Analysis of Students' Digital Activities via A Learning Analytics Approach

Lappeenranta–Lahti University of Technology LUT

Industrial Engineering and Management (GMIT) Master's Thesis

2021

Chao Jiang

Examiner(s): Professor Leonid Chechurin

Associate Professor Kalle Elfvengren

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Engineering Science

Industrial Engineering and Management (GMIT)

Chao Jiang

Analysis of Students' Digital Activities via A Learning Analytics Approach

Master's thesis

2021

92 pages, 37 figures, 15 tables

Examiner(s): Professor Leonid Chechurin

Associate Professor Kalle Elfvengren

Keywords: Learning Analytics, Social Network Analysis, Digital Activities, Discussion Forum, Visualization.

The trend of digitalization is unstoppably spreading in all industries and aspects of our life including education. Remote education was picked by most universities and institutions after the outbreak of Covid-19, however, the evaluation of students' performance in the virtual learning environments (VLEs) became a challenge to the teacher.

This research was carried out to investigate the feasibility of solving the problem with tools from Learning Analytics (LA). Network attributes, centrality, and positional analysis under Social Network Analysis (SNA) were selected as parameters for evaluation of students' behaviours in discussion forum.

Graph, and matrix were studied as media to present and communicate the result of analysis, techniques such as hierarchical clustering, Multidimensional Scaling, and Blockmodeling were tested for visualization of results.

ACKNOWLEDGEMENTS

Gratitude from the bottom of my heart to my supervisor, Leonid Chechurin, who pulled me out of the "one-man's prison" during my darkest days in the pandemic period. I am fortunate to have a such erudite mentor to guide me on my way to the end of master thesis project research. (Thank you so much for your time both for meeting and tea break)

Much obliged to Kalle Elfvengren for being my second supervisor.

To Vladimir Balin and his wife Anastasia, thank you for the company during my tough time, and the help you offered when I was in quarantine.

To Apurva Potdar, Arnob Khan, Tamara Ahmed, thank you for the altruistic support in my study and life.

I would also like to thank my family that raise me up into a robust grown-up, and the endless support and care in my whole life. I love you!

Last, I want to thank LUT university for giving me a precious opportunity to pursue knowledge in the land of the curious.

30.09.2021, Lappeenranta

Chao Jiang

SYMBOLS AND ABBREVIATIONS

Roman characters

N	node in the network
D	nodal degree
C_D	degree centrality
C_B	betweenness centrality
C_C	closeness centrality
P_D	degree prestige

Abbreviations

LA	Learning Analytics
VLMs	Virtual Learning Environments
LMS	Learning Management System
EDA	Educational Data Mining
CoI	Community of Inquiry
SNA	Social Network Analysis
MDS	Multidimensional Scaling
SP	Social Presence
CP	Cognitive Presence
TP	Teaching Presence
TRIZ	Theory of Inventive Problem Solving
LDA	Latent Dirichlet allocation

Table of contents

Figures

Tables

1 Introduction

Online learning became a new normality for university students after the outbreak of Covid-19 pandemic, all the teaching sessions are provided by the teachers remotely in a cyber world. Along with it came a head aching issue, teachers are feeling more difficult to recognize how well their students are performing in the classes. Quite often, the students don't even turn on the camera or microphone, how can teachers still having a grasp of the class without seeing and or hearing their students?

This master thesis research aims at helping teachers to retain the control over the big picture, in another word, be aware of students' engagement and performance in the online learning space via a Learning Analytics (LA) approach.

1.1 Background

As computer-related technologies advance, digital skills become unprecedently important in our life. Companies now require candidates who are applying to any positions in any industry to obsess proficient Information and Communications Technology (ICT) skills to perform daily work tasks. Schools and universities also constantly stay updated to the latest digital technologies to conduct teaching and researching activities. Moreover, entertainment, social services are also embracing the unstoppable trend of digitalization, music, and movie streaming services such as Spotify and Netflix have almost send traditional physical media such as CD and DVD to grave. In everyday conversation, term "digitization" and "digitalization" are always being used interchangeably. But as Théberge (2015, pp. 345–354)argued that term "digitization" emphasizes more on the technological process of converting streams analogy information into digital formats with binary values (0 and 1) to present information, while the term "digitalization" focusses less on the technology sides, but more about the impact of digitization to our society, life, and changes to the stakeholders involved. In general, digitalization is more about WHAT (the content), while digitization highlights HOW (the technology).

Nowadays, learning has intensively cuddled with computer-related technologies to better complete its mission of knowledge diffusion. GOSA (no date)discussed digital learning as learning activity facilitated by technology to allow students to study over control of time, space, path, and or pace. External constraints such as time, space from traditional learning environment are no longer obstacles on students’ way to knowledge acquisition. Personalized adaptive learning becomes possible as individuals do not need to fellow the pace of the entire class with designated learning path prepared by instructor which is more popularly recognized as a matter of fact not suitable for everyone. As the saying goes, there are no two identical leaves in the world, all students should be considered as unique. Digitalization of learning makes it possible for students to find the way suit them most for personal development with support of diverse learning activities.

As discussed above, digital learning is a quite broad concept which can be characterized by the adoption of technology in learning activities. Digitalization could happen to both education infrastructure and learning content. For instance, new classrooms in most developed countries or regions are equipped with computer, projector, multimedia system. Teachers and students interact with the help of digital equipment. Massive amount of literature, books, articles, keynotes are now available in digital format. We can now access to almost unprecedentedly abundant of information than ever before with just one clock on the mouse at home without physically visiting to library or places required. Many learning strategies can be classified as digital learning including but not limited to blended learning, online learning, gamification, personalized learning. One good example in real life is the current education move from conventional classroom based to online remote teaching during the Covid-19 pandemic situation. The implement of remote learning plays a significant role in the battle where everybody needs to hold his/ her responsibility to fight against the virus.

1.1.1 Online learning

Online learning is a crucial component of digital learning, when the learning is moved to a online environment, both the digital infostructure and content will be applied into practical learning activities. In general, online learning can be categorized as synchronous online learning, asynchronous online learning and blended learning based on the means how students, tutors and the teacher participate the learning activity. Kaplan (2018) argued that in higher education, online learning can be categorized by two dimensions (time distance, and number of participants) into 4 sub-groups as shown in Table 1.

Table 1: 4 types of online learning in Higher Education

MOOCs (Massive Online Open Courses)	Unlimited number of students participate asynchronously
SMOCs (Synchronous Massive Online Courses)	Unlimited number of students participate synchronously
SPOCs (Small Private Online Courses)	Limited number of students participate asynchronously
SSOCs (Synchronous Small Online Courses)	Limited number of students participate synchronously

In this master thesis project, the type of online learning being studied is SPOCs, where the class size is relatively small comparing to MOOCs, and the way of students generating interaction is asynchronous in a form of text-based discussion forum post.

1.1.2 Synchronous online learning

In the synchronous scenario, all involved members (students, tutors and teachers) participate the learning activities as listed in Table 2 simultaneously on the internet. Platform such as Microsoft Teams, Zoom, Google Meet enable teacher and tutors to establish a virtual classroom, video call, real time chatting will replace face to face teaching and indoor communication. Students have almost the same learning experience as traditional classroom-based learning with extra freedom of space (not time in this case).

1.1.3 Asynchronous online learning

On the other side, asynchronous learning does not require students, tutor and teachers to interact simultaneously. The unobtainable freedom of time in synchronous online learning is realized in asynchronous online learning. Students have more freedom to control their pace of learning in most cases. Also, tutors and teachers can interact with learners more flexibly on schedule.

1.1.4 Blend/ hybrid learning

Blend learning or hybrid learning combines online learning and traditional off-line classroom-based learning. Except for participation of online learning activities, students are also required to involve in physical meeting with teacher, tutors and peers in form of seminar, workshop or other means. The course ***Research Methods for Master Students*** taught by Prof. Daria Podmetina and tutor Iryna Maliatsina in LUT University was conducted in the blend learning approach between September and December in 2019 at Lappeenranta, Finland. Asynchronous online learning and physical meetings were used for course delivery.

1.1.5 Learning activities in the online setting

Table 2: Difference between activities in synchronous and asynchronous context

	Synchronous	Asynchronous
Activities	Online lecture attendance, real time discussion, real time assignment, gaming	Video watching, asynchronous discussion, Assignment, Exam, Gaming

1.2 Research Objectives

This master thesis project was carried out with a primary expectation to unearth how would Learning Analytics (LA) help the teaching staff such as teachers to have better understandings regarding the students' behaviours when the teaching activities are carried out in SPOCs. What kind of outcomes can LA produced that could be further utilized as basis for teachers to make decision such as in-course mediation, final grading, and customized feedback giving.

1.3 Research questions

Guided by the research objectives, the following research questions came into view consequently:

1. What does LA do?
2. What tools are available in LA for teachers to use that can capture students' online performance?
3. How can teachers present and communicate the analysis results of students' online behaviour?

1.4 Research method

This master thesis project is an empirical case study with the expectation to understand how LA can benefit the teachers while making decision in scenarios such as offering guide during the course for students that are not closely followed or taking as basis for final grading and feedback giving.

1.4.1 Research Philosophy and theory development

The research "onion" diagram created by Saunders, Lewis and Thornhill (2016) as shown in Figure 1summarized the research philosophy into 5 types, in this thesis project, the guiding philosophy is pragmatism as the goal of this thesis project is to find out what tools from the learning analytics discipline can be utilized by teachers to distinguish students based on their performance on the online virtual learning environment. The focus of this research is practical problem driven, and the expected outcome of the study are practically usable solutions to a specific need. This work developed the theory via a deductive approach, a case study was done to test the Community of Inquiry (CoI) theory.

1.4.2 Methodological choice

The methods used in this research is mainly quantitative as the evaluation of students' activity are mostly statistical analysis. Case study was picked as guided by the chosen research philosophy, as it provides the "real-life" trial to test the theory, and it can be further easily converted to usable solutions or outcomes for solving problems in the real world.

1.4.3 Data Collection

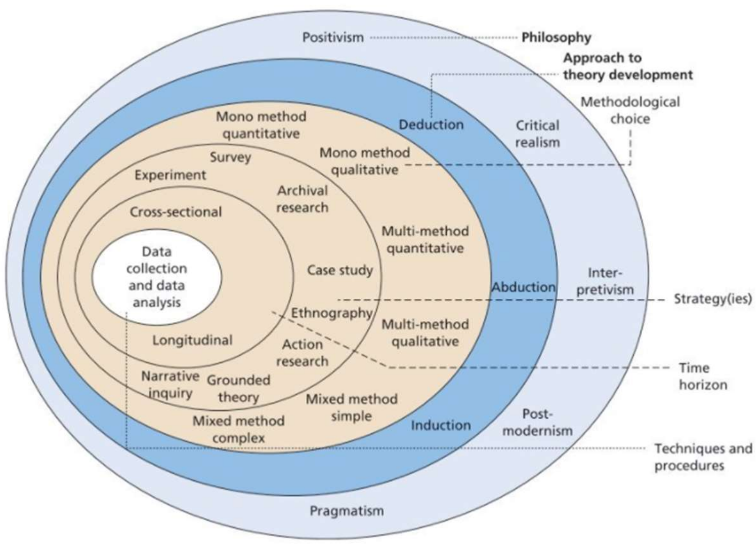

Figure 1: the research "onion" (Saunders, Lewis and Thornhill, 2016, p. 124)

This research used secondary data for a practical test of the selected LA tools and the CoI model, the data was a log file extracted from Disqus LMS which contains the information about the students' interactions records, and contents of posted in text format. Besides, a self-generated data was used to give the sample data a reference basis.

1.5 Structure of thesis

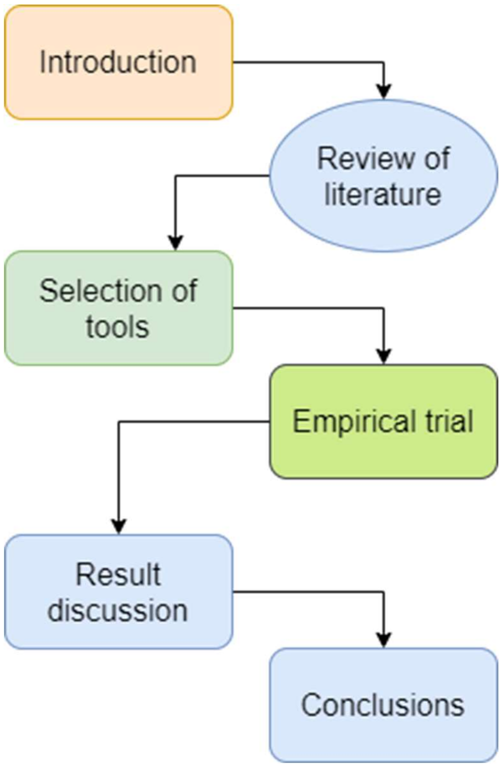

Figure 2: The structure of thesis work

As clarified in Figure 2, the first chapter of introduction will give the audience a general understanding of the background where this research lies, and where the possible outcomes will be applied.

To pursue the answer for the first research question, a review of literature was done where scientifically published papers and books were studied in the second chapter and third chapter, after this, the job of learning analytics should be recognized.

The fourth chapter presents the selected tools for the teachers to seize the students' behaviours and the mechanism inside the black box was attempted to be explained as well. A chapter for empirical trial of the tools selected were performed right in the following chapter.

The results of the practical test of the tools were discussed in the discussion chapter and accompanied with the conclusions chapter which summarized the background, findings, and limitation in a concise piece of writing.

2 Learning Analytics

LA is a relatively new discipline as it came to researchers' eyes about 1 decade ago. The number of recorded publications on the topic of Learning Analytics on Scopus database started to take off around 2011 when the 1st International Conference on Learning Analytics and Knowledge was held in Banff Alberta, Canada. As Figure 3 shows the number of recorded publications increase drastically after 2011 before reaching its peak in 2019 at 927 articles.

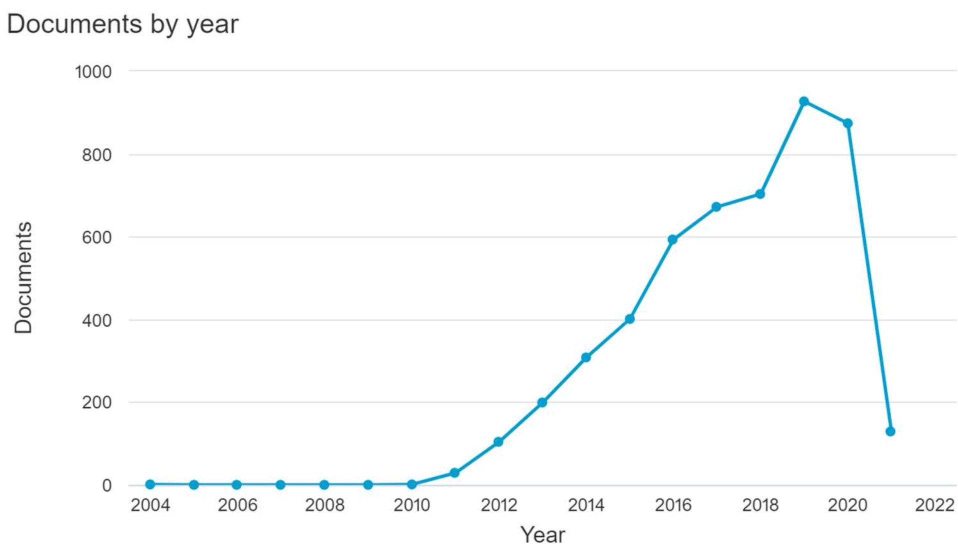

Figure 3:Number of documents on Scopus in "Learning Analytics" by year

Figure 4 tells that Computer Science and Social Science are the 2 largest study fields with significant contributions to research world on learning analytics with the ration of 43.6 percent and 26.7 percent, respectively. Following by Engineering (9.7%), Mathematics (6.8), and decision science (2.9%).

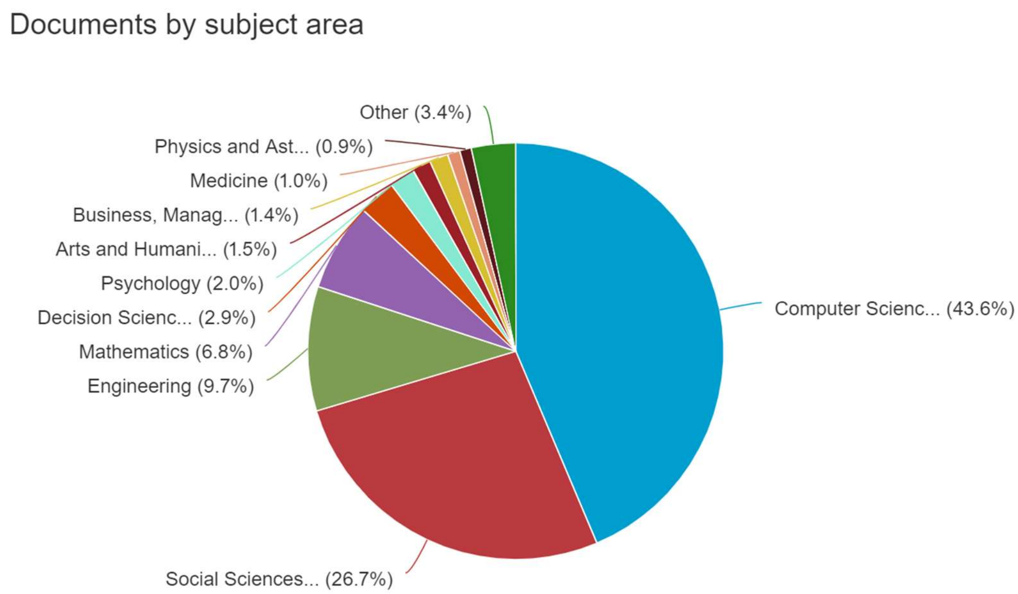

Figure 4: breakdown of LA articles by subject on Scopus

As visualized in Figure 5: US, Spain, Australia, UK and Germany are the top 5 countries that publish most of the research articles on Learning Analytics with indexability on Scopus by March of 2021.

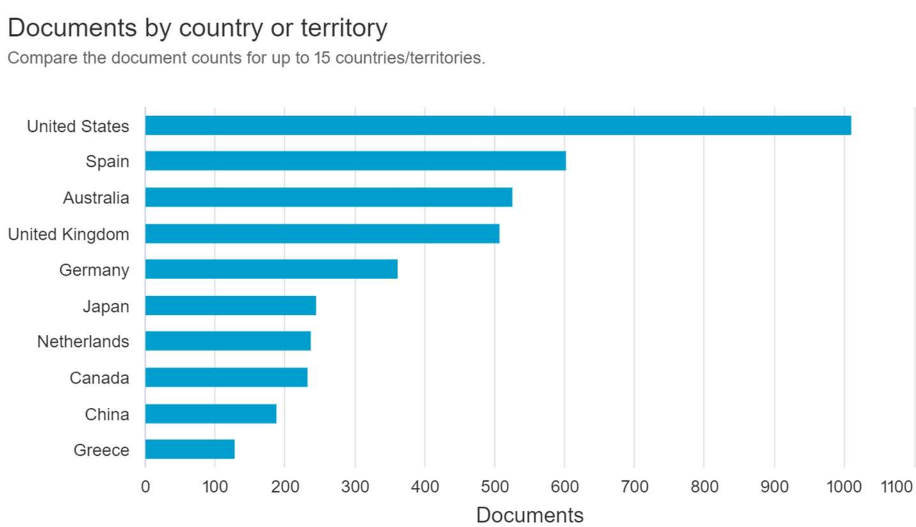

Figure 5: Documents counts of LA by country or territory.

2.1 Origin of LA

Way before Big Data, Online learning penetrating into the education sector, The Open University in UK already started to conduct research about students' distant learning by monitoring their progress course by course for ten years.(McIntosh, 1979, pp. 77–86) Tinto's research about students persistence and engagement and emphasize on integration of academia and society inspired people's attention on research about students drop out.(Tinto, 1998, pp. 167–177) In the early 21st century, the information read/write web application made data more easily to collect and analyse.(Berners-Lee, Hendler and Lassila, 2001, pp. 34–43)

Like many other scientific research fields, the definition of LA is not exclusive. As the community of LA develops, new definition will be brought to the table to update, correct, or replace the outdated ones. Siemens (2010)pointed out that the most widely accepted definition of LA was introduced in the 1st International Conference on Learning Analytics and Knowledge: 'Learning analytics is the measurement, collection, analysis, and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs.' (Siemens, 2010)

Another definition from the perspective of business intelligence was given by Cooper in 2012:

'Analytics is the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data.' (Cooper, 2012, pp. 1–10)

LA overlaps with many other disciplines such as Educational Data Mining (EDM), Academic Analytics. EDM and LA both use massively generated educational data to help researchers to understand and promote learning. However, they differ from priority of objectives. A lot of comparisons between LA and EDM have been made by researchers. Siemens and Bake (2012, pp. 252–254)discussed in the 2nd International Conference or Learning Analytics and Knowledge that EDM focuses more about the automation of discovery in big dataset, while LA emphasize on the influence of change to people's judgement about learning. LA study the entire system, while EDM works on components inside the system. LA has background in Semantic Analytics, and EDM is developed from software develop and modelling. In another word, EDM cares more about the Machines or technology, whereas LA sets people's need as priority. EDM and LA could develop together to help educator, students to solve the awaiting problems in learning.

Until 2005, the main research interest of EDM was about human-computer interaction, followed by prediction study. Baker, Yacef (2009, pp. 3–17) argued that DM and Machine learning (ML) could be used to help the educators to evaluation learners' performance and give learners better guidance and feedback.(Zaiane, 2001) Fer guson (2012, pp. 304–317) pointed out that the research started to move to learning focused perspective from 2003 onward.

The timeline of milestones of EDM and LA are listed chronologically as shown in Figure 6. The First International Conference of Educational Data Mining is 3 years earlier than the First International Conference of Learning Analytics and Knowledge.

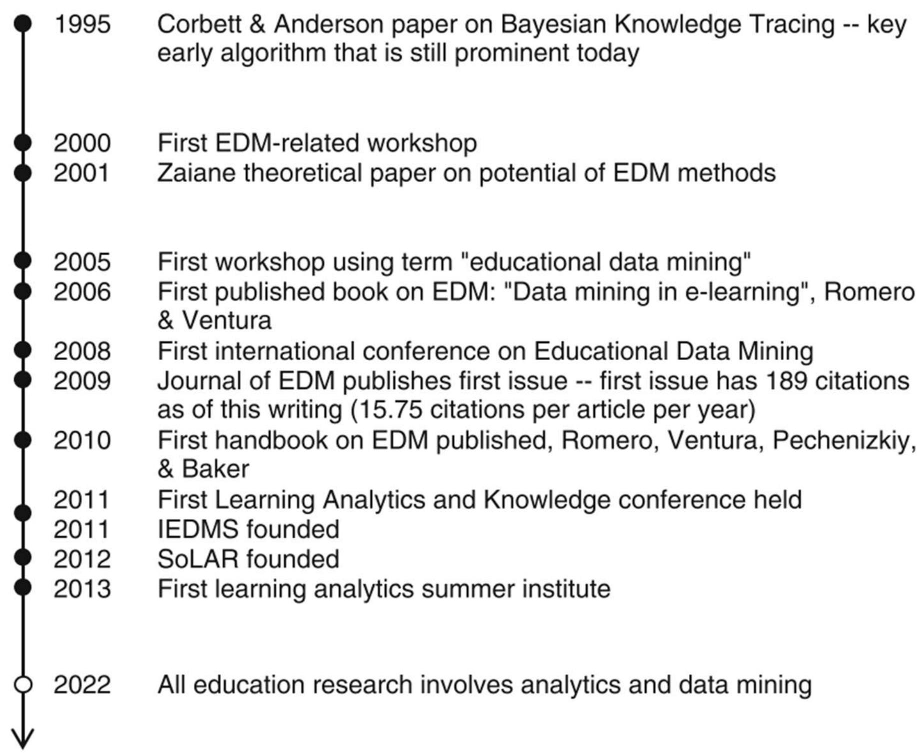

Figure 6: Important milestones of EDM and LA (Baker and Inventado, 2014, pp. 61–75)

2.2 Factors drive LA development

2.2.1 Big Data

Big data is one of the main drivers foster the development of LA. As Data from educational sectors is the key input for analytics. Business started to use big data to unearth patterns for extracting of valuable information to help executes make decisions, provide personalized content recommendation and advertisements. Big data became a hit, and other sectors also started to seek application of big data for improvement. In the education sector, the broadly use of VLEs, which is also named LMS such as Moodle enables institutions to collect massive amount of data with rich information. Moreover, Researchers can access to dataset generated from different platforms and settings. Longitudinal study and cross-sectional study are possible to be done with abundant learning-related datasets. The need of extracting useful information from the big datasets in the educational sector urges LA to develop.

2.2.2 Online Learning

The move of education from traditional classroom-based to Online environments can give students more freedom of space and or time to progress their learning possibly at their own comfortable pace. However, new problems also arise alongside the benefits. For example, students might feel being isolated because of the lack of active interaction with peers and teacher in the online environment. Teachers and tutors cannot provide needed cues on time to trigger students' reactions. Dringus, Ellis (2005, pp. 141–160)also mentioned that teacher might face unprecedent difficulty to evaluate and interpret students learning outcomes due to the large number of students and tons of discussion contributions produced in a relative long period (6 to 12 weeks). They might also get lost of direction and lose the engagement and motivation for study.(Mazza and Dimitrova, 2004, pp. 154–161)

2.2.3 Political/ Economic Challenges

Campbell (2007)addressed that many countries and territories demand institutions to measure, demonstrate and improve their performance. The US government clearly articulated the aims of increasing the overall educational attainment in the population and has been prepared to spend billions of dollars to realize it. (Norris *et al.*, 2008, p. 42) To promote and optimize learning outcomes at national or international level becomes another driver with a political/ economic viewpoint.(Ferguson, 2012, pp. 304–317)

2.3 The framework of LA

Chatti et al (2012, pp. 318–331) Proposed a framework for LA from four dimensions: WHAT, WHO, WHY, and HOW. As illustrated in Figure 7 below, WHAT refers to data and environment, WHY questions the objectives of learning analytics, HOW stands for the techniques could be applied for LA, and WHO includes all stakeholders who will receive benefits from LA.

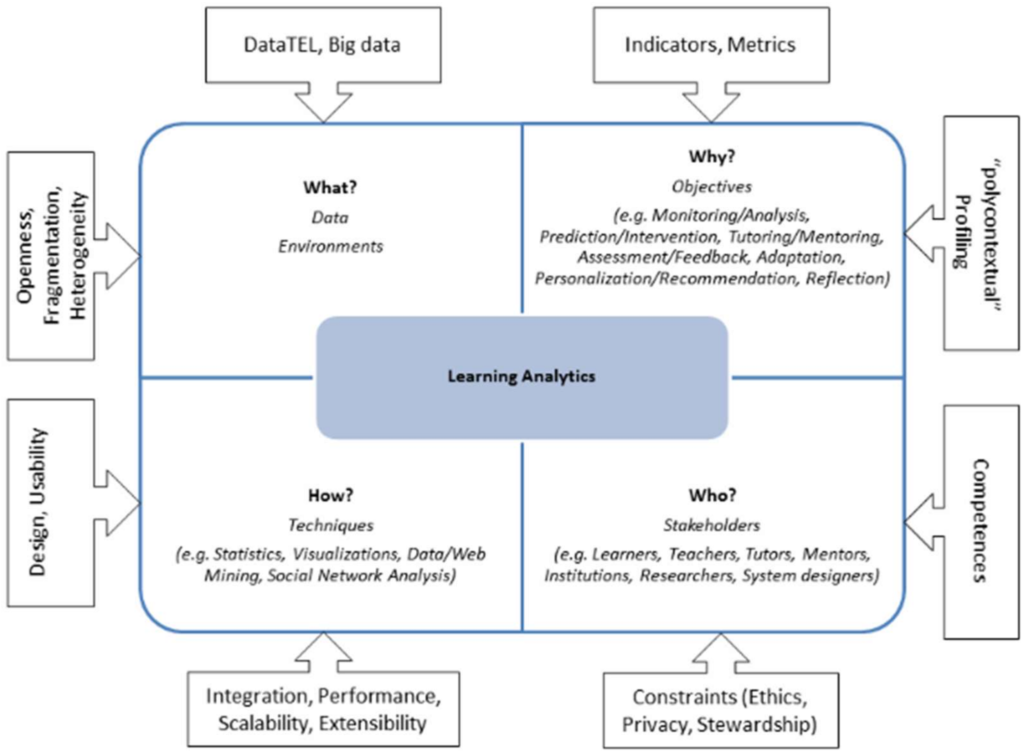

Figure 7: Learning Analytics reference model (Chatti *et al.*, 2012, pp. 318–331)

Greller and Drachsler (2012, pp. 42–57)added 2 more dimensions (Internal Limitation, External Constraints) to the framework of LA.

2.3.1 Stakeholders

Greller and Drechsler (2012, pp. 42–57)divided stakeholders into two main sub-groups: Data subjects as a group of learners, and data clients such as tutors and discussion moderator. Data subjects are supplier of learning-related data, students' interaction, personal information, learning progress could be collected easily in VLEs with implement of LMS such as Moodle or blackboard. Data clients could be roughly referring to people who will use the collected learning-related data for research or analysis. In the simplest case, the data client will be the teacher who collect and analyse educational data produced by students. But the clients could also be extending to higher level as institution, government. In some cases, Data subject and Data Clients could be the same group of people, as in the scenario students receive feedback form the

LA system, then the students are data subjects and data clients at the same time. When teachers and tutors participate in the learning activities, conduct trials of pedagogies, they also become the supplier of data. Chatti et al. (2012, pp. 318–331) proposed that intellectual tutoring system and system designer are also stakeholders of LA.

Students can get reflection of their study from the LA system and visualized feedback of their progress compared to the overall class performance. Teachers can monitor students' activities in the course, find students at risk during study and delivery intervention to help lagging students to complete the course. Institutions can obtain a finer picture of students drop out rate and graduation rate with possible manners to rearrange the course or curriculum. (Greller and Drachsler, 2012, pp. 42–57) LA tools should be able to provide reflection and feedbacks to all stakeholders, increase their self-awareness, assist the future decision making.(Chatti *et al.*, 2012, pp. 318–331)

Beyond the formal learning setting (school scenario), stakeholders could be expanding to more target groups, as lifelong learning prevails, the characters of data subject and data clients can be forgoing.

2.3.2 Objectives

The prominent objective is to extract useful but hidden information in the massive learning-related datasets to delivery values to all stakeholders in the LA system. Monitoring of information flow and interactions of learners inside the learning community can not only help learners to learn more efficiently, but also help the institution to optimize the organizational management process. (Butler and Winne, 1995, pp. 245–281) Objectives of LA could be classified into 2 groups: Reflection, and Prediction.(Greller and Drachsler, 2012, pp. 42–57)

Reflection

Greller and Drachsler (2012, pp. 42–57)described reflection as critical self-evaluation of a data client after receiving feedback based on their own dataset to obtain self-knowledge. However, reflection can also be seen as a client's self-evaluation based on other stakeholders' dataset. For instance, teachers can evaluate his/her own teaching performance by reflection of students learning outcomes.

Prediction:

Apart from benefiting the stakeholders in the end with delivery of reflection, LA can be used to predicting and modelling learners' behaviours and activities which could enable early intervention for "at-risk" students support or personalized adaptive learning. (Siemens, 2011)

2.3.3 Data

Greller and Drachsler (2012, pp. 42–57) stated that access to dataset is still a challenge for most researchers, as most dataset are not open to researchers. The simplest reason is the belongingness of the dataset, and concerns about privacy, ethnical issues. Lack of common dataset format, methods to anonymize and pre-processing the data for privacy and legal protection rights (Drachsler *et al.*, 2010, pp. 2849–2858). Standardized documentation of dataset informs user how to use it and license about to what extent can use and share/ distribute the datasets.

2.3.4 Techniques/ Instruments

By reviewing publications on the International Conference of Learning Analytics and Knowledge between 2011 and 2020. Research interests was classified as Figure 8 below. Most researchers focus on reflection (feedback), followed by modelling and prediction of students learning behaviours and activities. Many different technologies can be used for LA to manipulate the dataset for generating useful information for all stakeholders. The commonly used technologies including informational retrieval technology such as EDM, traditional statistics analysis techniques, social network analysis.(Siemens and Gasevic, 2012, pp. 1–2), Natural Language Processing (NLP).

Chatti et al. (2012, pp. 318–331) classified LA techniques into 4 main categories: Statistics, Information Visualization (IV), Data Mining (DM), Social Network Analysis. Different techniques should be adopted based on research objective. Now, we might have rich data but at the same time poor information, hence development of LA tools to convert raw data into meaningful information is one of the challenges of Learning Analytics from the technique Perspective.

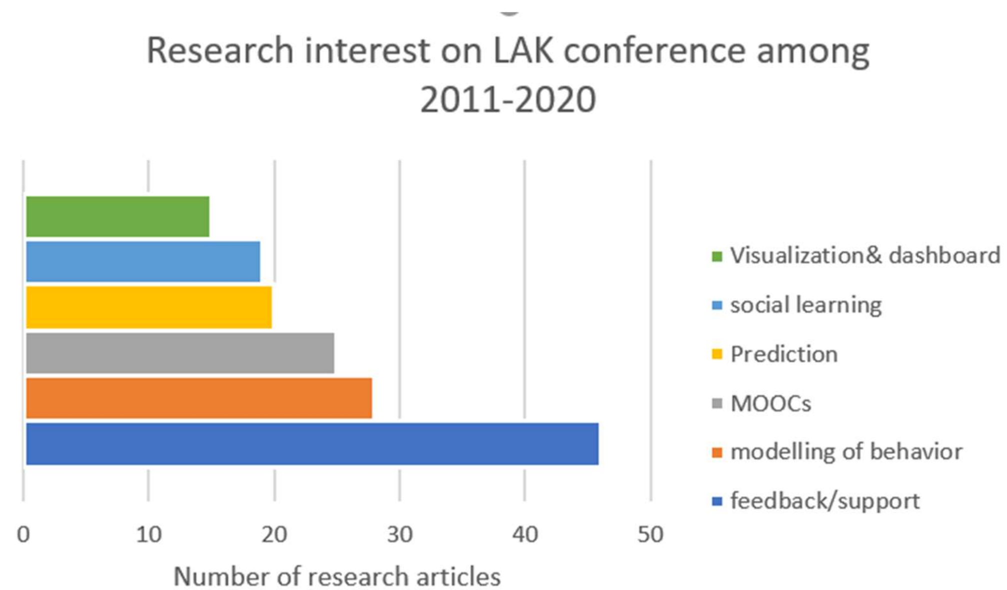

Figure 8: Research interest on LAK conference between 2011 and 2020

Internal Limitation

The internal limitations that restraint LA develop are mainly caused by competence, and acceptance. Reffay and Chanier (2003, pp. 343–352) warned that LA outcomes are algorithmically attained which requires researchers to have high level of competence to interpretative and presentative skills to make use of. Appropriate exploitation of educational data requires a lot of experts with qualified knowledge and skills to optimally interpret and report for further developing of practical action suggestions for learning improvement. Visualization as a powerful tool to present outcomes, widely used by researchers in LA, but the simplicity and attractive display of information might cause misleading. (Greller and Drachsler, 2012, pp. 42–57) And critical evaluation is also a big factor that crucially danger the development of LA. As evaluating a student based on a single source of data with limited parameters is objective, especially as nowadays the lifelong learning is rooted in people's mind, the learning context is extended to a much wider ecosystem.

The acceptance of LA is another player prevent the advance of this discipline. The goal of LA is to generate practical suggestions for action to optimize learning, without sufficient acceptance, the outcomes of LA will only be left as theoretical contributions. To increase the general public's trust on LA, new evaluating methods for LA tools should be given attention (Ali *et al.*, 2012, pp. 470–489)and on advanced technology acceptance models (Venkatesh and Bala, 2008, pp. 273–315)

2.3.5 External Constraints

External constraints come from two perspectives: Privacy and Ethnics. Because of these issues, learner or the data subjects might be a root obstacle for data generation. Legislation and policy need to be matured to help stakeholders to cope with privacy and ethnics issues in LA. As discussed in the Data section above, the data subjects' resistance to data collection could worsen the challenge for researchers to access to the dataset.

Privacy issue could be translated as collecting data from learners without "informed consent" (Ess and Jones, 2004, pp. 27–44)New technologies such as camera, LMS make students not even realizing that their data are being collected. In most cases, learning data and student personal data are managed separately by academic staff and IT faculty, this could be another obstacle for researchers to do comprehensive study as integration of learning data and personal data is necessary in some cases. Another cause of ethnic issue is the abuse of data, this could harm both students and teachers. Administrator might use the collected data for inappropriate purpose; therefore, institutions should consummate policies regarding the use of data.

Greller and Drachsler (2012, pp. 42–57) also pointed the potential harm to the student's development, as all the prediction and modelling techniques are providing analysis on average behaviours in the learning community, outliers are not managed which could cause deterioration of creativeness and innovation.

2.4 LA in CHINA

The investigation of LA development in China is mainly assessed with statistical information from 2 biggest Chinese Research Databases: CNKI (China National Knowledge Infrastructure, 中国知网), and WANFANG DATA.

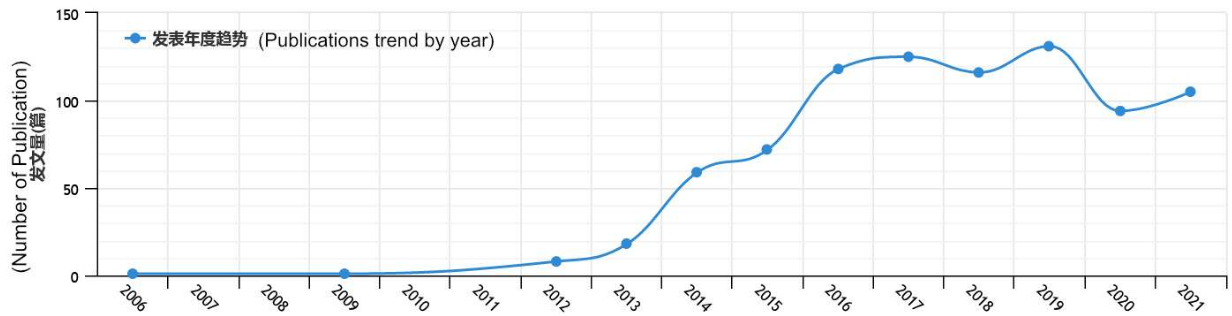

Figure 9: Publication within LA discipline in China

As Figure 9 tells, the number of publications started to climb after 2011 which is understandable as the first LAK conference was held in that year.

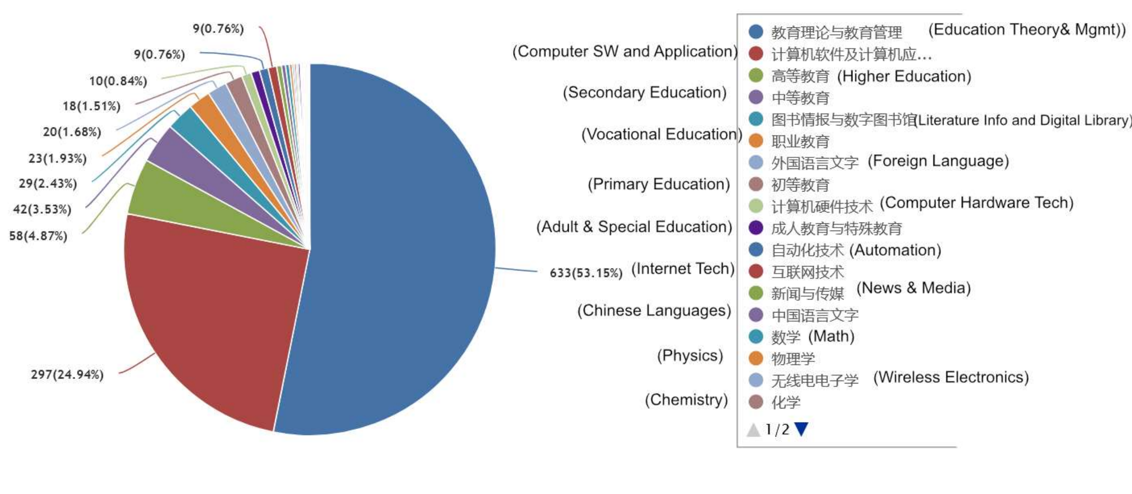

Figure 10:Research articles within LA in China breakdown by subject

One noticeable difference about LA study in China is that most publications are placed in Education Theory and Management subject, followed by Computer Software and Application, Higher Education, Secondary Education, etc. (Figure 10)

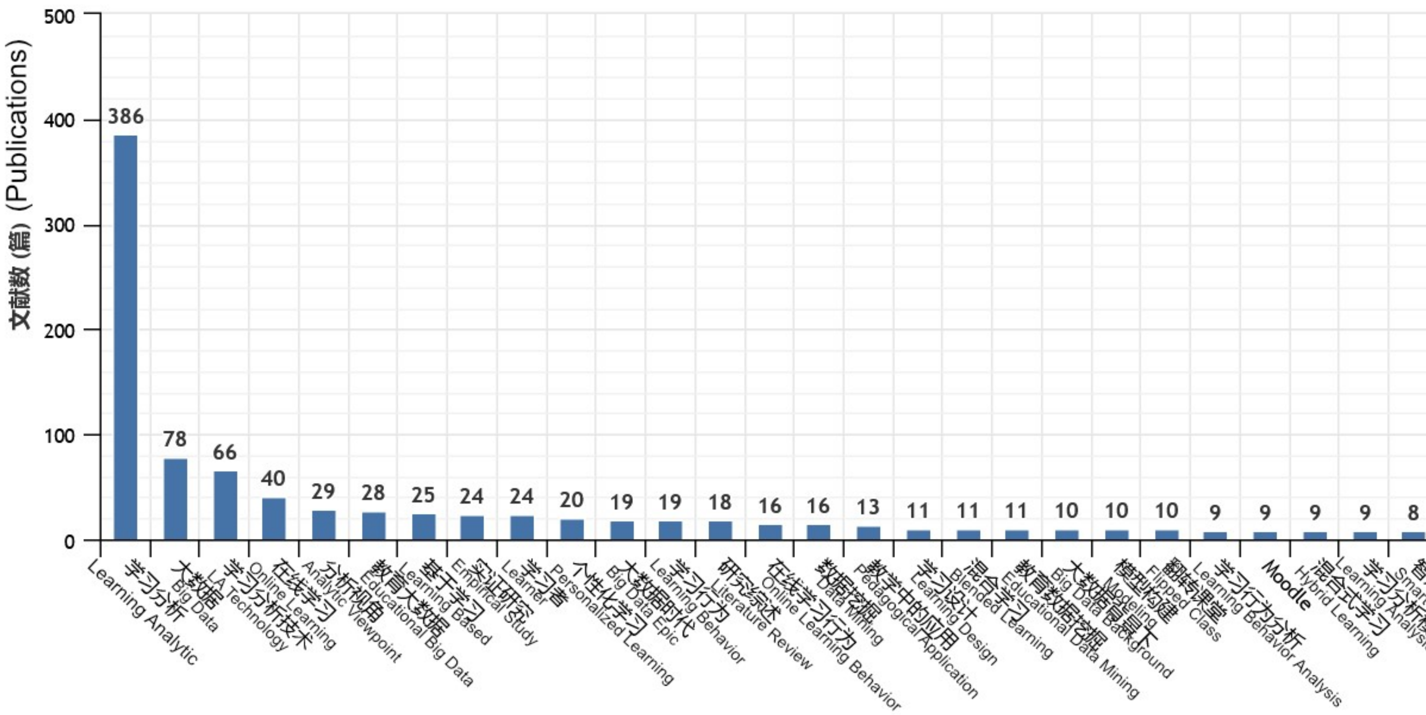

Figure 11:Publications count by sub-topic.

Figure 11 shows that Big Data, LA technology, Online Learning, Analytic Viewpoint, and Learning Based are 5 most popular sub-topics. And Figure 12 enlightens that most LA research are carried by normal universities in China with East China Normal University, Beijing Normal University, Central China Normal University, and Northeast Normal University dominating the top 4 rank.

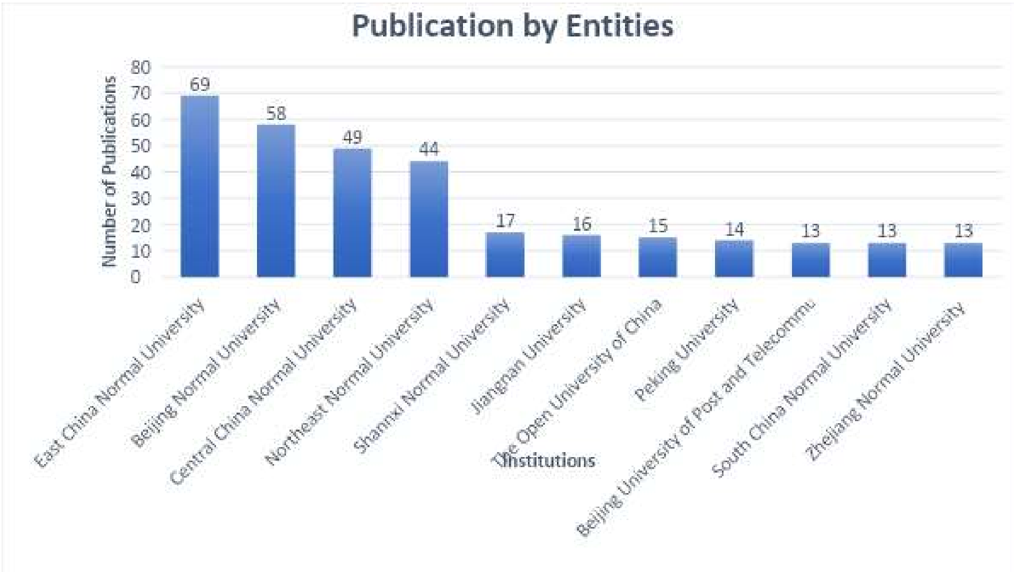

Figure 12:Top 10 Entities that produce LA publications in China

Modern Distance Education Research, Distance Education China, e-Education Research, China Educational Technology, and Open Education Research are 5 journals with high impact factors which relates to Learning Analytics discipline most. Visualized statistical description of research interests with the published articles in the 5 aforesaid journals in the past 10 years are charted in Figure 13, apparently MOOC is the hottest topic with 315 articles counted in the 5 journals from 2011 to 2021. Educational informatization and AI are followed with 272 and 164 publications respectively.

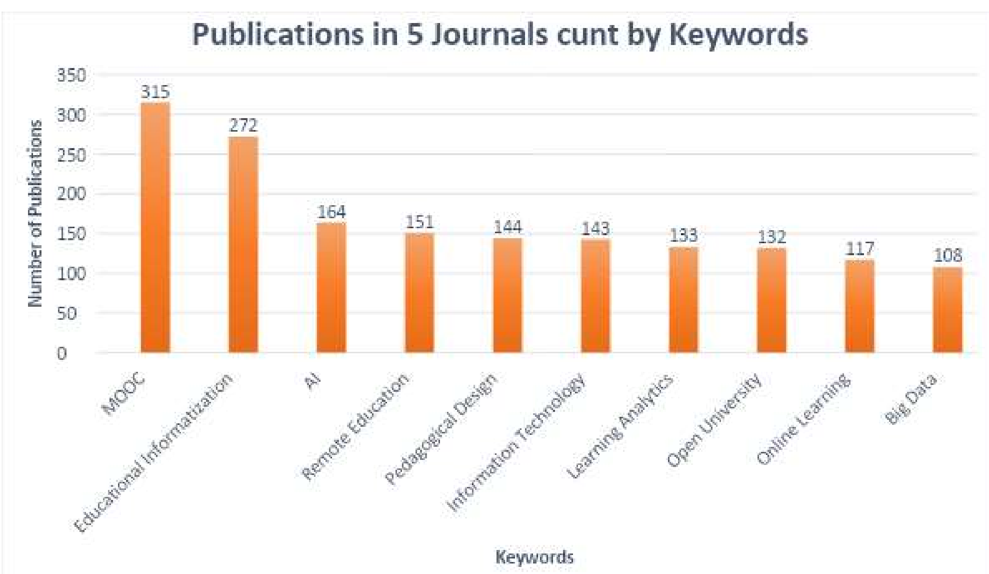

Figure 13: Popular topic of publication in the aforementioned 5 journals in last 10 years

Gu and Hu (2020, pp. 40–42) commented that the development of MOOCs had a significant influence on growth of LA as the previous interest were mainly MOOCs learners' behaviours analysis, Online learning quality and outcome estimation, and design for learning intervention.

As the blended learning evolve, there is a need for LA to be able to collect and process data collected from hybrid learning environment. And The focus of LA should return to people instead of technologies. LA and Learning Design should be treated as twins which means these two fields need to be more appropriately coordinated and harmonized.

Hu and Zhang (2018, pp. 184–191) noted that research interest of precision teaching in China took off around 2013 with increasing number of scientific publications published in CNKI onward. Most of publications in this filed are also from normal universities. The development of precision teaching in China could be classified into 4 stages. Stage 1 began with the introduction of "precise teaching objectives" in 2009, then the idea of "precise teaching content" and "precise teaching design" brought this discipline into the second stage. The introduction of precision teaching was delivered by Chunfeng Hua in 2016 with the inspiration of precision marketing, this is viewed as stage 3. The last stage is deemed as Zhiting Zhu and Hongchao Peng imported the concept of "precision teaching" from abroad with an emphasize on adoption of Information Technology on teaching activities. The research hotspots of precision teaching in China are listed below. (Hu and Zhang, 2018, pp. 184–191)

- Precision teaching objective exploration
- Content creation for precision teaching and application research
- Precision teaching model research
- Estimation Tool development for precision teaching
- Data-based research on decision making for precision teaching

3 Community of Inquiry (CoI)

3.1 Where does CoI come from

The CoI (CoI) was developed on the base of practical inquiry which emphasize on the students' ability to conduct critical thinking. The formation of the CoI theory could be traced back to beginning of the 21st century in University of Alberta, proposed by researcher Randy Garrison together with his colleagues. The establishment of the CoI framework was done between 2000 and 2009 which was named the first stage of CoI, since 2010 onwards, instruments were introduced to the CoI framework together with criticism. Till now, the CoI framework remains one of the most popular and extensively used framework in learning and teaching discipline (Castellanos-Reyes, 2020, pp. 557–560).

3.2 What does inquiry mean

Inquiry refers to the process of an individual pursuit of understanding on a specific interested subject. CoI is somewhere that allows individual learners to generate their own understanding of the subject, and actively share their opinions on the subject to others, with the purpose of correcting their misunderstanding and challenging the thinking initiation, which is also addressed "thinking out of the box". In order to keep the system operating properly, a balance between the learners and teachers should be realized. Neither the teacher, nor the students should be the true centre of the system. And diversity of environment is always appreciated. In an ideal scenario of CoI, critical thinking and creation of useful meaning are more likely to be produced. (D. R. Garrison, Anderson and Archer, 1999, pp. 87–105)

3.3 How does CoI look on learning

3.3.1 From a collaborative constructivism point of view

Learning can be viewed from many perspectives such as collaborative constructivist point of view and the perspective of transaction. In the first context, researchers see learning as collaborative, it is a dynamic constructive process. The crash between personal-subjective and social-objective is crucial. Dewey and Childs (1933, pp. 77–103) also distinguished the difference between collaboration and cooperation in their proposal of transactional communication in education, that the information/ ideas generated by individuals will be diffused into the society to construct knowledge. Personal contribution is not required in cooperation but asked in collaboration. Share of self-reflection and discourse among peers are the 2 components of an inquiry-based learning community. Learning is never a thing that can be accomplished by individual, external impact from the environment is inevitable, input is the essential requirement for any kind of learning, and before the prevail of internet, the communication was one-way only which leaves the learning quality questionable? The sport team could be viewed as a good example, as individual player contributes to win the game not only bring honour to the team but satisfaction to each team members. Constructive discourse requires both competition and collaboration. (Garrison, 2016)

3.3.2 The transaction perspective

Viewing the learning activities from the perspective of transaction can draw similar conclusions. The initial step of learning is the consecution of self-understanding of the question, and then during the stage of exploration, concepts or ideas get better defined and recognized among peers through discussion. The next step could be getting confirmation of group recognition which could be considered as the end of transaction. Garrison (2016) argued that the TP in a CoI is not only contributions from the instructors, the role of teacher and student is interchangeable, as in some cases, teachers can learn something new from their students. The responsibilities of roles on the teaching sides are very often complex as they are involved in almost all stages including course design and teaching activities during the course and evaluation of students' performance and feedback to the learner after completing the course in form such as grading to the student. Students are supposed to generate self-understanding of the topic and have discourse to exchange their ideas among others for better understanding of the question inside the community. The students have limited influence to determine the final grading of their learning result. (Garrison, 2016)

Lipman (2003) argued that a CoI is established with pre-set purposes and directions. The progress inside the collaborative community is not lead by conversation, which is viewed as casual and no well-structured, but continuous dialogues with logical support aiming at exploiting answers to questions bidirectionally. The main responsibility of an instructor in the class is not to dominate but to coordinate the inquiries among students as Lipman believes that learning is the process of exchanging experience.

The dialogue initiated by inquiry moves progressively with a goal to settle down in the form of conclusions, but those type of settlements are not permanent as they will be further progressed in future dialogues just like conducting scientific research. The sessional end of dialogues does not strive for any certain answers but could be a clearer definition of the question and a better understanding of the context among the students. (Lipman, 2003)

The CoI theory welcomes friendships and cooperation among students which can guide further questioning to discover the truth.

3.3.3 Three factors that determine the quality of learning

The three main factors that affect the quality of students' learning are the method of evaluating learning outcome, the arrangement of curriculum, and TP. First factor the evaluation of learning can affect students learning patterns, for example, if the final grading is 70% on the exam, then student's motivation of lecture attending might be relatively low compared to courses appreciate in-class activities and weight the result of exam 40% for the grade. the second consideration curriculum arranged is directedly determined by course design. In many cases, when the curriculum tends to cover more which consequently leads to a high workload, students' performance might be poor due to the unbalanced workload needed to be done and actual time they can spend on learning. Hence, the more in curriculum causes the less in students' performance. The last factor TP can be understand as the same subject taught by different lectures might be significantly different due to the way of interaction or the languages teachers use, and students from different classes might have very different level of knowledge understanding. The TP has a lot to do with teachers' knowledge in pedagogy. (Garrison, 2016)

3.3.4 The importance of independent thinking

In the CoI theory, independence and interaction is not a contradiction. As Lipman (2003)mentioned that to successfully operate the CoI, all the participants need to be able to think for their own, which emphasize the importance of independent critical thinking and evaluation of obtained information without too much external impacts from peers or the instructors.

3.4 Common weakness in contemporary education

3.4.1 One-way of information flow

Garrison (2016)discussed in his book which was titled *E-learning in the 21st century* that the flow of information in traditional education context remains passively which means the large amount of useful information to construct knowledge is received by students with a one-way transmission from the instructor. The efficiency of learning is largely constrained by the motivation of constructing new knowledge from the information receivers' side and simultaneously the ability of diffusing information pedagogically from the information senders' side. If we view the learning scenario as a system, then we will notice that the feedbacks from the students' side for teachers to improve or adjust their teaching activities are defectively missing with only countable number of feedbacks from the students in the end of course collected mainly via survey. Students in this case are not heavily engaged in the learning community, and how to attract students' interests to increase cohesion among peers become an extremely important responsibility for those teachers who want their students to comprehensively master some useful tings instead of just pass the exam for course credit.

3.5 The resource we can use

3.5.1 The infrastructure in the new age: Internet

In the mid of 90s, e-learning appeared into the crowd's horizon together with internet. Educators and researchers believed that it could be an alternative to the traditional learning path with better course design and teaching direction. From the traditional education perspective, individual and the society cannot be separated, as they won't exist independently from each other. The community is constructed with individuals, while the individuals all have their own independence to some extent. This can be understood as people are fundamentally social animals, which means people need to interact with peers and interdependent on each other, and this might be traced back to the ancient time when the environment was not so friendly to human beings, as the loners are less like to survive, the behaviour of social interaction just rooted into our gene as generations reproduce. (Garrison, 2016)

3.5.2 Text-based communication

In a learning community, the communication among the participating peers can be verbal, gestural, and text based. The verbal and gestural communications are always synchronous, while the communication in the text-based form is more likely to be asynchronous, hence most discussion-based learning platform will choose text-based communication as the media for students to exchange their ideas and feelings. Some studies stated that discourses done in the in the form of text-based asynchronous communication tend to be more formal and systematic, some explanations might be that the ideas made by students were delivered to the platform after careful rumination, which is a benefit of having abundant of time to think before reply. Besides, the content of communication done in the form of text is more feasible, economical to be archived for future review. The fundamental characteristics of text-based communication is its asynchronous property which allows peers to communicate with a more flexible timing schedule. This trait of high tolerance of delay for response might potentially improve the quality of communication as learners do not have to rush an answer with a specific tight time constraint. (Garrison, 2016)

Moving the learning activities from traditional classroom-based environment to computer-aided virtual environment does not give significant difference regarding the learning outcome as it is just a different approach/ medium of knowledge transfer. The essence of education is never heavily depended on tools, but engagement of students to proceed individual critical thinking reflection and discourse among peers.

3.6 Three presences in CoI

3.6.1 Social Presence (SP)

Mehall (2021, pp. 1–11)indicated in his paper that social interaction with purpose is a fundamental part of online learning. Purposeful interpersonal interaction has positive impact on students' satisfaction and cognitive presence (CP); however, a saturation interval does exits, which means over high level of SP will not improve CP and study satisfaction when it reaches beyond a pint. And the range of student's satisfaction is much wider than the range of their study outcomes.

The SP here mainly refers to the active effort the individual takes to establish connections with others inside the community and using communication to project their personal characteristics. as Rogers et al. (2009) discussed that the SP is not easy to obtain in a text-based communication context. This might be the cause of the missing of emotion in text-based communication. People do find that sometimes face to face communication works way much better than sending text to each other or talking on the phone as communicator can get extra information from such as body language and emotion of the person that they are having conversation with, therefore they can gather a richer understanding of the interlocutor.

The activities created for text-based asynchronous learning environment essentially lack gestural cues during the transaction of communication, which could add difficulty for students to established connection like the way how they interact with peer in a physical classroom.

A good amount of SP could create a comfortable climate for open communication, and people usually only talk openly to someone they are familiar with instead of stranger met on the street. Emotional communication is affective. Open communication and feeling of belongness to the community are the prerequisite for create SP. And SP is also dynamic in the community with influence from CP and TP. For example, when the learning goes to the later stage, learners might need to have deep understanding of the subject which requires more critical and in-depth reflection and self-generated understanding, hence they have focus more on the cognitive perspective, instead of social interactions among peers. (Garrison, 2016)

Without enough SP, open communication is hard to occur, but when the SP goes beyond as wanted, it might also constrain the development of CP, and more interaction between peers become interactive, or the fear of breaking friendship will reduce disagreement and criticization in the discourse stage (D Randy Garrison, Anderson and Archer, 1999, pp. 87–105)

Unorganized interaction is not enough, we must move our understanding of the relationship between collaborative learning and teaching one step further (Garrison and Cleveland-Innes, 2005, pp. 133–148) Simple social interaction is very limited, the purpose must exist before starting the interaction, the concept of purposeful interaction could be used to refine the boundary of SP. (Garrison, 2021)

Borup, West and Graham(2012, pp. 195–203)stated that video comments from teachers can increase the SP of teaching role, which can help to establish the connection between the learners and teacher which gives students the secure feeling to asking for help. 12 out of 15 students agreed that video comments made them feel like talking with real person instead of cold machines. 5 students addressed that they were more likely to tell their interlocutor's personality in the video comments. Feedback in form of video is more useful than ones in plain text, also video comments from each other in the community are beneficial. However due to the natural limit of asynchronous form of communication, students will still worry that their video comments will not be watched by peers. Some students reported that video comments can to some extent reduce unnecessary misunderstanding in communication with help of body language and facial impression.

3.6.2 Cognitive Presence (CP)

The CP indicates to which extent can the learners reflect, and construct understanding of the subject and develop discourse in the community (Garrison, Anderson and Archer, 2001, pp. 7–

23) Reflection is done on the base of critical thinking, while discourse could only occur when the individual learners inside the community have confidence and trust on others and sharing the same committee of understanding the commonly interested subject through communication. Cognition is achieved via practical inquiry which could be break down into 4 stages: Initiation, Exploration, Integration, and Solution. (Garrison, 2016)

CP is firmly connected with critical thinking. The purpose of CP is to verify previously obtained knowledge and create new knowledge. Independent thinking is the prerequisite for critical thinking, while when you learn something, you will naturally get some input from the external environment, and very likely to use the input to correct your understanding about something based on previous received or empirical experience.

In the field of cognitive activities, peers mostly perform practical inquiry, conduct reflection of the issue pointed, and share their own understanding among the community to exchange their ideas for discourse. In the initiation stage, questions, or problems were introduced and individual's interest and attention were also gathered. In the exploration stage, peers do brain storming, using systematic approach such as TRIZ to generate new ideas, five-whys to better define the questions, and try to trace the root of problem. The integration stage is when peers try to modify and combine their understanding with an intention to formulate a final universally acceptable settlement for the question or problem. And finally, used the created knowledge to verify the validity of their understanding in the resolution stage.(Garrison, 2016)

Vaughan and Garrison (2005, pp. 1–12)mentioned in their research that in the traditional face to face learning environment, triggers of event are higher than the cases observed in the online environment. And exploration seems to be both the dominant stage in the context of CP, however integration of the ideas generated by individuals is higher in the online learning group compared to the face-to-face offline group. While the final stage of cognitive, the resolute stage, seems to be both low in both studied environments. Hence, it is meaningful to let the TP to guide the community to the resolution stage.

Study conducted by Traphagan et al. (2010, pp. 923–936) revealed that students in text-based learning environment tend to have higher level of CP than the students who were interacting in a

3D virtual learning environment. Significant higher value of CP was found in the discourse stage in text-based team, and the capability/stability of team gets improved after several round of group tasks. And one thing worth addressing is that each team has unique way of working, therefore the teachers shall customize study for different teams to maximize the utilization of learning resource and final learning outcome. It is suggested to give students more time in the beginning stage (Table 3) to get familiar with the tools for online learning and practice them with others. Give more practical guidance regarding how to collaborate with others and conduct discourse in the community. (Traphagan *et al.*, 2010, pp. 923–936)

Table 3:Stages of Practical Inquiry (Garrison, 2016, p. 66)

Phase	*Descriptor*	*Indicator*
Triggering event	Evocative (inductive)	Recognize problem Puzzlement
Exploration	Inquisitive (divergent)	Divergence Info exhange Suggestions Brainstorming Intuitive leaps
Integration	Tentative (convergent)	Convergence Synthesis Solutions
Resolution	Committed (deductive)	Apply Test Defend

3.6.3 Teaching Presence (TP)

TP could be seen in forms of course design, implementation and guidance during the study course. And the design of course and implementation of activities are important to sustainably develop SP and CP among peers inside the community.

In some articles, researchers involved a new category of presence which usually describe as "learners' presence". However as aforementioned, in a healthy CoI, it is not suggested to distinguish the role of students and teachers and they are supposed to work, generate and share their flection and understanding of a specific commonly interested subject, and discourse for a conclusion ultimately. This is also addressed as balance between the learning side and the teaching side. (Garrison, 2016)

The role of TP can sometimes determine the stability of inquiry community. The teaching role is responsible to put all necessary resource at the right place on the right time to the right people. And in many cases, the role of teaching needs to boots transaction inside the community to accomplished expected result. The teachers need to design and organize, coordinate the course. As learners have little ability to determine the content of learning, the teachers should be carefully present suitable learning contents, and encourage individuals to have independent thinking of the topic, generate self-reflection and pass on through the community.

Only with learning tools is never enough, because students are not likely taking good utilization of those tools by themselves (Kovanović *et al.*, 2018, pp. 44–58). Students who have strong willingness to master new knowledge and consider discourse as a vehicle to reach their destination are more likely to participate the learning activities actively.(Wise *et al.*, 2013, pp. 323–343) Hew, Cheung and Ng (2010, pp. 571–606)articulated that student who lack motivation to participate in community discussion might need more in-detail instruction to guide them enter the discourse with peers, while the usage of LMS does not truly tell students CP.

TP also have the function of redirecting students who are not on the right track back to the correct learning track. Here the most important point to claim is that each member of the inquiry community could act as the teaching role. The philosophy of CoI is that all members are learners and teachers at the same time.

Akyol, Vaughan and Garrison (2011, pp. 231–246)elaborated the influence of course length on the development of inquiry community. The group cohesion was found stronger among students who enrolled in a course with short duration. Mastered knowledge and satisfaction of course were also observed higher in the short duration course. Regarding CP, the short course had higher score of the exploration stage than the long duration one, however, the individually generated ideas were better integrated in the long duration course. Besides, the long duration course also gained higher score in the resolution stage, and better affective communication

Practical application of CoI

CoI can be used to optimize course design. In the early stage, SP is more likely to dominate the CoI, but as course moves on, students tend to get more knowledge rather than understanding of their buddies in the community, so the CP and TP will gradually take over in charge of the community. (Garrison, 2016)

4 Tools of selection

Social Network Analysis (SNA) was integrated into the LA framework with a purpose to study the interactions among the learning individuals in a community within clearly defined boundary. Luke (2015) described network science as a broad approach to study biological, physical, social, and informational systems with relational lens. Three primary purposed are: (1) visualization of the network, (2) description of the characteristics of the overall network, as well as individual nodes, (3) build mathematical and statistical models of network structure and dynamics. In this thesis research project, the open-source R programming language will be used for practical data processing.

Knoke and Yang (2008)discussed in their books which described social network analysis as a method to measure and represent structural relations between individual entity and the rest actors (the network), how did such phenomenon occur, and the consequences. Social network analysis lies on 3 underlying assumptions: First, in social network analysis, structural relations play a more important role than other attributes such as age, weight, religion. Second, structural relation will affect entity's perception, belief, and actions via structural mechanisms. Third, structural relations are dynamic, which means all the actions performed by entity will have impact on their future structural relations.

Actor, or entity in the network could be natural person, a company, a state, or even a country and so on. The network is constructed with actors and the relations between actors. The relation is a kind of connection, tie between a pair of actors, or a dyad. The relation is not an attribute to the entity, but a joint dyadic property. Relation could be realized in all kinds of form, like when children play basketball, their relations are game cantered, relation of companies could be flow of capital or raw material. A network is generally a set of entities, and those entities relate to each other by one of more type of relations. Study of social network analysis mostly focus on one mode network which means all the actor/ entity are similar in the network work, such as all the students in a school, or cars in the street. There is also two-mode network, also known as bipartite network which contains 2 different type of entities such as students, and events. (Knoke and Yang, 2008)

4.1 Two approaches to represent a network

To represent the network, we can use graph to visualize the structural relations of the nodes. As a saying goes, a picture tells more story than a thousand words. Another way to illustrate the network is to use matrix, the classical mathematical language, to reveal the network. When the network is represented in a matrix, it is more flexible to manipulate for further analysis.

4.1.1 Graph

In social network analysis, researchers use social network diagram to visualize the structure of network, as well as most important attributes of the network such as the interactions inside the network represented via lines with arrows in directed network, and lines without arrows in undirected network. Points and lines are the 2 fundamental elements to construct a network graph, In social network, point is also called node or vertex, while line is called edge or arc, in this thesis work, those terms aforementioned are interchangeable. In a social network, the simplest motif to study is dyad, which literally means a pair of nodes. The situation could be binary for unweighted undirected network indicated by the presence or absence of the link between 2 nodes. In a directed unweighted network, the attributes of dyad could have 4 cases, first case is that no line between 2 nodes, second as line exits with arrow to 1 node, third case is the opposite of the second case, the arrow points to the other node, the last case is line with 2 arrows points to both nodes.

Triad consists of 3 nodes. The possible arrangements of nodes in a triad are more complicated than the situation in a dyad. In undirected unweighted scenario, the arrangement of 3 nodes of a triad has 8 possibilities, when directionality matters, the arrangement count get doubled to 16.(Hoffman, no date)

when the directionality matters, we call the network graph digraph as short for directed graph. In digraph, the width of edges is always used to indicate the attributes of the relation between two nodes such as frequency, strength and intensity. When the two nodes both initiate interaction to the other node, the relation could be visualized as a single line with 2 arrowheads on both tails and 2 individual lines with single arrowhead pointing to the receiver of the action. (Knoke and Yang, 2008)

A subgraph is simply a subset of the entire network consists of finite number of vertexes and edges. A connected graph requires connections exist among all pairs of nodes, either directly connected or indirectly connected. If exists at least one node that is not connected to any other nodes in the graph, the graph is defined as disconnected, and the node with connection to other nodes is call an isolate. Connection in graph is categorized into 3 main types: I. Strongly signify that all nodes in the graph are directly connected to each other mutually; II. Unilaterally connected, only one direction of edge exists among every pair of nodes, if a sends to b exist, then exclusively b sends to a shall not exit. III. Weakly connected requires a path exists for every two nodes disregarding its directionality. In graph, if a vortex being removed will cause the graph split into 2 graphs, such a vertex is called a cutpoint, and very similarly if a line is performing as a binder of "2 graphs", the line or edge is called bridge. Technically, graphs with too much cutpoints and bridges are fragile or vulnerable as they are more likely to cause break up under external interruption. In real world, cutpoints and bridges are channels that enable communication between different communities. (Knoke and Yang, 2008)

4.1.2 Matrix

The matrix used to describe network is called adjacency matrix, also known as *sociomatrix.* All the nodes are listed both in the column on the left and row on the top with the same order as demonstrated in Table 5. Therefore, the adjacency matrix is always a $n \times n$ square matrix. The intersecting cell of the row and column is filled with the relation value of the 2 interacting nodes.

In an undirected unweighted network, the value in the cell is binary, as 0 represents no relation, 1 tells relation exits. While in an undirected weighted network, the value is a scalar, normally the stronger connection between 2 nodes, the bigger scalar will be put in the intersecting cell. The cell along the diagonal has null value as it theoretically means the node's connection with itself, therefore in undirected network, there is $\frac{n^2-n}{2}$ useful entities in the upper triangle or lower triangle to describe the network, n^2-2 objects in directed network. The matrix notation is also applied to identify the value cell, for instance x_{ij} denotes the value of edge initialized by node *i* received by node j. (Knoke and Yang, 2008)

In directed networks, the values are also determined by the connection between nodes, however one big difference is the matrix is not symmetric along the diagonal, as the relation of Node *i* send something to Node j is different from Node j sends something to Node i. While in an undirected network, the roles of "sender" and "receiver" are indistinguishable. The Table 4 below represents how network types affect values in the cell, *C* is a scalar.

Table 4：Matrix cell values in different network

Network type	Nodes	Values in the cell
Undirected and unweighted	From A to B = From B to A	0 or 1
Undirected but weighted	From A to B = From B to A	0 or a scalar
Directed but unweighted	From A to B without From B to A	$X_{ab}=1$, $X_{ba}=0$
	From A to B with From B to A	$X_{ab}=1$, $X_{ba}=1$
Directed and weighted	From A to B without From B to A	$X_{ab}=C$, $X_{ba}=0$
	From A to B with From B to A	$X_{ab}=C$, $X_{ba}=C$

Table 5:A simple undirected and unweighted matrix with 4 nodes

	A	B	C	D
A		1	1	0
B	1		1	1
C	1	1		1
D	0	1	1	

4.2 Attributes to measure

measurement of social network could be realized from 2 perspectives, one perspective is to study the properties of the entire network such as size, density could generate a rough view about the social network entity we are studying; another approach is to analyse the network in micro level, we can also all its actor level, mainly based on centrality and prestige analysing tools.

4.2.1 On the network level

Some fundamental terms to describe a social network including size, density, diameter, clustering coefficient, and centrality.

Size is the simplest descriptive information a network can give. It tells the number of nodes or actors within the entire network. Size is also the number of row or column in the adjacency matrix expression of the network (in the case of one-mode network where all nodes represent a type of actor)

Density is another characteristic of network which could be understood as the proportion of observed actual edges in the network to the possible maximum ties that could occur theoretically. Therefore, the value of a network density could be range from 0 to 1. When the density approaches 1, it means the network is more ideally interconnected. The way of calculating network density is very straightforward, reported number of ties divided by the theoretical maximum number of possible ties.

The formulars to calculate network density in different scenarios in Table 6 were proposed by (Knoke and Yang, 2008)classified the density measurements into 4 cases. The base unit of the denominator is the maximum possible dyadic ties in an undirected network. *N* stands for the number of alters which equals the number of actors *n -1*, which means in a 5-actor small network, the number N will be 4. L is the reported number of ties in the network. The denominator needs to be multiplied by 2 and directionality will double the number of maximum possible ties when edge from A to B is not equivalent to edge from B to A. In valued undirected and directed network, the numerator is replaced with the summation of all reported ties with attached values. In unweighted network, the value of network density always falls between 0 and 1 but never exceed 1 as the number of reported ties and theoretical maximum number of possible ties will obey $L \leq C_N^2$ for all undirected network and $L \leq 2 \times C_N^2$ for all directed networks. In valued network, the density of network might exceed 1 as the numerator can be larger than the denominator with attached values. Ideally, we wish the network density value as big as possible, because it indicates the actors in the network are better interconnected with each other compared to networks with low density values.

$$C_N^2 = \frac{N!}{2! \times (N-2)!}$$

Table 6:Calculation of network density

Attribute	Formula for calculation
Binary undirected network density	$D = \frac{L}{C_N^2}$
Binary directed network density	$D = \frac{L}{2C_N^2}$
Valued undirected network density	$D = \frac{\Sigma L_w}{C_N^2}$
Valued directed network density	$D = \frac{\Sigma L_w}{2C_N^2}$

Components

In a big network, nodes are impossible to connect to each other, however there could be several sub-groups where the inside actors are all connected to each other directedly or indirectly. Those sub-groups could be viewed as components within the network. Component is a very simplified concept to study structure of a network, from the system engineering points of view, components are subsystems of the big entire system, researchers can put flexible restriction to control the boundary of components by adjusting criterion parameter, some strict defined components in network are defined as cliques and k-cores. in Programming language R, by default strongly connected components are returned as user perform the components () function. (Luke, 2015)

Diameter

Besides network size, diameter is another interesting characteristic to describe the network compactness. it measures the shortest path (Table 7) which is required to send information from node A to node B. For the entire network, diameter indicates the worst situation, in another word the longest of the shortest paths across all pairs of nodes in the network. This could be used to indicate the network efficiency. For example, a network with big size but small diameters tell that the information could be efficiently spreading at a relatively short period of time. While a network with small number of actors which creates a relatively large diameter is always inefficient for information diffusion, as it takes more time to deliver the information to other node compared to the same size networks with smaller diameters. (Knoke and Yang, 2008; Luke, 2015)

Table 7:Walk, Path and Geodesic in network analysis (Knoke and Yang, 2008)

Term	Definition
Walk	An alternating sequence of incident nodes and lines
Path	A walk with entirely distinct nodes and lines
Geodesic	The shortest path between two nodes

Clustering coefficient/transitivity

Clustering coefficient in network analysis could be examined by transitivity. Clustering refers to the tendency of closed triangle formation. For instance, when 2 actors share one common friend, the two actors also become friends to each other like density of the network, the clustering coefficient value is also between 0-1 (Luke, 2015)

4.2.2 On the individual level

Nodal degree

Nodal degree refers to the number of edges a node has, in directed graph, nodal degree has 2 types, nodal in-degree refers to number of edges with arrowheads points to the node, nodal outdegree refers to the number of ties sending out from the node. on the entire network level, we calculate the average value of nodal degree the degree must analyse to which extent are the actors in the network having interactions.

Calculation of nodal degree could also be done in 2 ways, summation or average. In weighted graph, this is easier to understand. Imagine the scenario as fellow, node i has 1 edge with weighted value3 and 7 edge of value 1, while mode *j* has 2 edges with value 5. On the point view of average, node *i* and *j* has same nodal degree summation 10, but node *j* has an average nodal degree of 5 which is much bigger than node *i* 1.25.

Geodesic distance

calculation of the geodesic distance in binary undirected graph is simply choose the shortest path, the number of edges along the path is the geodesic distance, in directed weighted graph, there are 4 steps to do it. Step 1 find all paths connect he dyad; step 2 finds lowest value of edge in each path; step 3 using the lowest edge value to be divided by number of edges along the path; step 4 choosing the path which returned the largest value in step 3 as the optimal one.

Centrality and prestige

Study of centrality in social network is used for eliciting important actors in the network. Nodes with high centrality is those possess more relation with other nodes either sending or receiving ties. The main study interest of network centrality could be roughly categorized into 2 level, node centrality, and group centrality. The node centrality is examined to understand performance of individual actors in the network, while the group centrality focusses on understanding of the entire network property. Knoke and Yang (2008) proposed the calculation of the degree centrality of individual actor as shown in Table 8, *g* indicate the number of actors in the network, while x_{ij} stands for the degree between node *i* and node *j,* the summation of all pairs of degree which node *i* involved will be the total degree centrality of node *i*. The degree centrality value could be further normalized by dividing the unnormalized degree centrality with *g-1* which stands for all other nodes in the network. A high out-degree centrality is an indicator of active participation in VLEs.

Table 8: Calculation of degree centrality

Attribute	Formula for calculation
Actor degree centrality	$C_D(N_i) = \sum_{j=1}^{g} x_{ij}\ (i \neq j)$
Normalized actor degree centrality	$C'_D(N_i) = \frac{C_D(N_i)}{g-1}$

The simplest measure of degree centrality is computing the direct ties a node has, in a simple context, a node with more direct connections has a stronger centrality over others with fewer direct edges, a more formal way of calculating a normalized degree centrality is listed in Table 8

Closeness centrality

Closedness centrality answers how close each node in the network is approachable to other nodes. In this scenario. Node that close to all other nodes are considered as situated in the prominent position inside the network. Table 9 demonstrated the method of calculating the betweenness centrality, the closeness centrality of a node *i* is the reciprocal of the summation of distance between node *i* and all other nodes in the network (Knoke and Yang, 2008)

Table 9: Calculation of closeness centrality

Index of actor closeness centrality	$C_c(N_i) = \frac{1}{\left[\sum_{j=1}^{g} d(N_i, N_j)\right]} (i \neq j)$
Normalized index of actor closeness centrality	$C_C'(N_i) = (g-1) \times C_C(N_i)$

Betweenness centrality

The nodes sit between paths of other indirectly connected nodes usually control the information flow in the network. Betweenness centrality measure to which extent a node is placed between paths of other pairs of nodes in this set-up, nodes with high betweenness centrality are prominent. Calculation of betweenness and its normalization is shown in Table 10. (Knoke and Yang, 2008) The numerator $g_{jk}(N_i)$ in Table 10 stands for number of geodesic paths between node *j* and node *k* that contains node *i*, while the denominator remains the total number of geodesic paths between node *j* and node *k*.

Table 10:Calculation of betweenness

Attribute	Formula for calculation
Actor betweenness centrality	$C_B(N_i) = \sum_{j<k} \frac{g_{jk}(N_i)}{g_{jk}}$
Normalized actor betweenness centrality	$C_B'(N_i) = \frac{C_B(N_i) \times 2}{(g-1) \times (g-2)}$

Prestige and power

In a directed network, node with more out-degree is viewed as actor with strong power, a node with more in-degree is considered as prestige. The more out-degree a node has, the more influence a node can achieve in a network, the more in-degree ties represent respect and fellowships from other nodes in the network. Table 11 demonstrated how the prestige value and its standardization could be done mathematically. (Knoke and Yang, 2008) As you can find in Table 11, the calculation of prestige is almost same as the calculation for degree centrality, but the directionality is changed, the prestige centrality measure edge from other nodes that pointing to the node which is measured for the prestige.

Table 11:Calculation of prestige

Attribute	Formula for calculation
Actor degree prestige	$P_D(N_i) = \sum_{j='1}^{g} x_{ji} \ (j \neq i)$
Standardized actor degree prestige	$P_D'(N_i) = \frac{\sum_{j=1}^{g} x_{ji}}{g-1} (j \neq i)$

Except above mentioned attributes on the nodal level, there are some other attributes such as eigenvector centrality which take a node's eigenvector centrality as a function of its neighbours' eigenvector centrality, simply means a node will be more important if its neighbours are important.

4.3 Detection of sub-groups

4.3.1 Cliques

Clique is one of the easiest cohesive sub-groups for its straightforward definition. A complete sub-graph of the network where all nodes in the subset have all possible ties among them. However, two disadvantages are inherently rooted from the definition of clique. All possible ties need to occur among the subset nodes, due to its conservative attribute, it is not easy to form.

The second shortcoming due to the fragility of clique definition, it is not common to increase the number of cliques as the size of the entire network grows. Even in a large network, the number and size of cliques remain limitedly small. (Knoke and Yang, 2008)

4.3.2 K-cores

To overcome the disadvantages of cliques caused by rarity, K-core is introduced as a variation of clique idea. A K-core is defined as a subset of the graph in which each vertex as at least connections to k other vertexes. (Luke, 2015)

4.3.3 n-cliques

n-clique is another way to overcome the difficulty of community detection caused by the strict boundary of clique. The nodes inside a n-clique must be achievable within n links by any other nodes. (Knoke and Yang, 2008)

4.4 Positional Analysis

Another important study discipline under social network analysis is to elicit positional information of the nodes. With positional analysis, we can find so called “competitors” of a particular node in the network. For online discussion forum in the learning scenario. It can also be used to viewed as a useful tool to allow people find “another me” for the learners. People are social animals, which inherently makes people want to interact with other, as discussed in the CoI chapter, the positional analysis in SNA can help peoples to find similar entities, which might improve learners’ engagement about the course and encourage transmission of opinions among the networks.

Structural equivalence has the strictest constrains to classify nodes into the same class, followed by automorphic or isomorphic equivalence, and regular equivalence.

4.4.1 Euclidean distance

Euclidean distance is used to quantify how far 2 nodes lie from each other. When the Euclidean distance of 2 nodes is 0, then the 2 nodes are perfectly structural equivalence. We can in general use Euclidean distance to describe the extent to which 2 nodes are similar to each other. (Knoke and Yang, 2008)

4.4.2 Structural equivalence

Structural equivalence has the strictest confinement, it says only when 2 nodes has exactly same relations with same nodes, they are structural equivalent. For example, the math teacher and history teacher of the same Class A in middle school could be viewed as structurally equivalent disregarding the subjects they teach. Here one limit is that the students have to be exactly from the same class and exactly the same number. In a scenario when the teaching attributes are not important, the relation between the math teacher and history teacher is substitutive, either of them can become the guardian of the students in the class. (Knoke and Yang, 2008)

4.4.3 automorphic/ isomorphic equivalence

Automorphic or isomorphic equivalence was proposed to make the boundary of similar class less restrictive. If 2 nodes have same patterns of ties with different sets of nodes that play the same roles in relation to that position. For example, a biology teacher who teach biology in another Class B which has the same number of students as Class A, we can say that the biology teacher is automorphically equivalent to the math and history teacher who serve for Class A (Knoke and Yang, 2008)

4.4.4 Regular equivalence

The definition of regular equivalence is somewhat mouthful, however the main characteristics could be viewed as

'Actors are regularly equivalent if they have the same kind of relations with actors that are also regularly equivalent.' (Knoke and Yang, 2008)

"Two nodes are said to be regularly equivalent if they have the same profile of ties with members of other sets of actors that are also regularly equivalent." (Robert A and Mark, no date)

More intuitively explain, 2 mothers who have different husband and different children are regularly equivalent as they both have husband and children. 2 managers that leads different size of teams are viewed as regularly equivalent as they both have subordinates even the quantity of subordinates is different,

4.5 Methods of reporting analysis

When the network is relatively small with limited number of nodes and vertexes, the visualization is generally not difficult to achieve. However, when the number of nodes goes beyond thousand, the number of edges in the network also arise respectively. I many cases, the visualization of the big complex network will provide some hairball like image, which pragmatically gives the graph readers very inadequate information about the structure of network and nodes relations with each other.

4.5.1 Hierarchical clustering

clustering gathers nodes who are structurally equivalent or approximately equivalent into subgroups. The clustering begins with setting every single node as a cluster by default, and progressively combine nodes that are accepted as similar to each other, and then clusters will be combined with each other based on distance into a parent class level. The final graph will be in a tree structure, such a diagram is called tree diagram or dendrogram. Researchers may choose the Euclidian distance of 2 nodes d_{ij} or the correlation coefficient of the 2 vertexes r_{ij} as the criteria for clustering, and studies had proven that choosing either criterion produce very similar outcomes. The threshold of group nodes $d_{ij} < a$. Three clustering criteria could be applied to further combine clusters; I. complete link require every pair of nodes from 2 clusters has Euclidean distance smaller than the threshold a. II. the average link requires the average Euclidian distance from 2 clusters less than threshold a. III. single link requires that shortest Euclidian distance exist between 2 clusters below the threshold. (Knoke and Yang, 2008)

4.5.2 Multidimensional scaling

multidimensional scaling using pair wised Euclidian distance to group nodes in the network, like hierarchical clustering, it uses MSD programs to detect similarity or dissimilarity among nodes. Reducing the dimension of network structure for visualization. The output of MSD is a spatial map where nodes with smaller distance are plotted closer than nodes with larger distance.

$$Stress = \sqrt{\frac{\sum\sum\left(f\left(x_{ij}\right) - d_{ij}\right)^2}{Scale}}$$

The Stress indicate the discrepancy across all pairs between the observed matrix and the computed matrix. The $f\left(x_{ij}\right)$ is a nonmetric, monotonic function, and Scale is a scaling factor to keep stress between 0 and 1. Ideally fitting requires Stress value stays below 0.1 and adequate fitting between 0.1 and 0.2, when the value of stress goes beyond 2, it is considered as deprived. (Knoke and Yang, 2008)

4.5.3 Blockmodeling

Blockmodeling uses matrix algebraic operations to perform matrix partitions, which literally means divide the single matrix into multiple subgroups based on pre-set criterion. The term block here refers to square submatrix where the nodes inside are structurally equivalent or similar that having similar relations with other nodes located in other blocks. 2 nodes will be compared by using the CONCOR program, if they have same ties with other nodes in the matrix, they are structurally equivalent, and their correlation coefficient is 1.0 while if 2 nodes have opposite correlation, their correlation coefficient will be -1.0, while most cell values will lie between -1 and 1 when it comes down to it. Therefore, the program will repeat the process above until all correlation coefficients in cells become either 1 or -1. Then the sociomatrix is separate into 2 homogenous blocks where all intra correlation coefficient is 1 and inter correlation coefficient is -1. And then program will repeat the process happen above to further separate the matrices according to researchers' needs. Hence it could be seen as applying continuous dichotomy to arrange similar actors into the same blocks to provide aggregated level of information. The output of blockmodeling is resulted in form of a density matrix and an image matrix. The word density here applies the same philosophy as in network density, it is the percentage of observed ties over maximum possible ties among the nodes in each block. the image block is simply to recode the density matrix value to 0 or 1 with an alpha density cut-off which is commonly the overall network density value. (Knoke and Yang, 2008)

4.5.4 Topic modelling

In addition to analyse of students' online behaviour quantitatively on the base of interaction log, it is also important to understand what the students are discussing with each other, we need to understand the theme of the posted content. The LDA technique is a useful tool that can help researchers to elicit the discourse topic from the forum data by text mining procedure approach. LDA sees a document as a collection of words which might contains multiple topics, the order of words in the text is not important, as the topics generated are based on probability theory, therefore based on the number of topics assigned, the algorithms will automatically put words that relevant to each other to construct a topic based on statistical models. (Blei, Ng and Jordan, 2003)

5 Empirical test of selected tools

The first dataset is a l log of students' online activity from a TRZI course stored on the Disqus platform. The raw data stored in the xml format was further wangling into a tabular format with R programming language and its IDE RStudio.

An edge list was created for social network analysis, simply it is a tabular data with 2 columns respectively represent the actor who received the comment, and the actor who posted the comment. The directionally is defined as starting from the comment poster and ended to the comment receiver.

Another column of values coded as "message" was utilized for the LDA topic modelling test, all the rows of "message" values were combined to a corpus for additional processing.

5.1 The big five insights of the TRIZ class network

As summarized in Table 12 the TRIZ class network generated by the sample data contains 63 nodes, and in total 356 weighted edges. The edge density of the network is around 0.09 which implies that the probability of 2 nodes inside the network to construct an edge is less than 10%.

The network has a diameter of 8 which means there is a pair of nodes inside the network that information flow from one node to another must go through 8 edges, more clearly explained as there are 6 other nodes placed on the path between these two nodes. A randomly generated network was created with same amount of node and edge as shown in Table 13 to offer a relative comparable reference, in that randomly generated network, the diameter is 5. In this case, as discussed in the last chapter, we shall state than the efficiency of information spread inside the network is low as the size of network is relatively small.

Table 12:Big Five insights of the TRIZ course network

Big Five insights of TRIZ course Network	
Size	63 vertexes, 356 edges
Density	0.09114183
Diameter	8
Components (weak)	1 weak component with 63 vertexes
Components (strong	11 strong components, 53 vertexes for one strong component, and 1 each for other 10 components
Clustering coefficient/ transitivity	0.3939005

1 component was detected with igraph package's default component algorithm while the mode was set as "weak", and 11strong components were discovered while mode was altered to "strong". Among the 11 strong components, 1 has 53 nodes, while other 10 components have 1 node each. While in the randomly generated network, only 1 strong component was sensed.

The transitivity value of our TRIZ course network is 0.39, which is almost 2 times bigger than the transitivity value reached by the randomly generated network. That means the nodes in the TRIZ course have 2 times the tendency to from sub-community compared to a random network with same number of vertexes and edges.

Table 13:Big five insights of a randomly generated network

Randomly generated network with same number of nodes and edge	
Size	63 vertexes, 356 edges
Density	0.09114183
Diameter	5
Components (weak)	1 weak component with 63 vertexes
Components (strong	1 strong component with 63 vertexes
Clustering coefficient/ transitivity	0.2006288

5.2 Centrality distribution

The degree centrality was measure for the TRIZ course network, the distribution of nodal degree property was visualized with scatterplot where the x axis indicates the index of node, y axis stands for the unnormalized degree of each node. Three histograms were also generated to cast the frequency of each level of degree that had be achieved.

As revealed in Figure 14, the degree centrality of the network resulted in 3 right skewed histograms. Most of the vertexes inside the network have total nodal degree below 20, with a few reached over 20. One node had the largest nodal degree value at 80.

The distribution patterns of in-degree and out-degree are positively correlated. There are more nodes whose out-degree lie below 5 compares to the histogram of in-degree diagram. Like the comparison did for big five insights in the last section, a scatter plot and histogram were also created for the randomly generated network. The degree in the random network is normally distributed as the outline of the frequency of different degree level is in a bell shape.

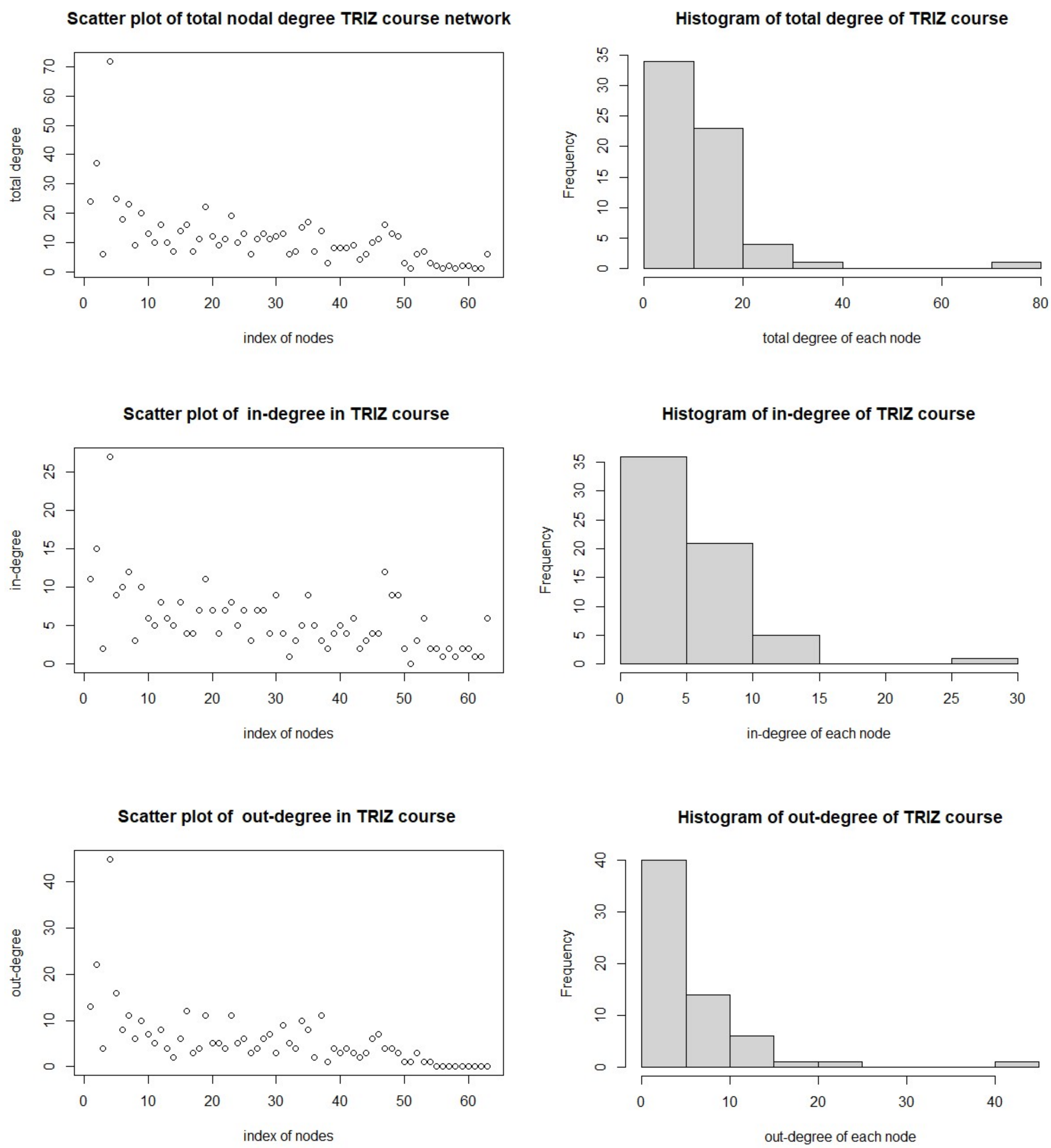

Figure 14:Centrality measure of TRIZ class network

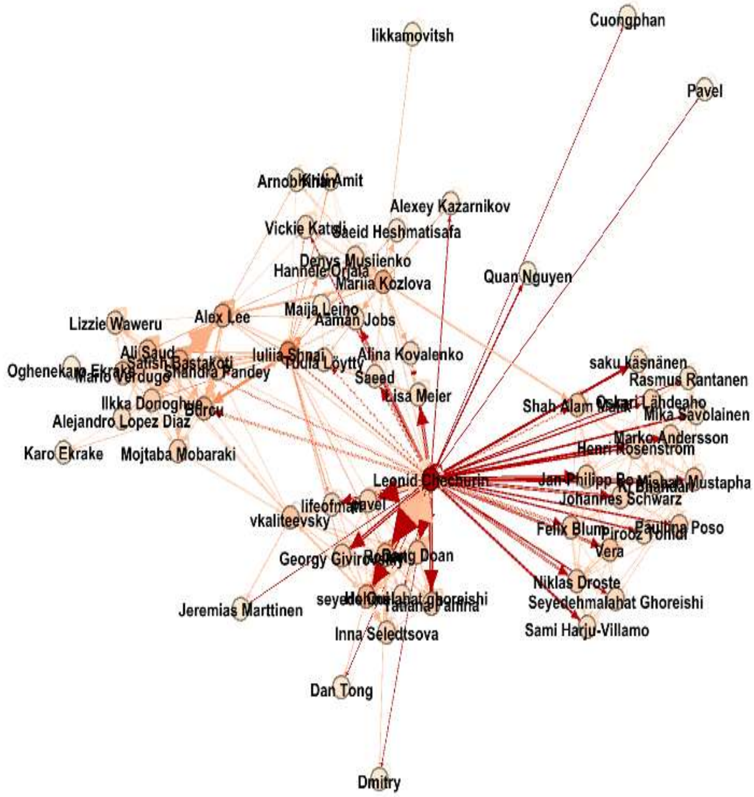

Figure 15:Ranking based on nodal degree

While ploting the network into a graph, the degree centrality can be used as a basis for ranking, the alpha value could be assigned based on the nodeal degree attributes, therefore, the researchers can swiftly identify the actors that are actively shareing their thoughts or those nodes posess prestige by receiving more comments from other actors in the network(Figure 16).

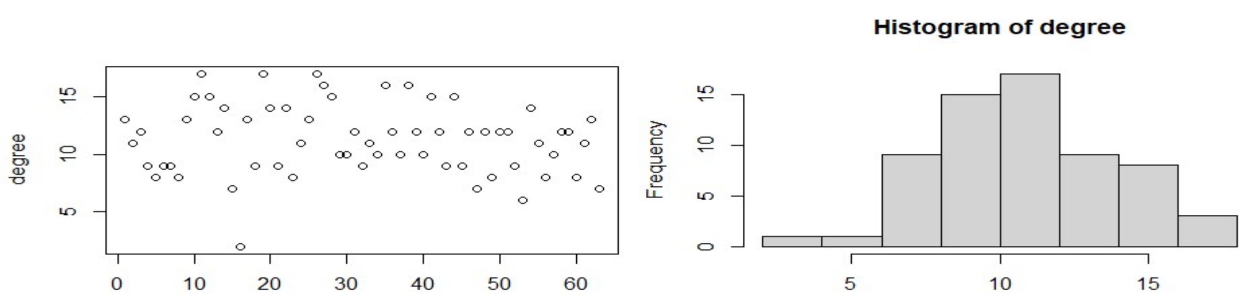

Figure 16:Degree distribution of randomly generated network

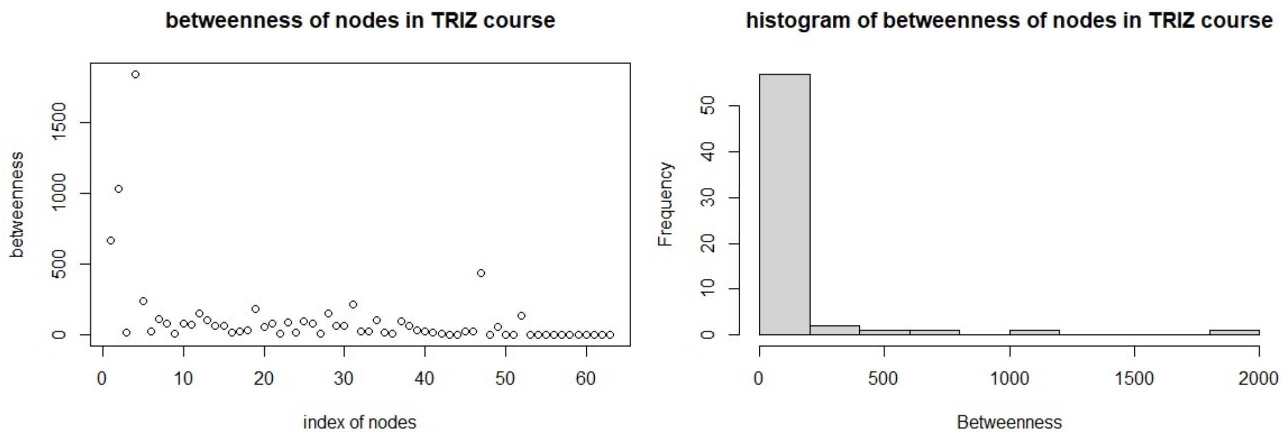

Figure 17:Distribution of betweenness centrality in the TRIZ course network

Betweenness

Like the scattering of degree centrality above, most nodes in the TRIZ course have relative low score of betweenness with a few exceptions. Which means most of the nodes in the network do not have control of the information flow inside the network.

Closeness

As discussed in the last chapter, the closeness centrality measure to which extent a node is approachable by other nodes, hence a node will have a high score of value of closeness if it is easily reachable by other nodes. The distribution of closeness centrality is comparatively closer to the normal distribution as indicated by its nearly bell shape histogram.

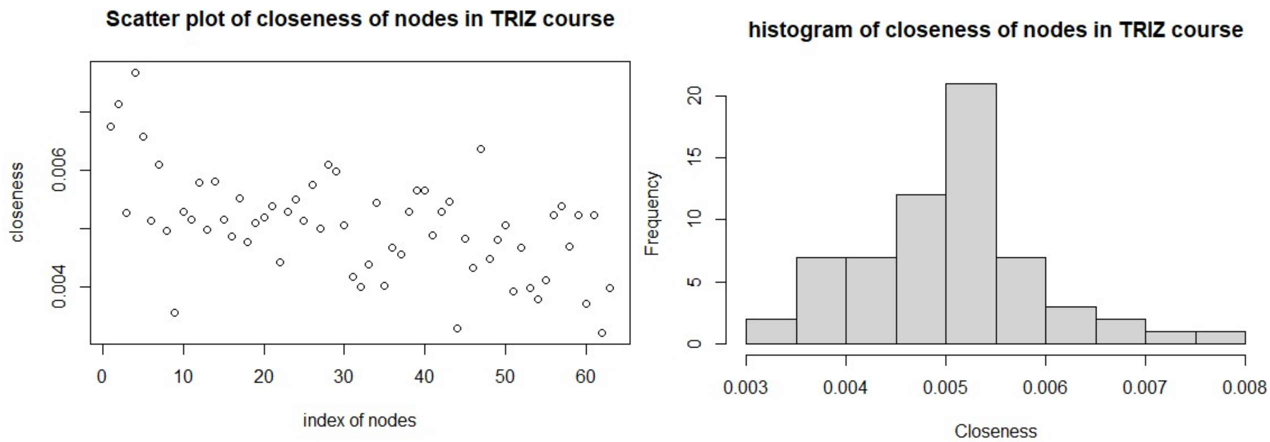

Figure 18:Closeness centrality of TRIZ course network

Eigenvector centrality

Eigenvector centrality is used to gauge the nodal prestige inside a network, as shown in Figure 19, most of nodes have low score of eigenvector centrality, with 6 nodes possess score over 0.2.

A randomly generated network with 3000 nodes and probability of 10% for edge formation between a random pair of nodes. The eigenvector of 3000 nodes network is also in a normal distribution (Figure 20).

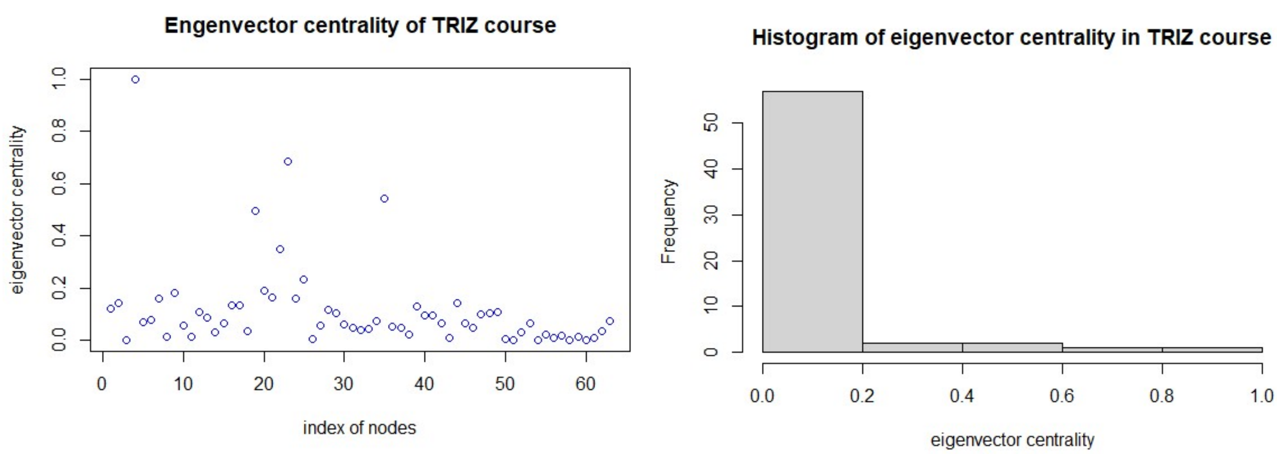

Figure 19:Eigenvector centrality of the TRIZ course network

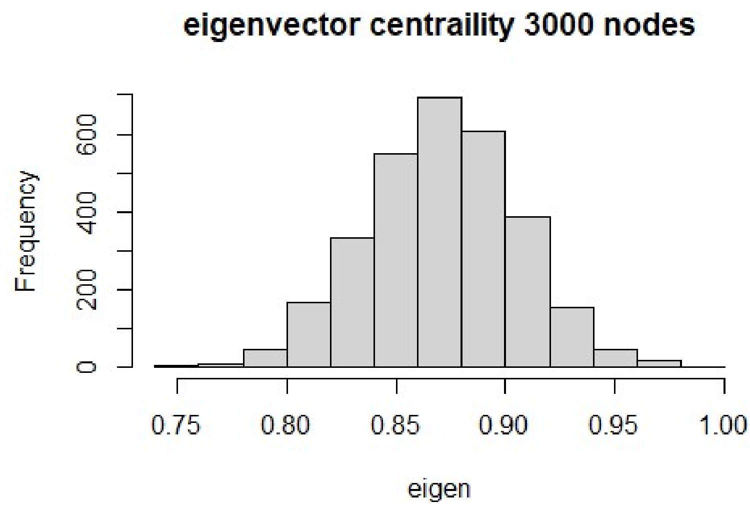

Figure 20:Closeness distribution of random network with 3000 nodes

5.3 Attributes correlation

As many attributes on the nodal level were measured, it is worthwhile to further study the correlation among those attributes. Heat map of the correlation of those attributes could be visually reveal whether an attribute has correlation with another attribute or not. As illustrated in Figure 21, the attributes are place on both the horizontal and vertical axis followed by the identical order.

Since nodal degree is consist of both in-degree and out degree. Hence total degree is highly correlated to in-degree and out degree， followed by betweenness centrality and page rank.

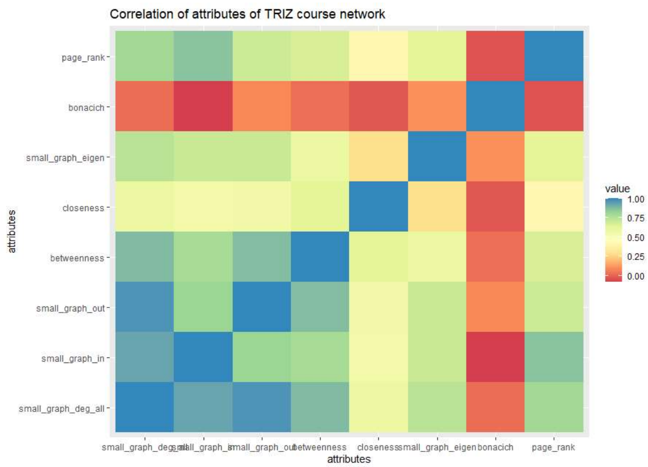

Figure 21:Correlation of nodal level attributes

Betweenness centrality has a stronger positive correlation with the nodal out-degree than in-degree. Closeness centrality has similar value of positive correlation with other attributes except for Bonacich. The eigenvector centrality shares similar correlation scores with degree attributes and the page rank, Bonacich seems not have solid positive correlation with other attributes. The page rank has a higher positive correlation with the nodal in-degree centrality compares to the out-degree centrality.

5.4 Subgroup detection

5.4.1 Cliques

211 cliques were found in the network with a minimal requirement of 5 nodes. The largest 3 cliques consist of 7 nodes as shown in table below.

It is noticed that the cliques are often overlapping with each other, as we can see in the Table 14 below, many actors have memberships of different cliques at the same time.

Table 14:Three biggest cliques

[[1]]
+ 7/63 vertices, named, from 62d9b9b:
[1] Leonid Chechurin Felix Blum Vera
[4] Henri Rosenström Kj Bhandari Johannes Schwarz
[7] Pauliina Poso

```
[[2]]
+ 7/63 vertices, named, from 62d9b9b:
[1] Leonid Chechurin Felix Blum       Vera
[4] Henri Rosenström Jan-Philipp Ro   Johannes Schwarz
[7] Pauliina Poso
```

```
[[3]]
+ 7/63 vertices, named, from 62d9b9b:
[1] Leonid Chechurin         Dang Doan
[3] Ho Qui                   Robert
[5] Georgy Givirovskiy       seyedehmalahat ghoreishi
[7] Tatiana Panina
```

5.4.2 K-cores of the network

K-core refers to a subset of a graph which requires the member of the k-core subset has at least k connections with other vertexes. Therefore, nodes that belongs to a k-core with relatively higher value k_i are inherently also member of other k-cores that holds a lower value k_j. Which could be further clearly explained that a node belongs to k-core with requirement of 9 connections is inherently members of k-core subsets where the requirements are 8 or less connections with other vertexes. The K-cores membership of several actors in the TRIZ class is displayed as Figure 22.

Alex Lee	Iuliia Shnai
10	10
Kiriti Amit	Leonid Chechurin
6	10
Mariia Kozlova	Ali Saud
7	10
Satish Bastakoti	Alejandro Lopez Diaz
10	8
Burcu	Ilkka Donoghue
10	10

Figure 22 K-cores membership in the network

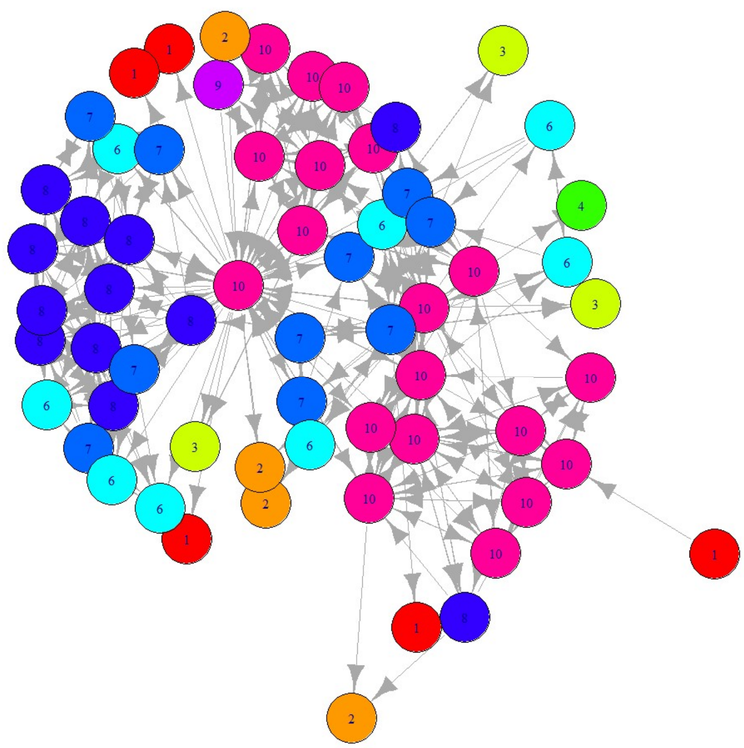

Figure 23:K-core structure of TRIZ course network

To understand how sociable a student is, we can use the k-core methods to uncover how many other community members they are interacting with. We can cluster the nodes based on k-core algorithm to visualize the entire network. In Figure 23 above, nodes were grouped based on the highest k-core subset they belong to and marked with different vertex colours for differentiation.

As we can see from the Figure 23 above, most of the actors in the network have at least 8 or even 10 connections with other vertexes. five of the nodes that only contains 1 connection with others and 4 nodes have 2 connections to different vertexes. Together with other 4 nodes who has 3 or 34 connections with other nodes, we might mark these 13 students inside the TRIZ course in a relatively “dangerous” situation, as the lack of communication might discourage them for idea generation and course discourse, which might affect their academic performance.

5.4.3 K-cliques

Table 15:an example of 3-clique subset of TRIZ course

```
3-cliques`[[4]]
 [1] "Alex Lee"          "Iuliia Shnai"
 [3] "Kiriti Amit"       "Leonid Chechurin"
 [5] "Mariia Kozlova"    "Shandra Pandey"
 [7] "Tuula Löytty"      "Mario Verdugo"
 [9] "Lisa Meier"        "Ho Qui"
[11] "Georgy Givirovskiy" "Robert"
[13] "Tatiana Panina"    "Arnob Khan"
[15] "lifeofmatt"        "Aaman Jobs"
[17] "Denys Musiienko"
```

Another approach to group students is using the k-clique algorithm, as the name says, all the member inside the k-clique subset shall be reachable less than 3 edges. Table 15 lists one of a 3-clique subset where the path is less than 3 edges for any pair of nodes that belong to this 3-clique subset.

5.5 Positional Analysis

Due to the strict confine of structural equivalence, the TRIZ course network does not have any nodes that possesses the same position and have identical interactions with the same set of other vertexes. By allowing approximation, similar nodes are marked with the same vertex colour for network visualization as presented in Figure 24.

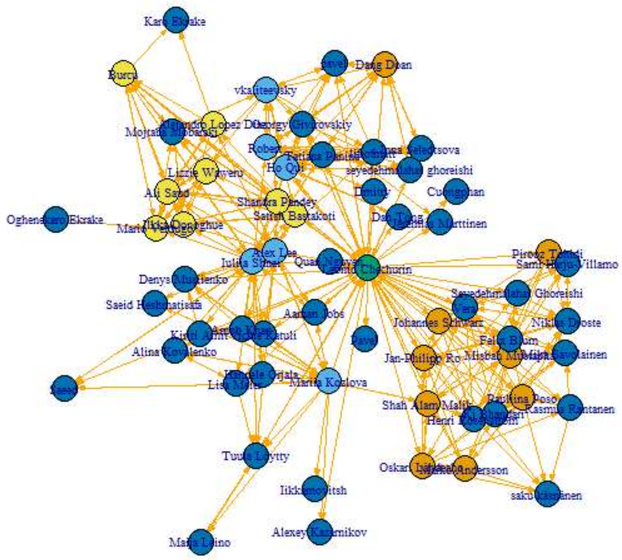

Figure 24:structural equivalence with approximation

Isomorphic equivalence

Figure 25 demonstrated how nodes inside the TRIZ course network are clustered with Louvain method, as the confinements is further reduced, we can find more nodes are marked as green and light blue.

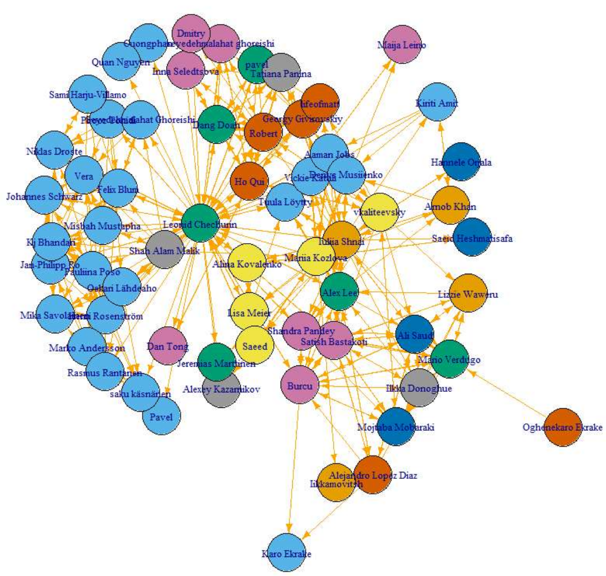

Figure 25: grouping based on isomorphic equivalence of nodes

Regular equivalence

The least strict similarity measure regular equivalence can put more nodes to one cluster as it only requires the nodes to have similar pattern of interactions with othes nodes (Figure 26).

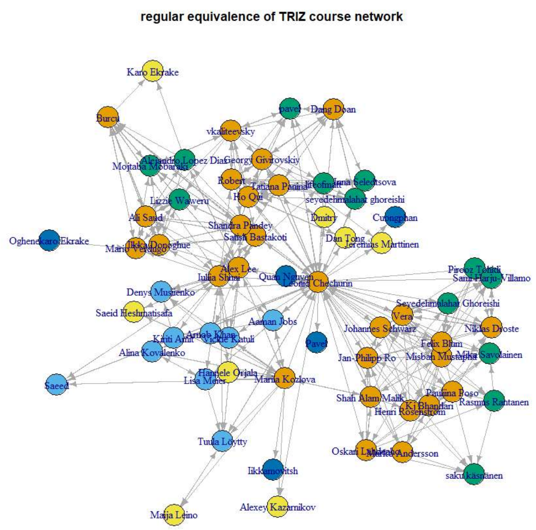

Figure 26:Regular equivalence in the TRIZ course network

5.6 Blockmodeling

Instead of representing the network in a network graph, the small network of the TRIZ course can be displayed in a matrix view. Figure 27 demonstrate how the nodes in the network can be allocated to different blocks. The easiest way to understand a block model matrix is to analyse the blocks sited on the diagonal line. Based on similarity, nodes inside the same block on the diagonal line is much similar to each other than nodes from outside the block. Depends on researchers' need, number of blocks can be customized for different purposes.

We can see, a less strict requirement of similarity will form less blocks, as the precision of similarity goes up, more blocks will be divided (Figure 27). The block-matrix is symmetric, hence, the upper triangle information is identical to the lower triangle. The blocks sited outside the analogical line can help us to analyse the similarity on the block level, when the non-diagonal block has lots of dark spots, it implies that the 2 corresponding blocks might be similar to each from some dimensions.

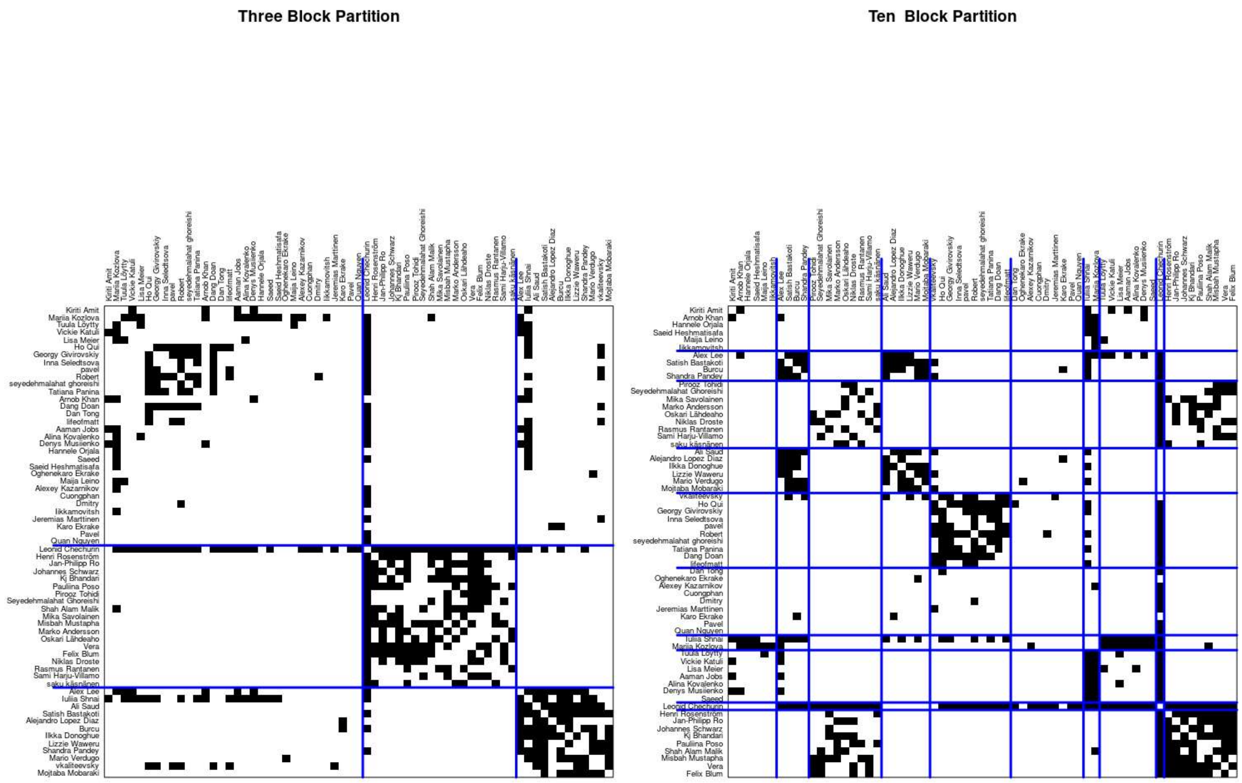

Figure 27:Blockmodeling of TRIZ course network

Of course, it is also feasible to plot the block model in a network format, as the essential work of block model is also clustering based on similarity. Figure 28 is a visualized 4-block model of the TRIZ course network. The nodes sit inside the same block are marked with the same vertex colour.

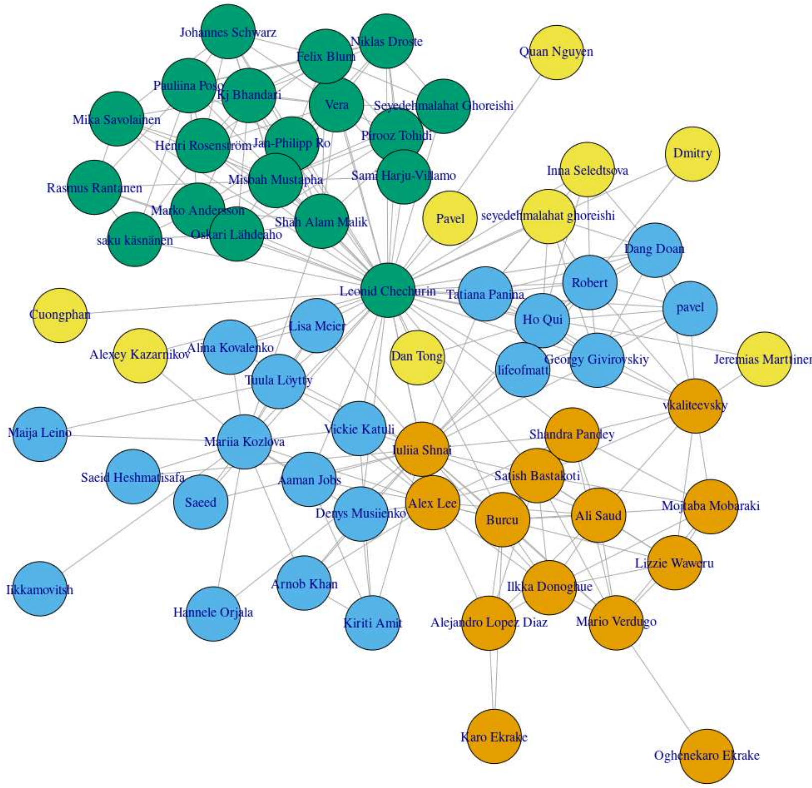

Figure 28:A visualized network by block division

5.7 Hierarchical clustering

Hierarchical clustering using a tree diagram, also known as dendrogram to put similar nodes together. To perform hierarchical clustering, 2 methods could be used: the agglomerative approach, and divisive approach. Figure 29 presents the network of the TRIZ course by agglomerative method of hierarchical clustering, which was accomplished via a “bottom-up” strategy, while Figure 30 shows the divisive approach of clustering the network through “top-down” scheme.

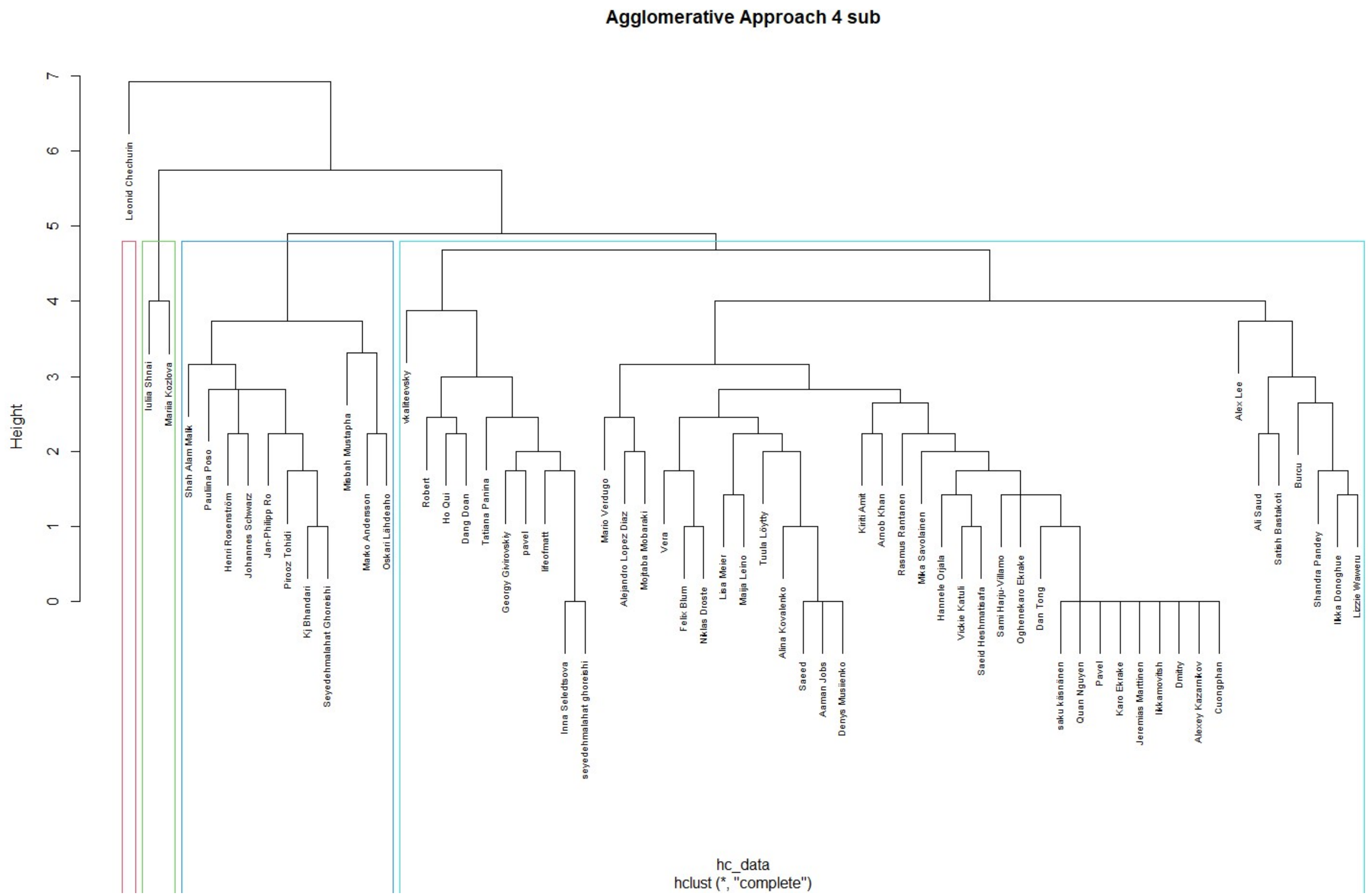

Figure 29:Agglomerative Hierarchical clustering of the TRIZ course class

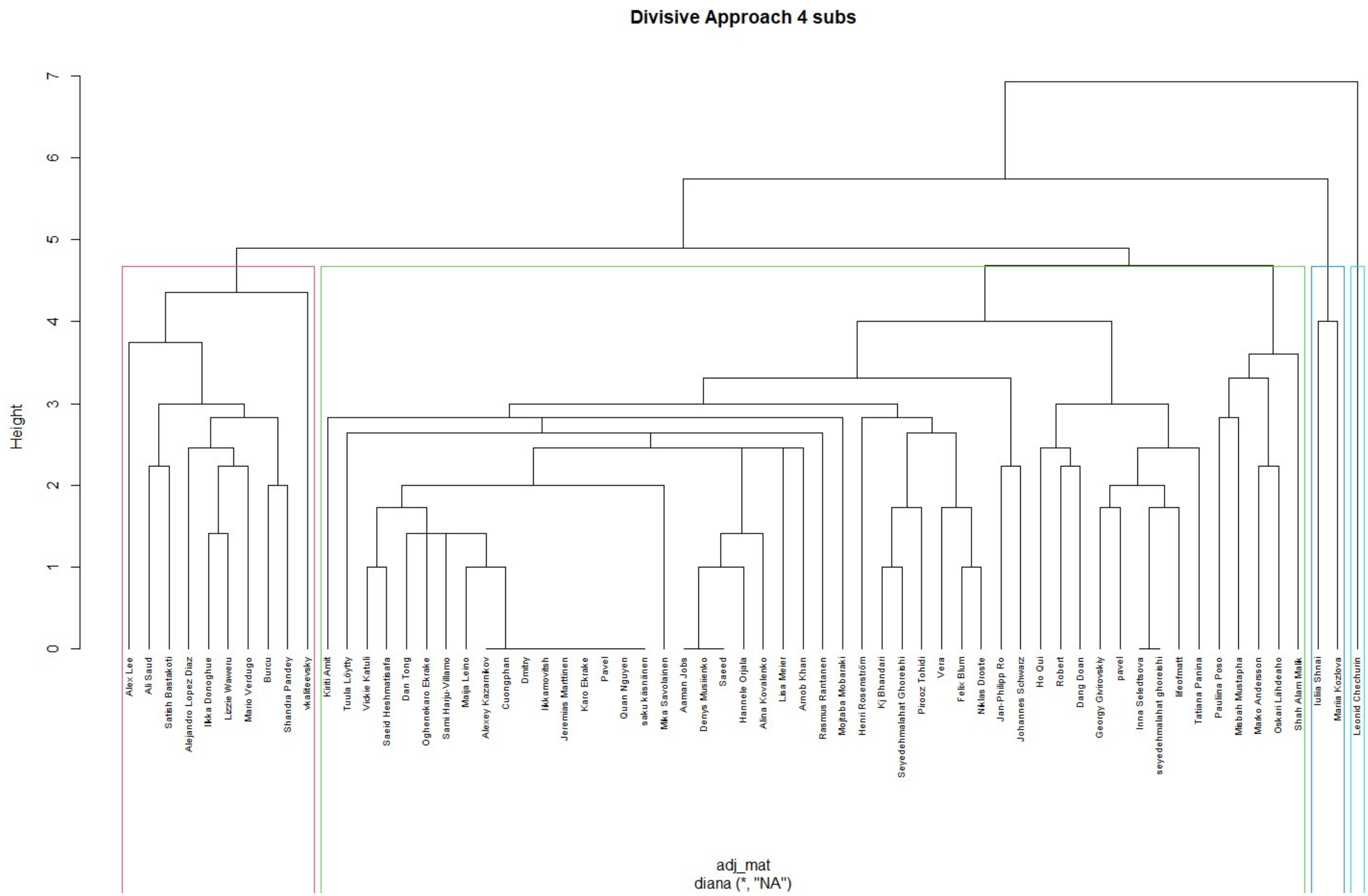

Figure 30:Divisive Hierarchical clustering of the TRIZ class

To better cater researchers needs, dendrograms can be further cut into multiple branches. As shown in Figure 29 and Figure 30, four branches were arbitrarily divided for ease the readability of the dendrogram. The clustering of the TRIZ course network through 2 different methods produced same clustering results, one branch with one actor only, and another branch that covers 2 nodes. The third branch contains 12 actors in the agglomerative approach of hierarchical clustering, while in the divisive method, only 10 actors are combined, the rest of the actors in the network are placed into the fourth branch in both clustering methods.

5.8 Multidimensional scaling

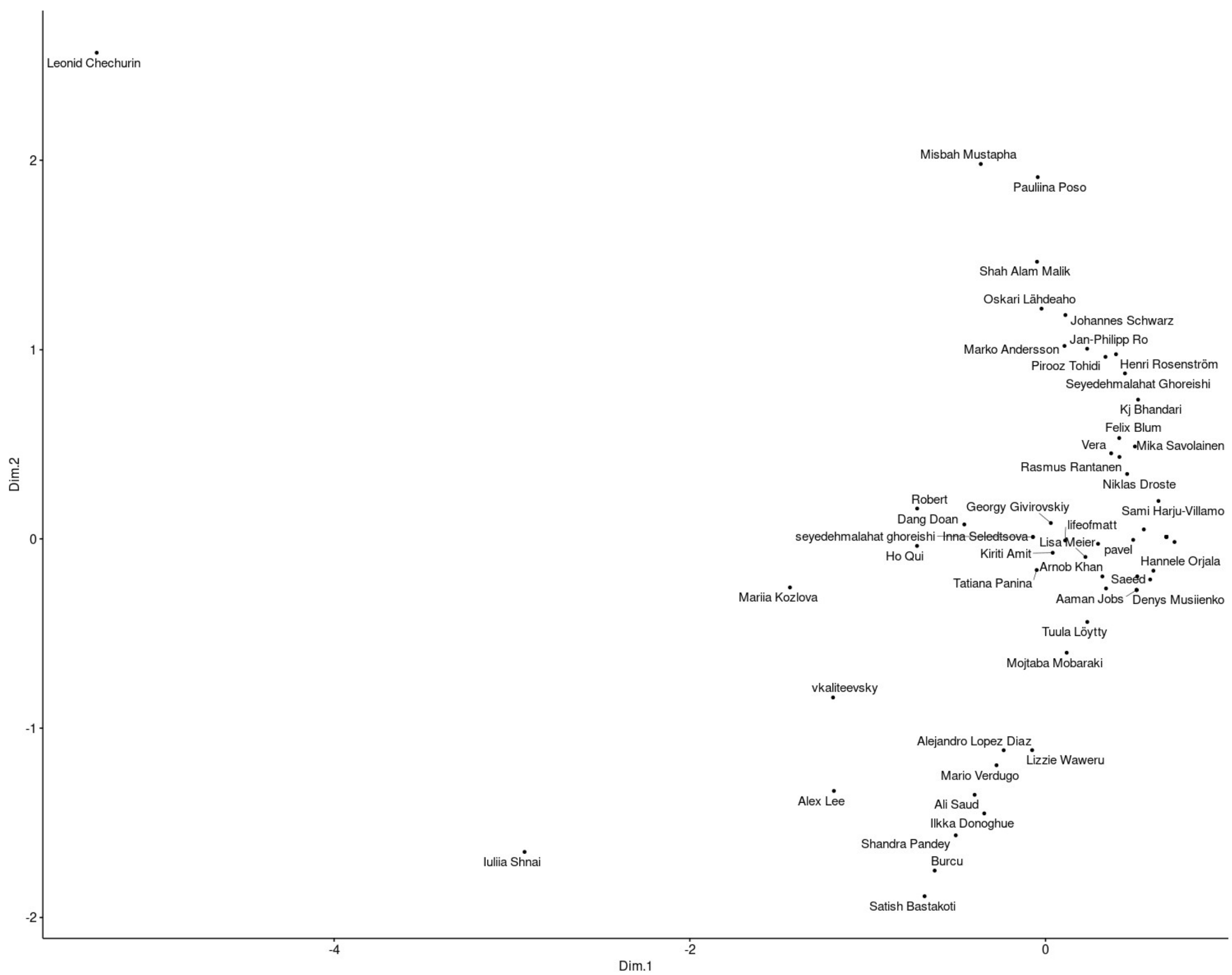

Figure 31: A simple MDS of TRIZ class

Apart from hierarchical scaling, multidimensional scaling is another technique to visualize the members of network based on the similarity between each other. The network of the TRIZ course was plotted in a 2-dimensional space as demonstrated in Figure 31. Like the hierarchical clustering approach, nodes that share a greater similarity are plotted closer to each other on the plane. As we can see from Figure 31, most of the actors are drawn closer to each other on dimension 1, while on dimension 2, actors are plotted more separately compared to the first dimension, which means from the perspective of the second dimension, actors are quite different from each other. Comparing to a randomly generated network as illustrated in Figure 32, the actors are much similar to each other on dimension 1, which might be rooted in their identities as students in the TRIZ course.

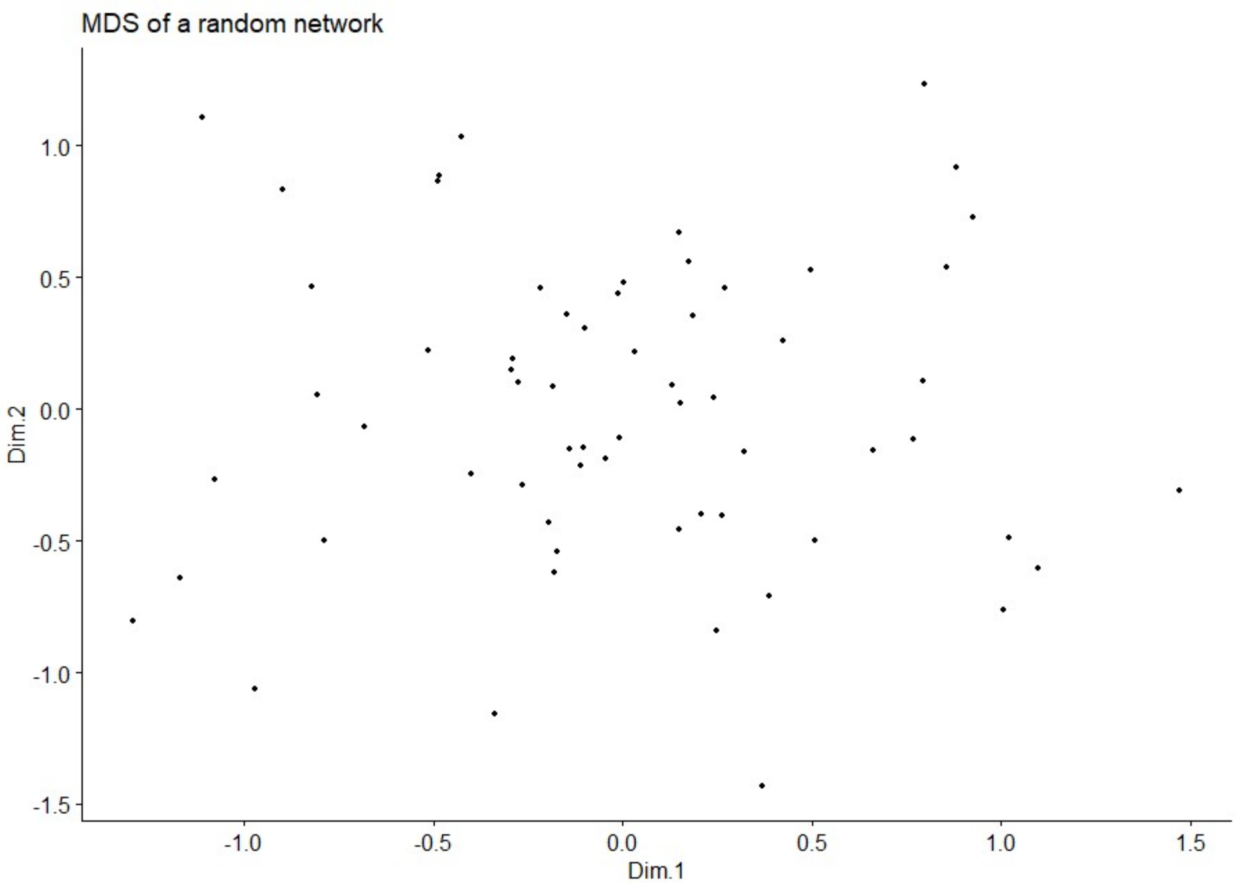

Figure 32:MDS of a random network

Further grouping can be helpful to help researchers understand how similar the actors are to each other on an aggregated level. Figure 33 exhibits the possibility to further divide the multidimensional scaling plot into subs with customized grouping parameters according to researchers' desire.

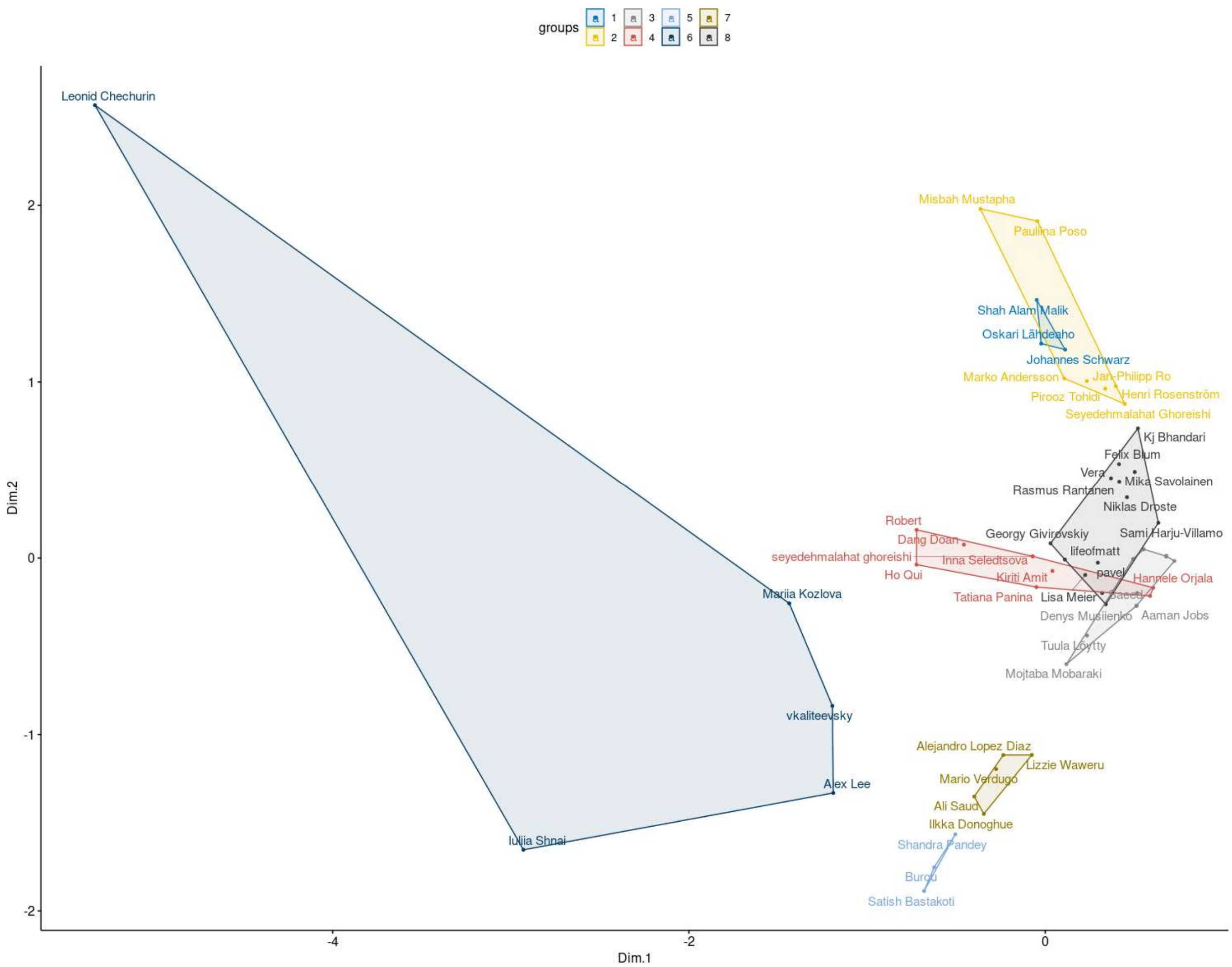

Figure 33:Further grouping based on MDS

Multidimensinal scaling could also be applied as a layout technique for network plotting. Positional analysis could be combined with MDS, for example, Figure 34 used MDS as the layout technique to plot the network while vertex's color are determined with isomorphic equivalence method. The actors drawned nearby might not be in the same role or position according to the isomorphic algorithm, however, they are similar to each other while analyzed while being projected to the selected dimensions chosen by the MDS algorithum.

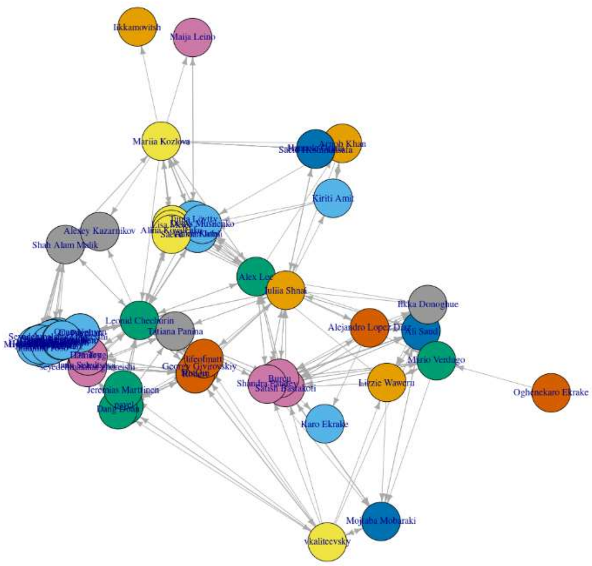

Figure 34:MDS as a layout form isomorphic grouping visualization

5.9 Random walks

Community detection can also be done by conducting random walks in a graph, the walktrap community finding algorithm published by Pons and Latapy (2005, pp. 284–293) can effectively detect subgroups as shown in Figure 35. The core of the walktrap algorithms is to calculate the nodes similarity with distance realized by random walks. The idea was that the shortest random walks tend to stay inside the network. (Pascal and Gabor, no date)

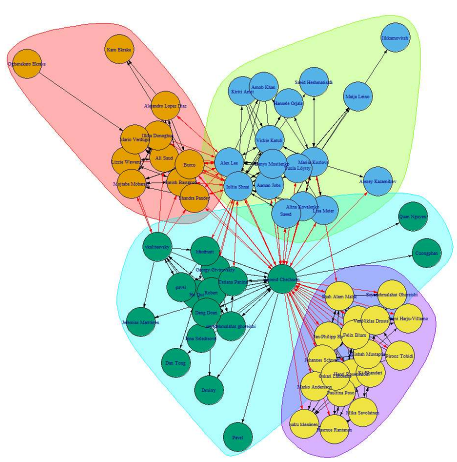

Figure 35:Community detection with walktrap

5.10 Topic modelling with LDA

Based on probability of words co-appearance, several topics were generated through the LDA unsupervised machine learning algorithm with the textual documents originally extracted from the xml documents from Disqus platform. Some words appeared in multiple topics generated by the unsupervised machine learning LDA algorithm as demonstrated in Figure 36. However, the weight of the same word can vary among different topics, for instance, the word “function” weights much less in topic 1 while benchmarking with other topics. Besides, the understandability of topics also varies, topics which contains more words with high value score are more easily understandable by human being from the linguistics point of view.

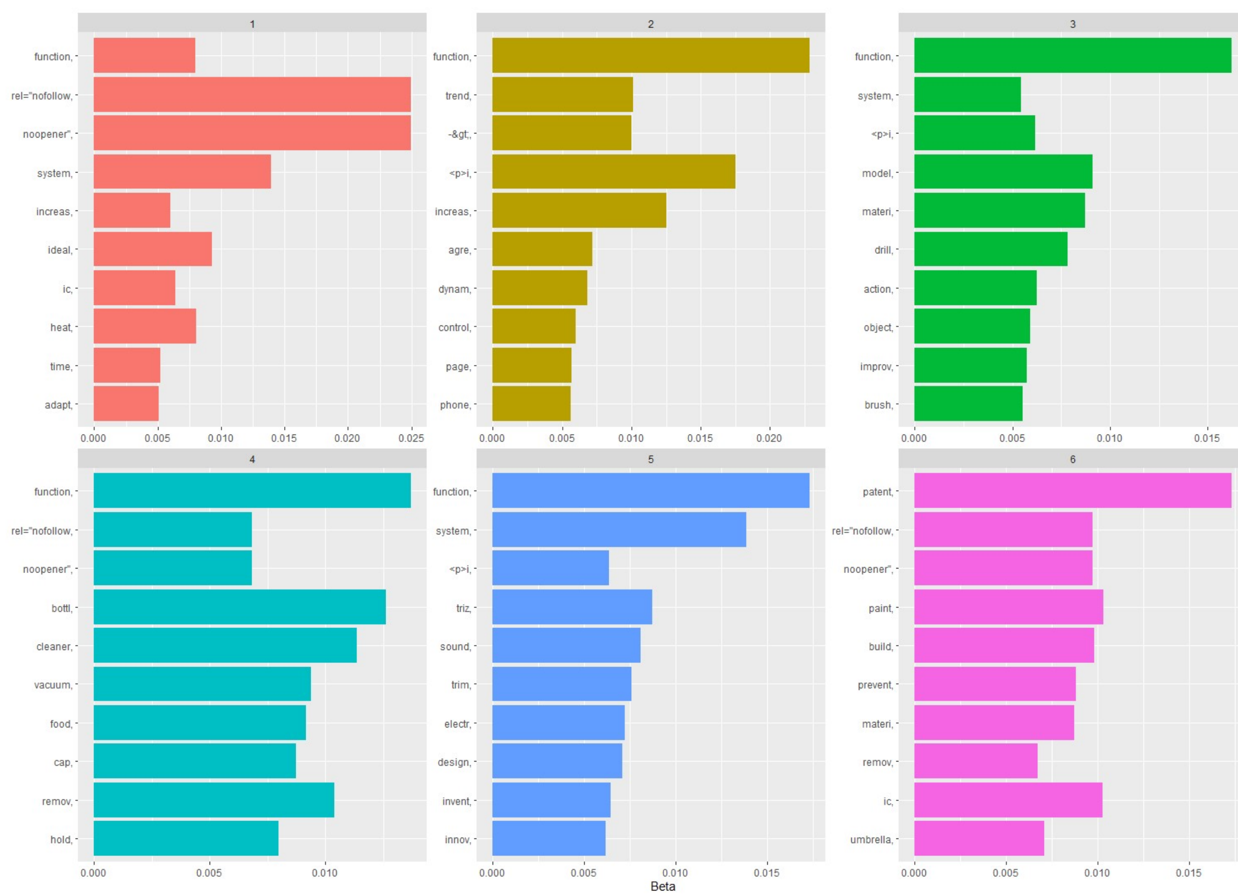

Figure 36:Topics produced by LDA unsupervised algorithm

6 Discussions

As previously discussed in the introduction and literature review chapter, the massive adoption of digital apparatus and the relocation of learning space from traditional offline classroom to the internet-based virtual learning environment have make it much flexible to collect and store students' interaction in various data formats. With learning management systems, tons of learning based bigdata are generated continuously, however, with unstructured data only, it is miserably difficult to understand how our learners are doing, and to provide feedback or evaluation. Just like the process of knowledge creation, it is indispensable to using tools to interoperate the mine of data to practical information that can be cognitively understandable, and then convert the useful information to well-structured knowledge that has been universally acknowledged.

Learning analytics has attracted numerous interests from scholars and experts that come from various disciplines such as education, computer science, and social science, etc. In the literature review stage, answer to the first research question was found: The two major tasks that researchers that are doing under the learning analytics discipline can be coarsely summarized as reflection, and prediction. Reflection is somewhat a descriptive analysis of the data being giving to study, intends to explain what has happened. Prediction used previous data to forecast the variable's magnitude interval in the future based on statistical and probability models.

This master thesis project focuses on the reflection perspective of LA. With an aim to explore students' online activities to give lecturers or examiners a clear view of how the leaners is acting in the learning community, which could be used as reference for in-course intervene as part of the responsivity of guidance for the TP role in the community of inquire model, and it is also applicable to use as support for grading and feedback giving in the end stage of the course.

The second research question in the introduction chapter aims to seek useful tools that instructors can take from the learning analytics discipline to get a richer understanding of their students' performance on an internet-based learning context. SNA was therefore noticed to be an appropriate technique that can be applied to generate useful information about students' performance on an entire community level, and further to an individual level.

Under the domain of SNA, several tools were tested to probe the internal situation of the learning community. Measurement of the network properties originated from the network theory was used to grab a primary awareness of performance of class in the global level. Then the centrality measurement from the graph theory and network analysis domain was conducted to throw a light to the analysis of learners' traits on an individual level. Sub-groups detections were accomplished through various practices such as analysis of actors' similarity based on actors' positions and roles in the network model. Moreover, topic modelling realized through LDA algorithm was carried out as an objective to help researchers understand the theme of discussion based on posted text from learner.

Regarding the third research question about how the results of students' online activities analysis

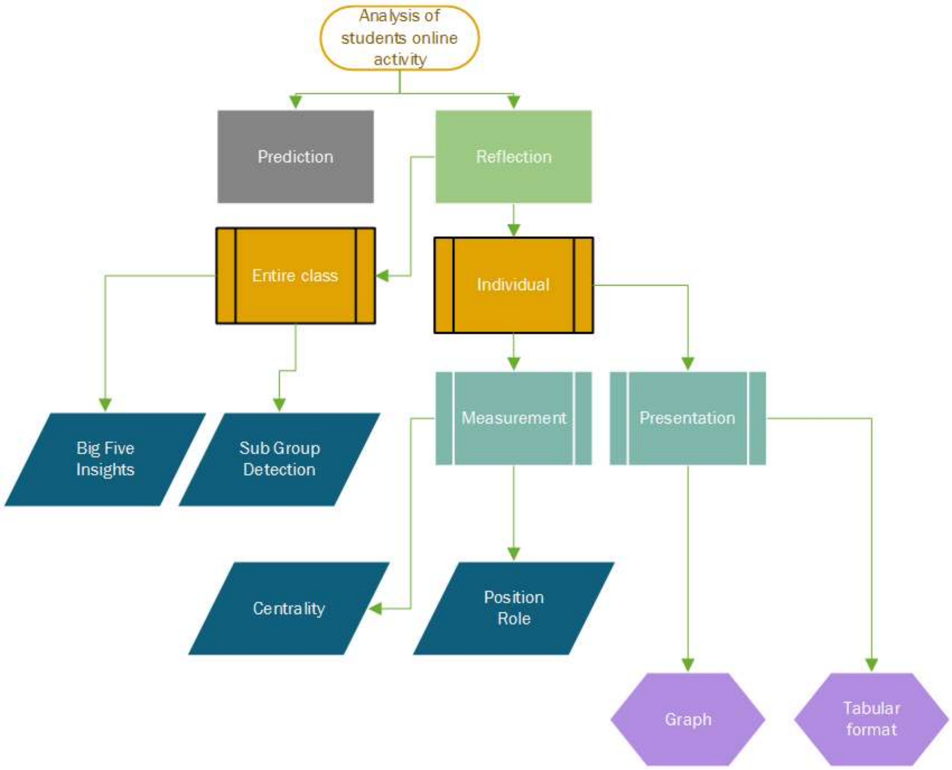

Figure 37:The reflection measuring toolkit

can be presented, it was found that the results could be presented either in a graph format, or in a tabular frame format such as matrices. In some cases, the results of analysis could be simultaneously visualized and encoded in a matrix format such as the blockmodeling section where all members in the learning community are places on both horizontal and vertical axe, the similarities between each other on the individual level and block level are displayed in an image matrix.

The big five insights measure of the TRIZ course network ended up as a table that describe the properties on the global stage, as Table 12 tells, the size of the network is 63 with an edge density of 0.09 at two decimals. Which could be interpreted as the possibility of a pair of actors randomly picked from the above-mentioned network to form a connection is around 10 % which is relatively low.

To understand how different the specific network is compares to random reference, a network with same number of actors and edge density was randomly generated. The diameter of network was gauged at 8 from the TRIZ course network, while in the randomly generated network, the diameter was 5, which means the efficiency of information spread is lower in the TRIZ course network than the randomly generated network. And the TRIZ course network seems to be less compact than the randomly generated network as more strong components were detected in the TRIZ course network. (11 strong components detected in TRIZ course network, and only one was detected in the randomly generated network) The last measuring attribute on the network level was the clustering coefficient, which is also known as transitivity, the transitivity value from the TRZI course network and randomly generated network were 0.39 and 0.20 respectively, which implies that members in the TRIZ course network are more likely to form sub-groups than the ones from the randomly generated network.

Centrality measurement of the TRIZ course network was also compared to the randomly generated network, the degree distribution including both in-degree and out-degree was found to be a right skewed shape in the histogram from the TRIZ course network, while a normal distribution was found from the randomly generated network on its degree attribute.

The members K-cores structure was visualized in the sub-group detection section. members who share the same number of connections to other members was marked into the same vertex colour for visualization, and it was found very straightforward for the instructor to find the students who lack discussion mate, which can be used for in-course intervene to help the students who are "in danger" from the CoI point of view.

The blockmodeling technique was noticed to be a convenient tidy format to plot the relationship among the actors in the network, it is applicable for the instructors to foster actors from different blocks to have interaction

Clustering is mostly based on similarity of actors in the network, however the isomorphic equivalence technique and the MDS seems have different internal criteria as Figure 34 illustrated, the actors who are isomorphic equivalent are clustered into same vertex colour, however, those isomorphic equivalence are not plotted near each other with the MDS algorithm as reference for layout. This might be explained as actors in the same network might not be in the same role or position, however they share some likeness on one or more underlying dimensions.

The results discussed above is only a case study with a particular dataset collected from a discussion forum under a university level TRIZ course. Hence the results discovered from the data used in this master thesis project are not universally applicable to other scenario such as adult working skill training course.

The findings of this master thesis project based on the input data seems not very encouraging. The entire class is not actively having interaction as indicated by the network's edge density and diameter. Also, a higher tendency to form sub-groups can also leads to a negative impact on the spread of information. The centrality distribution indicates that the teacher and teaching assistants were dominating the discussion which was strongly not recommended by the CoI theory.

This research attempts to discern how instructors can evaluate students' performance with tools from the learning analytics discipline. Corresponding to the three presences from the CoI theory, SP could be studies via social network analysis and presented in multiple ways. However, the CP and TP cannot be practically measured by SNA. For this reason, it is severely appreciable the future researchers could develop framework or bring useful tools to help understand the CP of individual learners in an internet-based learning context, such as text mining techniques. And it would be stimulating to study the relation between students' SP and CP.

7 Conclusions

This master thesis project examined how teachers can utilize LA to get a better understanding of students' performance in an internet-based VLEs, which can be used as reference for grading, and feedback giving, as well as providing in-course mediation to encourage students "in danger" to establish more connection with peers and sharing self-generated idea and further discourse.

SNA was found to be a useful toolkit study the students' online interaction mostly on the SP perspective. And graph and matrix were found to be feasible to represent the network structure of a class. Analysis results of students' activities based on some criteria such as centrality, similarity analysis on the individual level and sub-group detections on the class level could be further visualized in network graph, block-based image matrix, MDS, dendrogram.

The performance of the entire class was examined by gauging the network global properties such as density, diameter, components, and clustering coefficient. An edge density of 0.09 was found in the sample dataset, which signifies that the open communication in the TRIZ course was not well achieved due to the lack of interaction among random pairs of actors in the network. And the relatively large diameter (with the network size as reference) of the network also gives a signal that the efficiency of information diffusion is not so favourable. The transitivity of the class network (0.39) is also higher than the value measured from the matched random network.

The right skewed histograms of centrality distributions indicates that the polarization is very critical in the class network. The roles of TP such as instructor and teaching assistants seems to dominate the activities in the network, therefore a balance of learning and teaching was not achieved in the duration of that TRIZ course.

Several sub-group detections methods were used to discover cohesive groups in the TRIZ class. Nodes were clustered into a subset of the entire network based on similarity, or accessibility measurements such as clique and K-cores and n-clique. The visualization of findings was done in different techniques, such as hierarchical clustering, multidimensional scaling, and blockmodeling.

The results of this study as summarized in Table 16 suggest that instructors can use SNA as a tool to measure the students' online activities, and the result of analysis can be easily visualized in multiple ways. This research made SNA becomes a useful toolkit for the teaching staff to keep track on students' progress both on a class level and individual level practically easy to apply. Clustering of the students can offer the teaching staff an opportunity to identify the similar students, which can be used during the course period for mediation, which could be positive for encouraging inter-cluster communication as recommended in the CoI theory in the format of communication. However, the CP and TP were not covered in the SNA toolkit, therefore evaluation of the students' overall learning must be accompanied with other attributes from other learning analytics tools to unveil the qualitative reality. For the future, more focus should be paid to the cognitive perspective and TP and possible integrate them into a new toolkit together with the SNA techniques that had been examined in this research, thus, teachers can get more even-handed foundation to guide students and giving evaluation and feedback based on the reflectional analysis of students' online footprints.

7.1 Summary of conclusions

Table 16 A summary of conclusions

Q1 What does LA do?	LA is used for reflection and prediction study
Q2 What tools are available in LA for teachers to use that can capture students' online performance?	Under SNA and NLP, network level attributes such as density, transitivity could be used to get a grasp of entire class. Centrality measurement could be used to unveil individual level information such as engagement in interaction, prestige, Positional/similarity analysis to find cohesions Topic Modeling with LDA could help teachers to obtain themes of the online discussion
Q3 How can teachers present and communicate the analysis results of students' online behavior?	Network graph and matrix could picture the structure of class; HC, MDS, blockmodeling could help to give interpretation of similarity among individuals.

References

Akyol, Z., Vaughan, N. and Garrison, D. R. (2011) 'The impact of course duration on the development of a community of inquiry', 19(3), pp. 231–246.

Ali, L. *et al.* (2012) 'A qualitative evaluation of evolution of a learning analytics tool', 58(1), pp. 470–489.

Baker, R. S. and Inventado, P. S. (2014) *Educational data mining and learning analytics*. Springer (Learning analytics), pp. 61–75.

Baker, R. S. and Yacef, K. (2009) 'The state of educational data mining in 2009: A review and future visions', 1(1), pp. 3–17.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) 'The semantic web', 284(5), pp. 34–43.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) 'Latent Dirichlet Allocation'.

Borup, J., West, R. E. and Graham, C. R. (2012) 'Improving online social presence through asynchronous video', 15(3), pp. 195–203.

Butler, D. L. and Winne, P. H. (1995) 'Feedback and self-regulated learning: A theoretical synthesis', 65(3), pp. 245–281.

Campbell, J. P. (2007) *Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study*. Purdue University.

Castellanos-Reyes, D. (2020) '20 Years of the Community of Inquiry Framework', 64(4), pp. 557–560. doi: 10.1007/s11528-020-00491-7.

Chatti, M. A. *et al.* (2012) 'A reference model for learning analytics', 4(5–6), pp. 318–331.

Cooper, A. (2012) 'What is analytics? Definition and essential characteristics', 1(5), pp. 1–10.

Dewey, J. and Childs, J. L. (1933) 'The underlying philosophy of education', 8, pp. 77–103.

Drachsler, H. *et al.* (2010) ‘Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning’, 1(2), pp. 2849–2858.

Dringus, L. P. and Ellis, T. (2005) ‘Using data mining as a strategy for assessing asynchronous discussion forums’, 45(1), pp. 141–160.

Ess, C. and Jones, S. (2004) *Ethical decision-making and Internet research: Recommendations from the aoir ethics working committee*. IGI Global (Readings in virtual research ethics: Issues and controversies), pp. 27–44.

Ferguson, R. (2012) ‘Learning analytics: drivers, developments and challenges’, 4(5–6), pp. 304–317.

Garrison, D. R. (2016) *E-learning in the 21st century: A community of inquiry framework for research and practice*. Taylor & Francis.

Garrison, D. R. (2021) *PURPOSEFUL AND SOCIAL INTERACTION*. Available at: http://www.thecommunityofinquiry.org/editorial30.

Garrison, D Randy, Anderson, T. and Archer, W. (1999) ‘Critical inquiry in a text-based environment: Computer conferencing in higher education’, 2(2–3), pp. 87–105.

Garrison, D. R., Anderson, T. and Archer, W. (1999) ‘Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education’, 2(2), pp. 87–105. doi: 10.1016/S1096-7516(00)00016-6.

Garrison, D. R., Anderson, T. and Archer, W. (2001) ‘Critical thinking, cognitive presence, and computer conferencing in distance education’, 15(1), pp. 7–23.

Garrison, D. R. and Cleveland-Innes, M. (2005) ‘Facilitating cognitive presence in online learning: Interaction is not enough’, 19(3), pp. 133–148.

GOSA (no date) *What is Digital Learning?* . (https://gosa.georgia.gov). Available at: https://gosa.georgia.gov/about-us/what-digital-learning.

Greller, W. and Drachsler, H. (2012) ‘Translating learning into numbers: A generic framework for learning analytics’, 15(3), pp. 42–57.

Gu, X. and Hu, Y. (2020) ‘Learning Analytics: Opportunities and Challenges’, (26(2)), pp. 40–42.

Hew, K. F., Cheung, W. S. and Ng, C. S. L. (2010) 'Student contribution in asynchronous online discussion: A review of the research and empirical exploration', 38(6), pp. 571–606.

Hoffman, M. (no date) *11 Transitivity, structural balance, and hierarchy | Methods for Network Analysis*. (https://bookdown.org/markhoff/social_network_analysis/transitivity-structural-balance-and-hierarchy.html). Available at: https://bookdown.org/markhoff/social_network_analysis/transitivity-structural-balance-and-hierarchy.html (Accessed: 29 September 2021).

Hu, Y. and Zhang, J. (2018) 'Present Situation and Trend of Precision Teaching in China—Based on the Bibliometric Analysis of CNKI'', in. Guangzhou (Proceedings of the 22nd Global Chinese Conference on Computers in Education (GCCCE 2018)), pp. 184–191.

Kaplan, A. (2018) 'Academia goes social media, MOOC, SPOC, SMOC, and SSOC: The digital transformation of Higher Education institutions and universities'.

Knoke, D. and Yang, S. (2008) *Social network analysis*. 2nd edn. Los Angeles, Calif.] ; SAGE.

Kovanović, V. *et al.* (2018) 'Exploring communities of inquiry in Massive Open Online Courses', 119, pp. 44–58. doi: 10.1016/j.compedu.2017.11.010.

Lipman, M. (2003) *Thinking in education*. 2nd edn. New York: Cambridge University Press.

Luke, D. (2015) *A User's Guide to Network Analysis in R*. Cham: Springer International Publishing AG.

Mazza, R. and Dimitrova, V. (2004) 'Visualising student tracking data to support instructors in web-based distance education', in. *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pp. 154–161.

McIntosh, N. E. (1979) 'Barriers to implementing research in higher education', 4(1), pp. 77–86.

Mehall, S. (2021) 'Purposeful interpersonal interaction and the point of diminishing returns for graduate learners', 48. doi: 10.1016/j.iheduc.2020.100774.

Norris, D. *et al.* (2008) 'Action analytics: Measuring and improving performance that matters in higher education', 43(1), p. 42.

Pascal, P. and Gabor, C. (no date) *Community strucure via short random walks*. (igraph.org).

Pons, P. and Latapy, M. (2005) 'Computing communities in large networks using random walks', in. *International symposium on computer and information sciences*, Springer, pp. 284–293.

Reffay, C. and Chanier, T. (2003) *How social network analysis can help to measure cohesion in collaborative distance-learning*. Springer (Designing for change in networked learning environments), pp. 343–352.

Robert A, H. and Mark, R. (no date) *12. Network positions and social roles: The idea of equivalence*. Available at: https://faculty.ucr.edu/~hanneman/nettext/C12_Equivalence.html (Accessed: 2021).

Rogers, P. L. *et al.* (2009) *Encyclopedia of distance learning*. IGI Global.

Saunders, M., Lewis, P. and Thornhill, A. (2016) *Research methods for business students*. Harlow, Essex: Pearson Education.

Siemens, G. (2010) *Learning Analytics & Knowledge: February 27-March 1, 2011 in Banff, Alberta*. Available at: https://tekri.athabascau.ca/analytics/.

Siemens, G. (2011) *Learning Analytics: EDUCAUSE* . Available at: https://www.slideshare.net/gsiemens/learning-analytics-educause.

Siemens, G. and Baker, R. S. d (2012) 'Learning analytics and educational data mining: towards communication and collaboration', in. *Proceedings of the 2nd international conference on learning analytics and knowledge*, pp. 252–254.

Siemens, G. and Gasevic, D. (2012) 'Guest Editorial - Learning and Knowledge Analytics', 15(3), pp. 1–2.

Théberge, P. (2015) *Digitalization*. Routledge (The Routledge Reader on the Sociology of Music), pp. 345–354.

Tinto, V. (1998) 'Colleges as communities: Taking research on student persistence seriously', 21(2), pp. 167–177.

Traphagan, T. W. *et al.* (2010) 'Cognitive, social and teaching presence in a virtual world and a text chat', 55(3), pp. 923–936.

Vaughan, N. and Garrison, D. R. (2005) 'Creating cognitive presence in a blended faculty development community', 8(1), pp. 1–12.

Venkatesh, V. and Bala, H. (2008) 'Technology acceptance model 3 and a research agenda on interventions', 39(2), pp. 273–315.

Wise, A. F. *et al.* (2013) 'Broadening the notion of participation in online discussions: Examining patterns in learners' online listening behaviors', 41(2), pp. 323–343.

Zaiane, O. (2001) 'Web usage mining for a better web-based learning environment'.