in myocardial infarction, the effect of hypoxia on tumour cell biochemistry, and, most recently, the discovery of a genetically based mechanism of amino acid metabolism regulation.

4 CONTRIBUTION

To obtain the survey network data, we created a google form and shared it with all our batch mates, asking a student about their group with whom he studied together during the course. Since many students didn't fill-up the form, hence after getting the student relationships in raw format, we filtered raw data to useful data by removing null entries and created a network, where nodes represent the students and edges represent the existing relationship between students. Similarly, to obtain the correlation network, we ask one of professor to share each student's academic performance within our branch. After getting the information, we created the correlation network, where the inputs were the marks of each student in quizzes, assignments, mid-term, and end-semester examinations.

To find the relationship between the survey network and the correlation network, since the survey network was an undirected graph and the correlation network was a weighted graph, we considered the degree distribution in the survey network and the strength distribution, which is equivalent to degree distribution in weighted networks. After that, we compare each student's degree distribution and strength value to get the relation between the surrey network and the correlation network.

To experiment with the Information Diffusion Model, we use the survey network as an input. We assumed that active nodes will be the students who regularly attend the classes (since the students who take notes or attend class regularly have more information than other students, hence we refer to those students as seeds). We also assumed diffusion probability would be a random value close to 0.

We considered the correlation network to observe the temporal changes to figure out how much a student's mind changes over time. This is done by partitioning the data into two parts (since the course length was around 2.25 months, we considered quizzes and exams that happened within the first month and other exams within the second month). After dividing the data, we used correlation to build an undirected network for both inputs, where the edge represents that the correlation value is greater than 0.8 between the nodes. Then we used the community detection algorithm over the graph, on both partitions, to figure out the same mindset students, where one community, within a network, will represent a group of students having a similar mindset. After that, we took two clusters from both partitions having one node(student) common and used the Jaccard coefficient to figure out the temporal changes of that node(student).

After that, we took the whole mark distribution as input to find the unique group of students whose mindsets differed from other students. And then, we built the correlation model, where a node represents the student, and an edge represents that the correlation value between endpoint nodes is greater than 0.8. After building the model, we filtered the nodes(students) having connections less than 2, which means they are less likely to be correlated with other students.

5 EXPERIMENTS AND RESULTS

5.1 Data Collection

To obtain the survey network data, we created a google form and shared it with all our batchmates, asking a student about their group with whom he studied together during the course. After collecting data, we got the table in following format:

Table 1. Study Group Information

Email Address	Mark the Students you have Studied with Prior to the Pandemic When We were on Campus
email1	name1, name2, name3, ...
email2	name6, name19, name33, ...
email3	name23, name34, name12, ...

And to obtain the correlation network, we ask one of professor to share each student's academic performance within our branch. After getting the marks distribution, we got the table in following format:

Table 2. Marks Distribution Information

Roll Number	Email	Name	Quiz1	Quiz2	Quiz3	Quiz4	Quiz5	Mid term	End term	Assignment1	Assignment2
roll1	email1	name1	q11	q12	q13	q14	q15	m1	e1	a11	a12
roll2	email2	name2	q21	q22	q23	q24	q25	m2	e2	a21	a22
roll3	email3	name3	q31	q32	q33	q34	q35	m3	e3	a31	a32

5.2 Data Preprocessing

The collected data through the survey (google forms) is in the form of a .csv file, as shown in table 1, where the respondent's email is recorded against the names given by each respondent. Data was in an adjacency list format. To convert this into a graph, we will first need to assign every name an id, and the same id is allotted to the student who used that particular email. Once this is done, the data will be in a Map/Dictionary (key, value) where the key is the student allotted unique id and value is a list of ids corresponding to the student's connections.

For the marks distribution data, the data was already in the format of a simple table where all the scores of the students are present. The rows represent the student, and the columns are the names of different tests conducted. Depending on the task at hand, we take selected columns for all the students.

5.3 Experimental Setup

5.3.1 Relationship between the Survey network and the correlation network.

(1) We first try to understand to what extent the graphs are similar using a graph distance metric called DeltaCon[6].
(2) DeltaCon uses the similarity matrix of a given graph.
(3) For a given two graphs, we first calculate the adjacency matrix and diagonal degree matrix; using them, we obtain the similarity matrix.
(4) Once we obtain the similarity matrices for each graph, we use them to calculate the Matusita distance[4].
(5) Degree distributions/ strength distributions are also compared to evaluate the topology. (strength is taken for the weighted correlation graph)

5.3.2 *Experimenting ICM on the Survey Network.*

(1) We take a single node and simulate information diffusion[1] using the ICM model[9] for 100000 Monte Carlo[7] runs using the survey network.
(2) First, we assign a probability for each edge which will be the probability of information propagation.
(3) This probability value is a hyperparameter and can be tweaked to experiment with.
(4) We then choose a single node as a seed set.'
(5) This seed is used in ICM to estimate the number of activated nodes.
(6) Similarly, we do the simulations on the correlation network and infer the difference of the estimates.

5.3.3 *Temporal changes in the Correlation Network formation.*

(1) We will divide our filtered marks distribution data into two partitions. One part includes Quiz1, Quiz2, Quiz3, and Mid-term marks, while the second part includes marks distribution of Quiz4, Quiz5, and End-Sem marks.
(2) Then we will pass both the inputs to the correlation function, giving an NxN matrix, defining the correlation value between each pair of nodes.
(3) After getting the matrix, we will filter the matrix by considering the threshold value as 0.8. It will give us a covariance matrix, where a value less than 0.8 will be considered 0, and a value greater than 0.8 will be considered 1.
(4) Now, we have a covariance matrix with values 0 and 1. Using this matrix, we will plot our indirect graph where nodes will represent student and edge will represent that the covariance value between the two-node is greater than 0.8.
(5) After getting the two networks, on both inputs, let's say G1 and G2, we will pass both the graph to the community detection algorithm, which uses the *naive_greedy_modularity_communities()* [8] algorithm to find the number of communities in G1 and G2.
(6) After getting the list of communities, we will pick two communities, one from G1 and one from G2, where one node is common. And for that node, we will calculate the *Jaccard coefficient* [5], a similarity measure function that measures the similarity of given two communities.
(7) In this way, we can plot and observe temporal changes for all the nodes.

5.3.4 *Identifying the unique minded students through Correlation-Network Analysis.*

(1) We will take full marks distribution as input to figure out the uniques nodes in the network.
(2) Then we will pass the inputs to the correlation function, giving an NxN matrix, defining the correlation value between each pair of nodes.
(3) After getting the matrix, we will filter the matrix by considering the threshold value as 0.8. It will give us a covariance matrix, where a value less than 0.8 will be considered 0, and a value greater than 0.8 will be considered 1.

(4) Now, we have a covariance matrix with values 0 and 1. Using this matrix, we will plot our indirect graph where nodes will represent student and edge will represent that the covariance value between the two-node is greater than 0.8.

(5) After getting the network, we will make a list of nodes with a degree less than 2, i,.e, their correlation with others is very low. In this way, we will get those nodes which are unique in nature.

5.4 Results

5.4.1 Relationship between the Survey network and the correlation network : We observed that the Deltacon[6] distance between the survey network and the correlation networks gets minimized around the correlation coefficient 0.6 to 0.8. This can be observed in the Fig 3. The degree distributions are very different as we can see in the Fig 2. For the survey network we can observe that it follows scale-free property while the strength distribution does not. This shows that the actual relations are very different from the correlations in the academic performances.

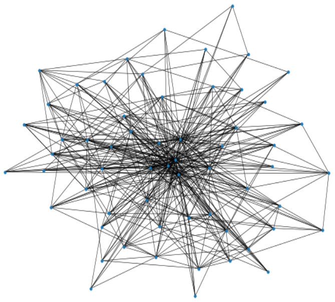

Fig. 1. Visualization of Survey Network

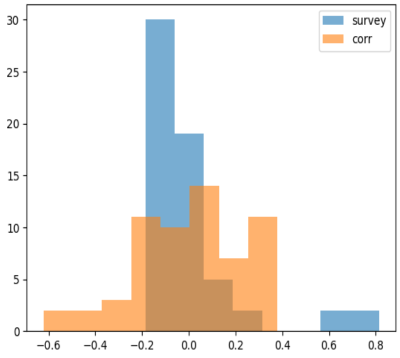

Fig. 2. Degree Vs Strength Distributions

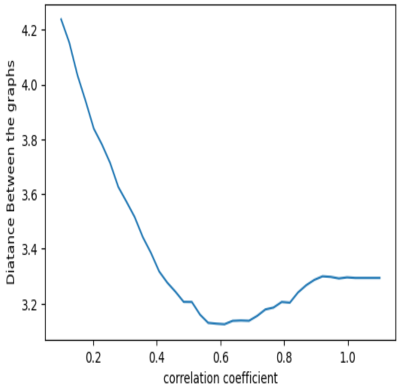

Fig. 3. Distance between the survey graph and correlation graph for a given coefficient value

5.4.2 Information Diffusion through ICM :. Using the survey network and a single node as a seed set, the average activations are around 20. In contrast, for the same seed and same diffusion probability, the activation was 26 in the case of the correlation network with a threshold of 0.8. This shows that even if the distance between the survey graph and correlation graph is very large, at a correlation threshold of 0.8, the correlation network is trying its best to mimic the survey network. It is the reason that we will be using 0.8 as our threshold value for further applications.

5.4.3 Temporal changes in the Correlation Network formation : After creating the correlation network and passing to the community detection algorithm, we got the result that for the first month, the **number of clusters** were **5** but in the case of the second month, the **number of clusters** are **2**. This shows that the size of the student community is increasing, however the number of communities is decreasing over time.

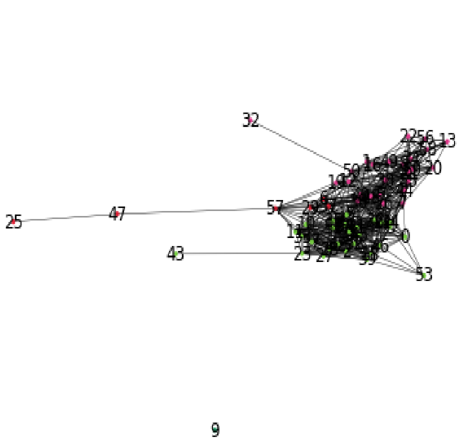

Fig. 4. Visualization of Correlation-Network Graph for first half Marks-Distribution data with threshold value as 0.8

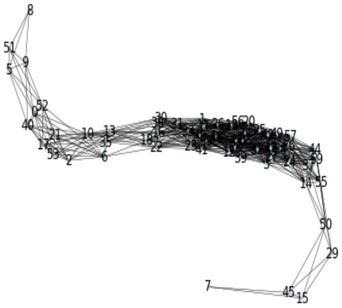

Fig. 5. Visualization of Correlation-Network Graph for second half Marks-Distribution data with threshold value as 0.8

As we can observe from Fig4 and Fig5, how the community shape is changing over time.

5.4.4 Identifying the unique minded students through Correlation-Network Analysis: After creating the correlation network for the whole mark distribution and printing the nodes having degree less than 2, we will get the nodes which are unique in nature. And the best part about this is that we legit got those nodes, which are either isolated or which are highly talented in our branch.

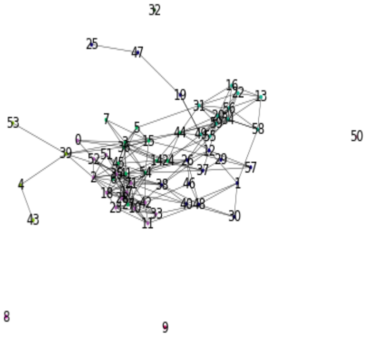

Fig. 6. Visualization of Correlation-Network Graph for whole Marks-Distribution data with threshold value as 0.8

6 CONCLUSION AND FUTURE WORK

Correlation networks are a bridge between traditional data science (where the data is in the simple format of tables) and Network science. Correlation-based network analysis provides us with the tools of SNA, which cannot be simply applied on data that is trivially non-network data. In this project, we used students' academic performances in different tests to build the correlation network. Once we have the network at hand, using thresholding (based on domain knowledge), we convert it into a social network of students based on the pairwise correlation coefficients. This provides us with a better understanding of students in an online class. Using CNA as the base framework, we extend it for the analysis of students through traditional SNA.

Online education has isolated students, and there is hardly any interaction between students. We conclude that this framework can provide us with essential insights such as which students are very different in terms of academic performance. The survey network shows the actual academic interactions among students. We experimented to what extent this network relates to the correlation network.

We wish to incorporate the negative correlation and utilize the network as a signed network for further analysis for future work. We saw that there are around 30% of edges are of negative sign. This can give us more insights for better understanding the relationships.

ACKNOWLEDGEMENT

We would like to thank the class of EE'18 for taking part in the survey and providing us with required information. Dr. Suman Kundu's guidance for this project is much appreciated.

REFERENCES

[1] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*. 519–528.

[2] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

[3] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*. Springer, 1–4.

[4] Lorenzo Bruzzone, Fabio Roli, and Sebastiano B Serpico. 1995. An extension of the Jeffreys-Matusita distance to multiclass cases for feature selection. *IEEE Transactions on Geoscience and Remote Sensing* 33, 6 (1995), 1318–1321.

[5] Lieve Hamers et al. 1989. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing and Management* 25, 3 (1989), 315–18.

[6] Danai Koutra, Joshua T Vogelstein, and Christos Faloutsos. 2013. Deltacon: A principled massive-graph similarity function. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 162–170.

[7] Christopher Z Mooney. 1997. *Monte carlo simulation*. Number 116. Sage.

[8] Michael Ovelgönne, Andreas Geyer-Schulz, and Martin Stein. 2010. Randomized greedy modularity optimization for group detection in huge social networks. *Proc. SNA-KDD* 10 (2010), 117–130.

[9] Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. 2008. Prediction of information diffusion probabilities for independent cascade model. In *International conference on knowledge-based and intelligent information and engineering systems*. Springer, 67–75.

[10] David Toubiana, Wentao Xue, Nengyi Zhang, Karl Kremling, Amit Gur, Shai Pilosof, Yves Gibon, Mark Stitt, Edward S Buckler, Alisdair R Fernie, et al. 2016. Correlation-based network analysis of metabolite and enzyme profiles reveals a role of citrate biosynthesis in modulating N and C metabolism in Zea mays. *Frontiers in plant science* 7 (2016), 1022.

A CODE

Codes of the proposed methodology and data collection is provided in the link [↗]