

**IMPACT OF STUDENTS' INTERACTIONS ON THEIR
ACADEMIC PERFORMANCE IN AN ONLINE LEARNING
ENVIRONMENT:
A DATA-DRIVEN APPROACH**

WANIGASINGHE ARACHCHILAGE PRAMODI PRASANJALI
RUKSHALA DE SILVA

14020157

WEDA ARACHCHIGE CHANDIMA IRESHANI IMALIKA
14020297

University of Colombo School of Computing
2018

DECLARATION

I, W.A.P.P.R. De Silva (2014/IS/015) hereby certify that this dissertation entitled Impact of Students' Interactions on their Academic Performance in an Online Learning Environment: A Data-Driven Approach is entirely my own work and it has never been submitted nor is currently been submitted for any other degree.

.....

Date

.....

Signature

I, W.A.C.I. Imalika (2014/IS/029) hereby certify that this dissertation entitled Impact of Students' Interactions on their Academic Performance in an Online Learning Environment: A Data-Driven Approach is entirely my own work and it has never been submitted nor is currently been submitted for any other degree.

.....

Date

.....

Signature

I, Dr. T.A.Weerasinghe, certify that I supervised this dissertation entitled Impact of Students' Interactions on their Academic Performance in an Online Learning Environment: A Data-Driven Approach conducted by W.A.P.P.R.De Silva and W.A.C.I.Imalika in partial fulfilment of the requirements for the degree of Bachelor of Science Honours in Information Systems.

.....

Date

.....

Signature

I, Dr. H.A.Caldera, certify that I supervised this dissertation entitled 'Impact of Students' Interactions on their Academic Performance in an Online Learning Environment: A Data-Driven Approach' conducted by W.A.P.P.R.De Silva and W.A.C.I.Imalika in partial fulfilment of the requirements for the degree of Bachelor of Science Honours in Information Systems.

.....

Date

.....

Signature

ACKNOWLEDGEMENT

This thesis could not have been written without the guidance of our supervisor, Dr. Thushani Weerasinghe and the co-supervisor, Dr. Amitha Caldera. We are really grateful for their commitment in encouraging us, clearing our doubts, providing feedback to improve the papers and guiding us from the beginning of the research. Next our gratitude goes to Prof. K.P.Hewagamage, the Director of University of Colombo School of Computing (UCSC), Dr. Ajantha Athukorala, the Deputy Director of University of Colombo School of Computing (UCSC), Dr. Damitha Karunaratna, Head of e-Learning Centre of the UCSC for permitting us to extract dataset from Bachelor of Information Technology (BIT) Learning Management System (LMS), aligning with ethical considerations. Thanking to all the lecturers who gave us comments and feedback to improve our research, especially to Mr. Prabhash Kumarasinghe, for the guidance given in data mining domain.

Further, we thank Mr. Geeth Hettiarachchi, the Admin of BIT LMS, for providing us the required datasets. The staff of the e-Learning Centre of the UCSC, especially Ms. Nisha Kumari Kariyawasm and Ms. Buddhika Dissanayake deserve our appreciation for providing the information on the BIT design structure in terms of course, forums and assessments. Additionally, our sincere thanks to Mr. Sivanandan Anjana, for providing the technical support. Also, we are grateful to the BIT students, especially Mr. Anuradha Shanaka who provided us more information on BIT forums and assessments in a student's perspective, which helped us to think more broadly. For supporting us with the proof readings, our sincere thanks to Mr. Siril Chandrasekara Gunawardhana.

This thesis would not have been possible without the personal support of our families. Therefore, our sincere thanks do not enough for their kindness.

ABSTRACT

Online learning has become a prominent practise among higher educational institutions due to its power to overstep time and cost constraints. Although it lacks face-to-face physical interactions among students and facilitators, current online learning platforms are designed to facilitate collaborative learning by enabling students to discuss subject related concerns in an online forum setting. However, when the student capacity grows learning management systems become incapable of providing deeper insights on each and every students' social behaviour. Hence, the course facilitators may feel difficult to gain a broader view of how actually the students interact with each other which makes them difficult to provide an informed intervention. For an instance, a facilitator may want to know how well students communicate with each other, are there any isolated students, are they gaining the maximum benefit out of discussions and are the discussions really help them in learning better. Through such understanding of students' social behaviour, a facilitator can make effective interventions to help the students in need. Therefore, current study evaluated the effectiveness of voluntary discussion forums on student's academic performance by examining the online forum interactions and assessment marks of hundred and fifty students. The study was conducted as a case study based on Bachelor of Information Technology (BIT) external degree programme offered by University of Colombo School of Computing, Sri Lanka and it followed a data driven approach. BIT is a fully online degree programme which uses MOODLE Learning Management System as their Virtual Learning Environment. Therefore, this study used interaction and assessment data extracted from the BIT Moodle database and analysed extracted data using social network analysis, correlation analysis and classification methods to provide valuable insights. The findings identified the factors which have affected students' engagement in discussion forums. Overall, the results showed that students who participated in the forum tended to have better performance in the assessments. Most importantly, the study could reveal importance of considering the social capital of the students when evaluating their engagement in online discussion platforms rather than just considering number of messages posted by each student. Findings from this study contributed to development of a tool which helps the facilitators to assess and monitor the students in online discussion environment in a more effective and efficient manner.

Keywords: *Collaborative Learning, Computer-Supported Collaborative Learning, Educational Data Mining, Learner Analytics, Performance, Social Network Analysis*

Table of Contents

Declaration	i
Acknowledgement.....	i
Abstract	iii
List of Tables.....	vi
List of Figures	vii
Abbreviations	viii
CHAPTER 1 INTRODUCTION	1
1.1 Context.....	2
1.2 Research Questions.....	5
1.3 Research Field and Approach.....	7
1.4 Methods and Techniques	9
1.5 Delimitations	9
1.6 Outline of the Thesis.....	10
CHAPTER 2 BACKGROUND	12
2.1 Collaborative Learning	12
2.2 Collaborative Learning and Social Interaction	12
2.3 Computer Supported Collaborative Learning (CSCL)	13
2.4 Student Interactions and Performance in CSCL Environment.....	14
2.4.1 Social Network Analysis	15
2.4.2 Data Analytics	18
CHAPTER 3 METHODOLOGY.....	21
3.1 Step 1 and Step2: Data Collection	22
3.1.1 Step 1: Qualitative Data Gathering	22
3.1.2 Step 2: Quantitative Data Gathering	22
3.1.3 Ethical Considerations.....	23
3.1.4 Data Cleaning and Preparation.....	23
3.2 Step 3: Visualisation.....	25
3.3 Step 4: Statistical Analysis	27
3.4 Step 5: Feature Selection	28
3.5 Step 6: Results and Conclusions.....	30
3.6 Prototype Tool	30
CHAPTER 4 ANALYSIS AND RESULTS	31

4.1	Factors Affecting Students' Social Engagement in Discussion Forums (course - level)	31
4.2	Student Behaviour in Discussion Forums	37
4.2.1	Network Level social Parameters.....	37
4.2.2	Node (user) Level Social Parameters	38
4.3	Correlation of the social network parameters with students' assessment marks.....	42
4.3.1	Correlation Analysis.....	43
4.3.2	Feature Selection using Classification	45
CHAPTER 5 DISCUSSION		54
5.1	Factors Affecting Students' Behaviour in Online Discussion Forums.....	54
5.2	Effectiveness of Students' Social Engagement in Discussion Forums	55
CHAPTER 6 IMPLEMENTATION		58
6.1	Course Wise Students Ranking	60
6.2	Overall Student Ranking	62
6.3	Testing.....	63
CHAPTER 7 LIMITATIONS		64
7.1	Data Collection	64
7.2	Content Analysis.....	64
CHAPTER 8 CONCLUDING REMARKS		66
CHAPTER 9 FUTURE RESEARCH		67
References		68
Appendix A: Interview Questionnaire		73
Appendix B: SQL Data Extraction Scripts		77
Appendix C: Approval Letter for Data Extraction.....		79
Paper I		80
Paper II		102

LIST OF TABLES

TABLE 3.1: BIT COURSE	22
TABLE 3.2: DATA CLEANING PROCESS	24
TABLE 3.3: NETWORK-LEVEL PARAMETERS.....	26
TABLE 3.4: INDIVIDUAL-LEVEL PARAMETERS.....	27
TABLE 4.1: RESULTS NETWORK LEVEL SOCIAL PARAMETERS.....	38
TABLE 4.2: CORRELATION FOR THE COURSE: 'INFORMATION SYSTEMS AND TECHNOLOGY'	43
TABLE 4.3: CORRELATION FOR THE COURSE: 'COMPUTER SYSTEMS I.'	44
TABLE 4.4: CORRELATIONS FOR COURSE 'WEB APPLICATION AND DEVELOPMENT I.'	45
TABLE 4.5: SOCIAL NETWORK PARAMETERS SORTED ON CORRELATION	46
TABLE 4.6: CLASSIFICATION ACCURACIES FOR 'INFORMATION SYSTEMS AND TECHNOLOGY'	47
TABLE 4.7: CLASSIFICATION ACCURACIES FOR 'COMPUTER SYSTEMS I'	49
TABLE 4.8: CLASSIFICATION ACCURACIES FOR 'WEB APPLICATION DEVELOPMENT I'	51
TABLE 4.9: FILTERED FEATURE SUBSETS.....	53

LIST OF FIGURES

FIGURE 1.1: HOME PAGE OF BIT VLE	3
FIGURE 1.2 : STRUCTURE OF AN ONLINE DISCUSSION FORUM.....	4
FIGURE 1.3 : RESEARCH DESIGN.....	7
FIGURE 2.1 : SOCIOGRAM	16
FIGURE 3.1 : METHODOLOGY.....	21
FIGURE 3.2 : FEATURE SELECTION PROCESS.....	29
FIGURE 4.1: THIS GRAPH SUMMARISES ALL THE INTERACTIONS OF COURSE 'INFORMATION SYSTEMS AND TECHNOLOGY'...	33
FIGURE 4.2: THIS GRAPH SUMMARISES ALL THE INTERACTIONS OF COURSE 'COMPUTER SYSTEM I.'	34
FIGURE 4.3:THIS GRAPH SUMMARISES ALL THE INTERACTIONS OF COURSE 'WEB APPLICATION AND DEVELOPMENT I'	35
FIGURE 4.4: INFORMATION GIVING NETWORK OF COURSE 'INFORMATION SYSTEMS AND TECHNOLOGY'	39
FIGURE 4.5:"INFORMATION GIVING NETWORK OF COURSE 'COMPUTER SYSTEM'	40
FIGURE 4.6:INFORMATION GIVING NETWORK OF COURSE "WEB APPLICATION AND DEVELOPMENT I."	41
FIGURE 4.7:CLASSIFICATION ACCURACY VARIATION IN FEATURE REDUCTION FOR 'INFORMATION SYSTEMS AND TECHNOLOGY'	48
FIGURE 4.8: CLASSIFICATION ACCURACY VARIATION IN FEATURE REDUCTION FOR 'COMPUTER SYSTEMS I'	50
FIGURE 4.9: CLASSIFICATION ACCURACY VARIATION IN FEATURE REDUCTION FOR 'WEB APPLICATION DEVELOPMENT I' ..	52
FIGURE 6.1: EIGENVECTOR CENTRALITY ALGORITHM.....	59
FIGURE 6.2: INTERFACE - SELECT 'COURSE WISE RANKING '	60
FIGURE 6.3: INTERFACE - DISPLAY COURSE WISE STUDENTS RANKS	61
FIGURE 6.4: BUILT SOCIOGRAM FOR A COURSE.....	61
FIGURE 6.5: STUDENT RANK - BEST E-LEARNERS.....	62

ABBREVIATIONS

BIT	Bachelor of Information Technology
CL	Collaborative Learning
CMC	Computer-Mediated Communication
CMS	Computer-Mediated System
DM	Data Mining
EDM	Educational Data Mining
ID	Instructional Designer
LMS	Learning Management System
MCQ	Multiple Choice Questions
MS	Microsoft
OADF	Online Asynchronous Discussion forums
SME	Subject Matter Expert
SNA	Social Network Analysis
TEL	Technology Enhanced Learning
TEL	Technology Enhanced Learning
UCSC	University of Colombo School of Computing
VLE	Virtual Learning Environment

CHAPTER 1 INTRODUCTION

Most of the academic institutions are attracted more towards the online learning concept due to its power to overstep the limitations of space, time and cost. Although, the students do not get a chance to interact with each other physically, they can communicate through forums, chat messages, activities in online collaborative learning platforms. According to ‘Connectivism’, a learning theory for digital age, by Siemens [1], learning is no longer an individualistic activity. With the advancement of digital social technologies, learning has become much more complex and it should occur through connections in a social network setting by sharing knowledge. Learning through forming connections is one of the key concepts in collaborative learning where students are supportive for peers’ learning and responsible for their own learning. Therefore, the success of one student aids for another one’s successfulness [2].

Asynchronous online discussion forums play a major role in replacing physical learning interactions with online collaborative learning interactions. Compared to synchronous interactions, asynchronous interactions provide more time for students to reflect, think, and search for extra information before contributing to the discussion [3], [4]. Also, they facilitate students to learn from ideas, shared resources, and experience of each other. Thereby, forums provide an environment to create learning communities and inculcate team spirit. Therefore, discussion forums in online courses can support knowledge production more effectively.

Unfortunately, due to a large number of students in this kind of online courses, and as the built-in analytics of major LMSs including Moodle offer only limited insights on students’ social behaviour, facilitators are struggling to observe, and evaluate students’ learning behaviours in order to provide the facilitation in a more precise manner. For instance, the LMSs provide the students’ online behaviour mostly in terms of frequency of participation [5]. But insights on how actually the students interact with each other, whether they are linked properly or isolated, are they gaining the maximum benefit out of online forum discussions to complement the lack of physical interactions and are the discussions really help them to perform well in academics are not explicit. Also, facilitators are unable to decide whether the existing design of online discussion forums and assessments are sufficient and promotive for the students to interact with their peers and learn together.

In a context, where formal learning is completely virtualized, analysing students’ online interactions could reveal patterns on how the contrast of students’ behaviour affect their academic performance which may lead to meaningful interventions. This will help facilitators to identify the

weak and isolated students who are at risk of failure and provide them with additional personalized support through simplified learning content and necessary instructions [6]. Furthermore, with the understanding gained through monitoring of online learning behaviours can help students to enhance their networking skills as well as communication skills and social capital by rewarding the active online presence of them.

Fortunately, LMSs are capable of creating powerful online courses which can motivate students to participate more. At the same time, these online courses are storing a vast archive of valuable data. Contemporary researchers in Educational Data Mining (EDM) have used different techniques to analyse and interpret these LMS data [6], [7], [8]. So, if a study can analyse the connection between the nature of students' interaction in an online collaborative learning environment with their performance in academics, by identifying those factors and their influence, that knowledge can be used to improve the way of delivering the course content and the facilitation. More importantly, it will also be helpful to minimise the adverse effects of distance learning and the lack of teacher involvement.

Considering the above requirement, this research followed a data-driven approach to identify how these students collaborate in an online learning setting and to analyse its impact on their academic performance in online assessments. The research was based on empirical study conducted with three different online courses that facilitate students following the Bachelor of Information Technology (BIT) degree programme at the University of Colombo School of Computing (UCSC), Sri Lanka.

1.1 Context

In the University of Colombo School of Computing (UCSC), Bachelor of Information Technology (BIT) external degree program also facilitates collaborative learning through their Learning Management System (LMS). BIT students are not receiving any face-to-face teaching from the UCSC. The curriculum, past examination papers, and the examination timetables are delivered through the website (<http://www.bit.lk/>). Also, the students meet the UCSC staff only at registration, examinations and award ceremonies. Therefore, to minimize the adverse effects of the learning without physical student-student, student-teacher interactions, a virtual learning environment (VLE) was introduced by the e-Learning Centre of the UCSC for all the courses of BIT degree on the Moodle platform (<http://moodle.org>)[9].

Especially in this kind of virtualized learning setting, student's mind tends to distract when they encounter a problem which cannot be resolved by themselves. Then, they start seeking the

external help from facilitators or peers to carry on their learning. Therefore, the discussion forums were created in the BIT VLE (<http://vle.bit.lk>) to cater this requirement [10].

At end of each course section, there is a discussion forum and it is used by facilitator and students to communicate with each other. A facilitator is well familiar with the course environment and the content. His/Her key role is to help students to find the solutions for course-related problems which might rise up during the course duration. If any inquires raised by students that is difficult to handle by the facilitator alone, he/she will seek further support from the relevant Subject Matter Experts (SME). Also, the students are intended to align with a set of netiquettes when communicating through the BIT VLE in order to prevent inadequate discussions and focus on learning [9]. Figure 1.1 shows the home page of the BIT VLE (<http://vle.bit.lk>).

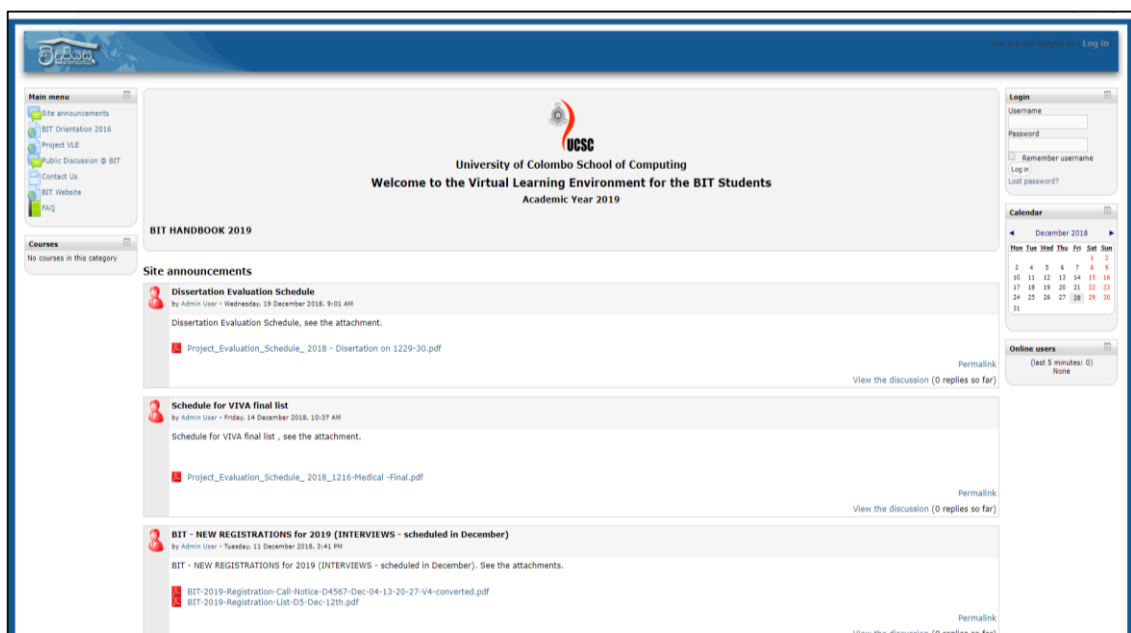


Figure 1.1: Home page of BIT VLE

After login to the system, students can navigate to the courses they are following. Under each course, there is a discussion forum after each lesson, to discuss the issues that the students might encounter. Generally, the facilitator asks to discuss on a particular topic or else students can initiate their own discussion topics. As depicted in Figure 1.2, students can answer or criticize the posts by their peers, which leads to expand knowledge while inculcating team spirit.

Home ► My courses ► IS4010 ► Analytics with RapidMiner ► Discussion forum for RapidMiner tasks ► How to obtain the comparison on accuracy?

Navigation

- Home
- Dashboard
- Site pages
- My courses
 - IS3003
 - IS3004
 - IS3005
 - IS3007
 - IS3018
 - IS4001
 - IS4004
 - IS4009
 - IS4010
 - Participants
 - Competencies
 - Grades
 - General
 - Overview of BI and analytics with BI life Cycle
 - Data Warehousing
 - Data usage in Projects & Operations
 - Review and Assessment
 - Predictive Analytics - Data Mining
 - Data Visualization
 - Analytics with RapidMiner
 - Task 1: Classification credit worthiness
 - Task 2: Association rule mining and recommender sys...
 - Task 3: Unsupervised learning - Clustering
 - Task 4: Text classification
 - Task 5: Anomaly detection
 - Discussion forum for RapidMiner tasks
 - How to obtain the comparison on accuracy?

Administration

- Forum administration
- Optional subscription
- Subscribe to this forum
- Subscribe to this discussion

Discussion forum for RapidMiner tasks

How to obtain the comparison on accuracy?

by - Wednesday, 7 November 2018, 1:25 PM

1 followed the steps of the Naive Bayes Classification pdf and created a process like in figure 1. Then I obtained results as in Figure 2 and 3. But I don't understand how to obtain a comparison on accuracy as illustrated in the pdf. And what do the results I have obtained really mean?

Display replies in nested form

Process

Figure 1

Figure 2

Figure 3

Re: How to obtain the comparison on accuracy?

by - Wednesday, 7 November 2018, 5:09 PM

To view the accuracy, connect the "performance node" to the result (fig. 1). This will give a confusion matrix, as per what it means check <https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>

Figure 1

Re: How to obtain the comparison on accuracy?

by - Wednesday, 7 November 2018, 9:17 PM

Thank you Malithi! It worked.

Error message when trying to execute the validation

Figure 1.2 : Structure of an online discussion forum

Not only the BIT Learning Management System (LMS) facilitates the students to discuss and work with other students but also it facilitates to evaluate their learning progress using quizzes. Mainly quizzes are of two types; practice quizzes and assessment quizzes. Practice quizzes can be attempted any number of times and they do not show the marks but provide the feedback to find correct answers. But assessment quizzes should be attempted only three times and basically it is

consisted of Multiple-Choice Questions (MCQ). Moreover, the students' learning performance are evaluated via assessment quizzes [9].

However, these online discussion forums are voluntary. Therefore, it's challenging to motivate the students to participate in online discussion forums. According to [10], although there are few thousands of students registered in BIT degree, the number of students participates in the forum is very poor. Therefore, to encourage the student participation by recognizing their engagement in forums, the e-Learning centre of UCSC introduced 'Best e-Learner of the Semester' award. Best e-Learners should not only engage in the forums well, but they should perform well in assessments also. Therefore, the selection process for best e-Learner considers both the active participation in forums and the performance in online assessments. The criteria used in the selection process is that, each student is assigned a weight based on the number of posts by each of them to the forum, and that weight is multiplied by their average online assessment mark. This is called the 'Forum Score (FS)' by each student for each course. Next, the average forum score (FS_{avg}) for all compulsory courses in the given semester is obtained. After that, the best 10% of the students are filtered out considering their average forum score (FS_{avg}). Finally, by qualitatively analysing that top 10% students' forum postings, the best e-Learner is selected who have posted rich, subject-related contents.

1.2 Research Questions

Although BIT Learning Management system is intended to support collaborative learning by enabling students to discuss with their peers, it is still a grey area that whether these discussions actually support the students to perform better in academics. The Moodle reports only provide the students' forum engagement in terms of the frequency of their participation (number of posts by each) [5]. However, there is not any mechanism to get a broad view of the collaboration among the students behind these online discussions. For instance, facilitators struggle to get insights on how do students interact with each other in forums, are they connected with peers as expected or isolated, are they struggling in the communication process, are they gaining the maximum benefit out of the discussions to complement the lack of physical interactions etc. and also to assess whether the online discussion forums are really supportive for their academic achievements. Additionally, to evaluate whether the existing design of discussion forums and assessments are sufficient or supporting the students to interact with peers to achieve their learning objectives. Hence, the interventions and design of these components can be improved to promote the student interactions and therefore support them to achieve their learning objectives.

Particularly, facilitators can consider changing curriculum, introducing new teaching methods, promoting social equity in student interactions, or fostering connections in learning communities [6]. Also, Instructional designers can plan how to optimize discussion forums in a way that it may improve students' academic performance.

Considering this necessity, a data driven approach was carried out to identify:

How do student interactions in online discussion forums affect students' academic performance?

In order to answer the above main question, following sub questions should be addressed first.

(a) What factors affect students' engagement in online discussion forums?

Due to the large number of students involved in online courses and the limited capacity of LMS to provide insights on students' social behaviour, it is difficult to have a clear picture of what is happening behind these online discussions such as what type of student and teacher networks are there, how they interact with each other, when do they interact more with each other etc. Thereupon, it is needed to identify what are the factors that have been motivated the students to collaborate more and what have disrupted their communication. That knowledge informs course facilitators to determine when to intervene and what to modify in order to enhance the productivity of discussions. So, after understanding the background of these discussion forums, the research next focused on:

(b) What social parameters best describes the students' behaviour in online discussion forums?

Through visualizing student interaction data in a social network point of view, the study could provide different perspectives on the social relationships within the network and reveal hidden patterns which describe the social behaviour of students in the forums. Also, visualization may help to identify the students who are isolated in the social network, students who have interacted well with their peers etc. Consequently, it would be able to highlight the level of collaboration among students, the strength of the interactions on the course and individual levels.

Additionally, those interactions should not be only quantitative, but it should also possess high quality. The ultimate goal of these online discussions is to promote learning, so that it is a necessity to identify what kind of impact these interactions have on student's learning. Therefore, finally the research focused on:

(c) *what social parameters best interpret a correlation with the students' assessment marks?*

Finding solutions to this question can shed light on further determining how much the social parameters can affect the academic performance of students. The factors affecting higher can be promoted more and the degrading factors can be eliminated as much as possible. Therefore, meaningful discussion forums can be provided with proper monitoring rather than just facilitating students to interact even without knowing whether it is successful.

1.3 Research Field and Approach

The research studies which are conducted with the primary objective of analysing data in an educational setting to resolve several educational research issues fall within the research field called 'Educational Data Mining (EDM)'. It is concerned with developing methods to explore the educational data to better understand students and the environment which they learn. The use of Internet in education has introduced a new paradigm known as e-learning or web-based education where large amounts of information about teaching–learning interaction are endlessly generated and available as gold mines. Exploiting various statistical, machine-learning, and data-mining (DM) algorithms over these gold mines of educational data, EDM focuses to better understand students and their learning, and to develop computational approaches that combine data and theory to transform practice to benefit students [11].

This research also falls under EDM as it focuses on analysing the data of online discussion forums and online assessments to explore the impact of student interactions on their academic performance in BIT Online Learning Environment. Moreover, according to the types of scientific researches, this research can be categorized as an 'explanatory research' since it investigates the correlation between student interactions and their academic performance [12].

The research design of an exploratory research is as below (See Figure 1.3).

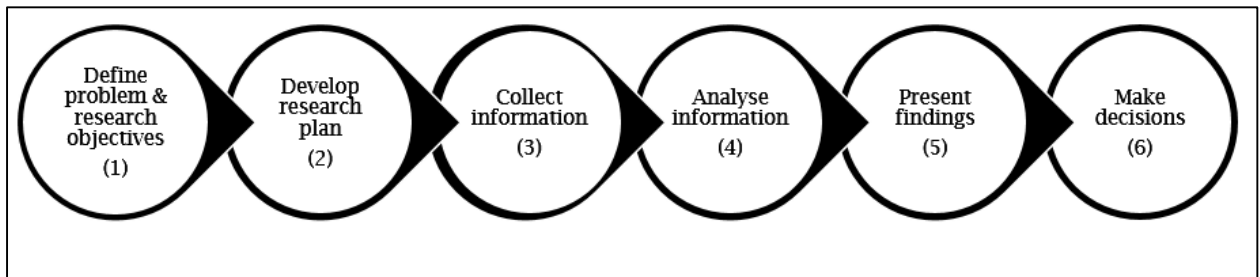


Figure 1.3 : Research design

First of all, it is needed to be familiar with the relevant domain by reading research papers and identify what work has been done so far. It helps to find gaps in existing knowledge and practices which can be answered through the research. Based on that, the research problem should be defined briefly and clearly. After that, the research objectives that leads to overcome the research problem should be clearly defined (Step 1 in Figure 1.3).

Next, the research plan should be structured as how to reach the previously defined research objectives. What type of research it should be (i.e. case study or explanatory), the sampling techniques (i.e. simple random sampling or convenience sampling), the technologies and tools to be used are explained in the research plan (Step 2 in Figure 1.3).

After having a clear idea on how to proceed, the information should be gathered first. Information might be qualitative or quantitative. However, altogether the gathered information should be integrated together to support a common analysis (Step 3 in Figure 1.3).

The gathered data is next pre-processed, and analysed using the techniques suggested by the literature to uncover the hidden patterns in students' behaviours (Step 4 in Figure 1.3).

The derived knowledge is then used for making a tool which assists facilitators in evaluating the students' behaviour (Step 5 in Figure 1.3).

Based on the findings and evaluations, facilitators can take important decisions to improve the productivity of online discussion forums to aid the better academic performance of BIT students (Step 6 in Figure 1.3).

Some related EDM researches which analysed forum usage data and performance data have followed quantitative approaches such as correlating the message frequencies with course performance [13], [14], while some others have used qualitative information with a content analysis approach [4]. Moreover, other approaches have considered the social network information as social aspect is one of the major features of online forums [6], [8], [15], [16]. Considering all, Romero has followed a mixed approach which is consist of quantitative, qualitative and social network information to provide a richer explanation on predictors of student performance in online discussion forums [7].

Therefore, this research is planned to conduct as a case study and follows a mixed approach which consists of both qualitative, and quantitative techniques to deeply investigate the social elements of online collaborative learning towards better performance. The first half of the study uses qualitative data gathering techniques to get an idea of how the online discussion forums and assessments have been designed by instructional designers and course facilitators in order to make students to achieve the learning objectives. Then this study will be acted as a case study that applies

data analytic methods to analyse student-student interactions and assessment marks in several courses in the BIT VLE, to investigate how the students' social structure has been built around their interactions, what type of network is there, whether the students have behaved as expected by the instructional designers, facilitators and whether there is a connection between student interactions and their performance etc.

1.4 Methods and Techniques

The study methods include conducting interviews, extracting forum and assessment data from Moodle database, visualisation of social networks and analysis of students' discussions in online courses. Initially, the interviews were conducted with Instructional Designers and facilitators of BIT degree programme to gather qualitative data on how the course, online discussion forums and assessments have been designed to align with the learning objectives of a course. Next, the interaction data in forums and assessment grades from BIT Moodle database were extracted using SQL queries by anonymising the true identity of students and facilitators. After that, the dataset was visualised and analysed using Social Network Analysis with the help of open source tool Gephi [6]. This revealed significant insights into the students' position in the social network structure and its social parameters. In the next step, the network parameters obtained from network visualisation were correlated with the assessment marks using statistical techniques such as correlation with the help of tool 'SPSS' [6], [13]. It was performed basically to determine how a student's position in the social network can affect how well they perform in their assessments. As there were several number of social network parameters that influence their performance, to identify the best subset of it which leads to more predicting accuracy, classification was used [7]. Therefore, based on the findings, the thesis discusses how the student's behaviour in online discussions can contribute their learning.

1.5 Delimitations

The present research is based on the area of student analytics and deals only with the students' interactions in Online Asynchronous Discussion Forum (OADF) due to the scope, cost and time constraints. Therefore, the research did not consider students' online interactions outside the online discussion forums such as activities, and private messaging facility in VLE. Also, the student's knowledge derived from other sources rather than online discussion forum (e.g., books) could not handle in this study. Moreover, the post (message) content of discussion forums could

not be analysed qualitatively due to the time constraints, however its impact could be reduced by initially removing the irrelevant discussion threads from the dataset and as the inadequate discussions were already prohibited through VLE netiquettes.

1.6 Outline of the Thesis

The thesis is divided into seven chapters, structured as follows.

Chapter 1 - Introduction

This chapter discusses the motivation for this research, specific research context and the specific research problems highlighting the importance of determining the impact of student interactions on their academic performance in an online collaborative learning environment. Additionally, it provides an overview to research field and approach, methods and techniques used and delimitations of the research. Each section in this chapter should be read sequentially since they are chronologically ordered with respect to the primary concepts of the research. Preferably this chapter should be read before the other chapters of the thesis.

Chapter 2 - Background

This chapter describes the relevant background studies in detail and their applicability and the drawbacks considering the various methods and data attributes in analysing students' online interaction data. This chapter discusses the importance of collaborative learning for online distance learning, how the social structure build behind the online discussions, how the position in the social network can impact on their learning performance and what types of parameters and techniques used to analyse that impact and what are the significant findings by the research community so far.

Chapter 3 - Methodology

The third chapter describes in detail the methodology followed by this research. The several methods such as data gathering methods (quantitative and qualitative), data visualization, statistical analysis, feature selection with classification along with the tools and technologies used, will be discussed under this chapter by furthermore providing the justification of our selection.

Chapter 4 – Analysis and Results

The observed results are provided in this chapter, answering each sub-question of the research; factors affecting students' social engagement in discussion forums, student behaviour in discussion forums, correlation of the social network parameters with students' grades.

Chapter 5 -Discussion

This chapter discusses the importance of the main findings and how and why the conclusions drawn from the findings are important. Also, it identifies how the results are consistent with prior knowledge of the domain or what are the unexpected findings. Additionally, it discusses the limitations or weaknesses of the approach used in this research.

Chapter 6 -Implementation

The basic idea and the design of the prototype implemented using the finding of this study included in this chapter.

Chapter 7 - Limitations

Here, the limitations and the weaknesses of the research approach being used is elaborated.

Chapter 8 - Conclusion

Here, the main findings of the research are concluded in this chapter.

Chapter 9 - Future Research

This chapter unfolds several new paths to further consider to enhance the present research in terms of quality and applicability.

CHAPTER 2 BACKGROUND

The online learning concept has been emerged a lot during past few decades. However, its history runs back to a time where instructors sent lessons via email and the students' completed assessments were returned back using email [17]. Modern online education has become a common practice at higher educational institutes due to its power to overstep the limitations of space, time and cost [6]. Learning Management Systems (LMS) are capable of creating powerful online courses. These online courses are storing a vast amount of valuable information, and this information is used by most of EDM researchers to discover hidden relationships between student interactions and performance [7].

2.1 Collaborative Learning

According to Laal and Ghodsi [18], Collaborative learning is an educational approach where students socially interact with other students, as well as facilitators to expand their knowledge on a particular subject or skill. According to Panitz [19], collaboration is a notion of interaction and interaction is the structure of the corporation designed to facilitate accomplishing of a goal or an end product by working together in groups. Roschelle and Teasley state that: “collaborative learning involves the “...mutual engagement of participants in a coordinated effort to solve the problem together” [20, p. 70]. The debate is still going on and it is beyond the scope of this article to state which definition or perspective is most appropriate. It is, however, important to note that the consideration of social component in all these researches. To support this further, a growing body of research has demonstrated that social network is a central element in collaborative learning environments [21], [22].

2.2 Collaborative Learning and Social Interaction

From the social network point of view, learning is a social and collective outcome achieved by means of seamless conversations, shared practices, social connections built from the social networks [23]. Knowledge constructed from these social networks is claimed as a component which is not only built through individual effort but also a collection of subcomponents constructed actively via ongoing social exchanges and interactions among multiple students embedded in

collaborative social networks [24], [25]. Hence, it seems social interaction is the key to collaboration. Also, this belief is shared by many (distance) educational researchers, reporting that social interaction is an indispensable condition for learning [26-31]. In Fact, with the growth of World Wide Web, the social networking has raised its concerns in many fields like commerce, communication and more importantly in education [32]. Many researchers have pointed out the increased adoption of social bookmarking, computers, Internet connectivity, and Internet access for teaching and learning [33], [34]. Moreover, Liu et al. [35] report that Technology Enhanced Learning (TEL) facilitates networked learning through computer-supported collaborative learning (CSCL) features that have been demonstrated to positively enhance learning when equipped with properly designed resources.

2.3 Computer Supported Collaborative Learning (CSCL)

Kirschner et al. [36] report that collaborative learning approach more applicable for online courses to achieve higher order learning outcomes. This confirms by J. Strijbos, P. Kirschner and R. Martens [37] by reporting collaborative learning in online environment enclose knowledge and skills which are difficult to acquire by learning individually. Therefore, all these educational researches have supported the concept of computer-supported collaborative learning. As reported in [22], interactions in CSCL learning environment are often remote, faceless, uncertain, and moderated by Computer-Mediated Communication (CMC) systems. Students' willingness to communicate in CMC discussion settings should affect their behaviours, especially how they build new social and learning relationships/networks with distributed, remote learning partners, who are often strangers [22]. Furthermore, asynchronous interactions made through these CMS systems benefit more compared to the synchronous discussions. Such benefits include getting more opportunities to interact with each other and more time to reflect, think, and search for extra information before contributing to the discussion [3], [4]. Online Asynchronous Discussion Forum (OADF) is a tool for CSCL which offers the opportunity for students to interact and cooperate in online communities. OADF is not just a tool to form student and facilitator interactions but also it allows both parties to shape the nature of the information exchange by reviewing posted content and analysing own ideas before responding since participants are not constrained to respond immediately in most cases [4]. In addition to that, in online learning forums, students can build on ideas posted by their peers and learn collaboratively [4], [38], [39] by presenting their ideas, decisively reading, analysing, judging, and evaluating others' posts, through writing replies to comments and appreciating each other [9]. Hence it highlights the importance of maintaining

student interactions in distance learning platforms. In fact, student interaction is the underlying driving factor of collaborative learning [9] and according to Garrison and Cleveland-Innes [40], this learning process requires self-directed learning and critical thinking skills. These skills are considered as essential factors for adult distance learning platforms [41]. In addition, students' interactions in a learning environment provides the opportunity for peer teaching and peer learning [42]. Peer-teaching concept is well accepted as it enables students to learn twice by teaching others [43], [44]. Hence Peer-teaching approaches are practised more often in higher educational contexts [42]. However, in order to support peer learning and peer teaching concepts, instructional designers of online courses require to pursue methods for encouraging and sustaining student interactions to reflect collaborative learning in online learning environments [9], [45]. Moreover, effective learning environments or courses can be designed “only by referring appropriate learning theories, instructional design theories and instructional design practices” [45, p. 243]. Also, it is essential that every learning activity designed by instructional designers to be aligned with the learning outcomes of the lesson [9].

According to Hewagamage et al. [10], Weerasinghe et al. [46], course content in BIT online learning environment which used in this study is designed in accordance with design guidelines and principles to enhance student interaction and collaborative learning. As Weerasinghe [42], [46], Usoof and Wikramanayake [47] report BIT online learning environment is equipped with student directed discussions providing students more opportunities to interact with each other and find solutions to course-related problems by themselves. Results of these studies imply that forums and assessments in BIT online LMS are designed according to the design principles, and student discussion forums and assessments are aligned with defined intended learning outcomes of each lesson. Having referred this as the hypothesis, the present study aimed analyse the impact of student interactions on their performance in online assessments.

2.4 Student Interactions and Performance in CSCL Environment

Discussion forums provided by CSCL environments can be categorized into several types. For instance, one such type of forum is where, facilitator provide an initiative for discussion or debate over a given topic. Participation in this type of forum is usually a part of the course requirements; students are either given credits for participating or graded according to their level of contributions. Another type of forum is where, it is implemented as a medium for students to ask question or discuss anything related to the course but is not directly associated with any grades. This type of a forum provides opportunities for students to discuss questions about the course

materials and their concerns so that their peers may respond to them. These types of forums are usually used by higher educational institutes to compensate for the diminishing opportunity for interaction in a large-sized course or a web-based course where massive number of students are engaged. Student participation in this type of forum is usually voluntary and intrinsic. Students may not be given any external incentive or course credit for participating. Students participate in these types of forums because of their own willingness to learn. However, in some circumstances there may be indirect evaluation mechanism to reward the online presence of students in order to encourage them to engage more in online learning context. For that, participation in online discussion forums could be a crucial factor. As mentioned in section 1.1, in BIT also they use such evaluation system to reward the “best e-learner “of a semester [10]. In order to select the candidates for this award, evaluators have used number of posts by each student along with his/her assessment mark because best e-learners are not only those who discuss things in the forums but also who score good in their assessments. Even though the current system proposed by past BIT researches works well when evaluating students, it does not consider the social capital a student may possess when interacting with his/her peers. Social capital is an important factor that arose along with emergence of social networking concepts.

When student discussions happen in CSCL environment it automatically forms a social network where it is possible to identify influence of each forum participant towards another. Many researchers have followed Social Network Analysis (SNA) (described in section 2.4.1) and data analytic techniques (described in section 2.4.2) to analyse student interactions in social networks and its effect on their performance [6], [8]. Conclusions derived from those researches inform that social network built within a CSCL community had a perceptible influence on student performance. Moreover, they demonstrated how the central positions of students within the emergent collaborative learning network resulted in higher levels of learning performance [6], [48]. However, in some researches it has given negative results [49] therefore it is needed to examine in which context SNA may act as a valid predictor. Moreover, social networking applications have become so popular and social networking concepts have established its power within a short span of time. Therefore, use of it to analyse interactions in online learning environment may give valuable insights which may otherwise not possible through traditional methods.

2.4.1 Social Network Analysis

With the emergence of social media like Facebook and Twitter, Social Network Analysis has become much popular as it generates patterns to discover the hidden relationships between

people, the types of those interactions, the causes for their presence and to measure their influence [48]. Today it is used in a wide variety of disciplines to investigate valuable information. For example, it is used in criminology to study the association between offenders, their criminal behaviour patterns. In organizational communication, it can be used to analyse the flow of information and the decision-making process. When considering the education field, one of the issues with the major learning platforms like Moodle and Blackboard is that the built-in analytics of them only offer limited insights to study student interactions. For example, Moodle offers instructors to view frequency of participation of students on courses while lacking the ability to deeply study the structure of the communication and student interaction patterns [5]. Therefore, Social Network Analysis is applied in education field basically to analyse the students' participation level in courses, their level of cohesion, the active and inactive students, the flow of information, efficiency of group work etc [6], [8]. As cited in [6] one of its major strength over other traditional analysis methods is its speed in producing information and easiness in interpreting results.

The visualization of the social network is depicted by a graph called '*sociogram*' which consists of *nodes* and *edges* (See Figure 2.1). An actor in the network (here, a student or a facilitator in the learning context) is depicted by a node and the interactions between those actors are depicted by the edges/lines between nodes.

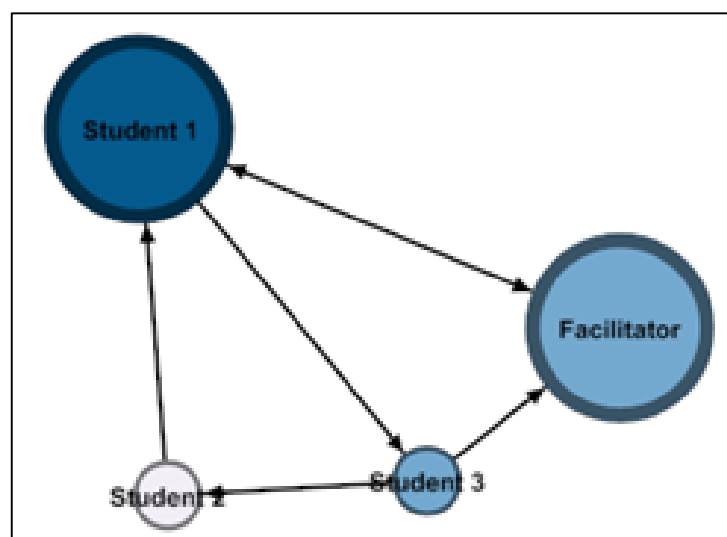


Figure 2.1: Sociogram

The centrality measures calculated using these sociograms reflects the behaviour of students in the online collaborative learning environment. For example, *degree centrality* is a measure of how active a student is in the network, *betweenness* centrality is a measure of how a

participant connects the unconnected peers and promote the discussion, *closeness centrality*, how close a student is to his peers and *information centrality* reflects the importance of a node in information flow etc [6]. Considering the previous researches related to student interactions and performance, Camilo et al. [8] found that the students who have posted and answered more questions have achieved high scores at the end of the course. Gašević et al. [15] found closeness centrality has a positive correlation with higher grades. Moreover, Ángel et al. [49] found a correlation between centrality measures and performance in some courses and negative in others which reported that further researches should be carried out to identify in which contexts SNA behaves as genuine predictors. Additionally, Camilo et al. [8] came to a conclusion in their research that rather than only using activity attributes (total time spent in the course, total number of sessions performed, average time spent and number of sessions performed per week, number of messages read and written in email and forums) it can be achieved improved results in performance and dropout prediction when combined it with SNA attributes (*degree, indegree, outdegree, betweenness, authority, hub, top3*).

SNA offers a very important view of the whole network which is invisible through general LMS reports. By monitoring an education system using SNA, the facilitators can gain a better understanding of their students and therefore to help them to get better results by providing required interventions [8]. As cited in [6], a course instructor/facilitator who uses SNA to monitor an online discussion forum can have a big picture of student interactions and can improve an inactive discussion which shows a smaller number of interactions between students by promoting a collaborative dialogue to encourage discussing. Also, for the isolated students who identified as in a risk of failure or dropouts who have few interactions with peers can be provided with inclusive online environments, well-designed scripts and rewarding collaborative learning to scaffold their learning and take them back to the safe zone. Moreover, if a facilitator dominated network was identified in the social structure where the students tend to reply to the facilitator rather than collaborate with peers, interventions should be provided to encourage facilitators to guide the students rather than dominating. When considering the power of a social element in analysing collaborative learning, Camilo et al. [8] have reported that a high accuracy has been obtained in predicting performance when SNA attributes were combined with traditional quantitative activity attributes like message frequencies. This has been confirmed by Romero [7] by reporting that SNA can expose the invisible sides of online collaborative learning by especially considering the social structure, relations and interactions which is not possible by traditional analysis methods.

2.4.2 Data Analytics

Through visualisation of forum interaction data, the derived social network parameters imply each student's position in the social network built behind the discussions. For instance, *degree centrality* and *betweenness centrality* interpret the students' behaviour, in terms of their participation, connectedness etc. In order to examine how these positions, impact the students' assessment grades, researchers have used various analytical methods [3], [21], [33], [34]. Also, prior to any complex analysis, several researchers have followed 'feature selection' process to select the most contributing and influential set of attributes from a given set of attributes which explain the variation of student performance.

2.4.2.1 Feature Selection

The literature revealed that the most appropriate social metrics that influence student performance could vary from each case study depending on the specific context [6], [7], [8]. Furthermore, Romero et al. [7] found that it is essential to filter out the most influencing parameters from all available parameters for better accuracy.

It is due to the reason that, there are various number of social network parameters which imply the students' position in the social network in various perspectives. *Degree centrality*, *in-degree centrality*, *out-degree centrality*, *eigenvector centrality*, *betweenness centrality*, *closeness centrality*, *authority*, *hub*, *pageRanks* are some of them. These parameters may variously impact on students' academic performance. Some parameters may be powerful predictors of performance while some are irrelevant. Therefore, when classifying students in to pass or fail using both these irrelevant parameters mixed with powerful predictors, it may lead to a negative effect. Hence, most important findings can be hidden. Furthermore, using only the relevant parameters may reduce the overall training time (reduce the curse of dimensionality), increase generalizability and defense against overfitting [50]. For example, Camilo et al. [8] used feature selection and found the most powerful predictors for their learning context as *betweenness centrality* and *authority* measures among the whole set of features; *degree centrality*, *in-degree centrality*, *out-degree centrality*, *betweenness centrality*, *authority*, and *hub*.

While some researchers have followed feature selection process, some researchers have excluded that in their investigation. However, regardless of the feature selection, all the researchers have used mainly two techniques to examine the association between the social attributes and student performance. Those are Statistical Techniques and Data mining Techniques.

2.4.2.2 Statistical Techniques

Among numerous types of statistical analysis methods, many researchers have used correlation, regression, standard deviation to investigate the impact of social attributes on student performance [6], [14], [48]. For example, Saqr et al. [6] have used Kendall's Tau-b test to measure the correlation between SNA attributes and performance. Results showed that attributes corresponding to the number of interactions (*degree centrality* and *out-degree centrality*) did not significantly correlate with student grade. However, *in-degree centrality* was moderately significantly correlated. All centrality attributes measuring the role in information relay were positively correlated with final performance. Also, several other researchers have analysed the impact of social presence towards student performance without necessarily considering the SNA aspect. For example, Palmer et al. [14] investigated the effect of online asynchronous discussions on student learning. Although the study did not explicitly mention the use of SNA, the data attributes they selected such as message frequencies (number of posts and replies) can be categorised under the defined criteria for SNA parameters (*out-degree centrality* and *in-degree centrality* respectively). They have used correlation (Pearson's linear correlation coefficient) to evaluate the relationship between the selected data variable pairs and the final unit mark of students. In addition to that, Cho et al. [48], have used the mean, standard deviation to assess forum participation. Linear regression analysis and correlation have been used to analyse the relationship between the numbers of posts and the standardised scores of each assessment measure. The results from both [14], [48] revealed a significant correlation between forum participation and academic performance of students.

2.4.2.3 Data Mining Techniques

Data Mining (DM) is the practice of scanning huge data repositories to discover new information and therefore to derive knowledge which is a valuable support for effective and timely decision making [51]. Application of data mining in the education field is known as Educational Data Mining (EDM). Here the raw data coming from Learning Management Systems (LMSs) are converted into useful information by applying the data mining process; preprocessing, data mining and post-processing. There is a wide variety of data mining techniques such as classification, clustering, association-rule mining, sequential mining, text mining. Also, there are several other techniques such as regression, correlation, visualisation which are not strictly considered as DM [11]. As cited in [52], these data mining techniques can be used for Data Analysis (explore data without any clear idea), Descriptive Modeling (provides models which show the relationship between different objects), Predictive modelling (prediction of unknown

values from different known variables), Discovering Patterns and Rules (spot behaviours like fraud detection). Among them, prediction plays a major role in Educational Data Mining. It can be accurately observed by identifying the key predictors of students' academic performance with the full use of online discussion forum [36]. Many types of research have focused on building models for predicting student performance in online collaborative learning. Cheng et al. [53] have analysed the frequency of access and the duration of sessions to categorise the students using cluster analysis. Using the number of messages written and read on the forum by students, Pena-Shaff and Nicholas [4] have built up a model for a student activity. Based on the number of discussions created, posts created, discussion and module course viewed and some other quantitative attributes, a new model was proposed by Widyahastuti et al. [13] to predict students' performance in the online discussion forum.

Moreover, using classification, Romero et al. [7] investigated that whether all the forum interaction data are relevant to predict their performance or can similar accuracy be obtained through just considering one subset of obtained data.

CHAPTER 3: METHODOLOGY

This research was conducted as a case study followed by a data-driven approach. The methodology followed in this study includes the tasks listed below (See Figure 3.1).

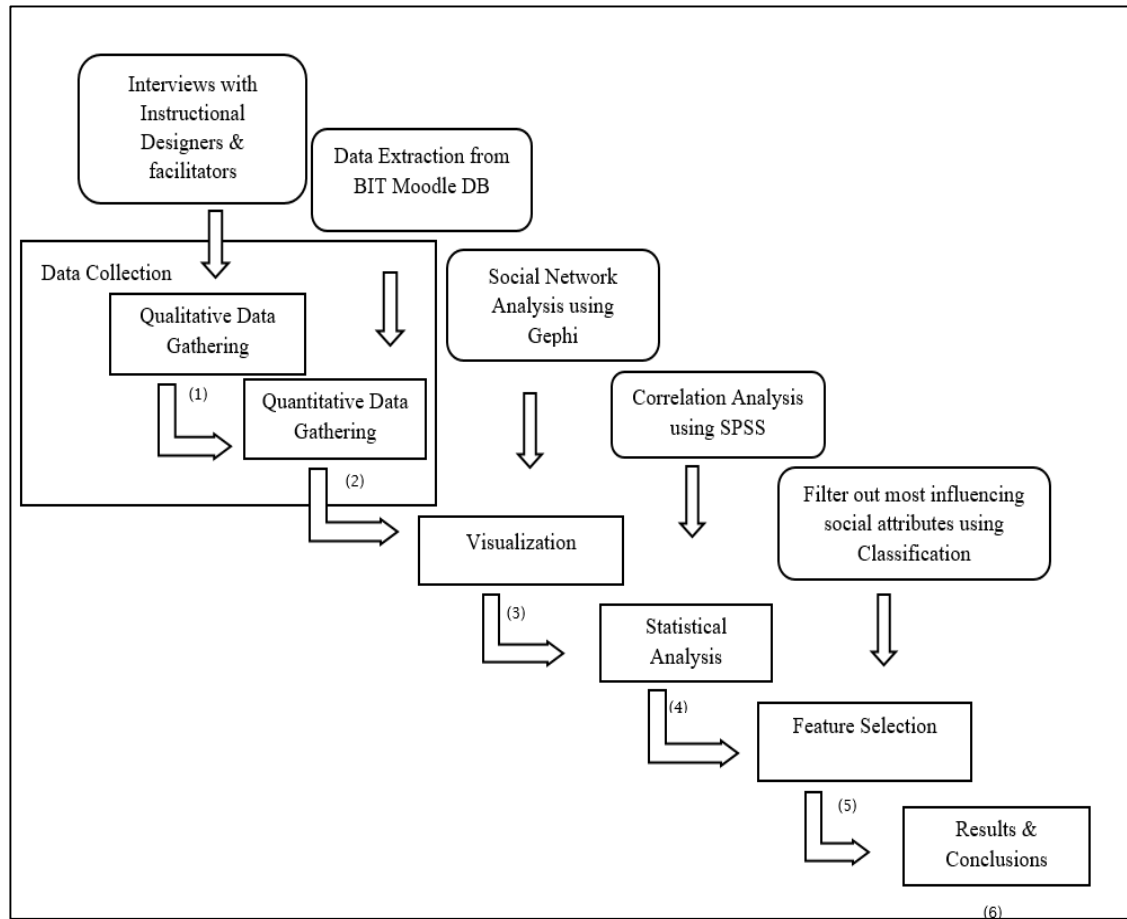


Figure 3.1: Methodology

Figure 3.1 graphically demonstrates the steps performed in the study. Steps are further described in the upcoming sections.

- 1 Data Collection: Qualitative Data gathering
- 2 Data Collection: Quantitative Data Gathering
- 3 Visualization
- 4 Statistical Analysis
- 5 Feature Selection
- 6 Results and Conclusions

3.1 Step 1 and Step2: Data Collection

This research followed a mixed approach to describe and analyse the impact of interactions among students in discussion forums towards their academic performance. The data gathered from interviews were qualitative data and they provided information about the BIT online learning environment describing the structure of the course, instructional design, course delivery methods and the student support. The data retrieved from BIT Moodle database can be treated as quantitative data to analyse interaction patterns and its relation towards students' performance in assessments.

3.1.1 Step 1: Qualitative Data Gathering

Course design is a crucial aspect for any research which focuses on analysing student behaviour associated with learning materials. In a learning environment where course materials and assessment materials designed by two different parties, it is essential to understand whether these two components were designed to align with the learning objectives of a particular course. Therefore, a face to face interview was conducted with two BIT facilitators in order to collect information on BIT course design, course delivery and how the online discussion forums and assessments have been designed to align with the learning objectives of a course (Step 1 in Figure 3.1). Questions that were used in the interview are attached in Appendix A.

3.1.2 Step 2: Quantitative Data Gathering

Structured Query Language (SQL) was used to extract interaction data and assessment grade data from the Moodle LMS database (Step 2 in Figure 3.1) and exported it to a table (spreadsheet). Extracted data consist Table 3.1 list all the courses from which the data were extracted for the present study.

Course Number	Course Name
IT1105	Information Systems and Technology
IT1205	Computer Systems I
IT1305	Web Application Development I

Table 3.1: BIT Course

The extracted dataset was then prepared into three datasets followed by the course name. Collected data contained records from January to end of May 2018. Finally, the gathered data were pre-processed in order to visualise students' social networks. The queries used to extract data from database are attached in Appendix B.

3.1.3 Ethical Considerations

Protecting Students' privacy is a major concern in any research which collects students' data. Since institutions are responsible to protect students' personal information, they are reluctant to provide it to outside parties without having fair reasons. Hence, authors of this paper prepared a letter requesting permission to get the required data extracted from the database of BIT VLE for this study. The letter clearly described and assured that prior to data extraction the data will be anonymized in order to preserve the confidential information of the students. Therefore, students' identities were anonymised at the point of data extraction. Also, the permission to extract data from the server was obtained by submitting a letter to the director of UCSC. The data used in this study was collected on permission and advice of the responsible authority. The letter prepared to seek permission to collect data from the BIT database is attached in Appendix C.

3.1.4 Data Cleaning and Preparation

First, forums which were categorized as 'announcements' were removed as those are only used by facilitators to send 'news and announcements' to the students. Second, discussion threads that are not related to subject context were removed. With respect to forum messages, several messages from all three courses had to be removed as those were not relevant with the subject context. Here the study only considered the messages and discussion threads which are more appropriate and relevant to the subject area.

Course	Number of records	Number of forums	Number of discussion threads	Number of posts	Number of students who have posted	Students' coding	Facilitator code	Forum Topic code
IT1305	Before the data	27	93	123	30	S1 to S29	F1	FT

	cleaning							
	After the data cleaning	25	84	114	29			
IT1205	Before the data cleaning	17	32	99	57	S1 to S57	F2	FT
	After the data cleaning	16	26	93	57			
IT1105	Before the data cleaning	26	127	233	64	S1 to S64	F3	FT
	After the data cleaning	23	115	219	64			

Table 3.2: Data cleaning process

The first course selected to analyse was ‘IT 1305 Web Application Development I’. Initially there were 30 records of students who have posted in the forum. However due to not having details of the target of the post one record was removed. Then data cleaning was carried out for other courses ‘Information Systems and Technology’ and ‘Computer Systems I’ respectively. After the data cleaning, students were coded into meaningful representation starting from ‘S1’. Also, the, facilitators for the courses are denoted by ‘F’ notation and Forum topic is denoted by ‘FT’ notation. All the details regarding data cleaning process is depicted in Table 3.2. In the next step, for each course two Gephi compatible excel sheets were prepared, one file containing nodes (forum participants) and other file containing edges (relevant interactions).

Finally, prepared MS Excel sheets were imported to Gephi tool in order to visualize the interactions of forum participants. Visualisation process is explained under the section 3.2.

3.2 Step 3: Visualisation

Visualisation is the process of analysing and depicting the nature of interactions between actors in a social network. By analysing the students' social networks, we can obtain several network parameters with respect to the social capital of a student. Then these network parameters can be used for further analysis to find a relationship between student's achievement [6]. Therefore, as the first step of the analysis, forum interactions of the participants were visualised using social network analysis techniques provided through various functions in Gephi tool (Step 3 in Figure 3.1). Gephi is an open source SNA application which can be used for interactive visual exploration of networks. It aids visualising all kinds of networks and complex systems with dynamic and hierarchical graphs [6]. Gephi, not only facilitates network visualisation but also provides calculations for SNA parameters, spatializing, node/edges partitioning, ranking nodes and filtering. [6], [23]. It has multiple inbuilt algorithms/layouts for network visualisation and among those 'Forced Atlas' layout is popular in the research field [6]. When compared to other SNA tools such as SNAPP, Meerkat-Ed or Forum Graph, Gephi inherits specific functionality; 'dynamic mode' which enables the user to visualise the network evolution by reflecting the changes in node position in real time [6], [54]. This technique helps to speed up the exploration and make it easy to work with complex data sets and produce valuable visual results [54].

Network Quantitative Analysis

When the previously made data-files were imported into Gephi, it provided two things as the output. First, a sociogram which represents the social network build behind the forum discussions. Second, it provided some quantitative parameters that calculated for each sociogram, which describes the position of each student in that social network. By using these parameters, we can quantify the connectedness and relations of students in a network. Centrality is the construct used to indicate how prominent a particular student is in a network or how important is that student to the communication of information [32]. The social network parameters that were calculated for each social network were mainly two types; network-level parameters, which is calculated for the entire network and user-level parameters that are calculated for each student. Network-level parameters that were calculated by Gephi are listed in Table 3.3.

Parameter	Description
Network size	The total amount of students in a network. If this is high, the level of student participation for the forum is high. [6]
Average degree	The average number of posts posted and received by each student. This measure implies the average level of interactivity of students in the forum. [6]
Network density	The ratio of actual interactions between students to the total possible. If this is high, students are participating efficiently and get the maximum benefit out of the forums. [6], [8], [55]
Diameter	The largest number of students needed to pass over to come from a particular student to another. Low diameter makes easy to interact with peers. [8], [55]

Table 3.3: Network-level parameters

Following are the other type of parameters; user-level parameters, which provides metrics to understand the statistical properties of the students in the graph (See Table 3.4).

Parameter	Description
In-degree centrality	The number of replies received by each student. Indicates the popularity/attractiveness of a student and peers are more likely to interact with this type of students. [6], [7], [8]
Out-degree centrality	The number of posts/messages posted by each student in the forum. An indicator of how active a student in the discussion. [6], [7], [8]
Degree centrality	The total number of messages posted and received. It is the sum of in-degree centrality and out-degree centrality. Implies how influential a student within the network. [6], [8], [15]
Betweenness centrality	The number of times a student comes in-between others. In this way, the participant connects the unconnected peers and thus facilitates communications and acts as a bridge or broker of information exchange. So, helps to identify which students and facilitators may spread the information quickly and effectively across the class. [6], [8], [15], [55]

Closeness centrality	The inverse of the distance between a student and all other peers. Indicates how close a student to his peers and therefore how easy to reach and interact with others. [8], [55], [56]
Harmonic Closeness Centrality	The sum of inverses of distances between a student and all other peers. Implies how close a student to his peers. [57]
Eigenvectors centrality	Estimates the social capital and the influence of one's ego network. Connections to well-connected or important students in the network bring high values. [6], [15], [55]
Eccentricity	How far a student from his peers (level of isolation). Higher values indicate less connectedness to peers. Therefore, difficult to reach. [6], [55]
Clustering Coefficient	The proportion of actual edges between a student and his neighbour peers to the total possible edges. High values impress the student more likely to work with peers in the group. [6], [55]
PageRank	Rank the students in the network according to their importance. [15], [55]
Hub	How many highly informative (important) students a particular student is pointing into. Students recommend each other based on the information they share. [7], [8], [15], [55]
Authority	The amount of valuable information a particular student holds, and it helps to identify the students with higher knowledge or skills [7], [8], [15], [55].

Table 3.4: Individual-level parameters

Several network measures imply several perspectives of students' behaviour. *In-degree centrality*, *out-degree centrality* and *degree centrality* indicate the number of interactions, where *betweenness centrality*, *closeness centrality* and *harmonic closeness centrality* implies the students' role in moderating the discussions. Furthermore, *eigenvector centrality*, *eccentricity* and *clustering coefficient* measure the connectedness of peers in the social network. Likewise, each student's position in the social network indicates his/her level of social engagement.

3.3 Step 4: Statistical Analysis

Social attributes obtained from social network analysis were then mapped against assessment marks of each student to measure correlation coefficient between ranked variables

(Step 4 in Figure 3.1). Since network data are prone to violate the traditional assumptions of conventional statistics (normal distribution and independence) [59] selecting the Pearson correlation might not give correct correlations as the output [60]. Because Pearson correlation requires data to be normally distributed, absence of outliers, linear, and homoscedasticity [58].

However, social network data in most circumstances may not follow all of the above requirements [59]. Instead, the relationship between the two variables can be better described by Kendall's Tau-B test which is a nonparametric equivalent to Pearson's correlation [58]. Kendall Tau-B test can be used to find association that exists between two variables measured on at least an ordinal or continuous scale [6]. Therefore, Correlation coefficients were calculated using Kendall Tau-B test supported by SPSS software which is widely used in social science researches to perform statistical analysis of data [6], [58], [61].

3.4 Step 5: Feature Selection

Using statistical analysis, the study could identify whether each student's participation, connectedness, and closeness in the social network are correlated with student's academic performance and to what extent. Next, the study further focused on what are the most powerful predictors of performance among them. As described in section 2.4.2.1, using a subset of these parameters instead of all available social network parameters might lead to enhance the accuracy of predicting their academic performance (Step 5 in Figure 3.1). So, a method was needed to filter out the best subset of social network parameters which describes the students' academic performance well. First, the derived social network parameters were sorted based on their correlation with the assessment marks from highest correlation to the lowest correlation. Then, by removing the attribute from the bottom of the list (lowest correlation) in each iteration, classification was performed as depicted in Figure 3.2. For instance, if there are altogether N number of social network parameters sorted on their correlation, First the entire set's classification accuracy is recorded as 'Classification Accuracy (N)'. Next, by removing the attribute in the bottom of the list, now the classification of the subset N-1 is recorded. Likewise, until only the highest social network parameter remains, the classification accuracies were recorded. This is to identify, how much the feature subset selected in each iteration can accurately predict the academic performance. The study used two classification algorithms; Naive Bayes and Random Forest, which have reported high classification accuracies in similar researches when classifying students' performance using forum data [7], [8]. Naive Bayes is considered as a classifier that quickly learn to use high dimensional features with limited training data compared to more

sophisticated methods. Also, it is able to extract more effective information even from a small data set [50]. Here, this case study used cross validation folds 10. That means, the Naive Bayes classifier randomly split the dataset into 10 partitions, fit the model into 9 partitions and use the remaining portion for validation. This iterates for 10 times and therefore, every single partition involves in both training and testing, therefore it allows us to utilize our data better. Random Forest is a collection of decision trees where finally all the results are aggregated into a single output. As it trains on different samples of data, the variance is reduced and overfitting is limited [62]. The number of trees used in Random Forest algorithm is its default value of 100. The tool used for the feature selection was 'Weka', which is widely used by Educational Data Mining field [7], [8]. After recording the classification accuracies obtained from two different algorithms, the results of the higher accuracies were considered. For instance, according to Figure 3.2, when removing the social network parameters one by one, classification accuracies have been increased, but when removing the bottom parameter from feature set N-2, suddenly the classification accuracy was dropped a bit. Therefore, it indicates, the feature set should include N-2 features. Likewise, when observing the results, the feature subset that has contributed to higher accuracies is taken as the powerful performance predictors of each course.

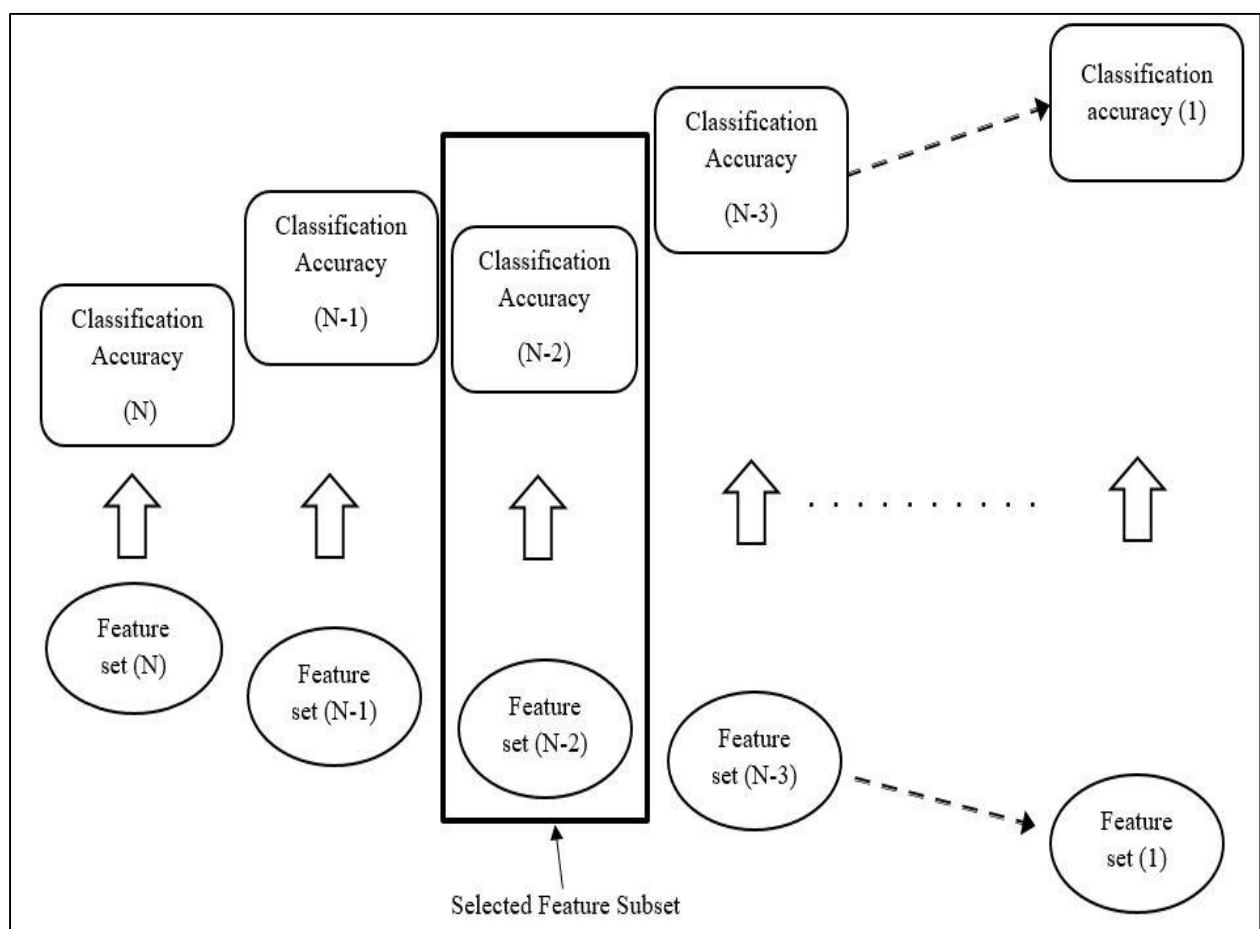


Figure 3.2: Feature Selection Process

3.5 Step 6: Results and Conclusions

Finally, considering all the results derived from each of the above steps, several conclusions (Step 6 in Figure 3.1) were drawn. The findings from this research may help instructors and instructional designers of the online courses to identify the gaps and pitfalls in the collaborative learning process.

3.6 Prototype Tool

As mentioned in Section 1.1 in order to select the best e-learner, BIT e-learning centre at UCSC use manual mechanism which considers the forum posts, marks along with a defined threshold value. Even though the criterion used by existing method is adequate it does not reflect the total value of a student collaboration with their peers. Therefore, we developed a prototype tool to evaluate and monitor student participation considering both student social capital and their assessment mark. The developed tool is able to rank the students in a course based on eigenvector centrality values calculated considering the messages posted by them

CHAPTER 4: ANALYSIS AND RESULTS

The data were drawn from discussion threads and assessments were visually and mathematically analysed on course level. Forum interactions were obtained from the number of posts, here a post was considered as either a message or a comment a student posted, or replied to, in the forum. Students' performance was measured by their marks for the online assessment component of a course. In total there were **426 posts** and **150 students** who have posted at least once in a forum. Forum interactions were analysed using SNA functionalities provided by Gephi tool. Results from SNA were organized into two sections (Section 4.1 and 4.2) considering two sub research questions. Then the social parameters obtained from SNA were correlated with students' assessment mark. Overall results revealed a positive correlation among social networking attributes and students' academic performance (described in Section 4.3). However, since correlation values were subjective to the course, classification algorithms had to be used to find the most prominent social networking parameters that influence one's social capital which in turns helps for better academic performance. Results showed *eigen centrality*, *degree centrality* as the most prominent social networking parameters which were common to all three courses. The analysis and results are detailly described in the next section.

4.1 Factors Affecting Students' Social Engagement in Discussion Forums (course - level)

Interactions in online discussion forums can be interpreted in a social network point of view considering different approaches [43]. One such approach to visualise the interactions of forum participants is considering their role in which they mark the presence in the forum. Therefore, forum interactions of students were mapped on two different sociograms. By analysing the sociograms, study could identify significant variations in the graphs and therefore it provided a notion to identify the factors which may have affected students' social engagement in online discussion forums.

The first set of sociograms (Figure 4.1, Figure 4.2, and Figure 4.3) were created considering all the posts of forum participants for each course. Each sociogram outlined the structure of the course and the patterns of interactions. Each node (circle) in the sociograms denotes to a forum participant, each edge (arrow) corresponds to forum interaction, the arrowheads represent the direction of the interaction. The size of each node is relative to its *degree centrality*, colour intensity relative to the *betweenness centrality*, and the thickness of edges

represents the frequency of interaction. Hence, larger the node it has higher *degree centrality* and darker the node its *betweenness centrality* is high.

When analysing each sociogram, we could identify a variation in prominent nodes in the networks. For instance, Figure 4.2 shows the instructor (F) being central to all interactions and receiving most connections (highest prestige). However, in the other two courses students have posted new messages to the forum topic more often so that 'FT' has been receiving the greatest number of interactions (see Figure 4.1 and Figure 4.3). By interpreting the role of the participant in these built networks, we could identify three main interaction types.

- I. Student - Student interactions
- II. Student - Facilitator interactions
- III. Student - Content interactions

By considering these identified types regarding quantity and influence, it helps to provide a general idea about the course structure. Moreover, it could be observed that the quantity of interactions for each respective type and influence from students were varying depending on the forum design and the interventions of the facilitator. From the sociograms, for two courses (Figure 4.1 and Figure 4.3) forum topic being the largest node it was obvious that student - content interactions were dominating than Student - Facilitator interactions and Student - Student interactions. One possible reason for having a large number of Student - Content interactions, could be the forum structure defined by the Instructional Designers. Since the design of a forum relies on interactions usually started by a student, it was intended that when a facilitator introduced a topic for the discussion, students may initially post to the forum topic. Therefore, the sociograms were well aligned with the instructional design of the selected courses.

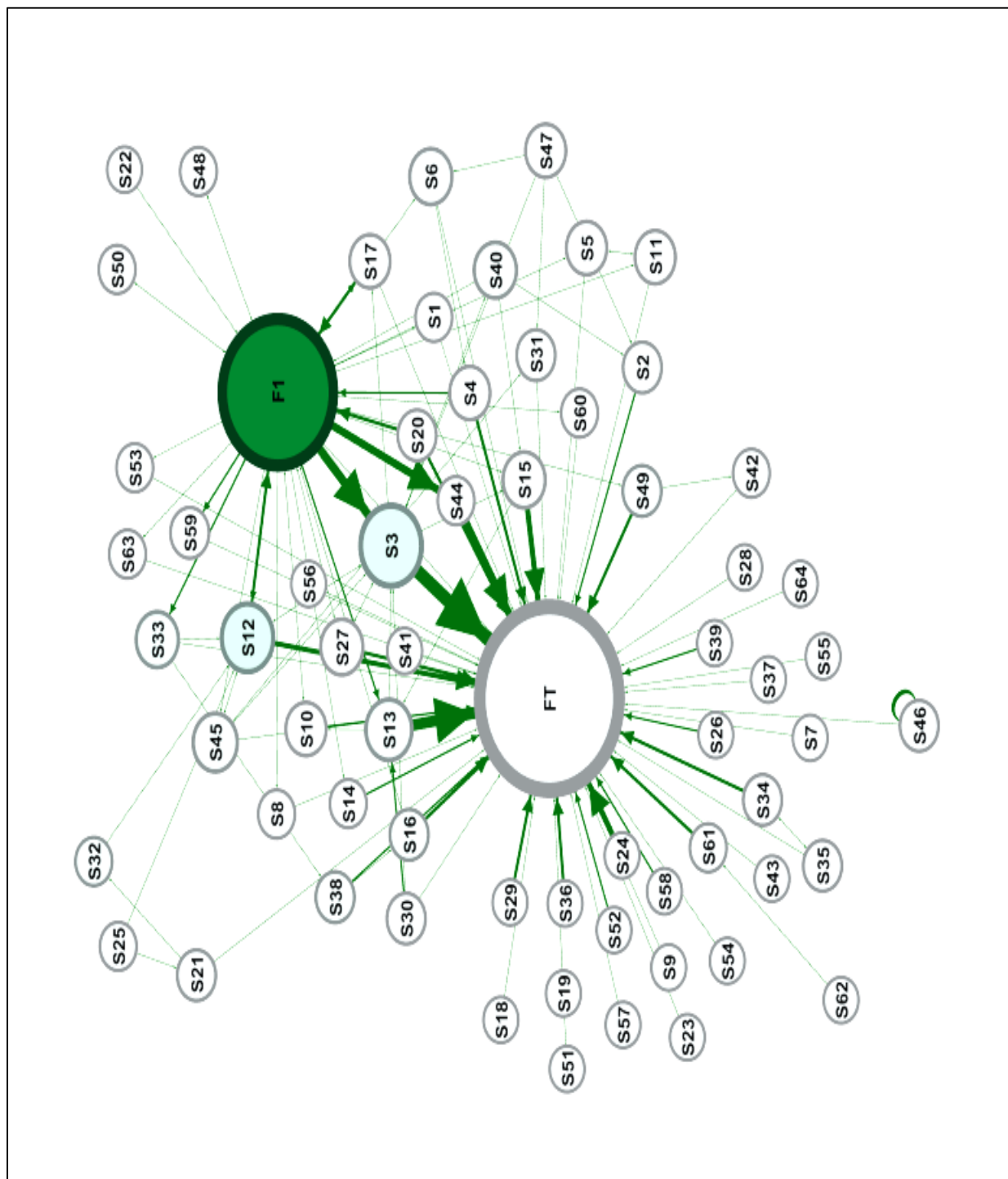


Figure 4.1: This graph summarises all the interactions of course “Information Systems and Technology”

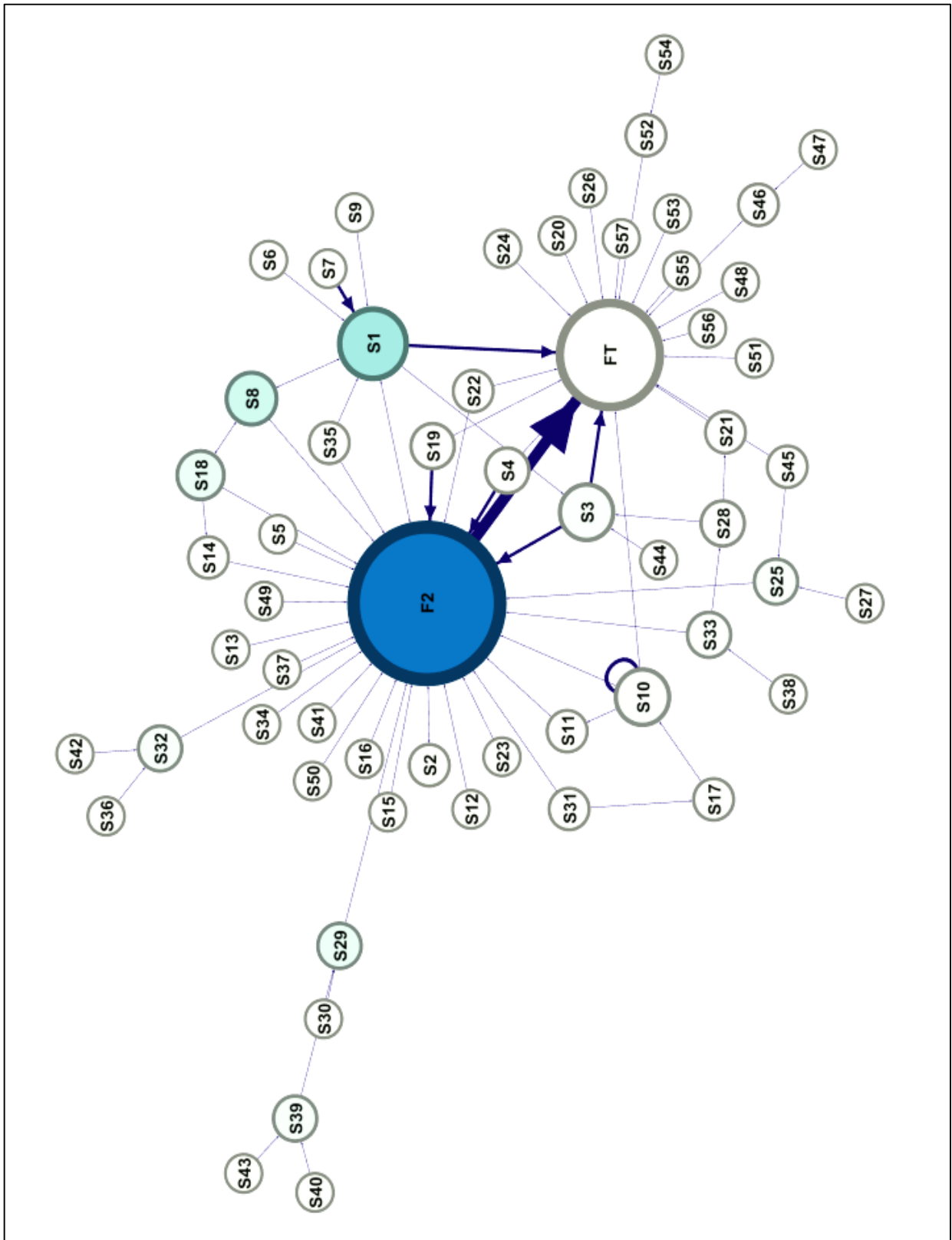


Figure 4.2: This graph summarises all the interactions of course “Computer System I.”

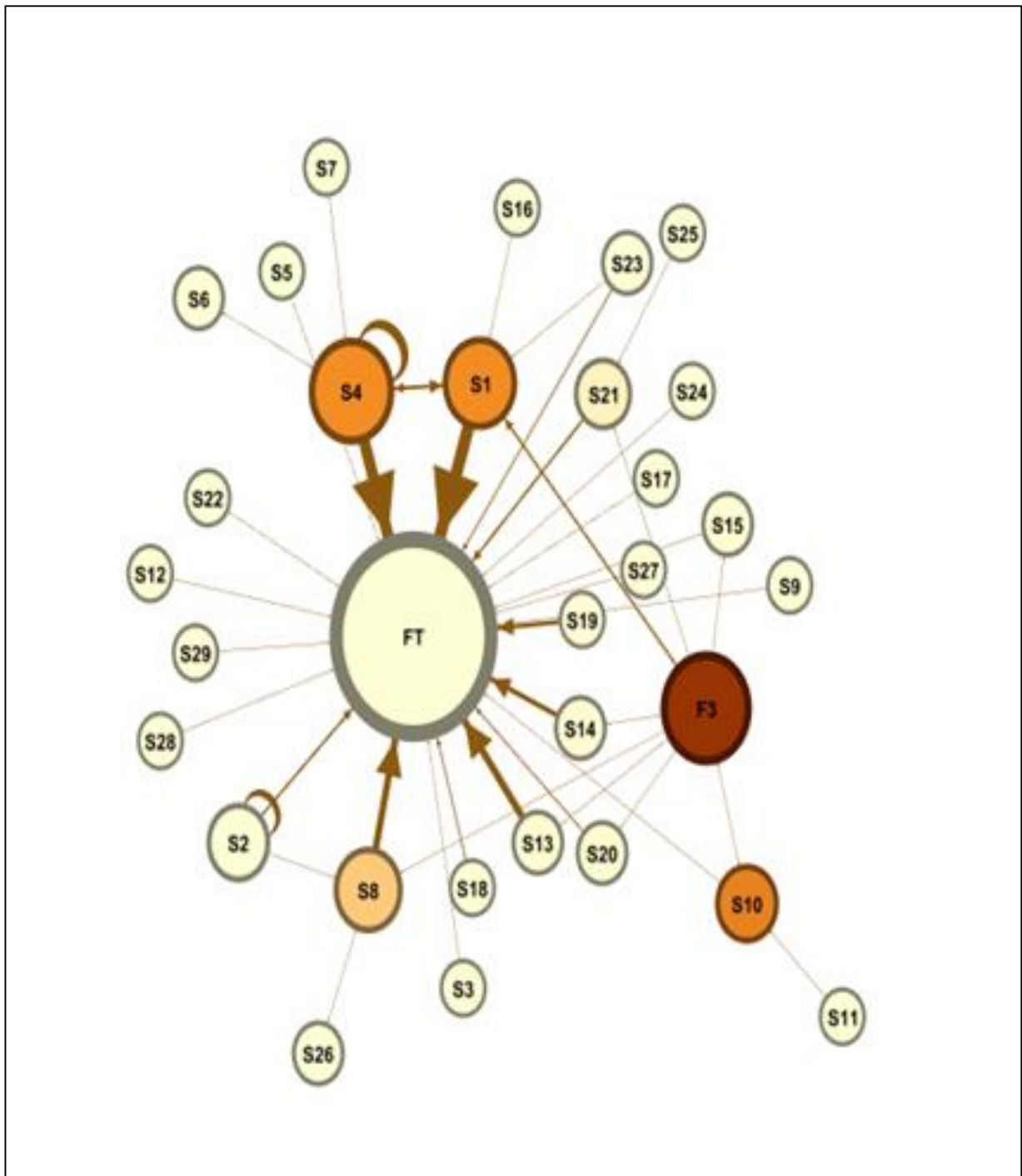


Figure 4.3: This graph summarises all the interactions of course “Web Application and Development I”

Interventions of the facilitator could be another factor that caused variations in the interactions which was visible in the sociograms. For an instance, the facilitator in “Computer Systems I” (see Figure 4.2) course has guided the students well throughout the discussion context. Most edges pointing to him and node colour being the darkest, it indicates that he is receiving most of the interactions (*high degree*) and has become the centre to the network (*betweenness centrality*). Therefore, in this course facilitator was receiving most of the messages and at the same time he was actively participating in the entire discussion environment. Therefore, for that course, there was a higher number of Student - Facilitator interactions. In fact, a facilitator in an online course context should be able to acknowledge, motivate, advice, provide instructions and guide students appropriately. Therefore, it seemed in this course context facilitator has actively involved with students which in turns has helped to keep the flow of information and encourage active engagement of students.

Apart from the forum design and interventions of the facilitator, course context also could have influenced to cause such variation in the graphs. For instance, even though there are a smaller number of students’ engagement in discussions for the course “Web Application and Development I” (see Figure 4.3), the information passing among students seemed to be high (larger, darker nodes). This is an indication where students have played an important role in moderating information between their peers by continuously communicating with them rather than just starting a new discussion thread in the forum every time when they post a message. Also, this kind of similar behaviour can be observed for the course “Computer Systems I” (see Figure 4.2) However, for the course “Information Systems and Technology “(see Figure 4.1) which is a theoretical course most of the students have replied to the forum topic by starting a new discussion thread rather than replying to the existing discussions threads of the facilitator and their peers. These findings indicate that students in theoretical courses were reluctant to participate actively in forums and collaborate with their peers to keep the flow of information in the forum. Given that courses “Web Application and Development I”, and “Computer Systems I” are more into practical aspect it showed students in practical courses were actively participating and moderating the discussions than in theoretical courses. Therefore, it seems Peer-learning and Peer-teaching concepts have been widely practised in practical courses than in the theoretical course. This is interesting finding which Instructional designers should consider when designing the learning materials.

Finally, by considering the time stamps of the posts, a time-lapse video (attached in the CD submitted) was created to visualise the evolution of the network over the whole duration of the course. The time series analysis done using the tool ‘Gephi’ was basically conducted in two methods. The first method is the fixed version of the social network evolution by means it shows

the presence of the actors once they joined to the network, and until the semester ends, they stay in the network. But in the second method, the dynamic version of the social network evolution, after a particular actor joined to the network, if he /she doesn't continuously engage in the forum, he/she get disappeared. Again, when a new interaction occurs later, again that actor appears in the network. Visualising the network regarding time revealed that most of the connections initiated during mid of January boomed till April, then gradually started to decrease and finally, disappeared during the end of May. As usually the first semester of BIT programme commences on January and students have not engaged much in the first half of the month. However, the network tended to grow wider in the mid of the semester. As the examinations were starting from May, the interactions were diminishing at the end of the semester. Moreover, it could be observed that in course 'IT 1305- Web Application Development I', facilitator has guided the students throughout the semester where in other two courses facilitator's involvement has been decreased towards the semester end. Therefore, time could be another prospect to have variations of interactions among the graphs.

4.2 Student Behaviour in Discussion Forums

In order to provide further insights from the social network analysis of the three datasets, social networking measures were calculated to find the social cohesiveness and the centrality in the built social networks. These measures were obtained considering the whole network of students as well as considering an individual student.

4.2.1 Network Level social Parameters

A network analysis was done to describe the overall linkage between the participants. The intensity of the engagement of participants in the network was measured using several measures.

Course	Network size	Average Degree	Network Density	Diameter
'Information Systems and Technology'	66	2.061	0.032	6
'Computer System I'	59	1.39	0.024	7

'Web Application and Development, I'	31	1.484	0.049	5
--------------------------------------	----	-------	-------	---

Table 4.1: Results Network level social parameters

When considering the *network size* and *average degree* measures, “Information Systems and Technology” course tend to has the highest values. This means that majority of the students of that course have participated in the forums. Although there were a smaller number of students’ participation in forums (indicated by *Network size*) for “Web Application and Development I” course, its *density* value was significantly high when compared to other two courses. Since the *density* measures the efficient participation among the participants in the forums, it seems students involved in other two courses were reluctant to actively post messages in the forums. When considering *diameter* measure course “Computer System I” tend to has the highest value. This could happen due to facilitator being the central node in this network. This means that majority of students have connected through the facilitator node and not with a fellow student. Since high *diameter* implies that students have to pass a large number of students to come from a particular student to another, it makes difficult to interact with their peers.

4.2.2 Node (user) Level Social Parameters

Although it is important to have an overall view of the status of collaborative learning in a course, it is more important to find prominence of a node (student) in order to interpret its link with the performance. Therefore, social parameters for each student were calculated considering the criteria defined in the ‘network quantitative analysis’ (Section 3.2). Figure 4.4, Figure 4.5 and Figure 4.6 show information-giving graphs indicating information spread between students and the facilitators. Node size was configured by *outdegree centrality* (outgoing interactions) to demonstrate the information giving participants, where students with more outgoing interactions have larger nodes. Colour intensity relative to the *eigenvector centrality* and the thickness of edges represents the frequency of interaction. Therefore, darker the node the eigenvalue of the node is high which means the node has well-connected neighbours who were the important nodes of the network. *Eigenvector centrality* is a good ‘all-round’ SNA measure, which is handy for visualising human social networks. Also, *out-degree centrality* is important to find well connected forum participants, and forum participants who are likely to hold most information or forum participants who can quickly connect with the wider network. Hence *out-degree* and *eigenvector* centralities are ideal for configuring the nodes in sociograms.

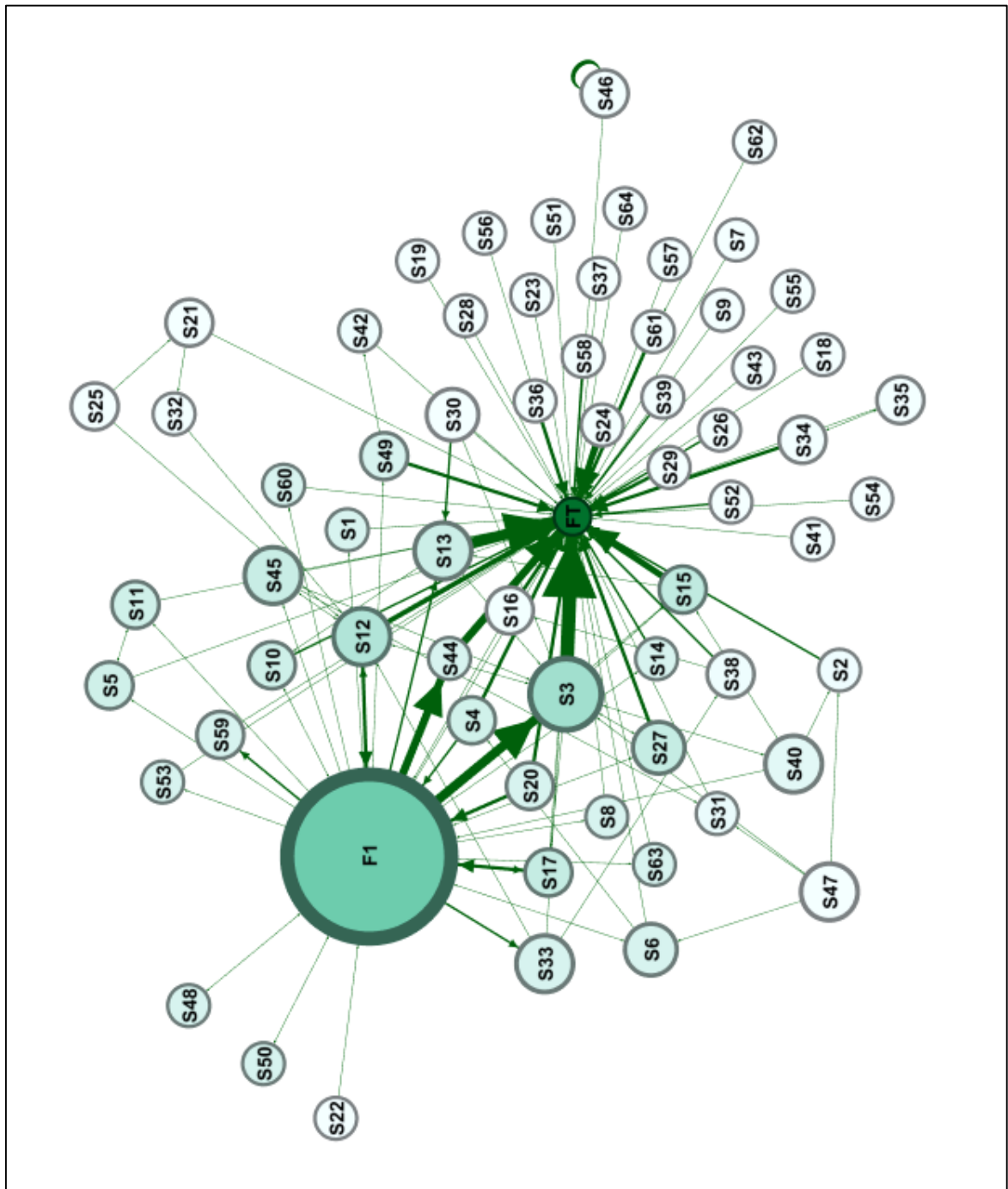


Figure 4.4: Information Giving Network of course “Information Systems and Technology”

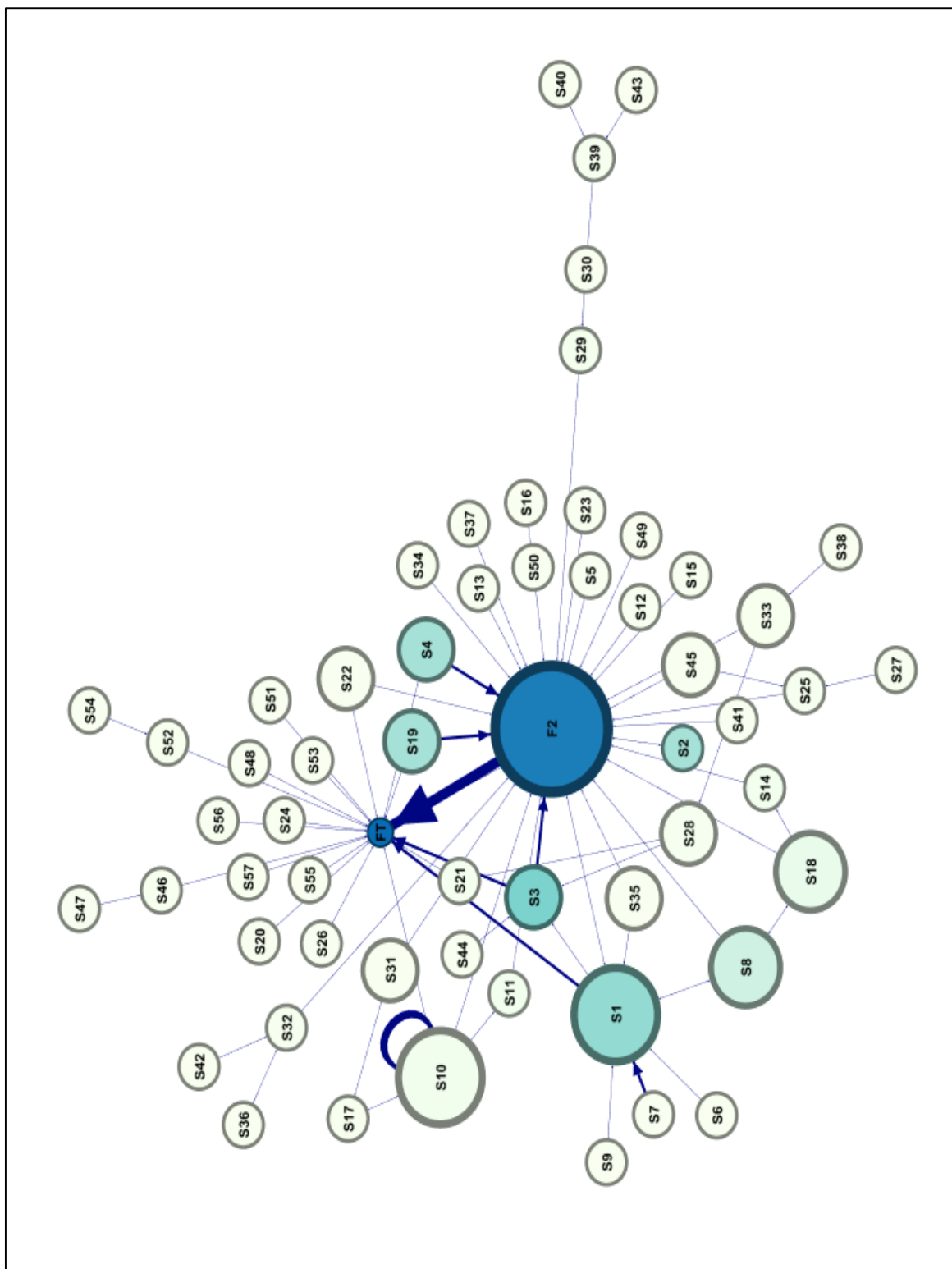


Figure 4.5: Information Giving Network of course "Computer System"

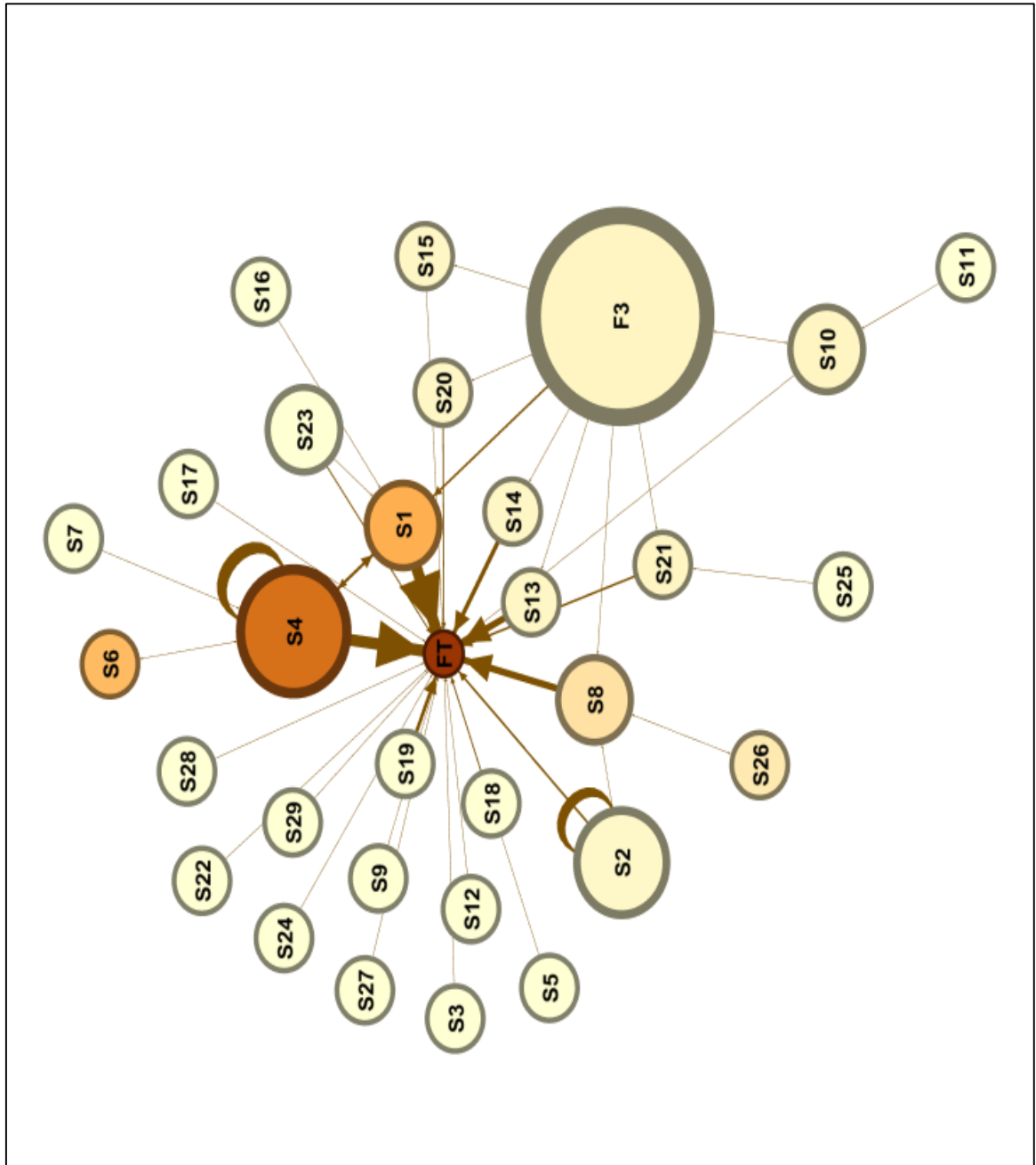


Figure 4.6: Information Giving Network of course "Web Application and Development I"

Information giving network graph of course “Information Systems and Technology” (demonstrated by Figure 4.4) shows that most students were actively participating in discussions, and it shows the interventions of the facilitator. Students S3, S12, S45 tend to have high *out-degree* values (large node size) and high *eigen centrality* values (dark node colour). Likewise, information giving network graphs for course ‘Computer System I’ (see Figure 4.5) and ‘Web Application and Development I.’ (see Figure 4.6) shows that, many students were actively participating in discussions. In Figure 4.5, it seems network was dominated by students like S1, S8, S10, S18 (larger node size) who had the highest prestige and it shows the higher number of interventions from the facilitator. Also, according Figure 4.5, S1, S3, S4, S19 tend to have higher social capital indicated by high *eigen centrality* (darker nodes). In Figure 4.6, network was dominated by students like S2, S4, S8, S10 who had the highest prestige (larger node size), and S4, S1, S6 were the students who had higher social capital indicated by high *eigen centrality* values (dark nodes).

When analysing all the information giving graphs (Figure 4.4, Figure 4.5 and Figure 4.6), it seemed S1, S3, and S4 students had higher centrality values for selected social parameters (*Out-degree and Eigen centrality*) which is an indication of active participation and existence of higher social capital. Also, these students were influential actors who were important to the flow of information across the network (*betweenness centrality*). Interestingly not only these students actively participated in the forum but also, they have scored higher marks for the assessments. Thus, with this analysis, we can conclude that students with a high number of interactions in forums and high social capital are likely to get good scores, a fact to be analysed thoroughly.

4.3 Correlation of the social network parameters with students’ assessment marks

Since visualisation provided a notion on the existence of the link between social network parameters and student’ assessment marks, it should be further clarified by correlating those social network parameters with assessment marks of students. Therefore, as the next step statistical analysis was performed using Kendall’s Tau-B test to obtain a correlation between the SNA parameters and students’ marks. Next, a feature selection using classification algorithms were done to identify the most powerful social network parameters among all available parameters which can be then used for gaining insights on their online behaviour and academic performance.

4.3.1 Correlation Analysis

Correlation was obtained considering pass mark (assessment mark) as the dependent variable and social network parameters (see Table 3.3) as independent variables. The analysis was done with the condition; if the sig. (2-tailed) value is less than or equal to 0.05, the correlation value is significant, or it can be proposed that there is a significant correlation. If the sig. (2-tailed) value is more than 0.05; the correlation value is less significant, so there is a relatively low correlation between the two variables [13].

As depicted in Table 4.2 in the course “Information Systems and Technology”, all the parameters except *eccentricity* show a positive relationship with the pass mark. Since *eccentricity* implies how far a student from his peers, it suggests that the less connected or isolated students tend to perform poorly in assessments. Moreover, it is visible that *eccentricity* (- 0.002 and 0.984) and *pageRanks* (0.006 and 0.951) have considerably low correlations with learning performance. Therefore, these two variables are not much suitable for explaining the variation of performance based on students’ behaviour. Moderate correlations were reported by *out-degree centrality* (0.135) and *hub* (0.134) which possess the largest values for this course. The remaining parameters’ correlations are decreased in order as *eigenvector centrality*, *authority*, *degree centrality*, *in-degree centrality*, *clustering coefficient*, *betweenness centrality*, *harmonic closeness centrality*, *closeness centrality*, *page ranks* and *eccentricity*.

		Assessment Mark		
		Correlation Coefficient	Sig. (2-tailed)	N
Kendall's Tau-b	Outdegree	.135	.173	64
	Hub	.134	.156	64
	Eigenvector centrality	.077	.409	64
	Authority	.069	.462	64
	Degree	.057	.543	64
	Indegree	.056	.567	64
	Clustering coefficient	.050	.602	64
	Betweenness centrality	.043	.663	64
	Closeness centrality	.024	.802	64
	Harmonic closeness centrality	.024	.802	64
	PageRanks	.006	.951	64
	Eccentricity	-.002	.984	64

Table 4.2: Correlation for the course - “Information Systems and Technology”

According to Table 4.3 in course, “Computer Systems I” also, all the parameters except *eccentricity* show a positive relationship with assessment mark. There are two significant correlations for *out-degree centrality*, *harmonic closeness centrality*. In a row, the values of correlation and sig. (2-tailed) of the variables are (0.249 and 0.023), and (0.193 and 0.047). That means the number of posts by each student and how many peers a particular student is pointing to has an impact on their performance in assessments. The next highest values reported by *closeness centrality* (0.186 and 0.059), *clustering coefficient* (0.185 and 0.088), *authority* (0.174 and 0.102), *eigenvector centrality* (0.145 and 0.164) and *degree centrality* (0.145 and 0.166) imply a moderate correlation with assessment mark. The rest, *pageRanks*, *hub*, *in-degree centrality* and *betweenness centrality* have a low correlation with the pass mark.

		Pass Mark		
		Correlation Coefficient	Sig. (2-tailed)	N
Kendall's Tau-b	Out-degree	.249	.023	57
	Harmonic closeness centrality	.193	.047	57
	Closeness centrality	.186	.059	57
	Clustering coefficient	.185	.088	57
	Authority	.174	.102	57
	Degree	.145	.166	57
	Eigenvector centrality	.145	.164	57
	PageRanks	.113	.275	57
	Hub	.101	.308	57
	In-degree	.095	.376	57
	Betweenness centrality	.062	.554	57
	Eccentricity	-.146	.152	57

Table 4.3: Correlation for the course - “Computer Systems I.”

As depicted in Table 4.4 in the course “Web Application and Development I”, *eccentricity* shows a positive relationship with the pass mark. It implies that, whether a student is far from the peers, it has not been a barrier to collaborate and learn. Additionally, in this course, *closeness centrality* and *harmonic closeness centrality* also show negative correlations. That means, even a student is close to the peers, that has not much affected on their learning performance. Altogether, this implies that, whether a student is reachable or not to others in the network is not a considerable matter in how they perform in this course. That means the distance between the students has not become a barrier for their collaboration. There are six significant values regarding the correlations

between assessment marks and *in-degree centrality* (0.409 and 0.006), *eigenvector centrality* (0.390 and 0.007), *degree-centrality* (0.386 and 0.009), *pageRanks* (0.381 and 0.009), *authority* (0.343 and 0.019), and *betweenness centrality* (0.330 and 0.027). This explains that not only the number of interactions but their connectedness and mediation also significant for their performance. The *out-degree centrality* (0.242 and 0.115) shows a moderate correlation. The rest of parameters *eccentricity*, *clustering*, *hub*, *harmonic closeness centrality* and *closeness centrality* have relatively low correlations with assessment marks.

		Pass Mark		
		Correlation Coefficient	Sig. (2-tailed)	N
Kendall's Tau-b	In-degree	.409	.006	29
	Eigenvector centrality	.390	.007	29
	Degree	.386	.009	29
	PageRanks	.381	.009	29
	Authority	.343	.019	29
	Betweenness centrality	.330	.027	29
	Out-degree	.242	.115	29
	Eccentricity	.138	.358	29
	clustering	.127	.405	29
	Hub	.119	.411	29
	Closeness centrality	-.111	.449	29
	Harmonic closeness centrality	-.114	.436	29

Table 4.4: Correlations for course - "Web Application and Development I"

Likewise, for different courses, the best correlating social network parameters also different. For instance, when the distance between students became a barrier for students to collaborate properly in one course, it does not matter for the other. So that the reason behind this might be the differences between course contexts and their design.

4.3.2 Feature Selection using Classification

Using the above correlations of social network parameters with academic performance, as mentioned in section 4.2.1, social network parameters were sorted from highest correlation to the lowest in all three courses as depicted in Table 4.5.

IT1105-Information Systems and Technology	IT1205-Computer Systems I	IT1305-Web Application Development I
Out-degree centrality	Out-degree centrality	In-degree centrality
Hub	Harmonic closeness centrality	Eigenvector centrality
Eigenvector centrality	Closeness centrality	Degree centrality
Authority	Clustering coefficient	PageRanks
Degree centrality	Authority	Authority
In-degree centrality	Eccentricity	Betweenness centrality
Clustering coefficient	Eigenvector centrality	Out-degree centrality
Betweenness centrality	Degree centrality	Eccentricity
Harmonic closeness centrality	PageRanks	Clustering coefficient
Closeness centrality	Hub	Hub
PageRanks	In-degree centrality	Harmonic closeness centrality
Eccentricity	Betweenness centrality	Closeness centrality

Table 4.5: Social network parameters sorted on correlation

After that, as mentioned in section 2.4.2.1, classification was performed using Naive Bayes and Random Forest to analyse how well these parameters can further describe or classify the students' grades and to what extent. The classification accuracies obtained for two different algorithms are depicted in Table 4.6, for the course 'IT 1105-Information Systems and Technology', Table 4.7 for the course 'IT1205-Computer Systems I' and Table 4.8 for the course 'IT1305-Web Application Development I' as below. Moreover, they are graphically depicted in Figure 4.7, Figure 4.8 and Figure 4.9. The x-axis depicts which social network parameter was removed in each iteration whereas y-axis shows the classification accuracies in percentages.

	Naïve Bayes Accuracy (%)	Random Forest Accuracy (%)
All parameters	54.688	54.688
Without Eccentricity	54.688	57.813
Without PageRanks	60.938	64.063
Without Closeness centrality	60.938	57.813
Without Harmonic closeness centrality	60.938	48.438
Without Betweenness centrality	59.375	51.563
Without Clustering coefficient	59.375	50.000
Without In-degree centrality	60.938	46.875
Without Degree centrality	57.813	48.438
Without Authority	60.938	53.125
Without Eigenvector centrality	57.813	62.500
Without Hub	59.375	46.875

Table 4.6: Classification Accuracies for 'Information Systems and Technology'

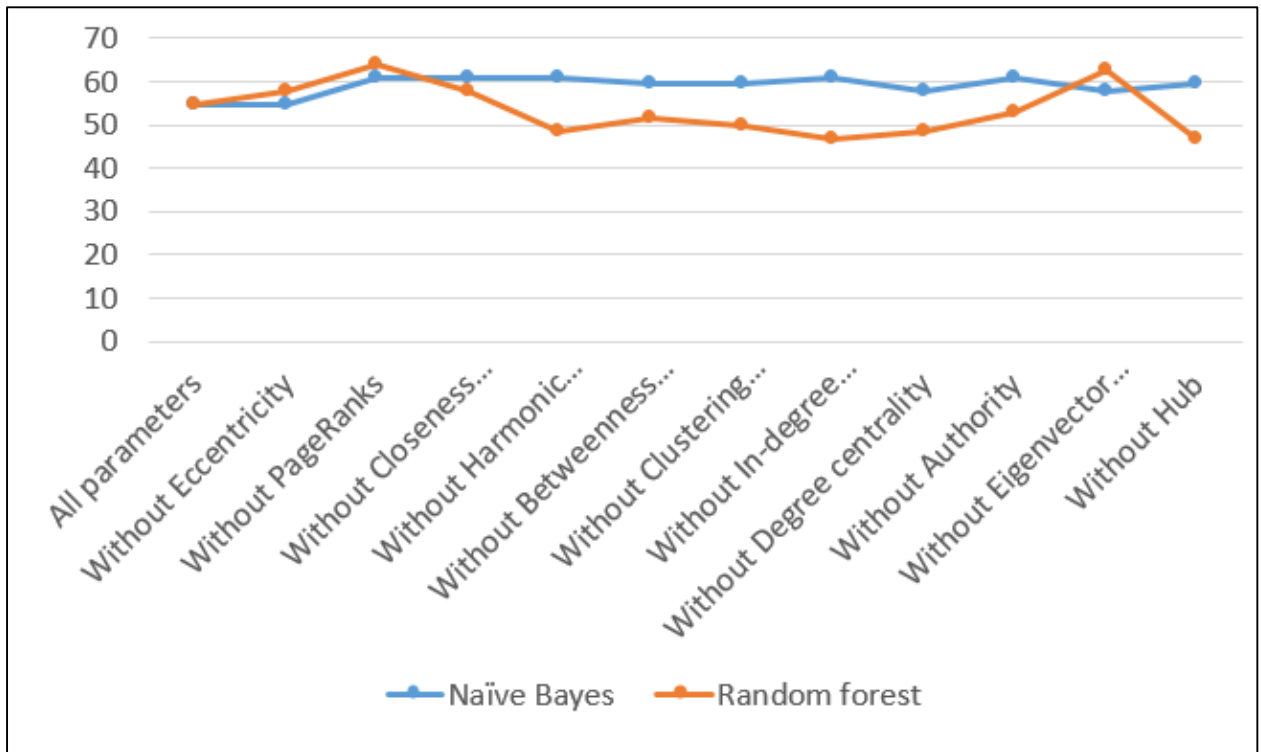


Figure 4.7: Classification accuracy variation in feature reduction for 'Information Systems and Technology'

According to Figure 4.7, the classification accuracies obtained by Random Forest algorithm was less than the results provided by Naive Bayes. Sometimes, its accuracy has declined beyond 50%. Therefore, for this particular course, only the classification pattern given by Naive Bayes was considered. When removing the least correlated attribute from the feature set, it depicted that classification accuracies have relatively increased and when the measure '*degree centrality*' was removed, the classification accuracy has suddenly dropped. So, for the course 'Information Systems and Technology', the best feature subset included top most features of the list, up to *degree centrality*. That means the subset is consists of *out-degree centrality*, *hub*, *eigenvector centrality*, *authority* and *degree centrality*, which are the best predictors of students' performance in online assessments. That means, not only the number of messages posted and received by each student (*out-degree centrality* and *degree centrality*), but also the student's social capital or how many knowledgeable peers the student is referring to (*eigenvector centrality*) also affects their learning. Moreover, the knowledge a particular student possesses (*Authority*) and how much of peers are recommended by each student(*hub*) also imply about their academic performance in this course.

In course ‘Computer Systems I’ also, as depicted by Table 4.7 and Figure 4.8, the classification accuracies of Random Forest algorithm were less than Naive Bayes algorithm in most of the points. So only the accuracies given by Naive Bayes were considered

	Naïve Bayes Accuracy (%)	Random Forest Accuracy (%)
All parameters	62.667	58.000
Without Betweenness centrality	62.667	56.000
Without In-degree centrality	62.667	58.667
Without Hub	62.667	56.667
Without PageRanks	61.333	62.667
Without Degree centrality	60.000	61.333
Without Eigenvector centrality	60.000	64.000
Without Eccentricity	60.667	63.333
Without Authority	60.000	58.667
Without Clustering coefficient	60.000	60.667
Without Closeness centrality	60.667	59.333
Without Harmonic closeness centrality	60.667	55.333

Table 4.7: Classification Accuracies for ‘Computer Systems I’

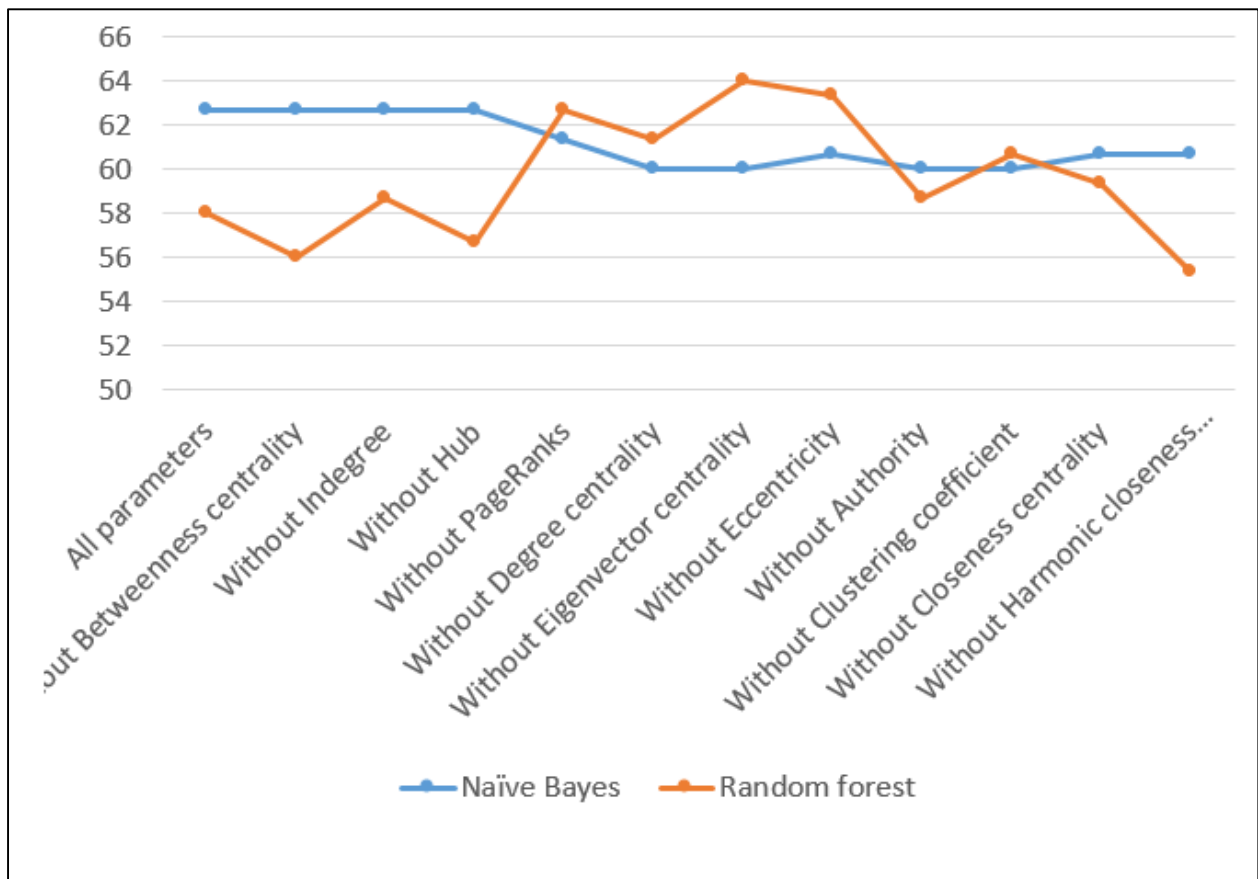


Figure 4.8: Classification accuracy variation in feature reduction for 'Computer Systems I'

So, when considering Naive Bayes classification accuracies, when removing the least correlated attribute from the feature set, according to Figure 4.8, it depicted that classification accuracies have been in the same level and suddenly has dropped when '*pageRanks*' measure was removed. That means, *pageRanks* also should be included in the subset and top most features of the list, upto *degree centrality* should be considered as the best predictors. So, for this course, the best feature subset included out-degree centrality, harmonic closeness centrality, closeness centrality, clustering coefficient, authority, eccentricity, eigenvector centrality, degree centrality, and *pageRanks*. It is in the sense, additional to the messages incoming and outgoing from students (*out-degree centrality*, *degree centrality*), the closeness of a student to peers (*harmonic closeness centrality*, *closeness centrality*), tendency to group with peers (*clustering coefficient*), the isolation level of student from the rest (*eccentricity*), the social capital (*eigenvector centrality*) and a student's importance in terms of skills he/she possess (*pageRanks* and *authority*) can be used to properly interpret the academic performance of students for this particular course.

In course 'IT1305 Web Application Development I' also, classification accuracies of Naive Bayes algorithm were larger than Random Forest as depicted in Table 4.8 and Figure 4.9.

	Naïve Bayes Accuracy (%)	Random Forest Accuracy (%)
All parameters	62.667	58.000
Without Closeness centrality	62.667	57.333
Without Harmonic Closeness Centrality	62.667	54.667
Without Hub	62.000	54.000
Without Clustering coefficient	62.000	52.667
Without Eccentricity	62.000	53.333
Without Out-degree centrality	62.000	54.000
Without Betweenness centrality	62.000	54.667
Without Authority	61.333	57.333
Without PageRanks	62.667	64.667
Without Degree centrality	60.000	66.667
Without Eigenvector centrality	57.333	56.000

Table 4.8: Classification Accuracies for 'Web Application Development I'

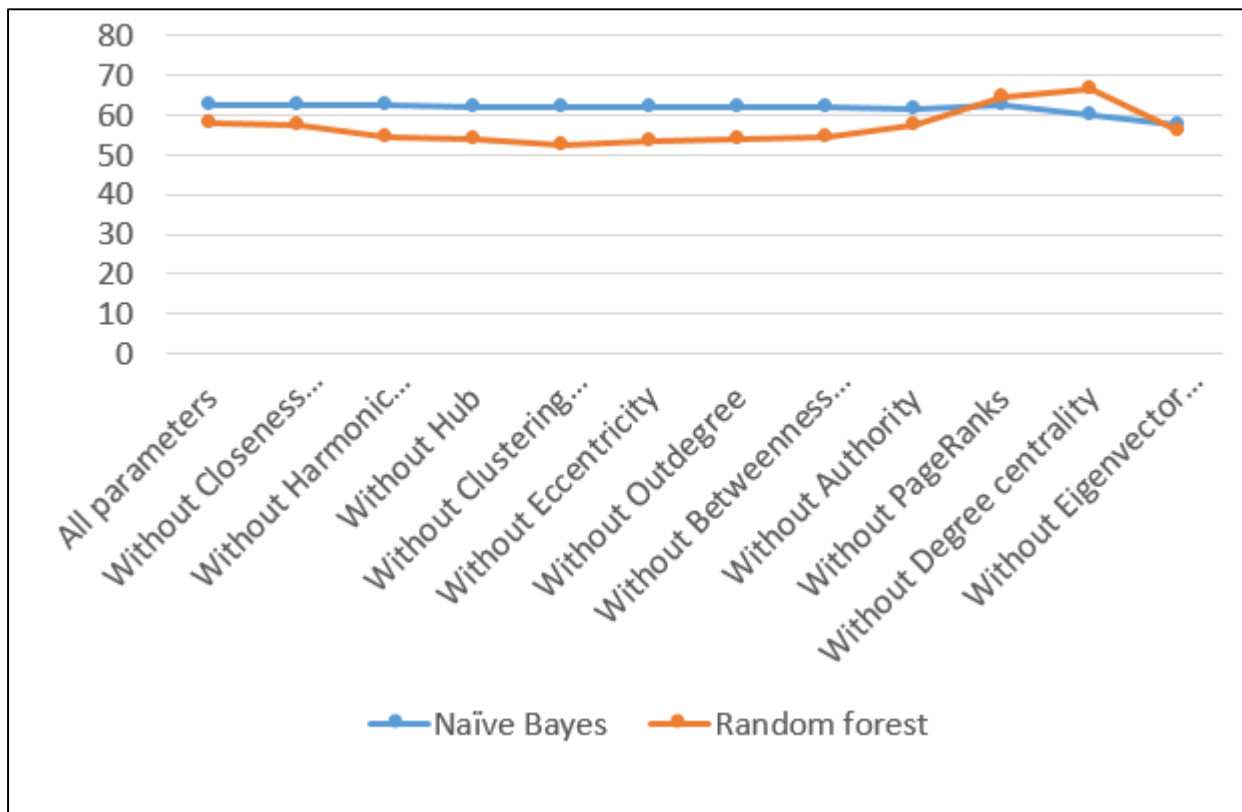


Figure 4.9: Classification accuracy variation in feature reduction for 'Web Application Development I'

For this course also, only the classification accuracies of Naïve Bayes were considered. Moreover, when *authority* parameter was removed, the accuracy has been suddenly dropped in Naïve Bayes results. Therefore, the feature set should include *in-degree centrality*, *eigenvector centrality*, *degree centrality*, *pageRanks*, and *authority*. (top correlated parameters including *authority*). That means, the number of messages by each student (*in-degree centrality* and *degree centrality*), the social capital built around (*eigenvector centrality*) and the information possessed by each student (*pageRanks* and *authority*) act as best performance predictors for this course.

Finally, the three best feature subsets filtered out for each course were compared together to find the final best feature subset which can represent all three courses (See Table 4.9) The reason for considering all three courses is that, as mentioned in section 1.1, the best e-Learner is selected considering the best forum engagement in all the courses each student follows.

“Information Systems and Technology”	“Computer Systems I”	“Web Application Development I”
Out-degree Centrality	Out-degree Centrality	In-degree centrality
Hub	Harmonic closeness centrality	Eigenvector centrality
Eigenvector centrality	Closeness centrality	Degree centrality
Authority	Clustering coefficient	PageRanks
Degree centrality	Authority	Authority
-	Eccentricity	-
	Eigenvector centrality	
	Degree centrality	
	PageRanks	

Table 4.9: Filtered feature subsets

It reported that eigenvector centrality, degree centrality, and authority (which are in bolded text) were among the best performance predictors common to all three courses. Additionally, as the parameter *out-degree centrality* has recorded the highest correlation with assessment marks in majority of the courses (IT1105 Information Systems and Technology and IT1205 Computer Systems), and a good correlation in the remaining course (IT1305 Web Application Development I), It also was taken into the final feature subset.

Therefore, the selected best parameter subset included altogether, *eigenvector centrality*, *degree centrality*, *out-degree centrality* and *authority*. That means, not only the quantity of interaction or the message count (degree centrality and out-degree centrality), but also the student’s social capital (eigenvector centrality), knowledge possessed by each student (authority) also affects their learning. This is a significant finding which again confirms the need for collaboration for the improvement of learning. If a student's performance is evaluated using all the courses, he/she has undertaken, the four social network attributes mentioned above can be used to evaluate the impact of his/her online behaviour to the performance in online assessments.

CHAPTER 5 DISCUSSION

A contemporary body of research in e-Learning has shown that students in CSCL setting make deeper and meaningful knowledge construction when they are engaged in learning environments where it facilitates interaction with other students. Discussion forums in such learning contexts facilitate students to improve their performance by providing the opportunity to discuss and communicate between other students as well as with the facilitator [1], [6], [17]. However, even though this concept has been discussed among the research community over decades, a few empirical demonstrations have been conducted considering social network component in the forums. Therefore, the current study provided insights on the importance of student-facilitator and student-student interactions in CSCL environments for students' academic performance by conducting an in-depth analysis of discussion forums using social network theories.

5.1 Factors Affecting Students' Behaviour in Online Discussion Forums

The results from social network visualizations showed that there are mainly three types of interactions exist in online discussion forums for all BIT courses. However, there was a significant variation in the number of interactions with respect to each identified type in the three courses. Moreover, regardless of the course, 'student-content' interactions were preceding among all the three types of interactions. Since the design of the courses relies on interactions usually started by the student, it was intended that students reply to the forum topic when trying to post a message with respect to a particular subject, so the sociograms were well aligned with the instructional design of this course. Results from sociograms revealed that some students have a high *degree centrality* which is a sign of higher activity. Furthermore, students who have a more influential role with higher *betweenness centrality* were the important nodes for the flow of information across the network. On the other hand, the presence of interconnections among students is a good indicator which shows there are considerable amount of debates and interactions among students trying to establish their cognitive and social presence [18], [21].

Also, depending on the interventions of the facilitator, network structure for each course showed slight variations. Nevertheless, results revealed that depending on the context of the course also, the students' interaction can be varied. It was observed that, the lack of motivation of students in the theoretical course to keep the flow of information in discussion threads. Given

that courses ‘Web Application and Development I’, and “Computer Systems I” are more into practical aspect it showed students in practical courses were actively participating and moderating the discussions than in theoretical courses. Therefore, changing the forum design considering the context of the subject could be a great method to enhance the participation of students in forums. Thus, Instructional designers should consider this when designing learning materials. For instance, increasing practical component of a subject would contribute to enhance peer-learning and peer- teaching among the students.

In addition, we conducted a time series analysis to investigate the evolution of the network over the whole duration of the courses. By considering the time-stamps of the posts, a time-lapse video was created visualise the social network in terms of time. This revealed about the time periods where students engage more in forums and when they seek guidance more from their peers and facilitators. This is another important finding which help facilitators to decide when is the right time to make an effective intervention so that most of the students in need will be benefited.

5.2 Effectiveness of Students' Social Engagement in Discussion Forums

In order to evaluate the effectiveness of students' social engagement in discussion forums, the study was focused on finding the relation between forum participation and course achievement. For this, calculated social parameters were categorised into three major areas in terms of quantity of interactions, each user's position of moderating information, and level of connectedness. Then each of these network parameters was correlated with student assessment mark. The final results from correlation analysis showed that the students who participated in the forum tended to perform better. This might suggest that participating in the discussion forums improves students' achievement in examinations.

However, one could argue that the students who participated in the discussion forum were the better performing students who were more engaged in the course and more likely to use the discussion forum to study and work harder in the course. Consequently, both forum participation and student achievement may entirely reflect this effect of social engagement of students. Thus, to better investigate the effectiveness of a discussion forum on student achievement, correlation analysis was conducted not only for one course but for the entire dataset obtained from three courses to better address the potential of an engagement confounds.

Since the study carried out with the several assumptions and limitations, it was needed to evaluate results from correlation with a caution. Firstly, although overall results suggest there a positive relation between forum interactions and students’ mark, most of the parameters did not

give significant correlations. However, students' performance may vary based on larger number of facts (e.g., intellect, study habits, study time, vocabulary), therefore it may not give significant correlations between students' performance and participation in the online forums [53]. Secondly, the negative correlations suggest that there is a considerable variation in the datasets. One of the main reasons for such variation may be, lack of motivation of students to participate in forums. Also, there is an enormous difference between the numbers of students and the posts among three datasets. One possible reason for this might be the difference in subject interest. The role of the facilitator, the structure of the forums may also be reasons to have such variations in those numbers. Likewise, for some courses, the sig. 2 tailed values did not represent significant correlations and *eccentricity* showed the negative correlations in all three courses. As *eccentricity* implies how far a student from his peers, it indirectly confirms that the less connected or isolated students tend to perform poorly in assessments. Only in course 'IT1305 Web Application Development I', all measures which imply the distance between each other (*eccentricity*, *closeness centrality* and *harmonic closeness centrality*) showed negative correlations. That means the distance between each other in the social network matters for student's academic achievement.

Due to these kinds of reasons, results for the correlation between social network parameters and student marks were different from course to course. As there are twelve social network parameters describing the students, according to literature, it was suggested that selecting a subset of them improves the classification accuracy [32]. That means, it can increase the descriptive and predictive power of students' academic performance using their online behaviours. Confirming that, the classification accuracies were increased when removing the least correlated social network parameters. Aligning with the past research evidence, this study also proved that the Naive Bayes algorithm reports higher accuracies compared to Random Forest when classifying the students' online interaction data [15], [32].

The overall results of the analysis showed a positive relationship between forum engagement and course performance. This suggests that extra opportunities for discussion and interaction provided by an online discussion forum can enhance learning and facilitate understanding of course materials which result in better academic achievement for the students. Finally, the best feature subsets for each course was derived and like in correlation, the number of best features filtered to the subset was varied from course to course. As mentioned in section 1.1, one student might undertake all three courses, therefore it is required to evaluate his/her performance common to all courses rather than checking for each course. Therefore, it is important to identify the most influential social network parameters considering all three courses to evaluate the impact of students' forum interaction on their performance in online assessments. The study discovered that the selected final parameter subset included altogether, *eigen centrality*,

degree centrality, *out-degree centrality* and *authority*. In BIT, the current best e-Learner selection process only considers the number of posts by each student (*Out-degree centrality*) as the only social network parameter [10]. Therefore, this study recommends the validity of the current selection process. However, the findings of this research further suggest to integrate the next highest correlated social network parameter, *eigenvector centrality*. Then, the evaluation may be more effective as it considers both the students' quantity of interactions and the social capital they possess. According to [24], [25] Knowledge is not built only through the individual effort, but also with a collection of sub components constructed via social exchanges. Therefore, this combined method provides an opportunity to evaluate the social skills of the students, which is the main intention of discussion forums in online learning environment.

CHAPTER 6 IMPLEMENTATION

As mentioned in section 1.1 in order to select the best e-Learner, BIT e-Learning centre at UCSC use manual mechanism which considers the forum posts, marks along with a pre-defined threshold value. Even though the criterion used by existing method is adequate it does not reflect the total value of a student collaboration with their peers. Furthermore, findings from this research also suggests that rather than considering the message frequency posted by the student it is better to use social capital possessed by him/her when evaluating his/her participation in online discussions. However, as mentioned in [10] one cannot totally rely on student participation in forums to evaluate his/her performance. Therefore, we developed a prototype tool to evaluate and monitor student participation considering both student social capital and their assessment mark. The developed tool is able to rank the students in a course based on eigenvector centrality values calculated considering the messages posted by them. These eigenvector centrality values are referred as the forum score in this article. Then forum scores are multiplied by their assessment mark to obtain the final scores. Based on the final score best students will be filtered and ranked. Messages posted by each student is organized considering a course, a forum and a discussion. This method can be used to choose the best e-learner in a more efficient and effective manner.

The tool is developed as a web application where it connects to the default Moodle database to fetch data. In order to calculate eigenvector centralities Eigenvector algorithm provided in Gephi modules was used [63]. Core part of the algorithm is presented in the Figure 6.1. Then a command line tool was developed to use by the PHP scripts for the execution.

Technologies used to develop the Eigenvector generation part:

- Java 8
- Maven building tool
- Gephi java plugins

Technologies used to develop the Web application:

To develop the backend of the web application PHP was used and the database used was MySQL. For the front-end development, we used CSS bootstrap framework, JavaScript, jQuery and jQuery data tables for displaying data in reports also to generate report in PDF/Excel. Additionally, SigmaJS library was used for graph plotting. The design and the development of the tool and the functionalities of the tool are explained further in the latter sections.

```

244 public double calculateEigenvectorCentrality(Graph graph, double[] eigCentralities,
245     HashMap<Integer, Node> indices, HashMap<Node, Integer> invIndices,
246     boolean directed, int numIterations) {
247
248     int N = graph.getNodeCount();
249     double sumChanged = 0.;
250     double[] tmp = new double[N];
251
252     for (int s = 0; s < numIterations; s++) {
253         double max = computeMaxValueAndTempValues(graph, indices, invIndices, tmp, eigCentralities, directed);
254         sumChanged = updateValues(graph, tmp, eigCentralities, max);
255         if (isCanceled) {
256             return sumChanged;
257         }
258
259         Progress.progress(progress);
260     }
261
262     return sumChanged;
263 }
264
265 /**
266  *
267  * @return
268  */
269 @Override
270 public String getReport() {
271     //distribution of values
272     Map<Double, Integer> dist = new HashMap<>();
273     for (int i = 0; i < centralities.length; i++) {
274         Double d = centralities[i];
275         if (dist.containsKey(d)) {
276             Integer v = dist.get(d);
277             dist.put(d, v + 1);
278         } else {
279             dist.put(d, 1);
280         }
281     }
282
283     //Distribution series
284     XYSeries dSeries = ChartUtils.createXYSeries(dist, "Eigenvector Centralities");
285
286     XYSeriesCollection dataset = new XYSeriesCollection();
287     dataset.addSeries(dSeries);
288
289     JFreeChart chart = ChartFactory.createScatterPlot(
290         "Eigenvector Centrality Distribution",
291         "Score",
292         "Count",
293         dataset,
294         PlotOrientation.VERTICAL,
295         true,
296         false,
297         false);
298     chart.removeLegend();
299     ChartUtils.decorateChart(chart);
300     ChartUtils.scaleChart(chart, dSeries, true);
301     String imageFile = ChartUtils.renderChart(chart, "eigenvector-centralities.png");
302
303     String report = "<html><body><h1>Eigenvector Centrality Report</h1> "
304         + "<br>"
305         + "<h2> Parameters: </h2>"
306         + "Network Interpretation: " + (isDirected ? "directed" : "undirected") + "<br>"
307         + "Number of iterations: " + numRuns + "<br>"
308         + "Sum change: " + sumChange
309         + "<br><h2> Results: </h2>"
310         + imageFile
311         + "</body></html>";
312
313     return report;
314 }
315
316 @Override
317 public boolean cancel() {
318     this.isCanceled = true;
319     return true;
320 }
321
322 @Override
323 public void setProgressTicket(ProgressTicket progressTicket) {
324     this.progress = progressTicket;
325 }
326
327 }
328 }

```

Figure 6.1: Eigenvector centrality algorithm

6.1 Course Wise Students Ranking

Students in a particular course are ranked according to the forum score (eigenvector centrality value). Therefore, students who are in the top of the list are the students who have participated more in the forum as well as they are the students who have considerable influence on other students. Figure 6.2. demonstrated the designed interface to show the student participation in a course. In here the students are ranked using eigenvector centrality values. Using this interface, a facilitator can gain an idea on to which level students have collaborated with each other and who are the isolated students.

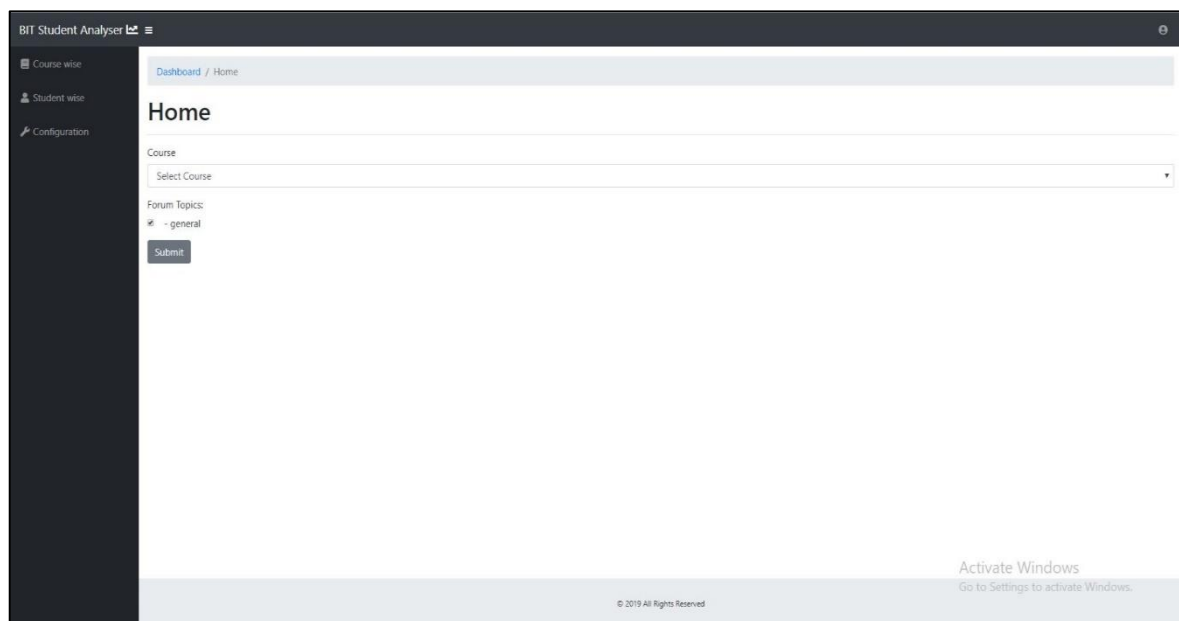


Figure 6.2: Interface - Select 'Course wise Ranking'

In the first Interface user (facilitator) has to select the desired course for which he/she want to see the student involvement in forums. After user submitting the selected course it will generate the resulting eigenvector centrality values (forum score), (see Figure 6.3) per student along with his/her ID, name, assessment mark and the forum ranking.

Additionally, by 'show graph' function facilitator can view the social network built in the course when students interact with each other. Therefore, it is easy for a facilitator to get a quick snapshot on how students collaborate in the course (see Figure 6.4)

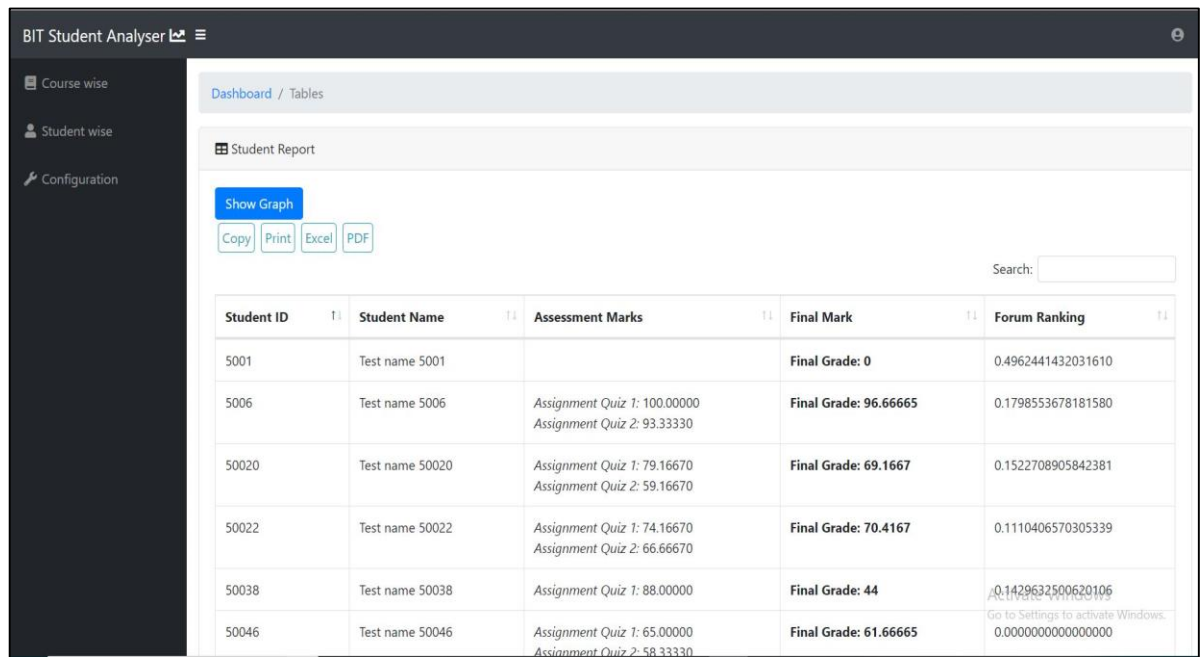


Figure 6.3: Interface - display course wise students ranks

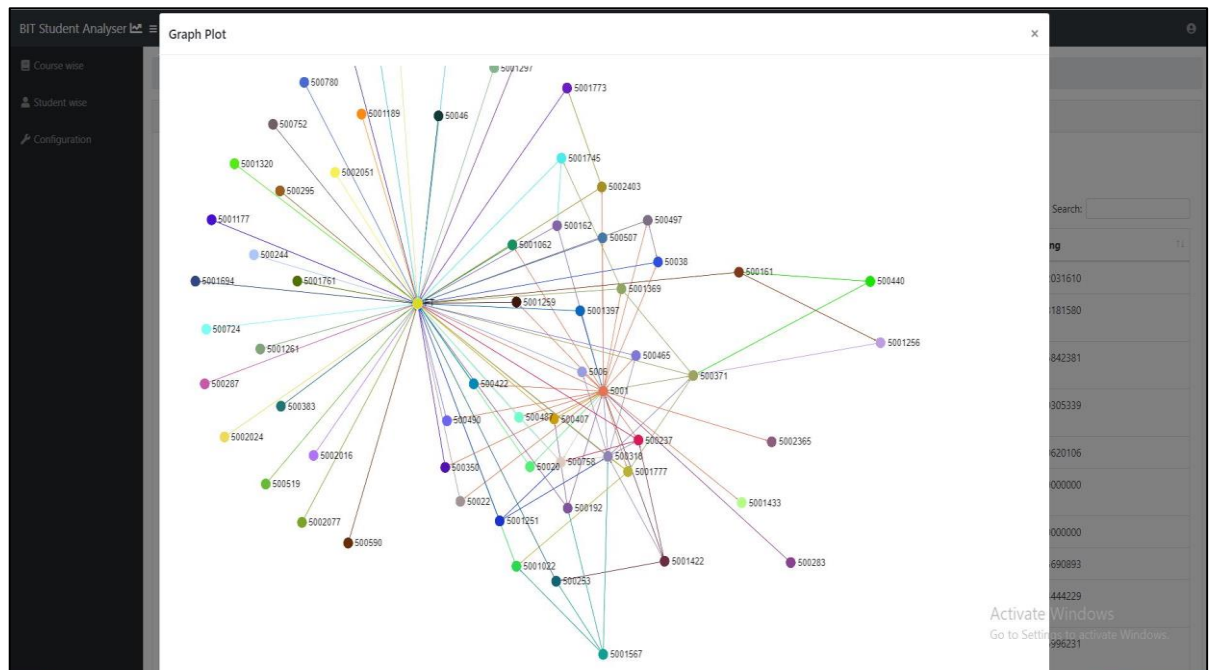


Figure 6.4: Built sociogram for a course

6.2 Overall Student Ranking

After obtaining forum scores those were combined with students' final assessment mark to obtain the final score for the evaluation. This is the overall student rank that can be obtained for considering several courses at once. This overall ranking can be used to choose the best e-Learners. The criterion used for this is as following.

- Course Forum Score (CFS) = Eigenvector centrality (forum score) of all forums in the course * Course final assessment mark

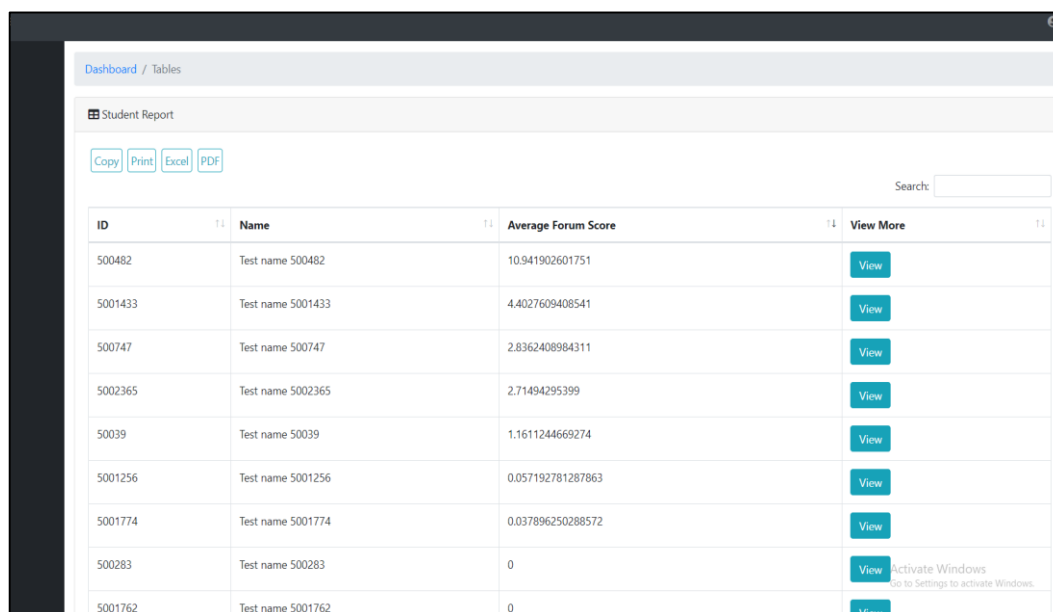
for all compulsory courses (i.e. CFS1, CFS2)

- Final Score == average (Course Forum Score)

If number of compulsory courses = 4:

Then, Average = Sum (CFS1 + CFS2 + CFS3+CFS4) /4

Eigenvector centralities and assessment marks of all the students in compulsory courses were considered. Only a best set of students were selected to show for the evaluation. Figure 6.5 demonstrated the interface designed to show the eligible students for best e-learner award. Here the students are displayed and ranked from highest 'final score' to the lowest. Most importantly, by clicking the 'view more' option with respect to a particular student, the facilitator can route to the messages posted by that student. Messages posted by the student are organized by the course, forum and a discussion. Thus, best e-Learner can be chosen by considering the richness of the message content posted by the student.



ID	Name	Average Forum Score	View More
500482	Test name 500482	10.941902601751	View
5001433	Test name 5001433	4.4027609408541	View
500747	Test name 500747	2.8362408984311	View
5002365	Test name 5002365	2.71494295399	View
50039	Test name 50039	1.1611244669274	View
5001256	Test name 5001256	0.057192781287863	View
5001774	Test name 5001774	0.037896250288572	View
500283	Test name 500283	0	View
5001762	Test name 5001762	0	View

Figure 6.5: Student Rank - best e-learner

6.3 Testing

The implemented system can be validated and verified in several ways. The main methods are to check whether the best e-Learner selected through the manual process can be visible in the top 10 students suggested by the system. This could not be done due to not having selected the best e-Learner for 2018 in BIT first year first semester students. Therefore, this validation is recommended as a future work (see chapter 9).

Another method is to check whether the post content of the top ten students are content related and rich. This could be validated, and the top students selected by the system was actively participating by both directly posting to the forum topic and giving feedback to the peers' posts also.

One more method is to ensure that the prominent students visible in the visualization of the social network (see section 4.2.2) are among this top student. This also could be validated.

CHAPTER 7 LIMITATIONS

Data used in this study were collected following two approaches. Firstly, the qualitative data were collected by conducting interviews with BIT facilitators. Secondly, quantitative data were collected by extracting students' forum interactions from BIT database. Data collected from the database included data records of students' discussions from the online discussion forums as well as their assessment marks which were retrieved from online assessment quizzes. Some limitations were encountered during the data collection and analysis process.

7.1 Data Collection

As stated in the previous sections the students in the BIT programme are distance learners. Usually they do not come to the UCSC for their studies. Most of the Students from various parts from the country are taking BIT degree program as external educational qualification and some of them do their studies part-time while working in their jobs. Due to these kinds of reasons, student participation in forums were much less and they were less motivated to participate in forms. Therefore, data records of discussion forums were less when compared to the assessment records. In fact, even though thousands of students had completed their assessment, most of them have not participated in discussion forums. Hence limited number of records were available for the study to analyse how learner interactions might have affected for their performance and also to provide more insights on how peer learning and peer teaching was exercised in this leaning context.

7.2 Content Analysis

The study only considered whether having more interactions in the discussion forums mattered students to perform better in their assessments. Therefore, the study does not consider the relevance of the message content posted by each student to the 'Discussion Topic' and how it would have affected other students for their learning purpose. The study only considered the forum posts which were contained in subject related forums there by it excluded any forum that categorized under the news and announcement tag. Although study considered subject related forums there could be messages posted by students which were not concerned with the subject discussion content. Moreover, those messages may have misled other students. Therefore,

considering relevance of each and every message in the forums would provide more insights to the students' online behaviour.

CHAPTER 8 CONCLUDING REMARKS

There are no sufficient researches conducted in collaborative learning to interpret the role of an individual student in a social network perspective. Also, there is no direct path to build new instructional technologies, tools or techniques which support collaborative learning. However, by considering suitable social structures and practices, it is possible to arrive for an outcome which may lead to obtaining desired student interactions and therefore to achieve performance goals. Therefore, student behaviours should be thoroughly analysed in several perspectives in order to discover hidden relationships between students, the causes for their active presence and to measure their influence on others.

By using Social Network Analysis, this study could uncover the hidden social network behind the online forum discussions. By visualising the social network, the study provided a broad view to course facilitators and instructional designers, whether the students have behaved as they expected, who are the influential students in the course and what are the pitfalls in collaborative learning process etc. When extended by the calculated network parameters, it offered a more precise image of the entire course network and each user's position, with respect to their participation and level of connectedness. Furthermore, a subset of social network parameters was filtered in order to enhance the descriptive power on students' performance. Finally, for each course and for all three courses, common social network parameters could be chosen which reflects the performance of students in online assessments. Using those findings, a prototype of an evaluation tool was implemented to select the best e-Learner in the BIT programme more effectively. It used *eigenvector centrality*, an all-rounder which better interpret the students' social learning behaviours.

These many insights are not possible using traditional mechanisms which only consider the counts of messages, but ignore the significance of the structure and social relationships. Therefore, this study is a fine example where SNA exposes the invisible sides of online collaborative learning and its impact on learning.

CHAPTER 9 FUTURE RESEARCH

Further research can be conducted to address the limitations of this research. Firstly, it would be more effective if the contents of the forum posts could be analysed to obtain high validity of the results rather than just considering the quantity of interactions. It could be done manually or automated, but automating the process of evaluating students' messages might be efficient as manual evaluation is a tedious and time-consuming task for instructors. As text mining has become prominent in EDM research community, those techniques can be applied. Secondly, future research can focus on using the power of SNA to improve online learning through monitoring and evaluation of students' behaviour in real time, early prediction of isolated students and therefore scaffold their leaning. This might be done through integration of SNA to Moodle platform. Moreover, adding real time performance prediction would be another useful innovation. Then, based on the students' forum interaction data, the system may be able to predict the performance level of each student, which helps the students to adjust themselves by working more harder to achieve.

In terms of the design of online discussion forums, a study can be conducted to deeply investigate what modifications in the design lead to more student interest and engagement, therefore to perform better. Then, the facilitators can be facilitated more while minimizing the pain points. Moreover, the validation of the suggested tool can be conducted through an extended research. The best e-Learner selected by the current selection process should be among the list of students suggested by the tool implemented in this research. Additionally, the post contents by the students who selected from the tool should be subject related, rich content.

Likewise, there are several new paths which this research is pointing to, ultimately to make the online collaborative environments more productive.

REFERENCES

- [1] G. Siemens, "Connectivism: A learning theory for the digital age", *International Journal of Instructional Technology and Distance Learning*, 2005.
- [2] Laal M, Ghodsi SM. Benefits of collaborative learning. *Procedia Soc Behav Sci*. 2012; 31:486–90.
- [3] B. De Wever, T. Schellens, M. Valcke and H. Van Keer, "Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review", *Computers & Education*, vol. 46, no. 1, pp. 6-28, 2006.
- [4] J. Pena-Shaff and C. Nicholls, "Analyzing student interactions and meaning construction in computer bulletin board discussions", *Computers & Education*, vol. 42, no. 3, pp. 243-265, 2004.
- [5] Borgatti SP, Mehra A, Brass DJ, Labianca G. Network analysis in the social sciences. *Science*. 2009;323(5916):892–5.
- [6] M. Saqr, U. Fors and M. Tedre, "How the study of online collaborative learning can guide teachers and predict students' performance in a medical course", *BMC Medical Education*, vol. 18, no. 1, 2018.
- [7] C. Romero, M. López, J. Luna and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums", *Computers & Education*, vol. 68, pp. 458-472, 2013.
- [8] C. Palazuelos, D. García-Saiz and M. Zorrilla, "Social Network Analysis and Data Mining: An Application to the E-Learning Context", *Computational Collective Intelligence. Technologies and Applications*, pp. 651-660, 2013.
- [9] T. Weerasinghe, "Designing Online Courses for Individual and Collaborative Learning: A study of a virtual learning environment based in Sri Lanka", Department of Computer and Systems Sciences, Stockholm University, 2015.
- [10] K. Hewagamage, K. Nishakumari, T. Weerasinghe and G. Wikramanayake, "'Motivating Student Discussions'-A Strategy to Develop Online Learning Community in the BIT Virtual Learning Environment", *Distance Learning and Education--International Proceedings of Computer Science and Information Technology*, 2011.
- [11] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601-618, 2010.

- [12] A. Bhattacharjee, *Social Science Research: Principles, Methods, and Practices*. 2012.
- [13] Widyahastuti, F., Riady, Y., & Zhou, W. (2017). Prediction Model Students' Performance in Online Discussion Forum, *Proceedings of the 5th International Conference on Information and Education Technology - ICIET '17*
- [14] S. Palmer, D. Holt, S. Bray, "Does the discussion help? The impact of a formally assessed online discussion on final student results", *British Journal of Educational Technology*, vol. 39, no. 5, pp. 837-858, 2008.
- [15] Á. Hernández-García, I. González-González, A. Jiménez-Zarco and J. Chaparro-Peláez, "Applying social learning analytics to message boards in online distance learning: A case study", *Computers in Human Behavior*, vol. 47, pp. 68-80, 2015.
- [16] M. Morzy, "On mining and social role discovery in internet forums", *Proceedings of the 2009 International Workshop on Social Informatics-SOCINFO '09*, 2009.
- [17] F. Alsaaty, E. Carter, D. Abrahams and F. Alshameri, "Traditional Versus Online Learning in Institutions of Higher Education: Minority Business Students' Perceptions", *Business and Management Research*, vol. 5, no. 2, 2016. Available: 10.5430/bmr.v5n2p31.
- [18] M. Laal and M. Laal, "Collaborative learning: what is it?" *Procedia - Social and Behavioral Sciences*, vol. 31, pp. 491-495, 2012.
- [19] T. Panitz, "Collaborative versus Cooperative Learning: A Comparison of the Two Concepts Which Will Help Us Understand the Underlying Nature of Interactive Learning", vol. 1999, 1999.
- [20] J. Roschelle and S. Teasley, "The Construction of Shared Knowledge in Collaborative Problem Solving", *Computer Supported Collaborative Learning*, pp. 69-97, 1995.
- [21] L. Harasim, S. Roxanne Hiltz, L. Teles and M. Turoff, *Learning networks*. Cambridge, Mass.: MIT Press, 1995.
- [22] K. Renninger, *Building virtual communities*. Cambridge [u.a.]: Cambridge Univ. Press, 2010.
- [23] J. Brown and P. Duguid, "Organizational Learning and Communities-of-Practice: Toward a Unified View of Working, Learning, and Innovation", *Organization Science*, vol. 2, no. 1, pp. 40-57, 1991.
- [24] J. Strijbos, P. Kirschner and R. Martens, *What we know about CSCL and implementing it in higher education*. Boston, Mass.: Kluwer Academic Publishers, 2004, pp. 245-259.
- [25] I. Nonaka and N. Konno, "The Concept of "BA": Building a Foundation for Knowledge Creation", *California Management Review*, vol. 40, no. 3, pp. 40-54, 1998.

- [26] L. Gilbert and D. R. Moore, "Building Interactivity into Web Courses: Tools for Social and Instructional Interaction.", *Educational Technology*, vol. 38, no. 3, pp. 29-35, 1998.
- [27] C. N. Gunawardena, "Social Presence Theory and Implications for Interaction and Collaborative Learning in Computer Conferences", *International Journal of Educational Telecommunications*, vol. 1, no. 2, pp. 147-166, 1995.
- [28] C. Gunawardena, C. Lowe and T. Anderson, "Analysis of a Global Online Debate and the Development of an Interaction Analysis Model for Examining Social Construction of Knowledge in Computer Conferencing", *Journal of Educational Computing Research*, vol. 17, no. 4, pp. 397-431, 1997.
- [29] S. Liaw and H. Huan, "Enhancing Interactivity in Web-Based Instruction: A Review of the Literature", *Educational Technology*, vol. 40, no. 3, pp. 41-45, 2000.
- [30] P. Northrup, "A Framework for Designing Interactivity into Web-Based Instruction", *Educational Technology*, vol. 41, no. 2, pp. 31-39, 2001.
- [31] E. Wagner, "Interactivity: From Agents to Outcomes", *New Directions for Teaching and Learning*, vol. 1997, no. 71, pp. 19-26, 1997.
- [32] Siemens G, "Learning and Knowing in Networks: Changing roles for Educators and Designers", *ITFORUM Discuss.*, vol. 27, pp. 1-26, 2008.
- [33] S. Wunsch-Vincent, *Participative web and user-created content*. Paris: OECD, 2007.
- [34] G. Salaway, J. Borreson Caruso and M. R. Nelson, "The ECAR study of undergraduate students and information technology", *EDUCAUSE Center for Applied Research*, vol. 6, 2007.
- [35] Q. Liu, W. Peng, F. Zhang, R. Hu, Y. Li and W. Yan, "The Effectiveness of Blended Learning in Health Professions: Systematic Review and Meta-Analysis", *Journal of Medical Internet Research*, vol. 18, no. 1, p. e2, 2016.
- [36] F. Kirschner, F. Paas, P. Kirschner and J. Janssen, "Differential effects of problem-solving demands on individual and collaborative learning outcomes", *Learning and Instruction*, vol. 21, no. 4, pp. 587-599, 2011.
- [37] J. Strijbos, P. Kirschner and R. Martens, *What we know about CSCL and implementing it in higher education*. Boston, Mass.: Kluwer Academic Publishers, 2004, pp. 245-259.
- [38] Erlin, N. Yusof and A. Rahman, "Students' Interactions in Online Asynchronous Discussion Forum: A Social Network Analysis", *International Conference on Education Technology and Computer*, 2009.
- [39] M. Conde, Á. Hernández-García, F. J. García-Peñalvo and M. Séin-Echaluze, "Exploring Student Interactions: Learning Analytics Tools for Student Tracking", *Lecture Notes in Computer Science*, pp. 50-61, 2015

- [40] D. Garrison and M. Cleveland-Innes, "Facilitating Cognitive Presence in Online Learning: Interaction Is Not Enough", *American Journal of Distance Education*, vol. 19, no. 3, pp. 133-148, 2005.
- [41] B.Muirhead, , "Insights for Teachers and Students.", *International Journal of Instructional Technology and Distance Learning*, no. 2003: pp.1–145, 2005
- [42] T. WEERASINGHE, R. RAMBERG and K. HEWAGAMAGE, "Designing a peer-teaching activity to promote inquiry-based learning", in *Mathematics and Computers in Contemporary Science*, 2013.
- [43] N. A. Whitman and J. D.Fife,Peer Teaching: To Teach Is To Learn Twice, ASHE-ERIC Higher Education ReportNo. 4, 1988, ASHE-ERIC Higher Education Reports, The George Washington University, One Dupont Circle, Suite 630, Dept. RC, Washington, DC 20036-1183, 1988.
- [44] T. Smith, Undergraduate Curricular Peer Mentoring Programs: Perspectives on Innovation by Faculty, Staff, and Students. Lexington Books, 2012
- [45] L. Cifuentes, "The Perfect Online Course: Best Practices for Designing and TeachingAnymir Orellana, Terry L. Hudgins, and Michael Simonson, Eds.", *American Journal of Distance Education*, vol. 24, no. 3, pp. 171-173, 2010.
- [46] T. Weerasinghe, R. Ramberg and K. Hewagamage, "Designing online learning environments for distance learning", *INSTRUCTIONAL TECHNOLOGY*, 2009.
- [47] H. Usoof and G. Wikramanayake, "Improving student learning through assessment for learning using social media and e-Learning 2.0 on a distance education degree programme in Sri Lanka", *European Conference on Educational Research, Vienna, Austria*, 2008.
- [48] H. Cho, G. Gay, B. Davidson and A. Ingraffea, "Social networks, communication styles, and learning performance in a CSCL community", *Computers & Education*, vol. 49, no. 2, pp. 309-329, 2007.
- [49] Joksimović S, Manataki A, Gašević D, Dawson S, Kovanović V, de KerekiIF. Translating network position into performance. In: *Proceedings of the sixth international conference on Learning Analytics & Knowledge - LAK '16*: 2016. New York: ACM Press; 2016. p. 314–323
- [50] A. Ashari, I. Paryudi and A. Min, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool", *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 11, 2013. Available: 10.14569/ijacsa.2013.041105.

- [51] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works", *Expert Systems with Applications*, vol. 41, no. 4, pp. 1432-1462, 2014.
- [52] S. Aslam and I. Ashraf, "Data Mining Algorithms and their applications in Education Data Mining", *International Journal of Advance Research in Computer Science and Management Studies*, vol. 2, no. 7, pp. 50-56, 2014.
- [53] C. Cheng, D. Paré, L. Collimore and S. Joordens, "Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance", *Computers & Education*, vol. 56, no. 1, pp. 253-261, 2011.
- [54] M. Bastian, S. Heymann, M. Jacomy, Gephi: An Open Source Software for Exploring and Manipulating Networks. Third International AAAI Conference on Weblogs and Social Media, 361–362 (2009).
- [55] D. Khokhar, Gephi cookbook. 2005.
- [56] F. Bloch and M. Jackson, "Centrality Measures in Networks", *SSRN Electronic Journal*, 2016. Available: 10.2139/ssrn.2749124.
- [57] M. Tarkowski, P. Szczepanski, T. Rahwan, T. Michalak and M. Wooldridge, "Closeness Centrality for Networks with Overlapping Community Structure", *THE ASSOCIATION FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE*, 2016.
- [58] A. Field, *discovering statistics using ibmspss statistics +spss version 22.0*. [Place of publication not identified]: Sage Publications, 2014.
- [59] Isba R, Woolf K, Hanneman R. Social network analysis in medical education. *Med Educ*. 2017;51(1):81–8.
- [60] T. Bergin, *An Introduction to Data Analysis: Quantitative, Qualitative and Mixed Methods*. 2018.
- [61] S. Chowdhry, K. Sieler and L. Alwis, "A Study of the Impact of Technology-Enhanced Learning on Student Academic Performance", *Journal of Perspectives in Applied Academic Practice*, vol. 2, no. 3, 2014.
- [62] S. Bharathidasan and C. Jothi Venkataeswaran, "Improving Classification Accuracy based on Random Forest Model with Uncorrelated High Performing Trees", *International Journal of Computer Applications*, vol. 101, no. 13, pp. 26-30, 2014. Available: 10.5120/17749-8829.
- [63] "gephi/gephi", *GitHub*, 2017. [Online]. Available: <https://github.com/gephi/gephi/blob/master/modules/StatisticsPlugin/src/main/java/org/gephi/statistics/plugin/EigenvectorCentrality.java>. [Accessed: 31- Dec- 2018].

Appendix A: Interview Questionnaire

Interview Questions

Q1: What are the courses available in first semester of BIT?

A1:

- IT1105 - Information Systems & Technology - General Course
- IT1205 - Computer Systems 1 - General Course
- IT1305 - Web Application Development I - General Course
- EN1101- Communication Skills – Enhancement Course
- EN1201- Introductory Mathematics – Enhancement Course
- EN 1301- Personal Computing- Enhancement Course

Q2: How is the structure of these courses?

A2:

General Courses: For each section of the course, there is a forum and a practice quiz and for the whole course there are two assessments. For example:

Section

Forum

Practice Quiz

assessment1

assessment2

Enhancement Courses: For each section of the course, there is a forum and a practice quiz and for the whole course there are two assessments.

Section

Forum

Practice Quiz

assessment1

assessment2

Online Assessment (In introductory mathematics-MCQ, in other two file submissions)

Q3: What is the difference between practice quiz and assessment?

A3:

Table 1: Practice quiz Vs assessment

Practice Quiz	assessment
One for each section	Altogether two for the whole course
Any number of attempts	Only three attempts
Same questions at each attempt	Different questions at each attempt
Marks are not taken to final grade	Marks are taken to final grade (Highest grade, not average)
No time limit	No time limit
MCQ	MCQ

Q4: What is the impact of assessments for the final grade of the course?

A4:

Final assessment mark= 40% of assessment1 + 60% of assessment2

Pass mark=50%

Q5: What is the Moodle database table that store these assessment data?

A5:

- MCQ Type assessments-->'mdl_quiz' table
- File submissions (online assessments) -->'mdl_assessment' table

Q6: In the forum, does it always start by the facilitator and what kind of things discussed there?

A₆: No. There is no such rule as it should be started by the facilitator. It is there for mainly for students to interact with their peers by discussing the issues they have. If it doesn't seem to have any interaction, the facilitator puts an interesting subject related topic for students to discuss. Forum is not compulsory and no marks are given. Can't put questions in their assessments.

Q₇: What is the role played by the instructors of other institutes?

A₇: They are granted the role 'student', they can download the course materials and see the forum, but can't post in forums and can't see the quizzes.

Q₈: Whether the design of course, forum, assessments and quizzes are aligned with the learning objectives and outcomes of the course?

A₈: Yes. They are designed to make the students achieve learning outcomes.

Q₉: Who is responsible for making these quizzes and forums?

A₉: The quizzes are given by the course coordinators and the forum topics are put by the facilitators according to the guidance of course coordinator.

Q₁₀: How many facilitators are there for a particular course and what kind of support is provided by the facilitator?

A₁₀: There is only one facilitator and he/she is responsible for answering the questions of students, maintain the course structure, upload course materials etc.

Q₁₁: Do you think the forum helps the students to achieve their learning objectives?

A₁₁: Yes. If they behave as expected it should be.

Q₁₂: What kind of interaction do you expect from students?

A₁₂: We have designed the environment for them to interact with peers and solve their issues (Peer learning).

Q13: Is there any mechanism to encourage students to participate more in forums and interact with peers?

A13: Yes, there is an award as ‘Best e-Learner’s Award’, to the student who has most number of forum interactions (Number of posts) and who is eligible for the diploma level. One student for each semester is selected.

Appendix B: SQL Data Extraction Scripts

SQL scripts used to extract data from the database.

```
CREATE SCHEMA IF NOT EXISTS `student` DEFAULT CHARACTER SET utf8 ;
USE `student` ;

CREATE TABLE IF NOT EXISTS `student`.`student` (`id` INT) ;

#import student userid from moodle db
INSERT into `student`.`student`
SELECT DISTINCT `id` FROM `moodle`.`mdl_user`
WHERE `moodle`.`mdl_user`.`id` IN ( SELECT `moodle`.`mdl_role_assignments`.`userid` FROM `moodle`.`mdl_role_assignments`
WHERE `moodle`.`mdl_role_assignments`.`roleid` =
(SELECT `moodle`.`mdl_role`.`id` FROM `moodle`.`mdl_role` WHERE `moodle`.`mdl_role`.`shortname` = "student" ));

#map userid with auto increment dummyid
ALTER TABLE `student`.`student` ADD `dummyid` INT NOT NULL AUTO_INCREMENT PRIMARY KEY FIRST;

UPDATE `student`.`student`
SET `student`.`student`.`dummyid` = CONCAT(500, `student`.`student`.`dummyid`);

#to map teachers
CREATE TABLE IF NOT EXISTS `student`.`teacher` (`id` INT) ;

#import userid from moodle db
INSERT into `student`.`teacher`
SELECT DISTINCT `id` FROM `moodle`.`mdl_user`
WHERE `moodle`.`mdl_user`.`id` IN ( SELECT `moodle`.`mdl_role_assignments`.`userid` FROM `moodle`.`mdl_role_assignments`
WHERE `moodle`.`mdl_role_assignments`.`roleid` IN (
SELECT `moodle`.`mdl_role`.`id` FROM `moodle`.`mdl_role` WHERE `moodle`.`mdl_role`.`shortname` IN ('teacher','editingteacher') ));

#map userid with auto increment dummyid
ALTER TABLE `student`.`teacher` ADD `dummyid` INT NOT NULL AUTO_INCREMENT PRIMARY KEY FIRST;

UPDATE `student`.`teacher`
SET `student`.`teacher`.`dummyid` = CONCAT(2, `student`.`teacher`.`dummyid`);

#create and insert ids into a table
CREATE TABLE IF NOT EXISTS `student`.`user_mapping` (
`id` INT,
`dummyid` INT) ;

INSERT into `student`.`user_mapping`
SELECT DISTINCT id,dummyid FROM `student`.`student`;
INSERT into `student`.`user_mapping`
SELECT DISTINCT id,dummyid FROM `student`.`teacher`;
```

Fig i: Script 1

```
#to map courses
CREATE TABLE `student`.`mdl_course` SELECT * FROM `moodle`.`mdl_course`;

#copying forum tables
#import mdl_forum
CREATE TABLE `student`.`mdl_forum` SELECT * FROM `moodle`.`mdl_forum`;

#import table mdl-forum_discussions
CREATE TABLE `student`.`mdl_forum_discussions` SELECT * FROM `moodle`.`mdl_forum_discussions`;

#update userid column with dummy id in student table
UPDATE `student`.`user_mapping` AS u INNER JOIN `student`.`mdl_forum_discussions` AS fd ON u.id = fd.userid
SET fd.userid = u.dummyid;

#import mdl_forum_posts
CREATE TABLE `student`.`mdl_forum_posts` SELECT * FROM `moodle`.`mdl_forum_posts`;

#update userid column with dummy id in student table
UPDATE `student`.`user_mapping` AS u INNER JOIN `student`.`mdl_forum_posts` AS fp ON u.id = fp.userid
SET dfp.userid = u.dummyid;
```

Fig ii: Script 1

```

#copying mdl-quiz tables

#import mdl_quiz
CREATE TABLE `student`.`mdl_quiz` SELECT * FROM `moodle`.`mdl_quiz`;

#import mdl_quiz_attempts
CREATE TABLE `student`.`mdl_quiz_attempts` SELECT * FROM `moodle`.`mdl_quiz_attempts`;

    UPDATE `student`.`mdl_quiz_attempts` AS u INNER JOIN `student`.`mdl_quiz_attempts` AS ds ON u.id = ds.userid
    SET ds.userid = u.dummyid;

#import mdl_quiz_feedback
CREATE TABLE `student`.`mdl_quiz_feedback` SELECT * FROM `moodle`.`mdl_quiz_feedback`;

#import mdl_quiz_grades
CREATE TABLE `student`.`mdl_quiz_grades` SELECT * FROM `moodle`.`mdl_quiz_grades`;

    UPDATE `student`.`mdl_quiz_grades` AS u INNER JOIN `student`.`mdl_quiz_grades` AS ds ON u.id = ds.userid
    SET ds.userid = u.dummyid;

#import mdl_quiz_overrides
CREATE TABLE `student`.`mdl_quiz_overrides` SELECT * FROM `moodle`.`mdl_quiz_overrides`;

    UPDATE `student`.`mdl_quiz_overrides` AS u INNER JOIN `student`.`mdl_quiz_overrides` AS ds ON u.id = ds.userid
    SET ds.userid = u.dummyid;

#drop table with originalids
DROP TABLE IF EXISTS `student`.`student` ;
DROP TABLE IF EXISTS `student`.`teacher` ;
DROP TABLE IF EXISTS `student`.`user_mapping` ;

#export student db

```

Fig ii: Script 2

Appendix C: Approval Letter for Data Extraction

06.03.2018

Prof. K.P.Hewagamage,

Director,

University of Colombo School of Computing.

Request to Access the Server Logs in BIT LMS

Dear Sir,

We are 4th year IS students who hope to conduct a research on 'Evaluating the best approach for analysing data to interpret student's learning styles in Online Environment'. Therefore we hope to analyse the server logs and database of BIT LMS using two best algorithms suggested by literature and validate which method gives the most accurate results.

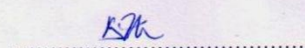
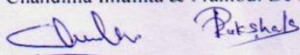
So we would be much obliged to you sir if you can grant us the permission to access the necessary data for our research.

We will be responsible of the ethical and secure aspects of data usage.

Thank You.

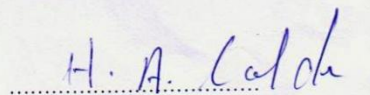
Yours Sincerely,

Chandima Imalika & Pramodi De Silva



Supervisor

Dr. T.A.Weerasinghe



Co-Supervisor

Dr. H.A.Caldera

Analytical Techniques to Investigate Online Learner-Learner Interactions

This is a literature survey conducted at the beginning of this research to identify which analytical technique would be more appropriate to analyse the LMS data. This is currently under the peer review process of the journal 'Technology, Pedagogy and Education'.

Analytical Techniques to Investigate Online Learner-Learner Interactions

W.A.P.P.R.De Silva

*University of Colombo School of Computing, University of Colombo
35, Reid Avenue, Colombo 7, Sri Lanka*

rukshidesilva90@gmail.com

Currently pursuing a Bachelor's degree program in Information Systems in University of Colombo School of Computing, Sri Lanka.

W.A.C.I.Imalika

*University of Colombo School of Computing, University of Colombo
35, Reid Avenue, Colombo 7, Sri Lanka*

cchandima2010@gmail.com

Currently pursuing a Bachelor's degree program in Information Systems in University of Colombo School of Computing, Sri Lanka.

T.A.Weerasinghe

*University of Colombo School of Computing, University of Colombo
35, Reid Avenue, Colombo 7, Sri Lanka*

taw@ucsc.cmb.ac.lk

PhD, Senior Lecturer at University of Colombo School of Computing, Sri Lanka

H.A.Caldera

*University of Colombo School of Computing, University of Colombo
35, Reid Avenue, Colombo 7, Sri Lanka*

hac@ucsc.cmb.ac.lk

PhD, Senior Lecturer at University of Colombo School of Computing, Sri Lanka

Analytical Techniques to Investigate Online Learner-Learner Interactions

Online collaborative learning has become a common practice in education due to its power to overstep the limitations of space, time and replace face to face interactions. Unfortunately, due to a large number of learners, the built-in analytics of major Learning Management Systems (LMS) offer only limited insights on learner behaviour. Fortunately, Learning Management Systems are capable of storing archive of data which reflects learner behaviour in online courses. However, it is not clear which analytical technique would be more appropriate to analyse these LMS data. This survey informed that mainly Social Network Analysis, Statistical techniques and Data Mining techniques have used to analyse learners' interactions. Also, the survey indicated that selecting a subset of all available online forum parameters aids for a better accuracy in deriving relationships between learners' interactions and their performance. Finally, this study proposes an aggregated analytical process to follow when analysing learner-learner interactions towards achievements.

Keywords: computer-supported collaborative learning; learner analytics; performance; social network analysis; educational data mining

1) Introduction

Exploring the methods to analyse the impact of learner interactions on learner performance has been a major research concern in the educational field. Yet, it seems we are still struggling to systematically study and explore learner-oriented data for better analysis. With the use of online learning, systems have come to a significant growth in the capture and analysis of learner interaction and performance data. The focus on systematic collection and analysis of learner-oriented data is also based in part on the drive to move from more subjective, anecdotally-oriented methods, to descriptive, data-driven research methods. These data-driven methods can help to evaluate proposed teaching methods or interventions in a more rigorous, higher-quality manner. Also, it may help to identify the significant factors that contribute to the observed learning outcomes.

Literature suggests that there has been a substantive rise in data collection for analysing learner interactions (learner-learner, learner-content, and learner-teacher) in Computer- Supported Collaborative Learning (CSCL) environments. However, little has been done to address the issues such as selecting effective methods to analyse these learner interactions and learning achievements and selecting a meaningful subset of data attributes for analysis. As such, the main objectives of the present study have been to: (1) systematically identify relevant research to find the effective methods to analyse learner interactions in online forums; (2) meaningfully categorize these studies; and (3) present how effective was the identified research in correlating learner interactions and achievements.

To address the above objectives, we identified and surveyed existing literature in this area to gain an understanding of the different approaches being used.

The driving questions for this literature survey were the following:

- What are the methods used to analyse learners' interactions and performance in CSCL environments?

- What information can be obtained from analysing learners' interactions in online discussion forums?

Categorization of literature was done to identify the critical commonalities and differences in the network visualization and data analytical methods being used to analyse data from online student discussion forums and the relationship of forum interactions with students' performance. Based on the findings of the literature review, the paper presents a more appropriate analytical process to investigate the relationship between learner-learner interactions and learning achievements in CSCL environments.

The rest of the sections in the paper is organized in the following way. Section 2 provides an overview of the role of collaborative learning in the online learning environment and the behaviour of learners in online discussion forums. The methodology followed by this survey is explained in section 3. The methods and data attributes used by researchers to analyse the learner interactions are presented in section 4. A new analytical process is proposed in section 5 by aggregating the strengths of identified analytical techniques and data types. Section 6 includes the discussion and finally, the conclusions are unfolded in section 7 as the major outcome of this survey.

2) Background

2.1) Collaborative Learning

According to Laal and Laal (2012), Collaborative learning is an educational approach where learners socially interact with other learners, as well as instructors to expand their knowledge on a particular subject or skill. According to Panitz (1999), collaboration is a notion of interaction designed to facilitate accomplishing of a goal or an end product by working together in groups. It is important to note that the Researchers who have conducted their researches on collaborative learning have highlighted social component as an underlying driving factor for the collaboration. To support this further, a growing body of research has demonstrated that social network is a central element in collaborative learning environments (Harasim, Hiltz, Teles & Turoff, 1995; Renninger, 2010). With the growth of the World Wide Web, social networking has raised its concerns in many fields like commerce, communication and more importantly in education (Siemens, 2008). Many researchers have pointed out the increased adoption of social bookmarking, computers, Internet connectivity, and Internet access for teaching and learning (Wunsch-Vincent, 2007; Salaway, Borreson, & Nelson, 2007). Moreover, Liu et al. (2016) report that Technology Enhanced Learning (TEL) facilitates networked learning through CSCL features that have been demonstrated to positively enhance learning when equipped with properly designed resources.

2.2) Computer Supported Collaborative Learning (CSCL)

Kirschner et al. (2011) report that the collaborative learning approach more applicable to online courses to achieve higher order learning outcomes. This confirms by Strijbos, Krischner, & Martens, (2004) by reporting collaborative learning in an online environment enclose knowledge and skills which are difficult to acquire by learning individually. Therefore, all these educational researches have supported the concept of CSCL. As reported in Renninger (2010), interactions in CSCL environment are often remote, faceless, uncertain, and moderated by computer-mediated communication (CMC) systems. Learners' willingness to communicate in CMC discussion

settings should affect their behaviour, especially how they build new social and learning relationships/networks with distributed, remote learning partners, who are often strangers (Salaway, Borreson, & Nelson, 2007). Furthermore, asynchronous interactions made through these CMC systems benefit more compared to the synchronous discussions. Such benefits include getting more opportunities to interact with each other and more time to reflect, think, and search for extra information before contributing to the discussion (De Wever, Schellens, Valcke, & Van Keer, 2006; Pena-Shaff, & Nicholls, 2004). Online Asynchronous Discussion Forum (OADF) is a tool for CSCL which offers the opportunity for students to interact and cooperate in online communities. Pena-Shaff, & Nicholls (2004) reported OADF is not just a tool to form students' and instructors' interactions but also it allows both parties to shape the nature of the exchange by reviewing posted information and analyzing own ideas before responding since participants are not constrained to respond immediately in most cases.

2.3) Learner interactions and Performance

Many researchers have analysed learner interactions and performance in online learning environments by using data from online discussion forums (Wunsch-Vincent, 2007; Saqr, Fors, & Tedre, 2018; Palazuelos, García-Saiz, & Zorrilla, 2013). Conclusions derived from these researches inform that learner interaction built within a CSCL community had a perceptible influence on individual performance. Moreover, they demonstrated how the central positions of students within the emergent collaborative learning network resulted in higher levels of learning performance (Saqr, Fors, & Tedre, 2018; Cho, Gay, Davidson, & Ingraffea, 2007).

3) Methodology

For this survey, we followed a light version of the guidelines for systematic literature reviews (Kitchenham, & Charters, 2007). Following sections describe steps carried out in searching relevant literature.

3.1) Identification of Relevant Literature

To search the literature, we started out with a set of keywords that are mostly used by researchers who analysed social interactions in CSCL environment. However, in the first round, search strings returned too many irrelevant papers. Therefore, restricting the domain to the online discussion forums, learner interactions and learner performance returned many relevant papers. To avoid a premature exclusion of potentially relevant papers, we adopted an inclusive approach; all undecided papers were included in the first step. We used the ACM and IEEE Digital Libraries, SpringerLink and Google Scholar web search engine to search papers.

3.2) Filtering

All included and undecided papers were collected into a file. After this step, the file contained fifty papers. In the next step, include/exclude papers were carried out based on their titles and abstracts.

- Papers with relevant titles to our domain and search questions were included.

- Papers with a better explanation in abstract relevant to our domain and search questions were included.
- Short papers, papers less than three pages were excluded.

3.3) Data Extraction

For data extraction, we developed a form based on the driving questions for the literature survey, described above. The final data extraction form contained the following main categories:

- Paper Title
- Year
- Methodological aspects of the research
- The context of the research
- Data that was collected and how it was collected and analyzed
- Overall results

The papers could be grouped into several categories. Also, there were some instances where one paper belonged to several categories. Figure 1 depicts the paper distribution across four major categories; Collaborative Learning, Social Network Analysis, Statistical Methods and Educational Data Mining with their percentages. Majority of the papers included details on Collaborative learning which 39.0% of all papers are. Social Network Analysis was used in 17.1% of all papers while Statistical Methods and Educational Data Mining categories included 14.6% and 29.3% papers respectively.

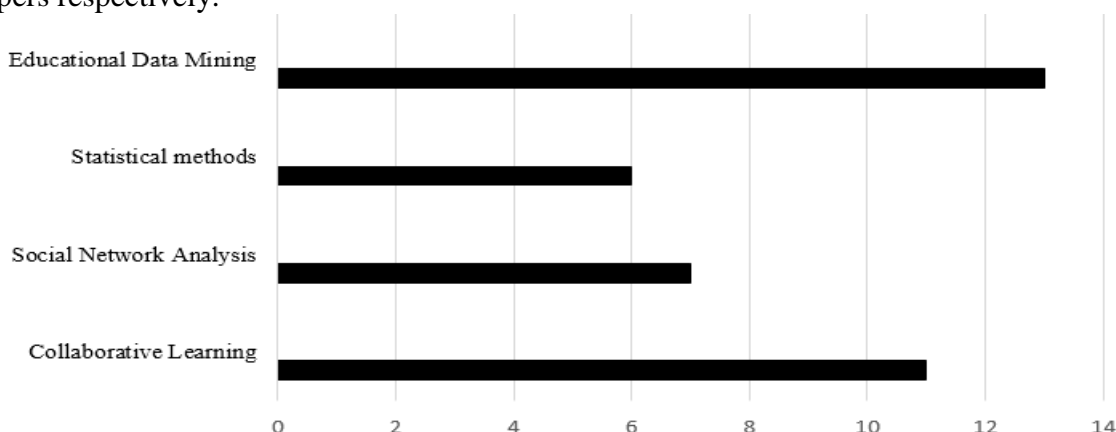


Figure. 1 Paper distribution

In the final document, there were thirty-three papers organized by the title, year and the author. Findings from these papers were explicitly used for this experimental survey.

4) Analysing learner interactions in CSCL environments

In this section, the analytical techniques used by past researches and the data obtained under each technique is presented.

In a CSCL environment where learners interact with peers, it is obvious that there is a social network structured automatically which stores learner behaviour. In order to explore these mines of information, Social Network Analysis is widely used in the education setting.

4.1) Social Network Analysis (SNA)

With the emergence of social networks like Facebook and Twitter, SNA has become much popular as it generates patterns to discover the hidden relationships between people, the types of those interactions, and the causes for their presence and to measure their influence [15]. Not only for social networks, today it is used in a wide variety of disciplines to investigate valuable information. As cited in the paper (Saqr, Fors, & Tedre, 2018), SNA is used in criminology to study the association between offenders, their criminal behaviour patterns etc. In Management, it is used to evaluate the organizational communicational hierarchies, the flow of information and the decision-making process. Medicine field uses the SNA method to identify the propagation of infectious diseases and examine the human gene networks while it further uses in an academic environment for citation analysis. On the other hand, Marketing seeks the help of SNA to explore the customer and supplier networks to find new business models (García-Saiz, Palazuelos, & Zorrilla, 2013; Cuéllar, Delgado, & Pegalajar, 2011). Likewise, there were some case studies in the literature which have applied SNA on Education to discover valuable information.

4.1.1) Education:

When considering the online education, one of the issues with the major learning platforms like Moodle and Blackboard is that their built-in analytics only offer limited insights to study learner interactions (Saqr, Fors, & Tedre, 2018). For example, Moodle offers teachers to view frequency of participation of students on courses while lacking the ability to deeply study the structure of the communication and student interaction patterns (Borgatti, Mehra, Brass, & Labianca, 2009). Therefore, Social Network Analysis is applied in education field basically to analyse the learners' participation level in courses, their level of cohesion, the active and inactive students, the flow of information, efficiency of group work etc (Saqr, Fors, & Tedre, 2018; Borgatti, Mehra, Brass, & Labianca, 2009). Supporting this task, Cuellar et al. (2011) proposed a method to formulate and interpret the learning management platforms as a social network to deeply explore the social structure between learners, teachers, and learning resources to uncover the hidden mines of valuable information to promote learning. This can be achieved by identifying the relevant actors and relations in the database tables according to the problem statement and transform these tables into a social network (Cuéllar, Delgado, & Pegalajar, 2011). As cited in (Saqr, Fors, & Tedre, 2018), one of the major strengths of SNA over other traditional analysis methods is its speed in producing information and easiness in interpreting results.

4.1.2) Visualization:

The visualization of the social network is depicted by a graph called 'sociogram' which consists of nodes and edges. An actor in the network (here, a student or a teacher in the learning context) is depicted by a node and the interactions between those actors are depicted by the edges/lines between nodes. The centrality measures calculated using these sociograms reflect the behaviour of students in the online collaborative learning environment. For example, degree centrality is the total of all incoming and outgoing interactions from an actor (Palazuelos, García-Saiz, & Zorrilla, 2013). It is a measure of how active a student is in the network and to what extent an actor is connected to others in a given social network. Likewise, these social attributes represent the broad base of psychosocial and social support necessary for high performance. In social network perspective, some learners surpass their peers due to its advantageous position than others in the social network. Likewise using SNA, it is possible to search for the optimal network positions

that strategically advantageous to individuals' performance (Cho, Gay, Davidson, & Ingraffea, 2007). Furthermore, social networks have a significant impact on CSCL setting as it is based on social interactions between distributed individuals. Therefore, SNA can discover the impact of individuals' investment in new social and intellectual capitals towards their performance (Nahapiet, & Ghoshal, 1998).

4.2) Attributes

When considering a social network of a learning management platform, there is a number of parameters that can be obtained which explain the different aspects of learner behaviour. According to Saqr, Fors, & Tedre (2018) and Palazuelos, García-Saiz, & Zorrilla (2013), these parameters can be viewed mainly as network-level (network parameters) and node-level (user parameters).

4.2.1) Network-Level Parameters:

As cited in Saqr, Fors, & Tedre, (2018), these parameters are calculated by considering the network as a whole. Network size is the total amount of nodes (actors) in a particular network. Average degree is the mean degree centrality of all learners which implies the average level of interactivity of participants. The extent of the learners' activity in the network is denoted by network density, the ratio of actual interactions between peers to the total possible. As more learners are participating in the online forum, this measure is increased. To provide a sense of group connectedness in a network, average clustering coefficient indicates the tendency of group members to interact together. As cited in Palazuelos, García-Saiz, & Zorrilla (2013), another useful network-level metric is the diameter of the network which is defined as the largest number of nodes needed to pass over to come from a particular node to another. Furthermore, there is reciprocity, that indicates the likelihood of occurring double links (with opposite directions) between vertex pairs. It is the ratio of the number of links pointing in both directions to the total number of links. If the reciprocity equals to 1 it is a purely bi-directional network where purely unidirectional when equals to 0.

4.2.2) Node-Level Parameters (User Parameters):

Basically, these metrics are in terms of individual nodes and these centrality measures can be calculated in different ways based on the context. Saqr, Fors, & Tedre (2018) classified these node-level parameters into three groups in their research as the measures which imply quantity of interaction, the role of moderation and the connectedness. First, under the quantity of interaction category, there are in-degree centrality, out-degree centrality, and degree centrality. In-degree centrality is the number of incoming interactions for a particular user which indicates the popularity or the prestige. Out-degree centrality is the number of outgoing interactions from a node that implies how active the particular learner in the network. Degree centrality is the number of links connected to a node ignoring its direction (Palazuelos, García-Saiz, & Zorrilla, 2013). In the second category, role of moderating, there are betweenness centrality the number of shortest paths from all nodes to all others that pass through such a node, information centrality that measures the importance of a node in information flow and network cohesion and closeness centrality, an indicator of how close a learner is to the other collaborators in a network. In the third category mentioned, metrics which implies the connectedness includes eigenvector centrality, eccentricity

and clustering coefficient. Eigenvector centrality considers how well connected the neighbours of the actor are to estimate the social capital and the influence of one's ego network. Eccentricity indicates how far a particular learner from peers which implies the level of isolation within the network. Clustering coefficient calculates the actual edges between a node and its neighbour peers to the total possible which deliver the tendency of a learner to work with peers. On the other hand, Palazuelos, García-Saiz, & Zorrilla (2013) used top3 parameter (if a node is ranked in the top 3 nodes in some of the previous centrality measures, top3 is true otherwise false). Moreover, Cho et al. [16] highlight another user parameter change propensity that observes how actively an individual has renewed social capital when participating in a new learning setting.

Previous researches have used the above parameters based on their specific context and problem statements. Saqr, Fors, & Tedre (2018) found in their case study that best parameters correlated with performance were in-degree, out-degree, clustering, eigenvector centrality and information centrality. Using these social metrics, this case study discovered precious findings that there is an instructor-centered network where learners tend to reply to the instructor rather than interact with peers which confirms the course has designed aligned with learning outcomes. Palazuelos, García-Saiz, & Zorrilla (2013) found top3, betweenness centrality, in-degree, out degree, and degree are most relevant performance predictors when selected by several feature selection algorithms. Moreover, here this study concluded that, for different predictor classifiers in data mining, different attributes are the most important ones.

4.3) Data Analytics

After visualization of data in order to identify the relationship with performance, network attributes are usually aligned with performance data. Mainly two methods have been used to analyse the correlation between these network attributes and learner performance. Those are,

- Statistical Techniques
- Educational Data Mining(EDM)

4.3.1) Statistical Techniques:

Several researchers have used statistical techniques such as correlation, regression, standard deviation to investigate the impact of social attributes on learner performance (Wunsch-Vincent, 2007; Saqr, Fors, & Tedre, 2018; Palmer, Holt, & Bray, 2008; Chowdhry, Sieler, & Alwis, 2014). In the following sections, we have categorized papers according to the statistical techniques followed by the respective researches.

4.3.1.1) Correlation and Automatic Linear Regression (ALM)

Saqr, Fors, and Tedre (2018) has used Kendall's Tau-b test to measure the correlation coefficient between ranked network variables. The test has been performed using permutation methods by PAST (Paleontological statistics software package for education and data analysis) and the permutation test was based on 9999 random replicates. A permutation test is a well-recognized statistical technique which aids in overcoming issues of analysing relational data using conventional statistical tests (Isba, Woolf, & Hanneman, 2016). Such issues arise due to the potential problems and statistically non-independent nature of relational data. However, in permutation testing, the obtained results are compared against random or quasi-random

permutations of the data so it omits inherent error-prone conditions of relational data. With the advancement of processing power, performing permutation testing has been easy for current researchers (Isba, Woolf, & Hanneman, 2016). Results for correlation between social attributes with performance showed that parameters corresponding to the quantity of interactions (degree and out-degree) did not significantly correlate with student grade. However, in-degree centrality was moderately significantly correlated. All centrality scores measuring the role in information relay were positively correlated with final performance. As for the statistical software, they have used SPSS software version 24 to perform ALM test. By using ALM they have checked if SNA parameters can be used to predict the final grade and to what extent variance of grade can be explained by learners' participation. ALM is an improved technique which is favourable in selecting a namely better variable, handling of extreme values (outliers), as well as merging of similar predictors and conducting ensemble methods [25]. As in [14], when used ALM for predicting midterm results with respect to learner's position, interactions, and relations in the network, it has given 71.6% as the accuracy. Important predictors for this were out-degree, in-degree centrality, and Eigenvector centrality. When predicting final results, the resulting accuracy was 70% and the important predictors were information centrality, out-degree, in-degree centrality and Eigenvector centrality.

4.3.1.2) Correlation and Multivariate Regression Analysis

Apart from SNA, it is important to mention that several other researches have been carried out without necessarily considering the Social Network Analysis aspect. For example, Palmer, Holt, and Bray (2008) have investigated the effect of online asynchronous discussions on student learning using the following data categories with respect to student usage of the online discussion area:

- the total number of forum messages read (or at least opened) by the student;
- the total number of new/initial discussion postings made by the student
- the total number of follow-up/reply discussion postings made by the student
- the final unit mark obtained by the student

Although the study has not specifically mentioned using SNA, the data attributes they have selected can be categorized under the defined criteria for SNA parameters. As for the statistical techniques they have used correlation (Pearson's linear correlation coefficient) to evaluate the relationship between the above mentioned data variable pairs and the final unit mark of the students. Also, multivariate linear regression has been used to find out how does each mentioned variable contribute to the dependent variable, final unit mark. As in (Palmer, Holt & Bray, 2008) by using Pearson's linear correlation coefficient, a significant correlation has been observed for the final unit mark and total number of new postings ($r=+0.49$). Note that in Pearson correlation, $-1 \leq r \leq 1$; where -1 means a strong negative correlation, 0 means no correlation and 1 means a strong positive correlation. In addition to that, scatter plots has been used to inspect variable pairs. Results in scatterplots revealed that the relationship between the final unit mark and the number of new postings plateaued after five new postings. The observed correlation between total number of new postings and the final unit mark was strongest for the number of new posts. When multivariate linear regression analysis was conducted with final unit mark as the dependent variable, the analysis model has predicted only 55.4% accurately of the variation on the final unit mark. An analysis of variance test suggests that the regression model is significant, although the

model predicts only 55.4% of the variation on final unit mark. The regression residuals were approximately normally distributed. However, Palmer, Holt and Bray (2008) specifically mention that the model explains only just over half of the variation observed in the final unit mark, hence there exist other factors with a significant influence on the final unit mark. Additionally, results on the regression analysis support the results of the data pair correlation analysis that the number of new postings contribute significantly and independently to the final unit mark.

4.3.1.3) Correlation, Mean Standard Deviation, and Linear Regression Analysis

Cho et al. (2017), has used mean, standard deviation to assess forum participation. Linear regression analysis and correlation had been used to analyse the relationship between the numbers of posts and the standardized scores of each assessment measure. The results have revealed a significant correlation between forum participation and performance of students.

Another significant research conducted by Chowdhry, Sieler and Alwis (2014) investigated the effect of number VLE visits on Learner performance. Although this study does not specifically mention the forum but they have used several forum data, i.e. forum view, adding posts on forum for their study. For statistical analysis, they have used Pearson's product-moment correlation to find the correlation between VLE visits and the final marks obtained by the students. Statistical software SPSS 20.0 (Academic version) has been used to perform the data analysis. The quantitative study has been carried out on the data of all the three modules (Law Module (LM), an Electrical Engineering Module (EEM) and a Mechanical Engineering Module (MEM)) taking the VLE visits as the independent variable and the final marks obtained by the corresponding students as a dependent variable. In addition, for data visualization purposes they have used techniques such as scatter plots with best fit line and bar graphs to obtain sufficient understanding of the data. The results obtained via Pearson's product-moment correlation coefficients showed that for both LM and EEM modules, the numbers of VLE visits are not correlated with the corresponding final marks of the students. However, for the MEM module, there was a moderate positive correlation between both variables. One of the conclusions of this study was that the students' academic performance may not necessarily be directly affected by the use of the VLE.

4.3.2) Educational Data Mining (EDM):

Data Mining (DM) is the practice of scanning huge data repositories in order to discover new information and therefore to derive knowledge which is a valuable support for effective and timely decision making (Peña-Ayala, 2014). EDM is the application of data mining in the education field. It can be considered as an interdisciplinary research field which inherits the properties of learning analytics, psychometrics, artificial intelligence, database management systems etc (Romero, López, Luna, & Ventura, 2013). With the support of statistical, machine learning and data mining algorithms, EDM focuses on resolving educational research issues by having a better understanding of students and their learning environment. Here the raw data coming from Learning Management Systems (LMSs) are converted into useful information by applying the data mining process. Romero et al. (2007) show that the EDM process follows the following four steps which are interactive and iterative. Most of the case studies have referred to this methodology (Saqr, Fors, & Tedre, 2018).

- Data collection- extract interaction data from LMS databases

- Data pre-processing- transform the data into a compatible format
- Data mining- apply data mining algorithms using data mining tools
- Data interpretation- interpret the results and support decision making in e-learning

There is a wide variety of data mining techniques such as classification, clustering, association rule mining, sequential mining, text mining etc. (Romero, & Ventura, 2007, 2009). As cited in Aslam & Ashraf (2014), these data mining techniques can be used for Data Analysis (explore data without any clear idea), Descriptive Modelling (provides models which show the relationship between different objects), Predictive Modelling (prediction of unknown values from different known variables), Discovering Patterns and Rules (spot behaviours like fraud detection) etc. Focusing on finding the impact of learner interaction data to the performance, to identify the relationship between parameters, Association rule mining, Correlation mining, Sequential pattern mining, Causal data mining can be used. In order to predict performance from known data attributes, Classification, Regression, and Density estimation are widely used (Aslam, & Ashraf, 2014). When analysing a CSCL setting to figure out its impact on performance using data mining techniques there is a large number of data parameters that can be obtained from online forums such as the number of posts created, number of messages read etc. Although some of them may be irrelevant for predicting students' performance. Therefore, to filter out the most suitable attributes correlated with performance, feature selection algorithms have been used by several case studies. Palazuelos, García-Saiz and Zorrilla. (2013) have used two feature selection algorithms named CfsSubSetEval and ClassifierSubSetEval in the data mining tool Weka to filter out a subset of attributes. Furthermore, Romero et al. (2013) have filtered five attributes from available nine data attributes using ten feature selection algorithms using Weka as the best attributes that should have a greater effect on learners' final performance. This study found that using a subset of attributes instead of all available attributes leads to more understandable and accurate classification and prediction models.

Researchers have used various types of data attributes such as quantitative, qualitative or social attributes in their case studies in order to find the impact of learner interaction on their performance.

4.3.2.1) Quantitative Attributes

Some researchers have used quantitative analysis methods that provide a systematic and powerful analysis based on quantitative data which can be measured and written down in numbers. Palmer, Holt and Bray (2008) have used quantitative data such as message frequencies (number of posts and replies, number of messages read, thread length and response time from previous messages) to find its correlation with student grades using multivariate regression analysis. Cheng et al. (2011) have analysed the frequency of access and the duration of sessions to categorize the learners using cluster analysis. Using the frequency of access and the duration of sessions, Khan et al. (2012) have established several categories of learners by cluster analysis. Based on a number of discussions created, post created, discussion & module course viewed and some other quantitative attributes, Widyahastuti et al. (2017) built a new model to predict students' performance in the online discussion forum.

4.3.2.2) Qualitative Attributes

Some other researchers have followed a qualitative approach which has mainly focused on content analysis (Pena-Shaff, & Nicholls, 2004). Content analysis can specify the intention of the students participated in online forums by reading their posts and validating their relevance to course content. Romero et al. (2013) have analysed on what messages are the best predictors; all the available messages or only the messages related to the course which finally concluded that content related messages improved the accuracy of prediction.

4.3.2.3) Social Attributes

Another set of researchers have applied data mining together with social network attributes in order to get a broader view of the structure of learner behaviour. For example, Palazuelos, García-Saiz and Zorrilla (2013) applied several data mining algorithms such as J48, Random Forests, Naive Bayes, Bayesian Networks, JRip and Ridor on social attributes such as degree, in-degree, out-degree, betweenness, authority and hub to predict learner performance and dropouts. This study reported that a high accuracy has been obtained in predicting performance when Social Network Analysis (SNA) attributes were combined with traditional quantitative activity attributes like message frequencies. This has been confirmed by Romero et al. (2013) by reporting some valuable insights which are not visible when only considering the count hits or replies, but ignoring the importance of structure, relations, and interactions. Considering all, Romero has followed a mixed approach which consists of quantitative, qualitative and social network information to provide a richer explanation on predictors of student performance in online discussion forums (Romero, & Ventura, 2013).

Also, several studies have focused on analysing and predicting performance not only at the end of the course but also before the end. For example, Saqr, Fors and Tedre (2018) and Cho et al. (2007) conducted their case studies using learner grades on both ends. Supporting this Romero et al. (2013) showed that performing prediction at both ends (in mid-term and end of the term) leads to more accurate performance classifications.

5) Results

Considering the above descriptions, Table 1 represents a summary of the identified analytical techniques and the data types obtained in each technique.

Table 1. Data types for each analytical technique

Analytical Technique	Types of data obtained
Social Network Analysis	Network-Level Parameters <ul style="list-style-type: none"> • Network size • Average degree • Network density • Average clustering coefficient

	<ul style="list-style-type: none"> • Diameter • Reciprocity
	Node-Level Parameters (User Parameters) <ul style="list-style-type: none"> • In-degree centrality • Out-degree centrality • Degree centrality • Betweenness centrality • Information centrality • Eigenvector centrality • Eccentricity • Clustering coefficient • Top3 • Change propensity
Statistical methods	Quantitative attributes <ul style="list-style-type: none"> • Total number of discussion messages read • Total number of new/initial discussion postings made by the student • Total number of follow-up/reply discussion postings made by the student • Author of the post • Target(reply) of the post • Post time created • Final grades/ midterm grades
	Social Attributes <ul style="list-style-type: none"> • Network size • Density • Centrality measures (in degree, out degree, degree, betweenness)
Educational Data Mining	Quantitative attributes <ul style="list-style-type: none"> • Number of posts • Number of replies • Number of messages read • Thread length • Response time from previous messages • Frequency of access • Duration of sessions • Number of discussions created • Number of posts created • Number of discussion • Number of module course viewed
	Qualitative attributes <ul style="list-style-type: none"> • Content of the post • Content of the reply
	Social attributes <ul style="list-style-type: none"> • Degree • In-degree • Out-degree

	<ul style="list-style-type: none"> • Betweenness • Authority • Hub
--	---

Considering above mentioned various types of analytical techniques and attributes, Figure 2 provides a big picture of the relationship between each analytical method (technique) and the attributes (parameters) used by each method.

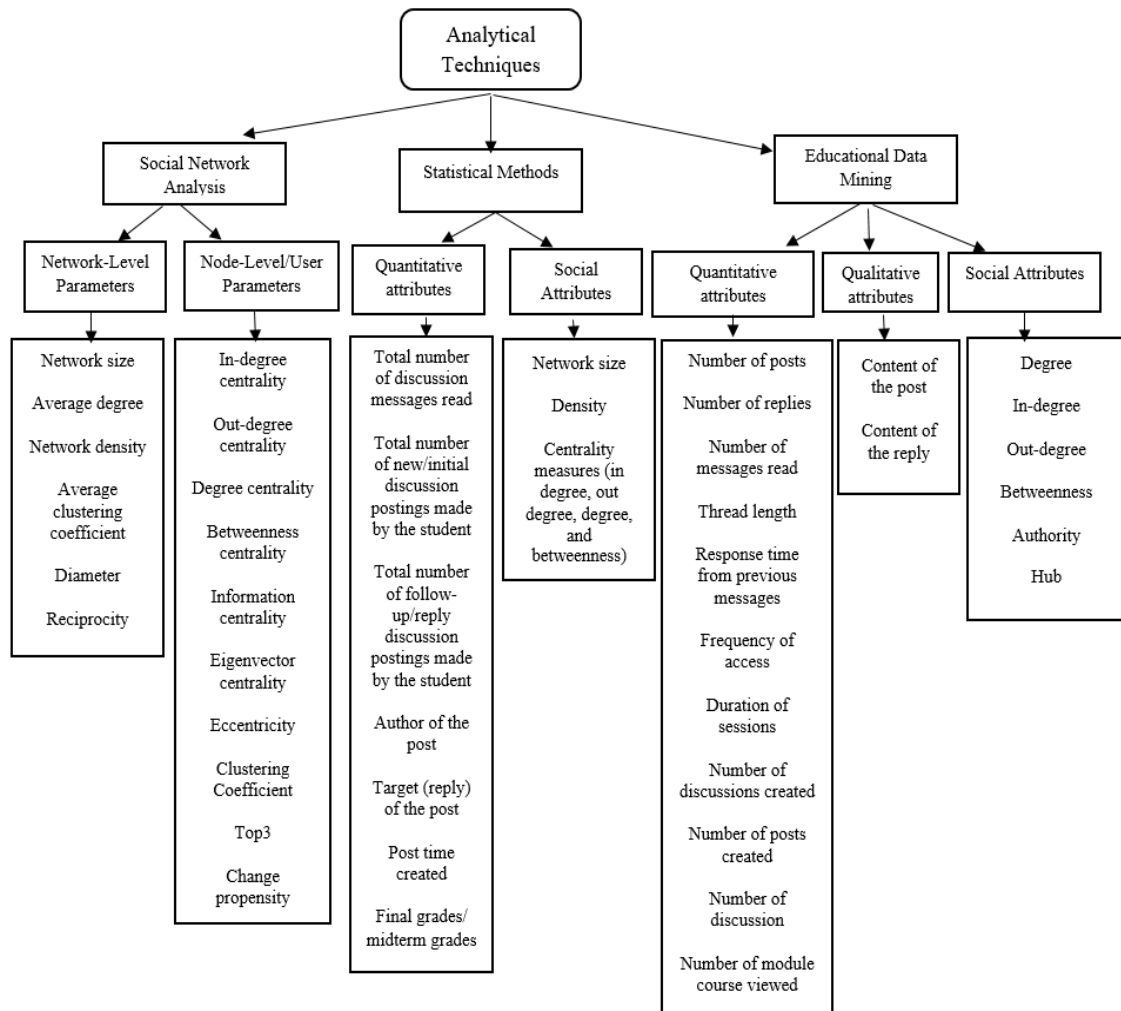


Figure 2. Overall view of analytical methods and attributes

When summarising the above major three analytical techniques, following Table 2 points out the major analysis purpose, features and issues regarding each technique.

Table 2. Summary of analytical techniques

Analytical Method (Technique)	Purpose of Analysis	Features	Issues
Social Network	Visualize the hidden social networks built	Measures are calculated for the	Privacy violation or can be harmful for individual

Analysis	<p>behind discussions, identify the types of interactions among learners and facilitators, measure and compare the positions and the performance of learners in a social network</p>	<p>overall network and for each individual.</p> <p>The optimal position of a learner in the social network is strategically advantageous to that individuals' performance.</p>	<p>standing as it reveals deeper information on each individual.</p>
Statistical Methods	<p>To identify what forum parameters are strongly related with performance and to what extent, how those parameters can be used to predict performance.</p>	<p>Uses techniques such as correlation, regression, standard deviation etc.</p> <p>Selecting only the significant parameters improves the accuracy when correlating interactions with performance.</p>	<p>Sometimes correlation coefficient can take negative values so better to consider the research context when choosing a data analytical method.</p>
Educational Data Mining	<p>Used for describing the dataset, predicting unknown variables from known variables & discovering patterns and rules etc.</p>	<p>Association rule mining, Correlation mining, Sequential pattern mining, Causal data mining can be used for correlating interactions with performance.</p> <p>Classification, Regression and Density estimation is used for performance prediction from interactions.</p> <p>Feature selection algorithms can be used to filter out the best parameters therefore improve the accuracy in prediction.</p>	<p>Not much suitable for small, homogeneous, structured data sets with few variables.</p>

Furthermore, considering the data analysis techniques, this study proposes a new analytical process as depicted in Figure 3 that would be suitable for analysing learner-learner interactions and their relationship to learning achievements.

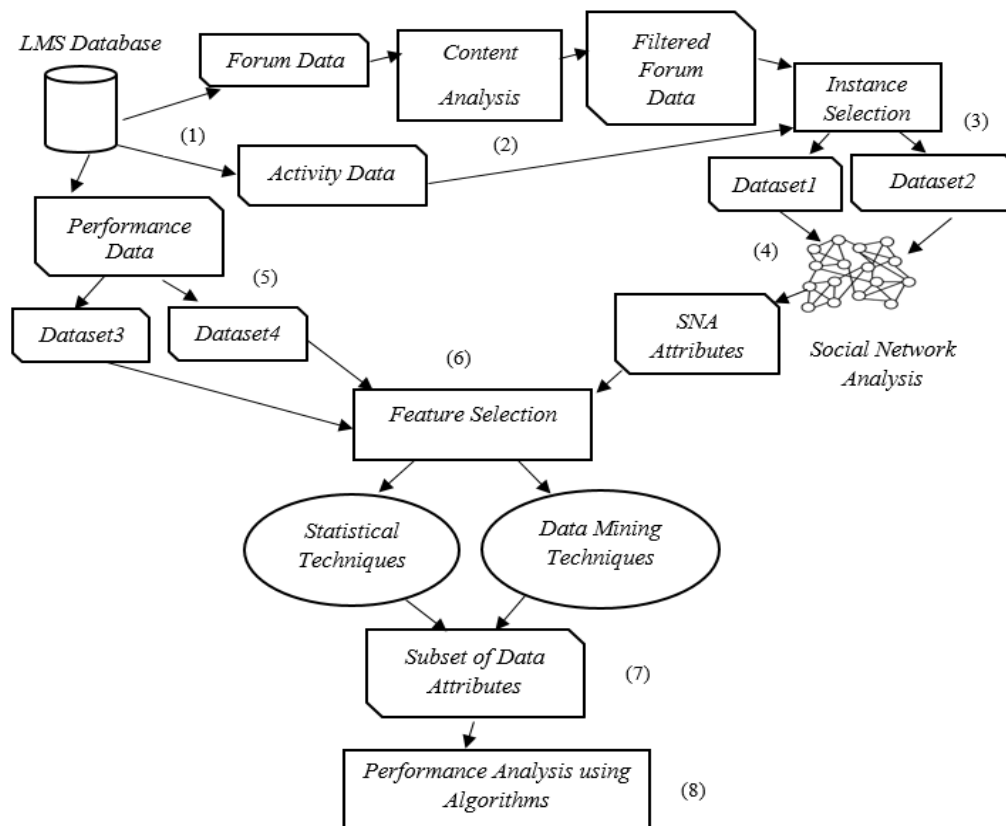


Figure 3. Proposed analytical process

First, the activity data, forum data and performance data (marks) are extracted from the LMS database (Step 1 in Figure. 3). Next, the forum data are analysed and filtered to get only the content related data such as posts which include the discussions related to the topic (Step 2 in Figure. 3). After that those filtered forum data together with activity data such as frequency of access, session durations are organized into two datasets as to consist of data up to mid of the course(Dataset1) and the other to include interaction data until the end of the course(Dataset2) (Step 3 in Figure. 3). Those two datasets are visualized using a SNA tool (i.e. Gephi) and SNA attributes are derived (Step 4 in Figure. 3). Also, the performance data (marks) are divided into two datasets as to contain marks up to mid-term (Dataset3) and until the end of the course (Dataset4) (Step 5 in Figure. 3). Next, the derived SNA parameters together with performance data are processed through feature selection techniques using statistical methods like correlation analysis or data mining techniques using feature selection algorithms (Step 6 in Figure. 3). Then a subset of appropriate data attributes will be derived (Step 7 in Figure. 3). Finally, the selected subset of data attributes is analysed using algorithms like classification, clustering or association rule mining in order to get the relationships between learner interactions and their performance (Step 8 in Figure. 3).

6) Discussion

Due to the novelty of social network analysis as a field, education-oriented SNA research has been very limited, and it has been mostly exploratory by nature (Saqr, Fors, & Tedre, 2018; Isba, Woolf,

& Hanneman, 2016). This was a negative effect of finding relevant papers related to the educational field. SNA plays a major role in visualizing the learner networks and provide deep insights which are not possible using traditional methods. For instance, it identifies the optimal network, risk positions that strategically advantageous to individuals' performance. The optimum use of SNA in evaluating online, collaborative learning should not separate centrality measures from visual analytics, but rather combine them to better understand the context and interpret the inferences of each indicator. Furthermore, SNA attributes are useful for improving the accuracy of both learners' performance and dropout prediction.

Considering statistical methods, Kendall's Tau-b test was effective in measuring the correlation coefficient between ranked network variables. In fact, it could positively correlate almost all centrality scores with the final performance of students. In addition, findings suggest that it is more appropriate to use Permutation test for the statistical analysis of network data due to its capability of omitting inherent error-prone conditions of relational data (Saqr, Fors, & Tedre, 2018). ALM (supported by SPSS 20.0) on the other hand, is a relatively new analytical tool for researchers who regularly use linear regression (Hongwei, 2013). Yet it has given effective results on analysing the impact of learners' interaction data on performance. ALM is effective particularly with very large data where manual handling of the data becomes too time-consuming or work-demanding to be practically accomplished. Also according to the findings reliability, accuracy and effectiveness of ALM are higher than using traditional regression models. According to the findings, Pearson's linear correlation model was another popular statistical analytical method used by many researchers to evaluate the relationship between learner's social interaction data with their performance (Cho, Gay, Davidson, & Ingraffea, 2007; Palmer, Holt, & Bray, 2008; Chowdhry, Sieler, & Alwis, 2014). It is shown that using Pearson's linear correlation model, most researchers could obtain a stronger relationship between social data and performance. However, it is shown that based on the context of the study, these results on correlation could be varied. In some researches, the coefficient has been negative indicating that social data are not correlated with the corresponding final marks of the students (Chowdhry, Sieler, & Alwis, 2014). Therefore we strongly advise to consider the research context when choosing a data analytical method. Finally, a multivariate linear regression analysis could predict performance in moderate accuracy.

While statistical techniques are used for relatively small, homogeneous, structured data sets with few variables, EDM focuses deeply on relatively large, heterogeneous, unstructured data sets with a large number of variables. With the use of complex Machine Learning and Artificial Intelligence, EDM develops computational approaches that combine data and theory to transform practice to benefit learners. EDM is an interactive and iterative approach which follows major four steps; Data collection; Data pre-processing, Data mining and Data Interpretation. There is a large number of parameters that can be obtained from analysing online discussion forums where some can be irrelevant for finding the influence on learners' performance. But some researchers have shown that only using a subset of attributes instead of all available attributes using feature selection algorithms leads to more accurate prediction results. Several types of researches have used several types of data where some used quantitative [22], [22], [31], another set used qualitative data (Pena-Shaff, & Nicholls, 2004; Romero, López, Luna, & Ventura, 2013) and rest used social metrics (Saqr, Fors, & Tedre, 2018; Romero, López, Luna, & Ventura, 2013). Anyway aggregating all three, Romero et al. (2013) showed the mix method leads to higher accuracy in predicting performance from learner interactions.

7) Conclusions

The findings of the study reported in this paper, informed, that it is not sufficient to build new instructional technologies or collaborative settings, but suitable social structures and practices that lead to desired interactions and therefore to achieve performance goals. Therefore the learner behaviours should be thoroughly analysed in several perspectives in order to discover hidden relationships between learners, the causes for their presence and to measure its influence. This survey identified three major methods that had been used to explore the learner behaviour in online learning environments; SNA, statistical methods and EDM.

SNA has become a powerful method to analyse the social structure of the online collaborative environment. As the literature shows, we can conclude that SNA attributes are useful for improving the accuracy of both learners' performance and dropout prediction. Furthermore, this survey highlights that the most appropriate social metrics that influence learner performance can vary from each case study depending on the specific context. Therefore it is important to filter out the most influencing parameters from all available data attributes. In that case, feature selection algorithms, correlation analysis can be used.

Considering the data analysis, the survey concludes to use statistical techniques like correlation analysis for relatively small, homogeneous, structured data sets with few variables. If the dataset consists of relatively large, heterogeneous, unstructured data with a large amount of variables, it is better to focus on data mining techniques like classification, clustering, regression mining etc.

For statistical methods, it seemed Pearson correlation has been used by many researchers. Also it has given good results correlating social data and performance. However, it seemed in several contexts it has given negative results. Therefore it is important to consider the context before choosing a particular data analysis method. More importantly, apart from ALM, all other statistical techniques are not appropriate to use when there is a large amount of data to process. Therefore data mining is preferred.

Furthermore, this survey reveals that performing prediction at both end of the course and before the end leads to more accurate performance classifications. By referring to various case studies which have used several types of data such as quantitative, i.e. number of post views, frequency of page accesses, qualitative, i.e. content analysis of posts and social components, i.e. degree centrality, betweenness centrality, we suggest it is better to follow a mixed approach to gain a richer explanation by aggregating the strengths of each method and eliminating the weaknesses.

Concluding all above suggestions, this survey presents a new analytical process that would be suitable for analysing learner interactions to evaluate its impact on their performance under section 5. This proposed analytical process can be used for obtaining a better understanding of the learner behaviours in online discussion-based learning environments.

References

- 1) Laal, M., & Laal, M. (2012). Collaborative learning: what is it?. *Procedia - Social And Behavioral Sciences*, 31, 491-495. doi: 10.1016/j.sbspro.2011.12.092
- 2) Panitz, T. (1999). Collaborative versus Cooperative Learning: A Comparison of the Two Concepts Which Will Help Us Understand the Underlying Nature of Interactive Learning. 1999.
- 3) Harasim, L., Hiltz, S. R., Teles, L., & Turoff, M. (1995). Learning networks. *Cambridge, Mass.: MIT Press*.
- 4) Renninger, K. (2010). Building Virtual Communities :(*Cambridge: Cambridge University Press*).
- 5) Siemens, G. (2008). Learning and Knowing in Networks: Changing roles for Educators and Designers. *ITFORUM Discuss.*, 27, 1-26.
- 6) Wunsch-Vincent, S. (2007). Participative web and user-created content. *Paris: OECD*.
- 7) Salaway, J., Borreson, C., & Nelson, M.R. (2007). The ECAR study of undergraduate students and information technology. *EDUCAUSE Center for Applied Research*, 6.
- 8) Liu, Q., Peng, W., Zhang, F., Hu, R., Li, Y. & Yan, W. (2016). The Effectiveness of Blended Learning in Health Professions: Systematic Review and Meta-Analysis, *Journal of Medical Internet Research*, 18(1).
- 9) Kirschner, F., Paas, F., Kirschner, P., & Janssen, J. (2011). Differential effects of problem-solving demands on individual and collaborative learning outcomes. *Learning And Instruction*, 21(4), 587-599. doi: 10.1016/j.learninstruc.2011.01.001
- 10) Strijbos, J., Krischner, P., & Martens, R. (2004). What we know about CSCL and implementing it in higher education. *Boston, Mass.: Kluwer Academic Publishers*, 245-259.
- 11) Renninger, K. (2010). Building Virtual Communities:(*Cambridge: Cambridge University Press*).
- 12) De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education*, 46(1), 6-28. doi: 10.1016/j.compedu.2005.04.005
- 13) Pena-Shaff, J., & Nicholls, C. (2004). Analyzing student interactions and meaning construction in computer bulletin board discussions. *Computers & Education*, 42(3), 243-265. doi: 10.1016/j.compedu.2003.08.003
- 14) Saqr, M., Fors, U., & Tedre, M. (2018). How the study of online collaborative learning can guide teachers and predict students' performance in a medical course. *BMC Medical Education*, 18(1). doi: 10.1186/s12909-018-1126-1

- 15) Palazuelos, C., García-Saiz, D., & Zorrilla, M. (2013). Social Network Analysis and Data Mining: An Application to the E-Learning Context. *Computational Collective Intelligence. Technologies and Applications*, 651-660.
- 16) Cho, H., Gay, G., Davidson, B., & Ingrassia, A. (2007). Social networks, communication styles, and learning performance in a CSCL community", *Computers & Education*, 49(2), 309-329.
- 17) Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.
- 18) García-Saiz, D., Palazuelos, C., & Zorrilla, M. (2013). Data Mining and Social Network Analysis in the Educational Field: An Application for Non-Expert Users, *Educational Data Mining*, 411-439.
- 19) Cuéllar, M., Delgado, M., & Pegalajar, M. (2011). Improving learning management through semantic web and social networks in e-learning environments. *Expert Systems With Applications*, 38(4), 4181-4189. doi: 10.1016/j.eswa.2010.09.080
- 20) Borgatti, S., Mehra, A., Brass, D., & Labianca, G. (2009). Network Analysis in the Social Sciences. *Science*, 323(5916), 892-895. doi: 10.1126/science.1165821
- 21) Nahapiet, J., & Ghoshal, S. (1998). Social Capital, Intellectual Capital, and the Organizational Advantage, *The Academy of Management Review*, 23(2), 242-266.
- 22) Palmer, S., Holt, D., & Bray, S. (2008). Does the discussion help? The impact of a formally assessed online discussion on final student results. *British Journal Of Educational Technology*, 39(5), 847-858. doi: 10.1111/j.1467-8535.2007.00780.x
- 23) Chowdhry, S., Sieler, K., & Alwis, L. (2014). A Study of the Impact of Technology-Enhanced Learning on Student Academic Performance. *Journal Of Perspectives In Applied Academic Practice*, 2(3). doi: 10.14297/jpaap.v2i3.111
- 24) Isba, R., Woolf, K. & Hanneman, R. (2016). Social network analysis in medical education, *Medical Education*, 51(1), 81-88.
- 25) Hongwei, Y. (2013). The Case for Being Automatic: Introducing the Automatic Linear Modeling (LINEAR) Procedure in SPSS Statistics, Multiple Linear Regression Viewpoints, 39(2).

- 26) Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems With Applications*, 41(4), 1432-1462. doi: 10.1016/j.eswa.2013.08.042
- 27) Romero, C., López, M., Luna, J., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums, *Computers & Education*, 68, 458-472.
- 28) Romero, C., & Ventura S. (2007). Educational data mining: A survey from 1995 to 2005, *Expert Systems with Applications*, 33(1), 135-146.
- 29) Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601-618.
- 30) Aslam, S., & Ashraf, I. (2014). Data Mining Algorithms and their applications in Education Data Mining, *International Journal of Advance Research in Computer Science and Management Studies*, 2(7), 50-56.
- 31) Cheng, C., Paré, D., Collimore, L., & Joordens, S. (2011). Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance, *Computers & Education*, 56(1), 253-261.
- 32) Khan, T. M., Clear, F., & Sajadi, S. S. (2012). The relationship between educational performance and online access routines: analysis of students' access to an online discussion forum. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, ACM*, 226–229.
- 33) Widyahastuti, F., Riady, Y., & Zhou, W. (2017). Prediction Model Students' Performance in Online Discussion Forum, *Proceedings of the 5th International Conference on Information and Education Technology - ICIET '17*

Impact of Students' Position in the Online Discussion Network on their Learning Performance

This is a conference paper accepted by '526th International Conference on Management and Information Technology (ICMIT)'.

Impact of Students' Position in the Online Discussion Network on their Learning Performance

¹PRAMODI DE SILVA, ²CHANDIMA IMALIKA, ³THUSHANI WEERASINGHE, ⁴AMITHA CALDERA

^{1,2,3,4} University of Colombo School of Computing Sri Lanka

Email: ¹rukshidesilva90@gmail.com, ²cchandima2010@gmail.com, ³taw@ucsc.cmb.ac.lk, ⁴hac@ucsc.cmb.ac.lk

Abstract: Online learning has become a prominent practice among educational institutions due to its power to overstep time, space and cost constraints. Although it lacks face-to-face physical interactions among students and facilitators, they are persuaded to communicate virtually through online collaborative learning platforms. More precisely, through online discussion forums students can interact with peers and expand their knowledge by building a social network among them. However, with the massive number of students in online courses and the limited capacity of Learning Management Systems, it is hard to get deep insights on students' social behavior. For instance, even the facilitators feel difficult to gain a broad image of how actually the students interact with each other, whether they are connected properly or isolated, are they gaining the maximum benefit out of discussions to complement the lack of physical interactions and are the discussions really help them in learning achievements etc. Therefore, this study focused on following an analysis of interaction and assignment data in online learning environment to identify how the position of a student in the discussion network impact his/her learning performance. The research followed Social Network Analysis combined with Statistical and Data Mining techniques. Finally, the study highlighted the importance of considering the connectedness among students rather than only considering the number of interactions by each in evaluating students' performance and productivity of discussions.

Keywords: Collaborative Learning, Computer-Supported Collaborative Learning, Social Network Analysis, Learner Analytics, Educational Data Mining, Performance

1. INTRODUCTION

Most of the academic institutions are attracted more towards the online learning concept due to its power to overstep the space, time and cost constraints. Although, there the students do not get a chance to interact with each other physically, they can communicate through forums, chat messages, activities in online collaborative learning platforms. According to 'Connectivism', a learning theory for digital age, by Siemens [1], learning is no longer an individualistic activity. With the advancement of digital social technologies, learning has become much more complex and it should occur through connections in a social network setting by sharing knowledge. Learning through forming connections is one of the key concepts in collaborative learning where students are supportive for peers' learning and responsible for their own learning. Therefore, the success of one student aids for another one's successfulness [2]. Asynchronous online discussion forums play a major role in replacing physical learning interactions with online collaborative learning interactions. They facilitate students to learn from ideas, shared resources, and experience of each other [3]. Thereby, forums provide an environment to create learning communities and inculcate team spirit. Therefore, discussion forums in online courses can support knowledge production more effectively.

Bachelor of Information Technology (BIT) which is an external degree program in the University of Colombo School of Computing (UCSC), facilitates collaborative learning through its Learning Management System (LMS). BIT students are not receiving any face-to-face teaching from the UCSC. Therefore, to minimize the adverse effects of the learning without physical student interactions, a Virtual Learning Environment (VLE) was introduced using Moodle platform. At the end of each course section, there is a discussion forum. The facilitator and students mainly communicate through these discussion forums. The facilitator is there to help students to find the solutions for course-related problems which might encounter during the course. Not only forums facilitate the students to discuss and work with other students but also it facilitates to evaluate their learning progress using quizzes [4].

Unfortunately, due to a large number of students in this kind of online courses, and as the built-in analytics of major LMSs including Moodle offer only limited insights on students' social behavior, facilitators are unable to observe, monitor and evaluate students' learning behaviours in order to provide the facilitation in a more informed manner. For instance, even the facilitators can't have a broad image of how actually the students interact with each other, whether they are connected properly or isolated, are they gaining the maximum benefit out of forum discussions to complement the lack of physical interactions and are the discussions really help them to perform well etc. In a context like this, where formal learning is completely virtualized, monitoring students' online interactions could reveal hidden patterns in how the variations in their interactions affect their learning performance. Past researchers have informed that the social network built within a Computer Supported Collaborative Learning community had a

perceptible influence on individual performance and the central positions of students within the emergent collaborative learning network resulted in higher levels of learning performance [3], [5]. Hence, identifying the position of a particular student in the social network and its impact to his/her learning performance is more beneficial in order to scaffold their learning. This will help facilitators to identify the weak and isolated students who are at risk of failure and provide them with additional personalized support through simplified learning content and necessary instructions. Furthermore, with the understanding gaining through monitoring of online learning behaviors can help students to enhance their networking skills as well as communication skills and social capital by rewarding the active online presence of them.

Therefore, this study is focusing on addressing the research question:

1. How do students' positions in the online discussion network affect on their performance?

In order to answer this, the study mainly focuses on answering the sub-questions;

- 1.1. What type of student and facilitator networks are built behind the online discussions?
- 1.2. What social network parameters best interpret the students' position in the online discussions?
- 1.3. How the students' positions in this network can affect their learning performance?

By understanding the students' position in the network and its impact to the performance, facilitators can use this knowledge to optimize the use of online discussion forums for knowledge production.

2. DETAILS EXPERIMENTAL

Learning is a collective outcome achieved by means of social connections built from the social networks [6]. Many researchers have followed Social Network Analysis, Statistical Analysis and Data Mining techniques to analyse student interactions and performance in online learning environments. Considering them, this study followed a methodology as depicted in Fig.1, which was a combination of best practices used by past researchers [3], [7], [8], [9].

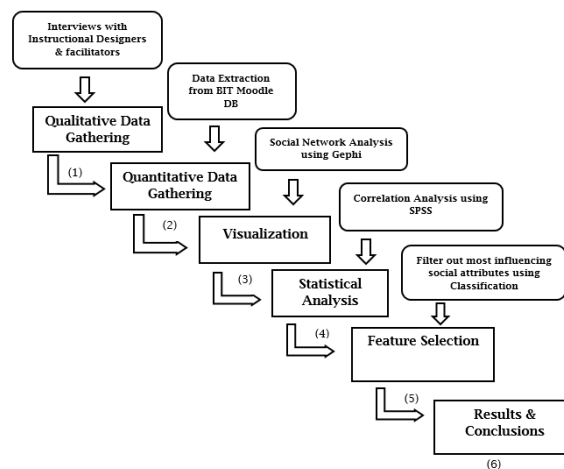


Fig.1. Methodology.

2.1. Qualitative Data Gathering (Step 1 in Fig.1)

A face to face interview was conducted with two BIT facilitators in order to collect information on BIT course, discussion forum and assignment design, course delivery, students' participation in forums, the structure of the course and the quizzes.

2.2. Quantitative Data Gathering (Step 2 in Fig.1)

Structured Query Language (SQL) was used to extract the data related to forum interactions and assignment grades from the database of the BIT VLE for one course for a one full semester. The study only considered the messages and discussion threads which are more appropriate and relevant to the subject area (content filtering). The dataset for the selected course 'Computer Systems I' initially included 17 forums, 32 discussion threads, 99 forum posts, 57 students and 1 facilitator. After data pre-processing by removing the missing targets and irrelevant posts to the forum topic, there were 16 forums, 26 discussion threads, 93 posts, 57 students and 1 facilitator.

2.3. Visualization (Step 3 in Fig.1)

Forum interactions of the participants were visualized using Social Network Analysis techniques provided through various functions in Gephi tool. Gephi has used in similar researches for analyzing the social networks built behind when interacting with each other. For instance, in criminology, it has been used to study collaboration between offenders. Gephi has multiple inbuilt algorithms/layouts for network visualisation and among those 'Forced Atlas' layout was used which is popular in the research field [3]. It also provided calculations for social network parameters in terms of network-level and user level. They describe the position of each student in a social network. They are further explained in section 2.3.1 and 2.3.2.

2.3.1. Network-level parameters

One of these network-level metrics is the *network size*, the total amount of students in a network. If this is high, the level of student participation for the forum is high [3]. The average number of posts posted and received by each student is indicated by *average degree*. This measure implies the average level of interactivity of students in the forum [3]. The *network density* is also a useful measure as it is the ratio of actual interactions between students to the total possible. If this is high, students are participating efficiently [3], [7], [10]. Other meaningful network-level metric is the *diameter*, the largest number of students needed to pass over to come from a particular student to another. Low diameter makes easy to interact with peers [7], [10].

2.3.2. User-level parameters

Rather than considering the overall interactions of students, focusing on each student might reveal much more significant insights. Therefore the user-level parameters play an important role.

In-degree centrality is the number of replies received by each student. It indicates the popularity/attractiveness of a student and peers are more likely to interact with this type of students [3], [7], [8]. And, the number of posts/messages posted by each student in the forum is measured by *out-degree centrality*, an indicator of how active a student in the discussion [3], [7], [8]. The sum of both above measures is given by *degree centrality*, and it implies how influential a student within the network. *Betweenness centrality* is the number of times a student comes in-between others. In this way, the participant connects the unconnected peers and thus facilitates communications and acts as a bridge or broker of information exchange. So helps to identify which students and facilitators may spread the information quickly and effectively across the class. Next measure, *closeness centrality* gives the inverse of the distance between a student and all other peers indicating how close a student to his peers and therefore how easy to reach and interact with others [3], [7], [10]. Moreover, the sum of inverses of distances between a student and all other peers is calculated by *harmonic closeness centrality* which implies how close a student to his peers [11]. The *eigenvector centrality* estimates the social capital and the influence of one's ego network. Connections to well-connected or important students in the network bring high values [10], [12], [13]. Another important measure, *eccentricity* implies how far a student from his peers (level of isolation). Higher values indicate less connectedness to peers, therefore difficult to reach [3], [5]. The *clustering coefficient* is the proportion of actual edges between a student and his neighbour peers to the total possible edges. High values impress the student more likely to work with peers in the group [3], [10]. There are another three most popular measures *pageranks*, *authority* and *hub*. Using *pageranks*, the students in the network are ranked according to their importance [10], [13]. It uses *hub* or how many highly informative (important) students a particular student is pointing into. Students recommend each other based on the information they share [7], [8], [10], [13]. Also *pageranks* use *authority*, the amount of valuable information a particular student holds helps to identify the students with higher knowledge or skills [7], [8], [10], [13].

2.4. Statistical Analysis (Step 4 in Fig.1)

Social attributes obtained from social network analysis were then mapped against assignment grades of each student to measure correlation coefficient between ranked variables. Since network data are prone to violate the traditional assumptions of conventional statistics (normal distribution and independence) [14] selecting the Pearson correlation might not give correct output [15]. Instead, the relationship between the two variables can be better described by Kendall's Tau-B test which is a nonparametric equivalent to Pearson's correlation [16]. Correlation coefficients were calculated using SPSS software which is widely used in social science researches to perform statistical analysis of data [3], [16], [17].

2.5. Feature Selection (Step 5 in Fig.1)

Next, the study further focused on whether these SNA parameters can be used to classify the students as pass or fail correctly, and to what extent (Step 5 in Fig. 2). According to research done by Romero et al. [8], reported that using a subset of these attributes instead of all available social attributes leads to improving classification accuracy. Therefore the study used two classification algorithms; Naive Bayes and Random Forest, which had been showed

high accuracies by similar researches in classifying students' performance using forum data[7], [8]. The tool used for feature selection was 'Weka', which is widely used by Educational Data Mining field. First, all the derived social attributes were sorted from higher correlation value to the lowest. Then using all attributes, classification accuracy was recorded. Gradually the features were removed one by one from the bottom of the list, which was ranked from higher correlation to the lower correlation values and classification accuracies were recorded for each iteration. Then, the attribute subset of highest classification accuracy was selected as the best feature subset for the course.

2.6. Results & Conclusions (Step 6 in Fig.1)

Finally, the results and conclusions are derived.

3. RESULTS AND DISCUSSION

First, the data drawn from online discussion threads were visually analyzed on the course level to gain an idea on how the social network has been structured behind online discussions and how each participant positions in it.

3.1. Types of student and facilitator networks in online Discussions

Forum participation was assessed by the number of posts by each student and facilitator. The interactions among them represents their communication through replies to others' posts. Therefore, the graph outlines the structure of the course and the patterns of interactions.

As depicted in Fig.2, each node (circle) in the graph denotes to a participant, FT represents the initial Forum Topic that students are discussing about, F2 is the facilitator and rest are students. Each edge (arrow) corresponds to an interaction, the arrowheads represent the direction of the interaction. The size of each node is relative to its degree centrality, colour intensity relative to the betweenness centrality, and the thickness of edges represents the frequency of interaction.

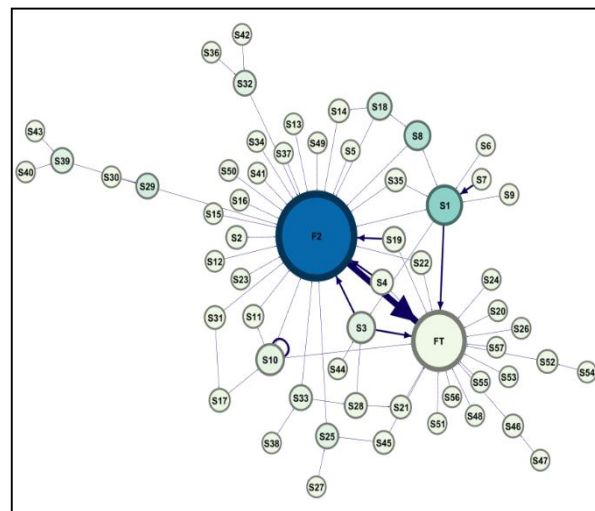


Fig.2. All the interactions of course 'Computer Systems I'

By interpreting the role of the participant in these built networks, we can identify three main interaction types.

- I. Student - Student interactions
- II. Student - Facilitator interactions
- III. Student - Content interactions

By considering these identified types regarding quantity and influence of those helps to provide a general idea about the course structure. According to the sociogram in Fig. 2, facilitator has the highest degree centrality (larger node) and betweenness centrality (darker node) values. Therefore, in this course, facilitator is receiving most of the messages and at the same time he is actively moderating the entire discussion environment. Therefore we can inform that Interventions of the facilitator might be another parameter that causes such variations in the graph. Furthermore, it shows that facilitator has properly done his duty by guiding, mediating the discussion to promote interactions within the discussion as students tend to communicate via facilitator rather than directly posting to the Forum Topic (FT). It shows that, the most outstanding student in this course is S1 who has highest degree and

betweenness centrality compared to other students. That means he might be posted high number of messages while collaborating with peers. S3 also shows a similar behaviour. Moreover, it is visible that S7 frequently communicates or replies to S1. Therefore, it informs that S1 plays a major role in improving the collaboration of the discussion and we can suggest that, appointing S1 as a leader in peer learning activities may be more effective. Also it is visible that, though the students like S32 and S29 are far away from the majority, as they have larger node sizes (high degree centrality) and dark colours (high betweenness centrality), they have managed to keep collaboration and get most of the use of online forums. This may indicate that facilitator will not need to specially concern on these students. Another highlighting result is S10 has frequently interacting with himself. That might be due to continuously adding new information to the already posted message. Therefore, it may inform that S10 is searching new knowledge continuously and updating others also.

Likewise, analysis of the positions of students in the social network build behind the online discussions can reveal hidden behaviours of the students. This can be used by facilitators to plan the course activities. For instance, the students who initiate and mediate the discussions such as S1, S3 can be used for peer learning, while using students who update frequently such as S10 to inspire others.

3.2 Students' position in online Discussion forums

To provide further insights from the Social Network Analysis, centrality measures were calculated for the obtained dataset which numerically implies the position of the students in the social network built behind the online discussions.

3.2.1 Network level parameters

The centrality measures calculated for the entire network is depicted in Table1.

Table 1: Network level social parameters

Course	Network size	Average Degree	Network Density	Diameter
Computer System I	59	1.39	0.024	7

The results show that fifty-nine participants (including forum topic and facilitator) have involved in discussion forums. If this is the same as number of students registered for the course, we can elaborate that the discussion seems to be more useful and participants get the maximum out of it. The network density implies the efficient participation among the participants in the forums, here as it is a considerable value, we can say that students are actively collaborating in forums. When considering the network diameter, high diameter implies students have to pass a large number of students to come from that a particular student to another, it makes difficult to interact with their peers. As here the diameter is relatively low, it makes easy to interact with peers and therefore improve the productivity of online discussions.

3.2.2 User level parameters

Although it is important to have an overall view of the status of collaborative learning in a course, it is more important to find prominence of a node (learner) in order to interpret its link with the performance. Therefore, social network parameters for each participant were calculated. Node size was configured by *out-degree centrality* (outgoing interactions) to demonstrate the information giving participants, where students with more outgoing interactions have larger nodes. Colour intensity relative to the *eigenvector centrality* and the thickness of edges represents the frequency of interaction. Therefore, darker the node the eigenvalue of the node is high which means the node has well-connected to neighbours who are the important nodes of the network (who has more connections).

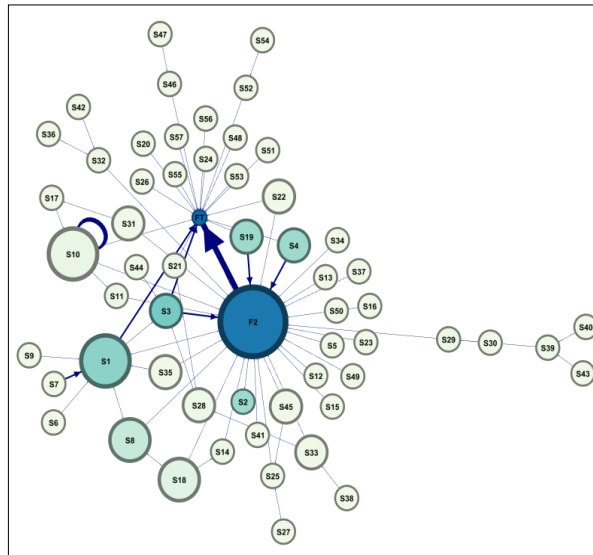


Fig. 3. Information Giving Network of course 'Computer Systems I'.

Fig.3. shows that most students are actively participating in discussions, and the network is dominated by students like S1, S8, S10, S18 who have the highest prestige (larger node size), and S1, S3, S4, S19 are the students who have higher social capital indicated by high eigen centrality values (dark nodes colour).

When analysing the graph, it seems that S3, S4, S19 students has higher centrality values for selected social parameters (degree, Out-degree and Eigen centrality) which are an indication of active participation and existence of higher social capital. Interestingly not only these students are actively participated in the forum but also, they have scored higher marks for the assignments. Thus, with this analysis, we can suggest that students with a high number of interactions and social capital in the forum are likely to get good scores, a fact to be analysed using data analysing techniques.

3.3 Correlation of the social network parameters with students' grades

Since visualisation provided a notion on the existence of the link between social parameters and student' grades it should be further clarified by correlating those social parameters with assignment marks using Kendall's Tau-B Test. If the sig. (2-tailed) value is less than or equal to 0.05, the correlation value is significant, if not there is a relatively low correlation between the two data [9].

Table 2: correlation for the course: 'Computer Systems I

Correlations				
		Pass Mark		
		Correlation Coefficient	Sig. (2-tailed)	N
Kendall's Tau_B	In-degree Centrality	.095	.376	57
	Out-degree Centrality	.249	.023	57
	Degree Centrality	.145	.166	57
	Eccentricity	-.146	.152	57
	Closeness Centrality	.186	.059	57
	Harmonic Closeness Centrality	.193	.047	57
	Betweenness Centrality	.062	.554	57
	Authority	.174	.102	57
	Hub	.101	.308	57
	Pageranks	.113	.275	57
	Clustering	.185	.088	57
	Eigen Centrality	.145	.164	57
	Pass Mark	1.000	.	57

As depicted in Table 2, in the course 'Computer Systems I', only *eccentricity* shows a negative relationship with the pass mark. This implies that, when a student is less collaborative and far from peers and seems to be isolated, then he/she tends to less perform in assignments. Additionally, there are two significant values regarding the correlations, *out-degree centrality* (0.249 and 0.023), *harmonic closeness centrality* (0.193 and 0.047). That means the number of posts by each student has an impact on their performance in assignments. The rest; *closeness centrality*,

clustering, authority, eccentricity, eigen centrality, degree centrality, page ranks, hub, in-degree centrality and betweenness centrality have a moderate correlation with the pass mark.

As there are altogether twelve social network parameters which describe each student's position in the social network, further investigation needed to identify what are the most appropriate social parameters that predict the student grade. Therefore, the study used two classification algorithms; Random Forest and Naïve Bayes to filter out which subset of the parameters gives higher accuracies. The results obtained are depicted in Fig.4. The x-axis depicts which social parameter was removed in each iteration whereas y-axis shows the classification accuracies in percentages.

As depicted in Fig.4, in this course, the classification accuracies obtained by Random Forest algorithm was less than the results provided by Naive Bayes. Sometimes, accuracy has declined beyond 50% in Random Forest. Therefore, only the classification pattern given by Naive Bayes was considered. When removing the least correlated attribute from the feature set, it depicts that classification accuracies are in a same level and suddenly it drops when 'pageranks' measure was removed. So, for this course, the best feature subset includes *out-degree centrality, harmonic closeness centrality, closeness centrality, clustering, authority, eccentricity, eigen centrality, degree centrality* and *Pageranks* (top correlated parameters including *Pageranks*).

That means, not only the quantity of interaction or the message count (*out-degree centrality and degree centrality*), but also the student's social capital (*eigen centrality*) also affects their learning. Moreover, how closely interacting the students are with their peers (*harmonic closeness centrality, closeness centrality, clustering*) is another predictor of performance. When the students are less collaborated and seems to be isolated (high eccentricity), study showed that it negatively affects on their learning performance. This is a significant finding which again confirms the need for collaboration for the improvement of learning [3], [8].

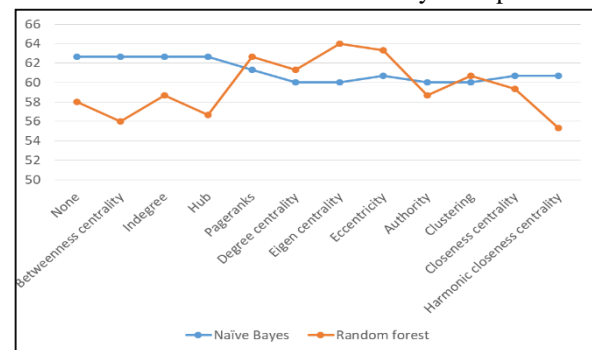


Fig. 4. Classification accuracies for course 'Computer Systems I'.

CONCLUSIONS

Using the Social Network Analysis and statistical methods, this study could provide useful knowledge on how a student's position in the social network build behind each online discussion forum impacts on his/her learning performance. Therefore, students' performance can be evaluated using the position of the student in the social network in terms of these nine social attributes. That means a student who possess high values for his/her *out-degree centrality, harmonic closeness centrality, closeness centrality, clustering, authority, eigen centrality, degree centrality, pageranks* and low values for *eccentricity*, compared to another student in the same discussion forum, the first mentioned student is more likely to perform better than the other in online assignments. Some student monitoring systems only consider the number of posts by each student (*out-degree centrality, in-degree centrality, and degree-centrality*) for evaluating students. Most of the LMSs also provide only that facility [8], [18]. Therefore, based on a student's position, facilitators can have insights on who are the influential students in the course and what are the pitfalls in collaborative learning process etc. This much insights are not possible using traditional mechanisms which only consider the counts of messages, but ignore the significance of the structure and social relationships. Therefore this study is a fine example where SNA exposes the invisible sides of online collaborative learning and its impact on learning.

As this case study focused on only one course, it can be tested to various types of courses with different subject areas. Therefore, future research can be focused on comparing the results derived for theoretical subjects and practical subjects to get more insights on how the course context affect their online behaviour. Also, the courses without a facilitator might uncover totally different behaviours. Moreover, it will be more beneficial to add a content analysis to assess the richness of posts in the forum and its contribution to improving learning performance.

ACKNOWLEDGMENTS

Our sincere thanks does not enough for the kindness of our supervisor, Dr. Thushani Weerasinghe, our co-supervisor, Dr. Amitha Caldera, Prof. K.P.Hewagamage and the staff of the e-Learning Centre of the UCSC.

REFERENCES

- [1]G. Siemens, "Connectivism: A learning theory for the digital age", International Journal of Instructional Technology and Distance Learning, 2005.
- [2] Laal M, Ghodsi SM. Benefits of collaborative learning. *Procedia Soc Behav Sci.* 2012;31:486–90.
- [3] M. Saqr, U. Fors and M. Tedre, "How the study of online collaborative learning can guide teachers and predict students' performance in a medical course", *BMC Medical Education*, vol. 18, no. 1, 2018.
- [4] T. Weerasinghe, "Designing Online Courses for Individual and Collaborative Learning: A study of a virtual learning environment based in Sri Lanka", Department of Computer and Systems Sciences, Stockholm University, 2015.
- [5]H. Cho, G. Gay, B. Davidson and A. Ingraffea, "Social networks, communication styles, and learning performance in a CSCL community", *Computers & Education*, vol. 49, no. 2, pp. 309-329, 2007.
- [6]J. Brown and P. Duguid, "Organizational Learning and Communities-of-Practice: Toward a Unified View of Working, Learning, and Innovation", *Organization Science*, vol. 2, no. 1, pp. 40-57, 1991.
- [7] C. Palazuelos, D. García-Saiz and M. Zorrilla, "Social Network Analysis and Data Mining: An Application to the E-Learning Context", *Computational Collective Intelligence. Technologies and Applications*, pp. 651-660, 2013.
- [8]C. Romero, M. López, J. Luna and S. Ventura, "Predicting students' final performance from participation in online discussion forums", *Computers & Education*, vol. 68, pp. 458-472, 2013.
- [9] F. Widyahastuti, Y. Riady and W. Zhou, "Prediction Model Students' Performance in Online Discussion Forum", *Proceedings of the 5th International Conference on Information and Education Technology - ICIET '17*, 2017.
- [10]D. Khokhar, *Gephi cookbook*. 2005.
- [11]"Centrality", *En.wikipedia.org*, 2018. [Online]. Available: <https://en.wikipedia.org/wiki/Centrality>. [Accessed: 03- Nov- 2018].
- [12]T. Weerasinghe, R. Ramberg and K. Hewagamage, "Designing online learning environments for distance learning", *INSTRUCTIONAL TECHNOLOGY*, 2009.
- [13] Á. Hernández-García, I. González-González, A. Jiménez-Zarco and J. Chaparro-Peláez, "Applying social learning analytics to message boards in online distance learning: A case study", *Computers in Human Behavior*, vol. 47, pp.68-80, 2015.
- [14]Isba R, Woolf K, Hanneman R. Social network analysis in medical education.*Med Educ.* 2017;51(1):81–8.
- [15]T. Bergin, *An Introduction to Data Analysis: Quantitative, Qualitative and Mixed Methods*. 2018.
- [16]A. Field, *Discovering statistics using ibm spss statistics +spss version 22.0*. [Place of publication not identified]: Sage Publications, 2014.
- [17]S. Chowdhry, K. Sieler and L. Alwis, "A Study of the Impact of Technology-Enhanced Learning on Student Academic Performance", *Journal of Perspectives in Applied Academic Practice*, vol. 2, no. 3, 2014
- [18]K. Hewagamage, K. Nishakumari, T. Weerasinghe and G. Wikramanayake, "'Motivating Student Discussions'-A Strategy to Develop Online Learning Community in the BIT Virtual Learning Environment", *Distance Learning and Education--International Proceedings of Computer Science and Information Technology*, 2011.