Using Learning Analytics to Understand and Support Collaborative Learning

Mohammed Saqr

Abstract

Learning analytics (LA) is a rapidly evolving research discipline that uses insights generated from data analysis to support learners and optimize both the learning process and learning environment. LA is driven by the availability of massive data records regarding learners, the revolutionary development of big data methods, cheaper and faster hardware, and the successful implementation of analytics in other domains. The prime objective of this thesis is to investigate the potential of learning analytics in understanding learning patterns and learners' behavior in collaborative learning environments with the premise of improving teaching and learning. More specifically, the research questions comprise: How can learning analytics and social network analysis (SNA) reliably predict students’ performance using contextual, theory-based indicators, and how can social network analysis be used to analyze online collaborative learning, guide a data-driven intervention, and evaluate it. The research methods followed a structured process of data collection, preparation, exploration, and analysis. Students’ data were collected from the online learning management system using custom plugins and database queries. Data from different sources were assembled and verified, and corrupted records were eliminated. Descriptive statistics and visualizations were performed to summarize the data, plot variables’ distributions, and detect interesting patterns. Exploratory statistical analysis was conducted to explore trends and potential predictors, and to guide the selection of analysis methods. Using insights from these steps, different statistical and machine learning methods were applied to analyze the data. The results indicate that a reasonable number of underachieving students could be predicted early using self-regulation, engagement, and collaborative learning indicators. Visualizing collaborative learning interactions using SNA offered an easy-to-interpret overview of the status of collaboration, and mapped the roles played by teachers and students. SNA-based monitoring helped improve collaborative learning through a data-driven intervention. The combination of SNA visualization and mathematical analysis of students’ position, connectedness, and role in collaboration was found to help predict students’ performance with reasonable accuracy. The early prediction of performance offers a clear opportunity for the implementation of effective remedial strategies and facilitates improvements in learning. Furthermore, using SNA to monitor and improve collaborative learning could contribute to better learning and teaching.

Keywords: Learning analytics, Social Network Analysis, Collaborative Learning, Medical Education, Interaction Analysis, Machine Learning.

Stockholm 2018

http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-159479

Department of Computer and Systems Sciences

Stockholm University, 164 07 Kista
USING LEARNING ANALYTICS TO UNDERSTAND AND SUPPORT COLLABORATIVE LEARNING

Mohammed Saqr
Using Learning Analytics to Understand and Support Collaborative Learning

Mohammed Saqr
To the exquisite roses
Carmen and Layla
Not knowing when the dawn will come
I open every door;
Or has it feathers like a bird,
Or billows like a shore?

Emily Dickinson
Last year, my father left in peace without having the chance to see what he always hoped for, I always wished he could witness this day. Without his hard work throughout his life that was dedicated to the family’s prosperity, I would not have made it. His encouragement, his faith in me and his assurances were the driving force that kept me going. Words cannot express the gratitude, thankfulness, and appreciation of the support my mother has offered me. It is just unfair that words are used to translate such a great feeling. I am immensely grateful for my main supervisor, Uno Fors. His scientific rigor, expertise, endless support, guidance, and continuous encouragement have paved the way for me to take every step. Uno has always been there for me, has always stood by me when I needed, and has provided the feedback and advice that helped me learn, progress and advance.

I am also vastly grateful to Matti Tedre, my supervisor who introduced me to the art and craft of academic writing, taught me how to think critically, and pushed me forward with his thoughtful, elaborate, and constructive comments. Although he left before I finished, his guidance and feedback still resonate until today. I extend my sincere gratitude to Jalal Nouri, my second supervisor, whose invaluable guidance, motivation and fine-grained feedback on every work made a huge difference. Jalal’s help extended to the way I manage time, study, readings, publications, and even social life. Jalal was always willing to help at all times and regarding everything. Without the steadfast support of my family, who endured difficult times until this work was completed; I would have never moved a step forward, for all what you have offered me, I am forever thankful and grateful.
Special thanks to all my wonderful colleagues and friends who helped, supported and were there for me along this journey, namely, Hazem, Josefina, Nina, Lisa Rolf, Melinda, Qi Dang. I also extend my thankfulness and gratitude to the people who helped make the administrative issues smooth and very swift namely, Tuija Darvishi, Britt-Marie, Irma, Amos, Eija, and Katarina.

محمد محمد صقر عبد الجليل
Mohammed Saqr
Stockholm, August 2018
Contents

List of tables .................................................................................................................. 1
List of figures ............................................................................................................... 2
List of articles .............................................................................................................. 4
Abstract ....................................................................................................................... 5

1. Introduction ............................................................................................................ 7
   1.1 Thesis structure ................................................................................................. 10

2. Problem and motivation .................................................................................... 11
   2.1 General problem definition ............................................................................. 11
      2.1.1 Research in collaborative learning settings .............................................. 13
      2.1.2 The need for efficient monitoring .............................................................. 15
      2.1.3 The need for studies on the impact of learning analytics ...................... 15
      2.1.4 The need for generalizability .................................................................... 16
      2.1.5 Learning analytics research in medical education .................................. 16

3. Research aim ......................................................................................................... 18

4. Definition and taxonomy .................................................................................... 20

5. Theory and philosophy ....................................................................................... 23
   5.1 A new paradigm? ............................................................................................. 23
   5.2 Inferring learning from data ............................................................................ 25
   5.3 Theory and learning analytics research ........................................................ 26
   5.4 Theoretical underpinning of my research ................................................... 28
      5.4.1 Self-regulation ......................................................................................... 29
      5.4.2 Engagement ............................................................................................. 32
      5.4.3 The theoretical basis of collaborative learning ...................................... 33
   5.5 My approach to the study of collaborative learning ...................................... 35
6. The learning analytics process ........................................................... 38
  6.1 Data capture .................................................................................. 39
  6.2 Preprocessing and Preparation ..................................................... 41
  6.3 Analysis & Interpretation ............................................................... 42
    6.3.1 Predictive modeling ............................................................... 42
    6.3.2 Clustering ............................................................................. 43
    6.3.3 Content analytics ................................................................. 44
    6.3.4 Social network analysis ......................................................... 45
  6.4 Insightful action ........................................................................... 51
  6.5 Feedback ...................................................................................... 53

7. Research on learning analytics ....................................................... 54
  7.1 An overview of previous research ................................................. 54
  7.2 Notable institutional implementations .......................................... 56
    7.2.1 Course signals ....................................................................... 56
    7.2.2 E2Coach ................................................................................ 58
    7.2.3 OU Analyse .......................................................................... 59

8. Challenges in the field of learning analytics .................................... 61

9. Collaborative learning .................................................................... 67

10. Methods .......................................................................................... 72
  10.1 The general context ................................................................. 72
  10.2 Research methodology ............................................................. 75
    10.2.1 Data capture ................................................................. 75
    10.2.2 Data preprocessing and preparation .................................... 77
    10.2.3 Data analysis, interpretation, and reporting ...................... 77
  10.3 Research ethics ............................................................... 82
    10.3.1 Data protection and privacy policy .................................. 82
    10.3.2 Consent ............................................................................ 83
    10.3.3 Ethical approval ............................................................. 83

11. Overview of the results ................................................................. 85

12. Discussion ....................................................................................... 92
  12.1 Limitations ............................................................................... 100

13. Methodological and theoretical contributions .............................. 102
14. Conclusions .................................................................................... 104
16. Future work .................................................................................... 106
References ............................................................................................ 108
List of tables

Table 1: Centrality measures of the classmates’ network 50
List of figures

Figure 1. The learning analytics process cycle 39
Figure 2. A graph of classmates’ network. 47
Figure 3. An LA dashboard showing a visualization of a student activity. 51
Figure 4. Another example of a teacher dashboard showing class engagement levels. 53
Figure 5. Course Signals at Purdue showing green (safe) and yellow lights (borderline safe). 58
Figure 6. A sketch of the E2Coach system. The system matches advice to students' information and personalizes the messages accordingly. 59
Figure 7. A schematic representation of the classic seven jumps approach. 73
Figure 8. A schematic representation of the approach adopted by Qassim University 74
Figure 9. A copy of the ethical approval 84
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial intelligence</td>
</tr>
<tr>
<td>ALM</td>
<td>Automatic linear model</td>
</tr>
<tr>
<td>AUC</td>
<td>Area under the curve</td>
</tr>
<tr>
<td>CART</td>
<td>Classification and regression tree</td>
</tr>
<tr>
<td>CL</td>
<td>Collaborative learning</td>
</tr>
<tr>
<td>CS</td>
<td>Course signals</td>
</tr>
<tr>
<td>CSCL</td>
<td>Computer supported collaborative learning</td>
</tr>
<tr>
<td>DLA</td>
<td>Dispositional learning analytics</td>
</tr>
<tr>
<td>EDM</td>
<td>Educational data mining</td>
</tr>
<tr>
<td>k-NN</td>
<td>K-nearest neighbors</td>
</tr>
<tr>
<td>LA</td>
<td>Learning analytics</td>
</tr>
<tr>
<td>LAK</td>
<td>Learning analytics &amp; knowledge conference</td>
</tr>
<tr>
<td>LMS</td>
<td>Learning management systems</td>
</tr>
<tr>
<td>NCBE</td>
<td>National committee of bioethics</td>
</tr>
<tr>
<td>OU</td>
<td>Open university</td>
</tr>
<tr>
<td>RQ</td>
<td>Research question</td>
</tr>
<tr>
<td>SNA</td>
<td>Social network analysis</td>
</tr>
<tr>
<td>TEL</td>
<td>Technology-enhanced learning</td>
</tr>
</tbody>
</table>
List of articles

The publications below are included in this thesis:


Abstract

Learning analytics (LA) is a rapidly evolving research discipline that uses insights generated from data analysis to support learners and optimize both the learning process and learning environment. LA is driven by the availability of massive data records regarding learners, the revolutionary development of big data methods, cheaper and faster hardware, and the successful implementation of analytics in other domains. The prime objective of this thesis is to investigate the potential of learning analytics in understanding learning patterns and learners’ behavior in collaborative learning environments with the premise of improving teaching and learning. More specifically, the research questions comprise: How can learning analytics and social network analysis (SNA) reliably predict students’ performance using contextual, theory-based indicators, and how can social network analysis be used to analyze online collaborative learning, guide a data-driven intervention, and evaluate it. The research methods followed a structured process of data collection, preparation, exploration, and analysis. Students’ data were collected from the online learning management system using custom plugins and database queries. Data from different sources were assembled and verified, and corrupted records were eliminated. Descriptive statistics and visualizations were performed to summarize the data, plot variables’ distributions, and detect interesting patterns. Exploratory statistical analysis was conducted to explore trends and potential predictors, and to guide the selection of analysis methods. Using insights from these steps, different statistical and machine learning methods were applied to analyze the data. The results indicate that a reasonable number of un-
derachieving students could be early predicted using self-regulation, engagement, and collaborative learning indicators. Visualizing collaborative learning interactions using SNA offered an easy-to-interpret overview of the status of collaboration, and mapped the roles played by teachers and students. SNA-based monitoring helped improve collaborative learning through a data-driven intervention. The combination of SNA visualization and mathematical analysis of students’ position, connectedness, and role in collaboration was found to help predict students’ performance with reasonable accuracy. The early prediction of performance offers a clear opportunity for the implementation of effective remedial strategies and facilitates improvements in learning. Furthermore, using SNA to monitor and improve collaborative learning could contribute to better learning and teaching.
1. Introduction

A chasm has been growing between the aspirations of the higher education system and what we can realize as educators. The strain between the rising number of learners and the tightening resources of the overtasked teachers has allowed the chasm to grow (Aragon, 2016). Under mounting pressure from learners, the public and governments alike, the search for solutions and opportunities has never been more intense. Sharing goals, objectives, and ambitions, many have turned to technology for a panacea.

Internet and technology at large have kindled the optimism that intelligent solutions may automate a broad array of resource-intensive tasks, thereby enabling teachers to pay closer attention to their pertinent teaching work (Reiser, 2001; Tamim, Bernard, Borokhovski, Abrami, & Schmid, 2011). Technology provides the potential to create and reuse materials for as long as they remain relevant, in theory facilitating remarkable savings in time and effort (Bichsel, 2013). Furthermore, the Internet enables the immediate distribution of materials to large numbers of students at a relatively low cost of production (Bichsel, 2013; Gulati, 2008; Mazoue, 2014). Such delivery can be consistent for all learners; nobody is provided better service than others, and more importantly not worse (Bichsel, 2013). Technology-enhanced learning (TEL) offers a plethora of possibilities and applications such as simulated experiments, virtual training and interactive multimedia that aim to boost engagement and motivation (Abrami et al., 2008; Bichsel, 2013; Cook et al., 2011; R. Ellaway & Masters, 2008; Wouters, Van Nimwegen, Van Oostendorp, & Van Der Spel, 2013). By enabling forms of communication that transcend the physical boundaries of time and place, the Internet has paved the
way for collaborative and interactive milieus that help students reflect on learning, share their experiences, and build new knowledge (Garrison, Anderson, & Archer, 2010; Hrastinski, 2008). Despite the unyielding rise in the rate of adoption and the scale of applications in technology-enhanced learning, concerns continue to be expressed regarding their effectiveness and the optimum approach (Kirkwood & Price, 2014; G. Wang, Gunasekaran, Ngai, & Papadopoulos, 2016). Successful online learning requires efficient learning design, along with specific learner qualities such as motivation, self-regulation, and learning strategies, as well as timely feedback and proper support (Burnette, O'Boyle, VanEpps, Pollock, & Finkel, 2013). The breadth of TEL offers limited personalized feedback for students and provides few insights regarding the online learning process (Gašević, Jovanović, Pardo, & Dawson, 2017; Graham, 2006; Pardo, Han, & Ellis, 2016).

Inspired by the fruitful application of data science in a multitude of industries and services, educators have used available data in the search for solutions (H. Chen, Chiang, & Storey, 2012; Siemens, 2013). The availability of massive data records regarding learners as well as the exponential surge in computing capabilities and the burgeoning of artificial intelligence (AI) models has rendered possible the handling of large volumes of data that can create a significant impact when used efficiently (Kitchin, 2014).

Learning analytics has emerged to explore the opportunities of sense-making of such data and create actionable insights. Such insights may embolden the provision of effective solutions to imperative issues in education such as attrition, personalized support, quality of learning experience, improved learning, and curriculum design, and informing decision makers. Learning analytics may also reduce gaps in personalized feedback in the resource-limited educational landscape (Lim, 2018; Papamitsiou & Economides, 2014; Siemens, 2013).

It was against such a background that this thesis was conceptualized to explore the potential of using students’ data in collaborative learning environments in order to understand and potentially improve learning and teaching. The thesis aims to use a theory-guided
approach to track students’ online activities to identify the indicators of better performance or the lack thereof to find opportunities for insightful proactive action. Moreover, the thesis aims to use learning analytics to guide educators and learners to identify gaps in collaborative learning and try to find opportunities for early support and intervention.
1.1 Thesis structure

The background section of the thesis is divided into nine chapters. The first chapter defines the problem and the main drivers behind this research, followed by the aim and research questions. I then go through the definition and taxonomy of learning analytics and closely related concepts, with the aim of shedding light on important terms and how they are related. I subsequently present a detailed chapter regarding the underlying theory, its role in analytics research (and learning analytics in particular), followed by the ways in which it is operationalized in the thesis. In the next chapter, I present the process of learning analytics from a methodological perspective. I place particular emphasis on the methods employed in this thesis, especially predictive modeling and social network analysis. I then explore previous research in the field with notable examples from research and institutional perspectives. This is followed by the challenges facing researchers and institutions, with a particular focus on ethics and privacy. Given that the context of this study is collaborative learning, I discuss the main ideas of this concept in Chapter 9.

Following the background section, the methods chapter is introduced with a detailed description of the context, data collection, ethics, analysis, and interpretation. The results of my studies are subsequently presented, followed by the discussion, the contributions of the studies, the limitations, the conclusions, and suggestions for future research.
2. Problem and motivation

In this section, I will cover the general problem, starting with the focus in the field of learning analytics on predictive models that rely on generic indicators with little emphasis on theory. I then highlight the importance of studying the structural and relational aspects of collaborative learning, which constitutes the central theme of the thesis. I will subsequently cover the challenges and research gaps that have been identified as motivating factors behind this research. These comprise linking research to theory and context, the lack of generalizability of research findings in collaborative learning settings, and the paucity of research on using learning analytics to create a real impact on learning and teaching. Owing to the significance of the issue of linking research to theory, I dedicate Chapter 5 to the discussion of the theory, significance and how it is operationalized in this thesis.

2.1 General problem definition

Driven by the data and the opportunities it may offer, most initial applications of learning analytics have focused on using students’ online trails to monitor their performance and to predict underachievers in need of support (Ferguson, 2012a; Leitner, Khalil, & Ebner, 2017; Shahiri, Husain, & Rashid, 2015; Tsai & Gasevic, 2017a). Researchers have used log data records from the counts of online activities, such as the count of logins or the frequency of clicks (generic indicators). However, these generic indicators have offered limited insights into the learning environment, have failed to contribute to the development of a valid theoretical model, have
only faint connection to actionable and meaningful intervention models, and have not provided suggestions for implementing a corrective strategy (Conde & Hernández-García, 2015; Nesbit, Xu, Winne, & Zhou, 2008; Winne, Nesbit, & Popowich, 2017). Predictive models that rely on generic indicators have been deemed challenging to replicate or generalize, especially in blended learning programs (Conde & Hernández-García, 2015; Tempelaar, Rienties, Mittelmeier, & Nguyen, 2018). Furthermore, this “one-shot” approach has been criticized for only considering the final product (e.g., the count of online activities and course views by the end of the course) and not the actual process, with little attention to time, connection to actual context and interactions with peers or communities (Winne et al., 2017).

Learners live in communities that shape their behaviors and enable them to interact with and influence one another. Traditional methods of analysis treat individuals as isolated, discrete units of analysis (Borgatti, Mehra, Brass, & Labianca, 2009; Burt, Kilduff, & Tasselli, 2013; Cela, Sicilia, & Sánchez, 2014), while overlooking the importance of interactions with teachers and peers and the influence of the social structure, which are particularly essential in collaborative learning settings. Advocates of social network analysis argue that a student’s social capital, interactions and position in the social structure have a significant effect on academic achievement, which tends to be overlooked by attribute-based researchers who focus on other factors such as age, gender, and performance (Gašević, Zouaq, & Janzen, 2013; Joksimović et al., 2016; Scott & Carrington, 2014). Such an argument has garnered considerable support from social network researchers. However, a lack of framework exists that could capture the relational or interactional construct of online learning. Furthermore, no conclusive evidence exists regarding which social network constructs are correlated with performance in different contexts, or how students regulate their collaborative learning (Dowell et al., 2015; Hernández-García et al., 2015; Joksimović et al., 2016). The methodological underpinnings of learning analytics research in collaborative learning should be aligned with theoretical assumptions and contextual conditions in order to ensure that results
can be empirically reproducible and research findings can be meaningfully useful (Fincham, Gašević, & Pardo, 2018; Hernández-García et al., 2015). Operationalizing interaction indicators, community structure, and relationships among collaborators in collaborative learning environments is both timely and significant. A model that can capture these indicators may help understand the patterns and behaviors of students in collaborative learning and better predict students’ performance. Furthermore, it may facilitate the creation of a monitoring mechanism for efficient collaborative learning and produce and evaluate informed data-driven intervention. Successful operationalization may also help create more accurate and reproducible models that can be replicated across courses that share a similar context. Accounting for the time dimension, self-regulation and engagement might boost accuracy and be considered highly aligned proxy indicator of learning relative to generic indicators (B. Chen, Knight, & Wise, 2018; Cicchinnelli et al., 2018; Molenaar & Järvelä, 2014). Thus, the importance of using pedagogically and contextually relevant approaches to collaborative learning cannot be overstated.

Therefore, I have identified gaps that provided motivations for this research, namely the paucity of research in collaborative learning, the lack of studies concerning the impact of using interaction indicators to guide a data-driven intervention, the difficulty of creating generalizable models, and the potential of learning analytics in medical education, which is of particular relevance to this thesis. These gaps will be discussed in detail in the following section.

2.1.1 Research in collaborative learning settings

In collaborative learning, students interact, exchange and communicate information. Generic indicators offer very little to understand the interactions, relationships and groups or communities where this occurs. Interaction and social network analysis indicators are more relevant to participatory learning settings and are deeply grounded in collaborative learning theories. For example, for the constructivist perspective of learning, SNA offers a bird’s-eye view of learners’ interactions and information exchange among
peers, highlights active and inactive actors, and identifies isolated and non-interactive students (Dado & Bodemer, 2017; Shum & Ferguson, 2012; Sie, Ullmann, Rajagopal, Cela, Rijpkema, & Sloep, 2012). For connectivism learning theory, SNA maps the connections and networks of information exchange, identifies the influential nodes of information and monitors the spread and diversity of perspectives (Mattar, 2010). For social capital theory, SNA can map the networks of resources and help identify different sorts of privileged access to connections, such as connections to powerful actors and bridging different communities (brokerage capital) (Dika & Singh, 2002; Kwon & Adler, 2014).

Researchers have used SNA to study interactions and relationships in online learning, with the most common research subjects comprising the study of patterns of interaction, knowledge construction, computer-supported collaborative learning (CSCL), and performance prediction (Cela et al., 2014; Dado & Bodemer, 2017; Sie, Ullmann, Rajagopal, Cela, Rijpkema, & Sloep, 2012). Studies addressing patterns of interaction have focused on the potential of the technique, but research has failed to address the impact of using the technique to monitor collaboration and guide intervention. Most studies that have attempted to investigate the role of SNA in the prediction of performance have used correlation statistics with different centrality measures. Although results have been encouraging and indicate a positive correlation, no uniform model to guide the operationalization of SNA indicators exists (Dado & Bodemer, 2017). A uniform model that captures the full gamut of the quantity, direction and significance of interactions among students and teachers is required. Such a model should capture the roles of collaborators and the structural properties of the collaborative groups and their communities. Consequently, it might help create a more efficient monitoring mechanism for collaborative learning, enhance the accuracy and generalizability of predictive models, and guide an informed intervention. These issues will be discussed in detail in the following section.
2.1.2 The need for efficient monitoring

Regardless of the power of CSCL as a medium for learning, it is insufficient as a guarantor of efficient collaboration. Research has shown that groups do not spontaneously engage in fruitful collaborative interactions (Fransen, Kirschner, & Erkens, 2011; Kreijns, Kirschner, & Jochems, 2003). Fruitful online interactions require careful group composition, efficient learning design, and a stimulating medium that motivates interaction. They may also require the management of roles, scaffolding by teachers, and scripting or orchestrating. Therefore, different analysis mechanisms have been developed to assess the efficiency of collaborative learning to ensure that it meets its intended pedagogical objectives (Fincham et al., 2018; Kreijns et al., 2003; Lambert & Fisher, 2013; Noroozi, Weinberger, Biemans, Mulder, & Chizari; Resta & Laferrière, 2007).

2.1.3 The need for studies on the impact of learning analytics

Considerable criticism has been voiced regarding the large number of studies concerning the potential of learning analytics, whereas few studies have examined the actual impact of intervention strategies (S. Dawson, Jovanovic, Gašević, & Pardo, 2017a). Research into social network analysis in education began in the 1990s and tended to evaluate the potential of using SNA visualization in mapping patterns of interactions or the roles of collaborators. However, according to systemic reviews (Cela et al., 2014; Dado & Bodemer, 2017; Sie, Ullmann, Rajagopal, Cela, Bitter, & Sloep, 2012), no research has been undertaken regarding the impact of intervention using SNA-based strategies. As such, little is known about effective strategies, the extent of benefits received by learners, or the optimal time for an intervention.
2.1.4 The need for generalizability

Some researchers have studied the potential of social network analysis centrality measures as predictors of learning (Cho, Gay, Davidson, & Ingraffea; Dowell et al., 2015; Hernández-García et al., 2015; Jiang, Fitzhugh, & Warschauer, 2014; Joksimović, Gašević, Kovanović, Riecke, & Hatala, 2015; Joksimović et al., 2016; Romero & Ventura, 2013). The results of these studies have indicated a possible positive correlation between centrality measures and academic achievement. For instance, Romero and Ventura (2013) have reported on the positive correlation between performance and number of interactions (degree centrality) and number of received interactions (indegree centrality). A similar finding has been reported by Hommes et al. (2012) and Joksimović et al. (2016). However, the results were not replicated between studies, and studies of multiple courses have failed to yield consistent results (Dowell et al., 2015; Hernández-García et al., 2015; Jiang et al., 2014; Joksimović et al., 2016). Researchers have suggested that context might constitute a defining factor and have called for a study to address contextual factors as a possible explanation of issues concerning generalizability and reproducibility. Another problem of these studies is the lack of a structured framework that can be used as a basis for a predictive model.

2.1.5 Learning analytics research in medical education

Learning analytics has been demonstrated as a potentially reliable method for identifying and supporting students at risk of underachieving in several fields (Kuzilek, Hlosta, Herrmannova, Zdrahal, & Wolff, 2015; Pistilli & Arnold, 2010; Wolff, Zdrahal, Nikolov, & Pantucek, 2013). Today, calls for the medical education community to harness these benefits can be heard (Doherty, Sharma, & Harbutt, 2015; R. H. Ellaway, Pusic, Galbraith, & Cameron, 2014). However, little research has considered the role or potential of learning analytics in tackling the issue of drop-outs in the medical education literature (Saqr, 2018). This apparent gap and requirement for research have stimulated prominent medical educationists
to advocate for the development of knowledge and capacity in the field. Ellaway, an editor of the journal *Advances in Health Sciences Education* and a prominent scholar, has summarized this issue in a position article on big data and analytics in 2014: “*If health professional education is to control its future, be accountable for the way its programs run, and thereby maintain its contract with society then health professional educators will need to be ready to deal with the complex and compelling dynamics of analytics and Big Data. We therefore need to explore, discuss, and critique these emerging techniques to develop a robust understanding of their strengths and limitations in the service of health professional education.*” (R. H. Ellaway et al., 2014, pp. 222).

The reasons behind the medical education community’s delay in following other educational disciplines may be connected to the challenging nature of LA research. LA requires a multidisciplinary approach that includes education, statistics, data science, and instructional designers. It also requires a clear ethical framework and data usage policies. Furthermore, it requires institutional readiness and stakeholders’ awareness, as well as funding and support (R. H. Ellaway et al., 2014).
3. Research aim

As a preparation step for this research project, a series of pilot studies were conducted and presented at international conferences (Al-Gadaa, Alkadi, & Saqr, 2014; AlGhasham, Saqr, & Kamal, 2013; Saqr, AlGhasham, & Kamal, 2014; Saqr, Kamal, & AlGhasham, 2014). These studies investigated the role of learning analytics in predicting students who are at risk of failing a course as well as the study of social network analytics in the context of collaborative learning. This practical research experience, along with the literature review and engagement with the education and learning analytics communities, helped formulate the initial research objectives.

Against a background of the paucity of empirical research on learning analytics in collaborative learning environments in general, and medical education in particular, this thesis aimed to explore the potential of using theory-guided learning analytics to understand and improve collaborative learning. I endeavor to use a model that captures the interactions, relations, self-regulating patterns and attributes of collaborators and their communities in order to help facilitate more effective monitoring of collaborative learning and better informed data-driven intervention, as well as to boost the accuracy of predicting underachieving students and most importantly to provide a more replicable learning analytics model.

The aim of this research was:
- To investigate the role of learning analytics in understanding and supporting online collaborative learning.
Two research questions were subsequently formulated:

- How can learning analytics and social network analysis reliably predict students’ performance in collaborative learning environments using contextual, theory-guided indicators?
- How can social network analysis be used to analyze online collaborative learning, guide a data-driven intervention, and evaluate it?
4. Definition and taxonomy

Learning analytics (LA) is a relatively new and rapidly expanding research area. The field was formally recognized following its first widely known definition during the inaugural International Conference on Learning Analytics & Knowledge (LAK) in 2011 (Siemens, 2013). The definition states that learning analytics is the “measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs” (Siemens, 2013, pp. 1382). Other definitions also exist, for instance, Elias (2011), has proposed that it constitutes “the selection, capture, and processing of data that will be helpful for students and instructors at the course or individual level.” (pp. 4). Cooper (2012) have defined LA as “the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data” (pp. 4). Erik Duval’s definition is arguably the simplest even though it is also quite comprehensive, claiming “learning analytics is about collecting traces that learners leave behind and using those traces to improve learning.” (Baker, Shum, Duval, Stamper, & Wiley, 2012, pp. 21). Although the definition of LAK is contested, and many have called for its revision, it seems to be the most used and endorsed concept by researchers in the field. Academic analytics and educational data mining (EDM) would appear to be related research endeavors, albeit with a different focus and stakeholders involved. Academic analytics uses business intelligence and data analysis at the institutional level to support decision making and hence help improve the operating efficiency of organizational processes (bin Mat, Buniyamin, Arsad, & Kassim,
At the institutional level, analytics are used to assess programs and offer inter-school, intra-school and institutional benchmarking. Its insights may enable institutions to understand their strengths and challenges and foster their growth. Academic analytics has the potential to boost the retention rates of enrolled students as well as the rates of completion of degrees and consequently graduation rates (bin Mat et al., 2013; Campbell et al., 2007; Ellis, 2013; Ferguson, 2012b; Siemens, 2013; Sin & Muthu, 2015). The term was popular during the early days of learning analytics, but over time lost ground to learning analytics, which is often used interchangeably.

EDM is a multidisciplinary research endeavor that uses data mining techniques to analyze the data collected or generated during teaching and learning (Bin Mat et al., 2014; Dutt, Ismail, & Herawan, 2017). Methods such as artificial intelligence, text analytics, information retrieval, and cognitive modeling are commonly used to find solutions to educational issues (Dutt et al., 2017; Slater, Joksimović, Kovanovic, Baker, & Gasevic, 2017). EDM has been defined by the International Educational Data Mining Society as “an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to better understand students, and the settings which they learn in” (Siemens & Baker, 2012, pp. 252).

EDM has been used to predict students’ future learning behaviors, create, or improve content models and instructional scenarios, as well as to study the impact of different learning support methods. Other uses include the personalization of learning, adaptive feedback and customized recommendations (Bienkowski, Feng, & Means, 2012; Falakmasir & Habibi, 2010; Papamitsiou & Economides, 2014; Peña-Ayala, 2014).

A growing overlap is evident among academic analytics, educational data mining and learning analytics as terms, given that the focus of these disciplines is the analysis of computer-generated data sets in the field of education research with the objective of enhancing the educational process. Researchers often use the same methods under different labels. It is possible that the future will
deepen or otherwise eliminate redundant distinctions in order to provide a clearer taxonomy (Siemens & Baker, 2012; Y. Y. Wong, 2016).
5. Theory and philosophy

As was described in the problem and motivation chapter, linking theory to context constituted a primary driver for this thesis. As such, this chapter is dedicated to discussing the subject in greater detail. First, I will address the debate surrounding the role of empiricism in analytics research and whether it represents a new paradigm or should be part of a continuum. A discussion of the subject of learning inference will be presented thereafter. I subsequently offer an overview of the position of theory in learning analytics research, followed by an overview of learning theories related to my thesis, and how recent research in the field is operationalizing theoretical models. The final section of the chapter discusses the theoretical underpinnings of my research, with a description of how I operationalized each theoretical model to answer the research questions.

5.1 A new paradigm?

Research in the field of learning analytics has been driven by the availability of unprecedented volumes of data generated via students’ interactions with technology (H. Chen et al., 2012; Siemens, 2013). Another driving factor has been the thriving of instrumented data science methods that have leveraged opportunities to address challenging, real-life problems and deliver meaningful insights at an unprecedented scale (G. Wang et al., 2016). Education has followed in the footsteps of rewarding industry models that have harnessed big data to boost their competitive advantage. The disrup-
...ive nature of big data analytics in the industry has led to a substantial debate regarding the radical paradigm shift that analytics has brought to the ways in which knowledge is defined and the methods conducted (Boyd & Crawford, 2012; Kitchin, 2014; A. F. Wise & Shaffer, 2015).

In traditional hypothesis-driven research, the scientific experiment investigates a preset premise that conforms to a theory or hypothetical model. The results are often explanations, revisions, and reevaluations of prior beliefs that lead to improvements in our understanding and the emergence of new questions (Kitchin, 2014; Siemens & Latour, 2015). Analytics is arguably a new data-driven approach to empiricism that seeks to identify novel patterns or rules through data mining and exploration (Anderson, 2008; Kitchin, 2014; Siemens & Latour, 2015). The proposition is that through the analysis of vast amounts of data, using inductive reasoning and a bottom-up approach, discoveries can be made (Kitchin, 2014). Advocates of the approach claim that researchers no longer need to make educated guesses or worry about constructing a hypothesis since the analysis of full resolution large data sets can enable the automatic discovery of valuable insights (Anderson, 2008; Kitchin, 2014). They further claim that the exploration of big data can lead to the discovery of patterns and correlations that were not expected and therefore might never have been discovered using traditional approaches (Anderson, 2008; Kitchin, 2014). Although these claims are bold, supporting examples have been demonstrated, in which the analysis of big data brought forth new and completely unexpected insights (Anderson, 2008; Boyd & Crawford, 2012; Kitchin, 2014). However, opponents have questioned the validity of this approach, citing numerous examples of spurious correlations resulting from the analysis of big data. Therefore, the risk of misinterpretation cannot be overruled on the basis of data size or the complexity of the algorithm (Boyd & Crawford, 2012).
5.2 Inferring learning from data

Learning is a complex concept that has over the years garnered considerable research across numerous disciplines such as neuroscience, physiology, psychology, education, philosophy, and computer science. Although the concept is commonly used, a consensus of how it should be defined does not exist (A. B. Barron et al., 2015; De Houwer, Barnes-Holmes, & Moors, 2013). Nevertheless, learning may be viewed as any process that leads to a change in capacity, behavior or adaptation due to experience or regularities in environment that cannot be exclusively explained by maturation or aging (Illeris, 2018). Our understanding of the concept of learning continues to evolve. Indeed, whereas the concept was traditionally understood as the acquisition of knowledge and skills, it is now believed to encompass an affective, sociocultural and ontogenetic dimension (A. B. Barron et al., 2015; De Houwer et al., 2013). Given that learning is an evolutionary, dynamic and developmental process that occurs over different spatial and temporal scales from seconds to minutes to lifetime, a valid concept of learning that captures its full complexity appears to be far from realizable (Horvath & Lodge, 2016).

As such, inferring learning from the analytics of students’ activities is confronted by the pursuit for an operative definition of learning (Greller & Drachsler, 2012; Lodge, Alhadad, Lewis, & Gašević, 2017). Using answers from other disciplines that have operationalized data insights has failed to help, as learning cannot be reduced to simple dichotomous possibilities as is true in e-commerce (purchase versus not purchase) or medicine (morbidity and mortality). Higher education institutions tend to measure learning according to performance, a practice that has been widely criticized for only capturing the momentary acquisition of knowledge or skills during or immediately following instruction, while failing to account for the acquisition of learning skills, long-term learning, and behavioral changes (Lodge et al., 2017; Soderstrom & Bjork, 2015).

Due to the challenging nature of defining learning or learning outcome, I have cautiously tried to use assessment and grades as indicators of course performance. Assessment and grades define how
students’ work and drives their learning strategies and approaches to course activities. While it is far from perfect as a measurement for learning, it is understandably aligned with students’ efforts, plans and learning strategies (Joughin, 2009). By cautiously I mean that in my studies, I have used the relative performance of students compared to one another, rather than using grades. Modeling relative achievement levels might represent a more contextually relevant method than setting a cut score, which tends to be arbitrary and judgmental (Kane, 2017).

5.3 Theory and learning analytics research

The early field of learning analytics was traditionally focused on creating predictive deficit models for the detection of underachievers, as well as predicting the indicators of poor performance. These models were not always explicit about theory (Clow, 2013; S. Dawson et al., 2017b; Tempelaar, Rienties, & Giesbers, 2015), and eventually attracted substantial criticism as a result (S. Dawson, Mirriahi, & Gasevic, 2015; A. F. Wise & Shaffer, 2015). The concern is that optimizing the educational system based on metrics that do not measure actual learning might risk optimizing learning away from learning, underscoring the need to build analytics around relevant metrics that are consistent with learning and based on theory (Clow, 2013; S. Dawson et al., 2017b).

The importance of theory in the development of learning analytics cannot be overemphasized; after all, learning analytics is about learning and should contribute to the development of learning sciences. A well-grounded theory would help advance knowledge in the discipline, promote research about crucial problems, and help advance effective instructional practices (B. Chen, 2015; S. Dawson et al., 2015; Reimann, 2016). The theory would guide research about study design, research methods, and the predictors and predictive models that should be included in the analysis. The theory would guide researchers concerning potential confounding factors, subgroupings, and possible data covariates. Moreover, the theory
would offer researchers a conceptual framework to interpret the results and conclusions in order to stimulate insightful action. Most importantly, such a theory would help researchers and stakeholders to generalize findings to broader contexts and population groups (Reimann, 2016; A. F. Wise & Shaffer, 2015). Some attempts at laying the theoretical grounds for the field have been made (Clow, 2012a; Colvin et al., 2015; Gašević, Kovanović, & Joksimović, 2017; Greller & Drachsler, 2012; Reimann, 2016). Nonetheless, substantial work still needs to be done (Gašević et al., 2017; A. F. Wise & Shaffer, 2015).

The widely accepted proposition among the learning analytics community is that if the field cannot realize an immaculate theoretical framework, it should at least aspire to do so (B. Chen, 2015; S. Dawson et al., 2015; Lodge et al., 2017; Reimann, 2016; A. F. Wise & Shaffer, 2015). Among calls to align research to learning theory, recent research has sought to improve data collection methods by embracing proxy indicators of learning that are more aligned with learning theories. Examples include constructivism and social capital theory in the context of collaborative learning (Agudo-Peregrina, Iglesias-Pradas, Conde-González, & Hernández-Garca, 2014; Chiu, 2014; Rizzuto, LeDoux, & Hatala, 2009; Saqr, 2018; Tervakari et al., 2013). In these studies, researchers have included interaction indicators, social network centrality measures of connectedness and information exchange, and collaborative network structure (Agudo-Peregrina et al., 2014; Chiu, 2014; Rizzuto et al., 2009; Saqr, 2018; Tervakari et al., 2013). Another emerging research trend comprises attempts to contextualize the digital traces recorded from students’ online activities to infer their self-regulation patterns and learning strategies (Barnard, Lan, To, Paton, & Lai, 2009; Cicchinelli et al., 2018; Gašević, Jovanović, et al., 2017; Schraw, 2010; Winne, 2017; Winne et al., 2017). Early reports suggest the validity of using these records of online activity as a representation of online self-regulation (Cicchinelli et al., 2018; Winne et al., 2017).
5.4 Theoretical underpinning of my research

Given that the context of this research was collaborative learning, its operationalization of data collection, analysis and inference builds on existing learning theories relevant to the collaborative learning context. The epistemological basis of collaborative learning comprises diverse educational philosophies and rich theoretical roots, “more like an arbor of vines growing in parallel, crossing and intertwining” (Macgregor, 1990, pp. 21). Constructivist learning theory is relevant to this study and the analysis methods implemented because it deals with students’ interactions, dialogue and shared co-construction of knowledge (Fosnot, 2013; Illeris, 2018; Schunk, 2012). Another theory that explores information exchange in networked learning is the connectivism learning theory, which is concerned with learning occurring in networked learning and the exchange of information (Bell, 2010; Siemens, 2004). Furthermore, social capital theory explains how individuals capitalize on their social relations and connections and benefit from their social attributes (Cloete, 2014; Kwon & Adler, 2014).

There is ample empirical evidence that indicates that strategies that aim to facilitate collaborative learning can benefit learning as a whole. For instance, a meta-analysis by Bernard et al. (2009) concluded that by increasing interactions among learners, the teacher or learning content demonstrates an adjusted average 0.38 significant positive effect on performance. Another meta-analysis found that courses designed to enhance collaborative learning and interaction among learners improved students’ engagement significantly and resulted in superior academic achievement (Borokhovski, Bernard, Tamim, Schmid, & Sokolovskaya, 2016).

Research has also illustrated that promoting certain practices can enhance knowledge construction in collaborative settings. Such practices include facilitating discourse (Garrison & Arbaugh, 2007; Lambert & Fisher, 2013) promotive interaction (Gillies, 2016; Kreijns et al., 2003; Resta & Laferrière, 2007), argumentation (Noroozi et al., 2012; Nussbaum, 2008; Wecker & Fischer, 2014), positive interdependence (Abrami, Bernard, Bures, Borokhovski, & Tamim; Resta & Laferrière, 2007) and individual
accountability (Gillies, 2016; Kreijns et al., 2003). In this context, teaching presence is an important determinant of a student’s sense of community, satisfaction, and perceived learning. (Bernard et al., 2009; Garrison & Arbaugh, 2007; Kreijns et al., 2003; Webb, 2009). Furthermore, interaction with teachers has been found to positively influence student academic achievement (Bernard et al., 2009) and predicts good performance (Cho et al., 2007; Joksimović et al., 2015; Romero & Ventura, 2013).

Two constructs are of prime importance here. First, the concept of engagement, as online learning requires students to be engaged, actively participating, and demonstrate considerable motivation, and persistence. Second, self-regulation, as online learning is a demanding process and requires successful strategies and regulation (Burnette et al., 2013; Reschly & Christenson, 2012; Winne, 2011). Each of these theories and constructs will be discussed in greater detail in terms of their relevance and a description of how they were operationalized.

5.4.1 Self-regulation

Self-regulation (SRL) is an effortful process that requires substantial cognitive resources to set goals, monitor progress, adopt learning strategies and acquire academic skills (Broadbent & Poon, 2015; Grossman & McDonald, 2008; Winne, 2011). The process of SRL can be viewed in terms of four loosely sequenced phases. In the first phase, the student researches aspects of the task in terms of constraints and affordances. Some of these aspects might be external, such as time allocated, access to learning resources and availability of support. Other factors might be internal, such as previous knowledge, motivation, available strategies to handle the task and estimated gains or penalties if not performed. The second phase involves planning and goal-setting based on the earlier task research. In the third phase, the learner engages in performing the task and evaluates his or her own performance. The feedback obtained enables the learner to adjust the approach. In the fourth phase, the learner evaluates the whole process and makes large-
scale adaptations that may require some of the previous phases to be repeated, or asking for help (Winne, 2011).

SRL may be defined as “an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment” (Pintrich, 2000, pp. 453). The definition conveys several intertwined elements. The first is the active, purposeful engagement of the learner in learning. The second element is the learner’s goal-directed focus on achievement. The third is the use of proactive learning strategies to regulate cognition and enhance performance. The fourth element is contextual and refers to the environment and how it might be either supportive or obstructive. Given that the process of self-regulation is taxing, a student must be motivated to perform, and be supported when this motivation is lacking (Pintrich, 2000; Winne, 2011; Zimmerman, 1990).

Various protocols have been developed to measure self-regulation. As a metacognitive process, it contains both an aptitude aspect and a temporal aspect. In the classroom, aptitude measures the metacognitive approach to learning, the learning strategies, and how a learner monitors his or her approach. The temporal aspects measure the events as they occur over a timeline of events (Winne, 2011; Zimmerman, 2008). As students work in online environments, their activities are recorded in the form of logs with time-stamps. The logs are observable indicators created by students during their engagement with the learning task. Researchers can contextualize these traces to build a description of the event (Baggetun & Wasson, 2006; Schraw, 2010; Winne et al., 2017). Certain self-regulation assumptions can be made according to the type of the trace recorded. For instance, access to course objectives, schedules, and requirements can be described as traces of planning (Cicchinelli et al., 2018; Winne et al., 2017).

5.4.1.1 Temporality

The time-stamps recorded for each online activity define the temporal aspect of self-regulation. The construct of time and temporality of students’ activities are central concepts in self-regulation.
learning theory (B. Chen et al., 2018; Winne et al., 2017). Students who efficiently use their time to manage their learning strategies are expected to perform better than their counterparts who do not. A large volume of empirical evidence seems to support this proposition (Broadbent & Poon, 2015; Burnette et al., 2013; Job, Walton, Bernecker, & Dweck, 2015; Wolters & Hussain, 2015). Procrastination in performing learning tasks provides a reliable indicator of the failure of self-regulation and consequently poor performance (Steel, 2007; You, 2015).

Self-regulation was operationalized as a theoretical basis for my first article in contextualizing data collection (Saqr, Fors, & Tedre, 2017). Data regarding early course login (before the start date), course orientation materials, course schedule and course booklet containing the intended learning objectives, course requirements and assessment methods were all used as indicators of task orientation and planning. Data regarding course view, logins, forum postings, access to course materials, and announcements were operationalized as indicators for engagement with the tasks. Data regarding formative assessment and access to grades were used as indicators of self-monitoring and self-evaluation. In order to reflect the continuous and purposeful engagement and motivation of the learner over the duration of the course, I calculated the engagement sub-indicators on a weekly basis; students who were constantly engaged received higher scores. Other aspects of self-regulation were also inferred from the collected data, and the proactive approach and goal-directedness were inferred from traces of early course access, regularity of course access, regularity of self-assessment, and access to course assessment materials and grades. Furthermore, time was an essential factor in the calculation of engagement indicators (both early participation and regularity).

The temporality aspect of self-regulation was further studied in detail in article 5 (Saqr, Nouri, & Fors, 2018), in which I mapped patterns of participation in online collaborative learning with emphasis on early contribution during different time constructs. An early pattern reflected proper time management skills, motivation to perform the required task and proactive engagement. Patterns of engagement with learning tasks even during college working hours...
were also used as indicators for regular and sustainable engagement.

5.4.2 Engagement

A considerable proportion of robust empirical evidence supports the assertions that engagement is a reliable indicator of learning, such that its value is “no longer questioned” (Reschly & Christenson, 2012, pp. 4; P. Trowler & Trowler, 2010). In addition, strategies that aim to increase engagement are likely to result in a desirable outcome for all students. Therefore, engagement has attracted a vast volume of research in higher education as well as in other fields (Reschly & Christenson, 2012; P. Trowler & Trowler, 2010; V. Trowler, 2010). Although there is considerable consensus regarding the significance of the construct in education, definition, elements and measurement techniques remain contested (Kahu, 2013). Nevertheless, elements of engagement can be inferred from students’ behavior in online environments (Ma, Han, Yang, & Cheng, 2015; Santos, Klerkx, Duval, Gago, & Rodríguez, 2014).

In study 1, I attempted to extrapolate online activities as proxy indicators for aspects of engagement (Saqr et al., 2017). For instance, commitment, involvement in learning, and rule-following were inferred from the regularity of logging to the LMS and followed up with course announcements and regular access to course materials. Students’ investment of quality efforts in learning, along with self-regulation of efforts, were inferred from the sustainability of learning efforts, time-on-task and purposeful access to formative assessments.

Engagement has a social aspect, and so interaction with peers and teachers, and participation in information co-construction, can all be deemed signs of engagement. Similarly, disengagement in participatory and collaborative environments could be manifested as a lack of interaction and social isolation. Therefore, students’ course interactions were operationalized as aspects of social engagement in the final three articles using interaction indicators (Saqr et al., 2017; Saqr, Fors, & Nouri, 2018; Saqr, Nouri, & Fors, 2018).
5.4.3 The theoretical basis of collaborative learning

5.4.3.1 Constructivist learning theories
Based on Jean Piaget’s ideas, constructivist theorists view learning as a contextual, goal-oriented and self-regulated process. The process is cumulative through the utilization of previous background knowledge along with current experiences to construct novel concepts or acquire new information. Learners are expected to take responsibility for constructing their own knowledge. Teachers are viewed as facilitators who stimulate learners’ curious investigation, motivate them to reflect, raise questions, explore answers, justify their own perspectives and debate with others. Learning occurs as a result of self-organization, reflective abstraction, dialogue with the community and challenging misunderstandings and erroneous conceptions (Fosnot, 2013; Illeris, 2018; Liu & Matthews, 2005; Schunk, 2012).

Social constructivists emphasize the dialectic aspect of learning. Knowledge is socially (rather than individually) created by groups of learners, leading to the active co-construction of meaning. Learning is shaped by the social and physical environment as well as in dialogue among peers. Vygotsky’s sociocultural theory, although constructivist in nature, emphasizes interpersonal influences and the social context as mediators for knowledge construction and learning (Liu & Matthews, 2005; Vygotsky, 1987). Vygotsky views peer interaction as a very useful tool for the development of knowledge and skills.

The zone of proximal development (ZPD) is a concept that compares what can be achieved by working with peers with that achieved by working alone. Vygotsky postulated that when a learner who lies inside the ZPD for a specific task is provided with the adequate support and scaffolding, the learner becomes empowered to complete the job efficiently (Liu & Matthews, 2005; Poddolskiy, 2012; Vygotsky, 1987). Thus, social constructivism emphasizes the role of group work and interactions with the commu-
nity to enable students to learn and build upon their ZPD. The process may be enhanced through scaffolding by a teacher or a knowledgeable instructor (Podolskiy, 2012; Vygotsky, 1987).

5.4.3.3 Connectivism learning theory
Connectivism is a learning theory that aims to address learning in the digital age and accounts for the societal changes where learning as a process is no longer an individual activity (Bell, 2010; Kop & Hill, 2008; Siemens, 2004). According to connectivism, knowledge, and learning exist in the multiplicity of opinions; learning occurs as a process of connecting and interacting with the sources of information. Moreover, connectedness is an important skill for acquiring information and remaining up-to-date in the long term. The theory of connectivism has important implications. First, it emphasizes skills for obtaining and evaluating information and staying up-to-date. Second, it underscores the value of collaboration in a functional community of knowledge and information sharing. Third, it lays the ground to develop a greater understanding of information exchange using modern analytic techniques such as social network analytics (Bell, 2010; Siemens, 2008; Siemens, 2004).

5.4.3.4 Social capital theory
According to social capital theory, the benefits one attains from having connections, relationships and position in the social structure are collectively known as social capital. Benefits include access to resources and opportunities as well as psychological and emotional support, enabling individuals to overcome challenging situations (Burt, 2002, 2004, 2015, 2017; Dika & Singh, 2002; Kwon & Adler, 2014). Moreover, relationships may provide happiness and motivation and act as a driver for the accomplishment of challenging tasks (Kwon & Adler, 2014; A. J. Martin & Dowson, 2009). Individuals who have ties to others in different groups have access to diverse opinions and fresh perspectives; these ties –although weak – provide a competitive advantage of being the gateway to external novel ideas, and offer a higher return of brokerage effort. Acting as
a broker between unconnected groups and bridging the communication gaps (social holes) brings another form of social capital: the social capital of structural holes (Burt, 2002, 2004, 2015, 2017).

5.5 My approach to the study of collaborative learning

For the study of collaborative learning, I used interaction indicators (numbers of posts, reads, views, replies) in study 1 (Saqr et al., 2017) and interaction indicators as well as social network analysis in the four subsequent articles 2, 3, 4, 5 (Saqr et al., 2017; Saqr, Fors, & Nouri, 2018; Saqr, Fors, & Tedre, 2018; Saqr, Fors, Tedre, & Nouri, 2018; Saqr, Nouri, & Fors, 2018). Given that SNA offers a quantification of interaction indicators, the term SNA will be used throughout this thesis to denote both interaction indicators and social network indicators. SNA may be a convenient as well as practical choice for the study of collaborative learning. SNA visual analysis and mathematical indicators (centrality measures and centralization) are highly aligned with participatory settings and are deeply grounded in collaborative learning theories. For the sake of operationalizing the SNA in a collaborative learning environment, I performed an extensive review of the literature to develop a representative framework that can capture the gamut of relationships, interactions and structural properties on individual collaborators and groups (Cela et al., 2014; Dado & Bodemer, 2017; Sie, Ullmann, Rajagopal, Cela, Bitter, & Sloep, 2012). Six main facets can be revealed by SNA, and are highlighted here with their theoretical basis:

- The visual mapping of communicational activities and the characteristics of the social structure (constructivist learning and connectivism);
- The quantity, direction, and significance of interactions (constructivist learning and social capital theories);
- Position in information transfer (connectivism, constructivist learning, and social capital theories);
• Connectedness and relationships (connectivism and social capital theories);
• Role in collaboration (Constructivist learning and community of inquiry);
• Community structure such as groups and classes (constructivism).

From the constructivist view of learning, SNA offers a bird’s-eye view of learners’ interactions, highlighting the participatory status of the group and the roles played by the individual participants. Individual participants may play an active role in the group, such as leaders who drive and stimulate others to collaborate; coordinators who coordinate and moderate interactions; or isolated who rarely participate. In study 2, mapping the interactions revealed a non-collaborative pattern of interactions, where information giving and receiving networks were dominated by the teacher. These findings stimulated the design of study 3 to explore an intervention solution (Saqr, Fors, & Tedre, 2018). In study 3, I operationalized different SNA centrality measures and interaction analysis indicators to map the roles of collaborators, revealing a non-participatory pattern on the side of the students and a dominating teacher. These non-collaborative patterns were a target of intervention that helped bring a meaningful improvement (Saqr, Fors, Tedre, et al., 2018). In study 4, the SNA centrality measures and roles played by the collaborators in the form of the interaction framework to predict performance and create a reproducible model proved reasonably reliable (Saqr, Fors, & Nouri, 2018).

Furthermore, the role and position in information exchange were indicated by SNA centrality measures. Three centrality measures are of particular relevance to information transfer: betweenness centrality (the times a user helped connect or broker information), closeness centrality (the ease by which a collaborator can be reached), and information centrality (the amount of information passing through the participant). A preferential role in the exchange or brokerage of information is a valuable source of social capital as it has greater access to resources as well as diverse perspective (Burt, 2014; Cloete, 2014; Dika & Singh, 2002; Isba, Woolf, &
Hanneman, 2017; Mattar, 2010). I used the measures of information transfer to monitor students’ roles in study 2, which revealed little participation on the side of students and required attention (Saqr, Fors, & Tedre, 2018). In article 3, the centrality measures were used to monitor improvements after introducing remedial measures (Saqr, Fors, Tedre, et al., 2018), and to understand the process of interaction as well as to predict performance in study 4 (Saqr, Fors, & Nouri, 2018).

From a constructivist and social capital learning theory perspective, SNA maps the connectedness of a learner, the strength of the ego network (own network) and the networks of resources, and helps identify different types of privileged access (Burt, 2007; Dika & Singh, 2002). The concept of connectedness and significance of relationships were operationalized using various centrality measures in my studies, in addition to the information transfer centrality measures mentioned earlier. In addition, prestige measures of range of connections (domain prestige), strength of peers (Eigen centrality) and rank prestige (power and performance of connections) were utilized. The centrality measures of connectedness were positively correlated with performance in studies 2, and 4, especially the centrality measures that were representative of strength of connections to high-performing and well-connected students (Saqr, 2018; Saqr, Fors, & Nouri, 2018; Saqr, Fors, & Tedre, 2018)

It is worth noting here that a single centrality measure is not enough as a representative for a specific concept. In my research, I used multiple centrality measures along with visual representation to infer the role played by collaborators, taking into account the situational context. As an example from study 3, for a student to be labeled as a coordinator, the student needed to have high to moderate out-degree centrality, betweenness centrality, indegree centrality and information centrality as well as average domain prestige (Saqr, Fors, Tedre, et al., 2018). A more detailed discussion of centrality measures and centralization is presented in Section 6.3.4.
6. The learning analytics process

The importance of the process of learning analytics stems from the fact that it provides the methodological basis of how learning analytics research is conceptualized, conducted, implemented, and evaluated. In this chapter, I introduce the process of learning analytics, followed by a detailed account of each stage, with examples where possible. Special emphasis is placed on data capture, analysis, and interpretation. The subject of social network analysis, a central issue in this thesis, is introduced in greater detail, covering the general concept, terminology, visualization, and mathematical analysis. An example of a social network is also added and analyzed mathematically and visually.

The process of learning analytics can be conceptualized as a sequence of interrelated steps or stages (see Figure 1). These comprise: 1) capturing data generated by learners; 2) preprocessing the data for analysis; 3) statistical and machine-learning methods for analysis; and 4) the insights are used to apply data-driven interventions or aid the decision-making process. Given that learning analytics is an iterative process, feedback can help improve and refine the methods (H. Chen et al., 2012; Clow, 2012b; Siemens, 2013).
6.1 Data capture

Every activity a student makes online can be recorded and used for learning analytics, including visiting the online management system in order to access course resources, searching for information in the online library, taking an online examination, writing an assignment, or communicating with a colleague. Hypothetically, there are no limits to what can be collected or used in learning analytics research as long as it conforms to ethical and privacy standards (Leitner et al., 2017; Sergis, Sampson, Leitner, Khalil, & Ebner, 2017).

The scope of learning analytics data collection is rapidly expanding in both breadth and scale due to improved learning technologies and increased awareness among stakeholders regarding the importance of learning analytics. The most common sources are the
logs that are recorded by learning management systems (LMS). Examples of data derived from LMS include frequency of logins, page views, hits on learning resources, chat records, formative assessments, time spent on tasks and work on assignments. These data may be extracted using analysis of the LMS log system, custom plugins that extract specific parameters such as time-on-task, or by using LMS built-in analytics dashboards (where applicable) (F. Martin & Ndoye, 2016; Misiejuk & Wasson, 2017; Na & Tasir, 2017; Sergis et al., 2017).

The gathered data may also include registration and demographic data such as age, gender, language proficiency, prior grades, socioeconomic status, scholarships, and residence (Arnold & Pistilli, 2012a, 2012b; Jayaprakash, Moody, Lauria, Regan, & Baron, 2014; Kuzilek et al., 2015; Rienties et al., 2016). In collaborative learning environments, the collected data encompass interaction indicators such as the number, size, and frequency of forum posting, relationships, connections, and learners’ positions. Interaction parameters may be captured via social network analysis techniques, while the content of interactions may be analyzed using data mining and content analytics (Chiu, 2014; Macfadyen & Dawson, 2010; Misiejuk & Wasson, 2017; Saqr, 2018; Shahiri et al., 2015; Tervakari et al., 2013).

Self-reported surveys concerning learning disposition, motivation, and engagement have also been used to account for affective, behavioral, and cognitive dimensions (S. P. Dawson, Macfadyen, & Lockyer, 2009; Shum & Crick, 2012; Tempelaar et al., 2015, 2018), an approach labeled dispositional learning analytics (DLA). Advocates of DLA argue that these data are more actionable and can trigger more informed data-driven interventions (Shum & Crick, 2012; Tempelaar et al., 2018). Temporal profiling of online activity using time series modeling has also been examined as a means of accounting for temporal patterns of variations in students’ activity (Brooks, Thompson, & Teasley, 2015; Goda et al., 2015). Multimodal learning analytics extends recording to neurophysiological, cognitive, physical, and digital data (Andrade & Danish, 2016; Blikstein, 2013; Blikstein & Worsley, 2016). Multimodal capture methods can potentially gather large volumes of fine-
grained information that extends beyond the limited scope of traditional learning analytics. Advocates of this approach postulate that such high-resolution, fine-grained physiological data can offer an unprecedented breadth of information about learning (Blikstein, 2013).

Whereas quantitative data collection methods capture the number of times a student uses a specific learning resource, they do not cover the rationale, motives or experiences of the users or stakeholders. The answers to such questions lie within the realm of qualitative research. Qualitative data collection methods include structured interviews and focus group discussions with students and teachers regarding their experience of the visualizations and insights through dashboards. Other common methods include surveys, open-ended questions, video recordings and analysis of reflections and feedback (Ali, Hatala, Gašević, & Jovanović, 2012; Knight, Tech, States, & Novoselich, 2016; Mazza & Dimitrova, 2004a; Ott, Robins, Haden, & Shephard, 2015; Silius, Tervakari, & Kailanto, 2013; A. Wise, Zhao, & Hausknecht, 2014).

### 6.2 Preprocessing and Preparation

At this stage, several processes are performed to prepare the data for analysis. The data are gathered from different sources and then compiled into a single data set. Thereafter, the cleaning of incorrect, corrupt or mislabeled data is performed to avoid erroneous conclusions. The final step of processing is data wrangling or transforming into a suitable format for statistical analysis. The process might entail text parsing, visualizations, aggregation, and calculations of certain metrics or feature engineering. The result of the process is a dataset ready for analysis. Mature learning analytics systems automate this process of integration and feature preparation through programmable modules (Mattingly, Rice, & Berge, 2012; Siemens, 2013; Slater et al., 2017).
6.3 Analysis & Interpretation

In tandem with the quest for relevant data sources, the exploration of data analysis methods has been relentless. Several techniques and methods have been used, with common strategies including statistical analysis and predictive modeling, clustering, content analytics and social network analysis. These methods will be reviewed with special emphasis on predictive modeling and social network analysis. Social network analysis will be covered in a separate chapter as it constitutes a principal method in the thesis (Misiejuk & Wasson, 2017; Nunn et al., 2016).

6.3.1 Predictive modeling

The predictive modeling of academic achievement may be the most common objective of modeling. The objective is the early identification of students at risk of low achievement or dropping-out, in order to offer them a proper intervention that might benefit their progress (Gardner & Brooks, 2018; Leitner et al., 2017). Other less common goals are to forecast course enrollments, predict the potential return on investment in intervention strategies, and to alert students’ advisors to potential disengaged students. Predictive modeling mainly uses data and digital traces recorded by learning management systems as well as other information about students’ demographic and registration data (bin Mat et al., 2014; Leitner et al., 2017; Nunn et al., 2016).

Early studies used simple bivariate correlations to identify factors that correlate with students’ greater performance (A. Y. Wang & Newlin, 2002a, 2002b). Later, regression models became increasingly widely used (Gašević et al., 2016; Macfadyen & Dawson, 2010; Romero & Ventura, 2013; Tempelaar et al., 2015). Recently, machine learning and artificial intelligence models have become more common such as clustering algorithms (Pardo, Han, & Ellis, 2017), classification algorithms (Romero & Ventura, 2013), decision trees (Spoon et al., 2016) and k-nearest neighbors (Kuzilek, Hlosta, Herrmannova, Zdrahal, Vaclavek, & Wolff, 2015).
The accuracy of predictive models has steadily improved with the expansion of data sources and recent developments in statistical and machine learning methods. Furthermore, different validation and quality assurance methods are being introduced to boost the reliability of prediction results (Gardner & Brooks, 2018).

Two types of statistical predictive modeling are frequently used. The first type (explanatory model) uses causal explanation by means of statistical modeling to build and test a predefined hypothesis. In this method, only the variables thought to cause a certain outcome or supported by a theoretical model are used as predictors (Brooks & Thompson, 2017; Shmueli, 2010). For example, a researcher interested in studying self-regulatory behavior would only include variables of task orientation, task engagement, evaluation, and adjustment as features within a predictive model for self-regulation. Recently, explanatory modeling has gained traction in learning analytics research driven by the growing interest in aligning methods with theory. The second type is predictive modeling, in which no causal assumptions are made. The main goal in this case is to use the available data to early predict a future event or an outcome such as passing a course or completing a degree (Brooks & Thompson, 2017; Shmueli, 2010). Predictive modeling requires no theoretical model and is not concerned with explaining the phenomena under study. Rather, the objective of predictive modeling is to flag a future event such as a student possibility of failing a course or dropping out (Brooks & Thompson, 2017; Shmueli, 2010). By flagging a future event, educators can work to reduce the possibility of it occurring, through remedial and supportive strategies.

6.3.2 Clustering

Clustering is concerned with identifying groups that share similar characteristics based on subsets of data points. Clustering is a widely used method with a varied spectrum of applications that includes the organization of course materials and assignments at different levels of difficulty, the automatic classification of different categories of course feedback, and provision of formative feedback
to different subgroups of students. Clustering is also used to offer personalized recommendations for subsets of students based on shared goals or performance. Personalized recommendations may be helpful resources according to choices of similar students or strategy recommendations according to intended goals (Avella, Kebritchi, Nunn, & Kanai, 2016; Romero & Ventura, 2010).

6.3.3 Content analytics

The focus of content analytics is the content, which includes the content of learning resources such as books, handouts, and web resources, as well as learning products such as discussion forums, assessments, assignments, and the content of social interactions. Content analytics may be defined as “automated methods for examining, evaluating, indexing, filtering, recommending, and visualizing different forms of digital learning content, regardless of its producer (e.g., instructor, student) with the goal of understanding learning activities and improving educational practice and research” (Kovanovic, Joksimovic, Gasevic, Hatala, & Siemens, 2017, pp. 78).

The field of content analytics includes a wide range of applications such as writing analytics, discourse analytics, and assessment analytics. Content analytics has been used for the creation of recommendation systems, provision of personalized feedback, knowledge and topic discovery in forum discussion, and analysis of learning outcomes (Kovanovic et al., 2017; Misiejuk & Wasson, 2017; Romero & Ventura, 2010).
6.3.4 Social network analysis

Social network analysis is an area of data analysis concerned with the mapping of social structures and relationships and the interactions among actors within these social structures (Borgatti et al., 2009; Burt et al., 2013). Through SNA methods, researchers can study the effects and influences of the structures and an individual’s positions and ties. SNA has been used across a wide array of fields and purposes, and new aspects continue to emerge. For instance, in economics, SNA has been used to analyze organizations’ communicative activities, management efficiency, and decision-making processes (Burt et al., 2013; Fang et al., 2015). In medicine, SNA is rapidly growing to include network analysis of genes, disease propagation, brain communicative activities, and enzyme interactions (Barabasi, Gulbahce, & Loscalzo, 2011; Martínez-López, Perez, & Sánchez-Vizcaíno, 2009). In academic authorship, it has been used to study citation analysis, publication trends, and scientific collaboration. The list also includes political, economic, historical, social and criminal sciences, among numerous other fields (Mingers & Leydesdorff, 2015).

6.3.4.1 Basics of social network analysis

The social structure is always referred to as a network, and a network is a group of actors who are connected through a relation. For instance, a group of school friends forms a network defined by the friendship relationship. The actors are usually referred to as nodes or vertices; nodes are members of the network, and ties (commonly referred to as edges) are the links that connect the nodes to form a social structure (a network). Actors can comprise humans, organizations, countries and any entity that can be grouped in terms of a relationship. Edges can also be diverse, with examples including marriage, kinship, friendship, business and any relationship that can link two entities (Borgatti et al., 2009; Burt et al., 2013). In this thesis, we will focus on human interactions and particularly learners’ interactions in academic environments. Networks are often analyzed via two methods: visual and mathematical analysis.
6.3.4.2 Visual analysis of social networks

The visual analysis of social networks displays relationships among network actors in terms of a graph, commonly known as a sociogram. In sociographs, nodes are displayed as circles, and edges as lines. In directed networks (where the target and the source of interaction are defined), the line is rendered as an arrow from the source to the target. The visual analysis offers an overview of the configuration of the network and the relationships among the actors. It can illustrate actors who are strongly connected, isolated, or who possess an advantageous position such as brokers who bridge communications. The possibility of summing all interactions into one graph can be helpful and informative (Borgatti et al., 2009; Burt et al., 2013).

In order to demonstrate the basics of social networks, I present in Figure 2 a group of 10 classmates who form a group of friends. Each pupil is depicted as a circle (node) and each relationship as an arrow (edge), starting from the source to the target. The whole group is referred to as a network and is visualized as a graph or sociogram. The network is based on nomination, for instance Sofia considers Jeanette, Sara, and Carmen as friends, and so there are two arrows representing this relationship, starting from Sofia. Similarly, two pupils (Sara and Janette) consider Sofia a friend. This visual representation (sociogram) is configured to show the node’s size proportional to the number of ties, hence we can see that Sara has the largest size and most connections. It is also configured to show pupils who are connecting and lying in between others in blue.
6.3.4.3 Mathematical analysis of social networks

Networks can also be mathematically characterized to quantify relationships, connectedness and actors’ importance in the network. Mathematical analysis can describe the whole network, the nodes, or the ties (Borgatti et al., 2009; Golbeck, 2013). Examples of network parameters are size, the number of nodes in the network (this is 10 in the classmates’ network presented in Figure 2), or average
degree (the average number of edges in the network). Density, which is the total number of interactions relative to the maximum possible (0.56 in the classmates’ network), reflects the degree of cohesion of a network.

At the individual level, nodes are characterized by what is known as centrality measures (prestige measures). Centrality is a mathematical expression of the node’s role, importance (prestige), connectedness, or position. Given that nodes may play different roles and that importance may be viewed differently according to the situation, different centrality measures exist that describe different roles. The centrality measures most relevant to the collaborative context will be discussed here with examples from the classmates’ network (Golbeck, 2013).

The quantity of ties

- Indegree centrality represents the total number of edges or links an actor receives. In the classmates’ network, it was the number of nominations by others as friends. Nodes with a high indegree centrality are considered popular, as influencers or leaders. For instance, in a social network such as Twitter, leaders and media stars have a large number of followers and are often seen as social influencers (Golbeck, 2013).
- Outdegree centrality represents the number of outgoing edges or links from an actor. In the classmates’ network, this was the number of nominations as own friends. Outdegree centrality is a measure of activity, sociability, and gregariousness (Golbeck, 2013).
- Degree centrality is the sum of both outdegree and indegree centrality measure and is used to quantify all of an actor’s connections or interactions (Golbeck, 2013; Rochat, 2009).

Role in relay brokerage of information

- Betweenness centrality represents the number of times an actor has been in between two other actors and hence connecting them. Actors with high betweenness centrality have an
important role in facilitating communications, connecting others and moderating interactions, and enjoy access to diverse perspectives. High scores of betweenness centrality are deemed brokerage social capital (Golbeck, 2013; Salter-Townshend, White, Gollini, & Murphy, 2012; Golbeck, 2013). In the classmate's network, Sara and Michael had the highest betweenness centralities, and so were most important in connecting others.

- **Information centrality** measures the importance of a node in information flow through the network. It is defined as the relative disruption of traffic through the network if the actor were to be removed, and as such it represents a measure of cohesion. Sara had the highest information centrality in the classmates’ network and was the most important person in the passage of information in the network (Latora & Marchiori, 2007).

- **Closeness centrality** measures how reachable an individual is to all other collaborators and members of the network (Golbeck, 2013; Rochat, 2009). Michael and Janette were the easiest to reach in the classmates’ network.

**Connectedness**

- **Eigenvector centrality** measures the connectedness of an individual and his or her network connectedness. A person connected to well-connected actors would demonstrate higher Eigenvector values (Golbeck, 2013). Sara and Michael were the most connected to a powerful network in the classmates’ network.

In order to demonstrate how centrality measures work, I calculated the centrality measures for the classmates’ network displayed in Figure 2. Table 1 displays the centrality measures of each person. Sara has the most nominations as a friend (indegree), and so she is the most popular pupil. Michael and Janette have nominated the highest number of friends, and so they are the most social. However, Michael has the highest degree and largest network of friends. Janette is the easiest to reach (closeness) by any pupil in the network, which may enable her to acquire information more
quickly than the others. Although Sara does not have the most connections, she is connected to the most powerful in the network (Eigen centrality), connects most people together (betweenness centrality), and most information passes through her. The interpretation of centrality measures depends on the situation. In the case of information exchange, Sara and Janette have better positions. However, in the case of spreading infection, they might be the most vulnerable, whereas Peter may be the safest. As such, isolating Sara and Michael might be the most effective solution to prevent the spread of infection to other classmates.

Table 1: Centrality measures of the classmates’ network

<table>
<thead>
<tr>
<th>Node</th>
<th>Indegree</th>
<th>Outdegree</th>
<th>Degree</th>
<th>Closeness</th>
<th>Betweenness</th>
<th>Eigen</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carmen</td>
<td>7</td>
<td>4</td>
<td>11</td>
<td>0.50</td>
<td>0.06</td>
<td>2.14</td>
<td>2.14</td>
</tr>
<tr>
<td>Layla</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>0.57</td>
<td>0.06</td>
<td>1.86</td>
<td>1.86</td>
</tr>
<tr>
<td>Carl</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0.57</td>
<td>0.02</td>
<td>1.43</td>
<td>1.43</td>
</tr>
<tr>
<td>Sofia</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>0.62</td>
<td>0.04</td>
<td>1.84</td>
<td>1.84</td>
</tr>
<tr>
<td>Sara</td>
<td>8</td>
<td>4</td>
<td>12</td>
<td>0.67</td>
<td>0.21</td>
<td>2.78</td>
<td>2.78</td>
</tr>
<tr>
<td>Michael</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>0.73</td>
<td>0.19</td>
<td>2.41</td>
<td>2.41</td>
</tr>
<tr>
<td>Janette</td>
<td>2</td>
<td>8</td>
<td>10</td>
<td>0.80</td>
<td>0.05</td>
<td>2.59</td>
<td>2.59</td>
</tr>
<tr>
<td>Oscar</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>0.67</td>
<td>0.05</td>
<td>2.39</td>
<td>2.39</td>
</tr>
<tr>
<td>Peter</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0.00</td>
<td>0.00</td>
<td>1.84</td>
<td>1.84</td>
</tr>
</tbody>
</table>
6.4 Insightful action

The ultimate goal of learning analytics is to offer students, teachers, and stakeholders actionable insights. A common way of presenting insights to students and raising their awareness is to use a web panel (dashboard) that provides information regarding activities in the form of graphs and visualizations. Prompts and personalized recommendations can be delivered to the dashboard or via customized messages or notifications. Dashboards (see Figure 3) offer students a visualization of their learning activities, the level of mastery of learning goals, and comparison with peers, in order to enhance their awareness and reflection and to track learning progression (Atif, 2013; Bodily & Verbert, 2017; Duval, 2011; Nussbaumer et al., 2012; Verbert et al., 2014).

Figure 3: An LA dashboard showing a visualization of a student activity. The image is available through Creative Commons Attribution-Non-Commercial License from the OU Analyse project.
Institutions may use learning analytics to monitor students’ activities and predict underachievers in order that they can be supported (Arnold & Pistilli, 2012; Misiejuk & Wasson, 2017). Learning analytics’ insights extend to tutoring, feedback, personalization, recommendations, and reflection (Moissa, Gasparini, & Kemczinski, 2015). At the teacher level, learning analytics have helped identify effective teaching strategies (Fritz, 2011, 2013), improve instructional design, and assess the effectiveness of pedagogies (Bakharia et al., 2016; Lockyer, Heathcote, & Dawson, 2013; Mor, Ferguson, & Wasson, 2015; Persico & Pozzi, 2015; B. T. M. Wong, 2017). It also supports informed decision-making (Ruipérez-Valiente, Muñoz-Merino, Leony, & Kloos, 2015; B. T. M. Wong, 2017), understand learners’ behaviors and class tendencies (Ruipérez-Valiente et al., 2015), and recognize disengaged students who might need support (Jayaprakash et al., 2014; Kuzilek, Hlosta, Herrmannova, Zdrahal, & Wolff, 2015). An example of a teacher dashboard showing students’ levels of engagement is displayed in Figure 4.
Figure 4: Another example of a teacher dashboard showing class engagement levels. The image is available through Creative Commons Attribution-Non-Commercial License from the OU Analyse project.

6.5 Feedback

Each cycle of learning analytics applications generates valuable insights into the process. These insights are expected to enlighten practitioners about improved ways of implementing solutions, avoiding mistakes, and refining strategies. Given that LA is an iterative process, greater knowledge of the students would improve the methods, refine the analytical models, and improve the process. As technology advances, new venues and further opportunities exploration and progress emerge (Clow, 2012a).
7. Research on learning analytics

7.1 An overview of previous research

Having introduced the process of learning analytics research in the previous section, in this chapter I provide an overview of how research in this field has progressed and yielded tangible results. I begin with exploring the early trials of inferring learning from students’ activities and how research progressed, with examples from landmark studies. I then present how learning analytics research materialized at the research and institutional levels, using three noteworthy examples (Course signals, E2Coach, OU Analyse). These examples were chosen for their historical, technological, and methodological importance.

The concept of studying trials of students’ online activities dates back a decade prior to the formalization of the field of learning analytics in 2011. For instance, Wang and Newlin (2000) reported that course page hits and forum postings correlated with the performance of online course takers (A. Y. Wang & Newlin, 2000). In a related study, they explored the potential of studying the data recorded by learning management systems (LMS) in informing teachers about students’ learning and ultimately predicting performance. The study included 121 students enrolled in a Research Methods in Psychology course, with the collected data comprising views of the course page, forum posts, and reads. They used a bivariate correlation to report the correlation between the sum of points earned by the students and their activities. Their results indicated a positive correlation between early page views and course performance (A. Y. Wang & Newlin, 2002b).

An example of the early implementations of visual dashboards displaying students’ activities were reported by Mazza et al. (2004).
They used visualization tracking to inform teachers about students who may require attention. The visualization helped teachers to gain an accurate view of the cognitive, behavioral, and social aspects of their students. The authors suggested that the tool helped the course instructors prevent course problems such as by identifying disengaged students at risk of dropping out. The teachers involved in the study commended the information offered by the visualization and described how it helped them track students’ activities (Mazza & Dimitrova, 2004b).

Ramos et al. (2008) investigated the role of various course components of online activities on the positive final course outcomes of 67 students enrolled in a Community Psychology course. They adopted a novel approach to data analysis by using a step-wise multiple regression model to predict the final outcomes from students’ tracking data. To further validate their findings, they performed cross-validation by using the equation generated from the “screening sample” to predict another test sample, “the calibration sample,” in a Psychopharmacology course with 52 students. Their results indicated that most of the variance in the outcome was explained by the number of clicked pages (Ramos & Yudko, 2008).

Dawson et al. (2009) studied the indicators of student motivation and correlated the findings with data derived from the LMS to test if patterns of online activity could indicate achievement orientation and therefore underlying motivation. Their findings suggested a positive correlation between participation in discussion forums and achievement orientation (S. P. Dawson et al., 2009). A year later, Macfadyen and Dawson (2010) published a proof of concept paper about how “Mining LMS data” can be used to develop an early warning system for teachers. The study broadened the scope of the data gathered to include network analysis, which included both correlations and multiple regression. Their findings indicated that the total number of forum posts, number of emails sent, and number of exams completed accounted for 30% of the variation in the students’ final grades. The regression model was able to correctly identify 81% of the students who failed. However, the first noteworthy implementation of an early warning system was described
by Kimberly E. Arnold (2012), who used “Course signals” at Purdue University to show safety levels as colored lights. A detailed discussion of course signals will be presented in the case studies in the next section.

With the introduction of the International Conference on Learning Analytics, research has substantially increased in volume and types of inquiry. Many institutions and countries are racing to harness the power of using data to support students, teachers, and stakeholders. Progress in the field of LA has been seen in three primary areas. First, the expansion of data scope by adding more digital sources, or adding innovative data capture methods such as digital sensors and physiological trackers. Second, improvements to data analysis methods by using advanced machine learning, data mining, or artificial intelligence methods. Third, exploring new objectives such as assessment of the relationship of physiological functions with learning, the exploration of high-resolution digital traces to infer self-regulation, or the study of the temporal aspects of learning (Blikstein & Worsley, 2016; Rajesh Kumar & Hamid, 2017).

7.2 Notable institutional implementations

The successful implementation of learning analytics to support students’ success and increase retention has proven to be fruit-bearing on the institutional level. Below is a discussion of three case studies that are either noteworthy for their historical importance (Course Signals), developed in-house (E2Couch), or for their scale of application (OU Analyse).

7.2.1 Course signals

Purdue University faced challenges of budget restraints, retention problems, and unprepared students. They sought solutions in developing an analytics system through a project titled “Course Signals” (CS) to help identify students at risk of underachievement (Arnold & Pistelli, 2012; Pistilli & Arnold, 2010). Course Signals
monitored students’ activities in real time and channeled their data to a predictive algorithm. The predictive algorithm received input from four sources:

- Student demographic data such as gender, age, or location;
- Prior academic history, including previous high school grades, academic preparation and examination grades;
- Students’ performance measured by the sum of grades earned so far in the course;
- Effort as defined by learning management system metrics compared to peers.

The algorithm calculated each student’s risk score and displayed a red, yellow or green signal on the student’s course page (see Figure 5). A red signal alerted a high likelihood of underachievement and thus an impending problem; a green signal indicated a high likelihood of passing the course; and a yellow signal indicated a borderline status (Arnold & Pistilli, 2012). Intervention would then be initiated by instructors according to each case; intervention starts by sending an email or an SMS, referral to an academic advisor or a meeting with the course instructor. The system was received positively by both students and staff members. Moreover, the university reported that using learning analytics through Course Signals improved the success rates of first- and second-year students, as well as retention rates (Arnold & Pistilli, 2012; Dietz-Uhler & Hurn, 2013).
Figure 5: Course Signals at Purdue showing green (safe) and yellow lights (borderline safe). Provided courtesy of Purdue University, image used with permission.

7.2.2 E2Coach

E2Coach is an adaptive advice system developed by the University of Michigan to provide personalized recommendations and guidance tailored to students’ goals. The advice was modeled on insights derived from peers in the form of motivation, recommendation of learning resources and feedback about performance. E2Coach (see Figure 6) creates a personalized dashboard for each student, containing a list of personal learning goals and objectives and delivering tailored recommendations. As the course advances, the system delivers custom personalized web updates and personal messages on the dashboard. The personal messages address students’ online regulation, deliver supportive messages, and contain comparative graphics and benchmarks. The dashboard
works as an awareness tool that prompts students to adopt a more effective strategy. E2Coach has helped motivate students to complete their courses and improve their study habits, and thus achieve more highly than their expected grades by 5%. Moreover, it offers students the advice they require in uncomfortable situations such as transitioning to new courses where they have little knowledge of what to do. Teachers have considered E2Coach a supportive tool that helps their students to focus on the important subjects (McKay, Miller, & Tritz, 2012; Wright, McKay, Hershock, Miller, & Tritz, 2014).

Figure 6: A sketch of the E2Coach system. The system matches advice to students’ information and personalizes the messages accordingly. The image is available from Educause.edu through Creative Commons Attribution-Non-commercial License.

7.2.3 OU Analyse

Open University OU Analyse is a modern learning analytics system, with the primary goal of early predicting underperforming students, whereby a significant intervention can have a meaningful
impact. OU Analyse receives data from two main sources: static (demographic) data and data derived from students’ interactions with the learning management system. Four predictive models are used to analyze the data: a Bayesian classifier for finding the most significant LMS resources, k-nearest neighbors (k-NN) and classification and regression tree (CART) for the analysis of demographic/static data and k-NN with LMS data. Results from the project indicated that identifying at-risk students was possible with 50% accuracy at the beginning of the semester, and accuracy increased to over 90% by the end of the semester (Kuzilek, Hlosta, Herrmannova, Zdrahal, Vaclavek, & Wolff, 2015).
8. Challenges in the field of learning analytics

The journey towards the development of learning analytics research has not been smooth. Indeed, numerous difficulties and challenges are associated with the journey, some of which are known, whereas others might emerge in the future. I present here a brief discussion of the most common and pressing challenges in the field faced by researchers, stakeholder, and institutions, with emphasis on ethical and privacy issues as they are currently most pertinent considering the European General Data Protection Regulation (GDPR).

As a relatively new research discipline, many challenges and open questions face researchers and institutions implementing learning analytics for the benefit of their students or institutions. These problems have been the subject of research and inquiry, and numerous international scientific communities are working together to find solutions (Avella et al., 2016; Chatti et al., 2014; Daniel, 2015; Ferguson, 2012a, 2012b; Nunn et al., 2016; Tsai & Gasevic, 2017a, 2017b).

8.1 Theory and connection to learning sciences.

One of the most pressing issues in learning analytics is how learning analytics is connected to learning, informed by theory, or can result in sensible development in learning sciences. A discussion of the theory and its significance was presented in Chapter 5.
8.2 Ever-expanding data sources
The diversity and scope of data sources will continue to grow owing to digitalization and innovation, bringing new challenges of usability, integration, storage, privacy, and relevance (Ferguson, 2012b; Siemens, 2013). Moreover, working with data from different sources and attempting to optimize and integrate these large heterogeneous data sets will require new skills, tools, infrastructure and most importantly research methods (Chatti et al., 2014; Ferguson, 2012b).

8.3 Student-centered analytics
Analytics has always focused on institutional goals and perspectives rather than the needs of the learners themselves, a problem that highlights the importance of a shift towards new areas, beyond grades and performance to include motivation, self-regulated learning, confidence, and meeting future professional objectives. Analytics can also help emphasize the role of formative assessments and students’ work during the course (Ferguson, 2012b).

8.4 Capabilities
Learning analytics requires trained personnel from different specialties and disciplines. Furthermore, it requires highly trained data scientists and statisticians. With current shortages in such specialties, it might prove difficult to build a learning analytics team. Similarly, deficiencies in leadership and expertise that can strategically build and maintain the implementation of learning analytics solutions currently exist (Nunn et al., 2016; Tsai & Gasevic, 2017b).

8.5 Validity
Validity points to how accurate is the analysis. An inaccurate prediction might wrongly mislabel students or expose them to undesired stress. Validity can always be boosted by choosing relevant sources of data, applying the correct algorithms, verifying results and predictions, and applying a feedback mechanism that permits evaluation of the entire process to avoid making conclusions based on erroneous correlations (Ferguson, 2012a).
The validation and replication of results remain a challenge in learning analytics as a field and most likely in education as a whole. The lack of replicability of research findings as well as the relatively small number of evidence-based approaches to analytics may hinder many institutions from adopting, investing, or funding learning analytics research (Pashler & Wagenmakers, 2012; Tsai & Gasevic, 2017b).

**8.6 Replication**

The validation and replication of results remain a challenge in learning analytics as a field and most likely in education as a whole. The lack of replicability of research findings as well as the relatively small number of evidence-based approaches to analytics may hinder many institutions from adopting, investing, or funding learning analytics research (Pashler & Wagenmakers, 2012; Tsai & Gasevic, 2017b).
8.7 Ethics and privacy issues
Although learning analytics has existed for some time, the volume of research in the field regarding ethical and privacy issues remains relatively low. Moreover, the available guidelines or policies regarding the ethics of conducting analytics research or user privacy had not matched the rapid pace of emerging research. Privacy issues can prove challenging because LA deals with data from multiple sources, potentially entailing different policies (Drachsler et al., 2015; Lang, Macfadyen, Slade, Prinsloo, & Sclater, 2018; Pardo & Siemens, 2014; Prinsloo & Slade, 2018; Rubel & Jones, 2016). The absence of a uniform legal or standard ethical framework has stimulated the LA community to negotiate these complex and relatively new challenges by relying on inspiration from other disciplines such as medical research and industry, as well as the relevant regulatory, cultural, and legal frameworks (Drachsler & Greller, 2016; Tsai & Gasevic, 2017a). The main principles to consider when conducting analytics research or implementing LA in education are summarized below.

8.7.1 Transparency
Transparency, openness, and clarity should be applied at all stages of learning analytics. Stakeholders should be adequately informed about the purposes of the captured data, collection methods, and processing. They should also be informed about the ways in which data will be analyzed if any third parties or companies will participate in data processing. Students should be offered clear guarantees that data will not be sold or transferred to other institutions or legal entities without proper consent (Bienkowski et al., 2012; Rubel & Jones, 2016; Tsai & Gasevic, 2017a; Tsai et al., 2018).

8.7.2 Control over data
Although challenging in practice, giving users (students or instructors) the right to access and correct the data collected is in the best interest of all parties involved in LA. Users’ control of data would boost the accuracy of results and enhance user trust in the system and its validity (Pardo & Siemens, 2014; Petersen, 2012; Slade &
Prinsloo, 2013). This can be critical with certain types of data, such as correcting the dates of enrollment and financial data, modifying the duration of calculation of their course access due to an excused leave or erroneously captured information (Pardo & Siemens, 2014).

8.7.3 Agency
Students should be encouraged to enjoy agency and voluntarily use the data for their own benefit and development, rather than merely for institutional profiling purposes or to decrease attrition rates. Agency emphasizes that students are participants in a learner-centric model, in contrast to an intervention-centric model that has the primary aim of improving general retention and attrition rates through intervention (Rubel & Jones, 2016; Slade & Prinsloo, 2013).

8.7.4 Consent
Obtaining informed consent is an established tradition in research practice, but in a rapidly evolving field with new tools and methods, it is difficult to maintain a form of consent that covers the current and future development of analytics research (Mantelero, 2014; Roberts, Chang, & Gibson, 2017; Slade & Prinsloo, 2013). Concerns have also been raised that the consent and ethical frameworks created in recent decades may not be suitable in the age of big data and machine learning, citing the vast volume and diversity of collected data and the complexity of procedures, which regular users may be unable to understand in any depth. This may necessitate an increased role for independent authorities and data protection bodies in order to enact the adequate necessary measures to protect users and ensure the correct balance between the benefits of the institution and user privacy (Mantelero, 2014; Prinsloo & Slade, 2018).
Consent in learning analytics does essentially share the same general principles as consent in other research disciplines. Users should be free to give consent voluntarily without being subject to any kind of negative consequences should they reject. Users should
be well-informed of the whole process, the types of collected information, storage, de-identification, future usage, and analysis. Details should be given transparently and in full (Kay, Korn, & Oppenheim, 2012; Pardo & Siemens, 2014; Slade & Prinsloo, 2013; Tsai & Gasevic, 2017a). Consent also needs to be revocable, so that users can opt out at any time, and the institution stops processing their information or the inclusion of their data in any processing. Full documentation of the entire process would be helpful in case of conflict (Gruschka & Jensen, 2014; Kay et al., 2012; Pardo & Siemens, 2014; Slade & Prinsloo, 2013).

8.7.5 Responsibility
Institutional policies detailing rules and responsibilities concerning data processing and the entire process of analytics (capture-reporting-analysis-feedback) should be clearly defined. A mechanism for complaint that would allow for the investigation of cases of abuse or breach of privacy should be in place, so that users can file concerns and request the removal of their information (Kay et al., 2012; Rodríguez-Triana, Martinez-Monés, & Villagrá-Sobrino, 2016; Roberts et al., 2017).

8.7.6 Minimizing Harm
Potential adverse impacts of LA have been identified in terms of learning analytics, such as labeling students, categorization, bias and promoting discriminatory practices. Efforts should thus be made to proactively guard against such issues (Rodriguez-Triana et al., 2016; Tsai & Gasevic, 2017a).
9. Collaborative learning

In this chapter, I present the contextual background of the thesis: the collaborative learning environment. The chapter begins with an introduction to collaborative learning. I then introduce computer-supported collaborative learning (CSCL) as a medium for collaboration.

One of the significant pedagogical developments in the last century was the introduction of collaborative learning, where a redefinition of the relationship between the student and the teacher was made. The teacher would no longer be regarded as a “sage on the stage” who instructs listening students, but instead a moderator who facilitates and scaffolds students to actively search, share, collaborate and most importantly assume responsibility for their learning (Johnson & Johnson, 2008; Laal & Laal, 2012; Macgregor, 1990).

Collaboration requires participants to mutually and intellectually coordinate their efforts to interdependently solve a problem, perform a task, or work together on a project. Collaborative learning typically occurs in groups that share a common learning objective, and it is often characterized by substantial dialogue, argumentation, and debate (Brindley, Blaschke, & Walti, 2009; Jeong & Hmelo-Silver, 2016; Johnson & Johnson, 2009).

9.1 What makes learning collaborative?

Successful collaborative learning requires five essential elements (Johnson & Johnson, 2008, 2009):

**Positive interdependence**: collaborators work together to accomplish a mutual goal. The success or failure of the group’s task is dependent on the work of each individual. Therefore, participants
have a genuine interest in supporting each other in performing their roles (Johnson & Johnson, 2008, 2009).

**Accountability:** positive interdependence increases participants’ feelings of responsibility to do their share of the task, and to facilitate other members’ work.

**Promotive interaction:** participants interact in a meaningful way by exchanging necessary resources and knowledge, providing support and constructive feedback to one another, and debating and refining conclusions in order to improve the decision-making process.

**Appropriate use of social skills:** for interpersonal interactions to be successful, collaborators should have social and small group skills that can help them to manage conflicts and stresses.

**Group processing:** collaborators evaluate the group’s performance and reflect on its successes and failures. The process is expected to improve future group performance and dynamics.

### 9.2 Benefits of collaborative learning

Efficient collaborative learning has several benefits that fit into three main categories: psychological, social, and academic. Psychological benefits include boosting learners’ self-esteem, reducing anxiety and promoting a positive student-teacher relationship. Socially, collaborative learning helps build a supportive environment that promotes cooperation and diversity of understanding, as well as strengthening learning communities. Academically, collaborative learning enhances the development of critical thinking and problem-solving skills. Collaborative learning helps students to retain information for longer, and motivates them to achieve and accomplish their learning objectives (Johnson & Johnson, 2008, 2009; Laal & Ghodsi, 2012).

### 9.3 Computer-supported collaborative learning

Ever since computers became part of mainstream education, educators have sought to utilize information technology effectively. With the introduction of the Internet, computer-supported collaborative learning (CSCL) has become an essential tool for collaborative learning (Jonassen et al., 1995; Stahl, Koschmann, & Suthers, 68
2006; Wasson & Mørch, 2000). CSCL offers learners the possibility to interact asynchronously regardless of time or location, thereby enabling collaborators to work at their own pace (Hew, Cheung, & Ng, 2009; Lou, Abrami, & d’Apollonia, 2001; Stahl et al., 2006). CSCL facilitates the process of creating and exchanging information through discourse and interaction with peers and experts (Jeong & Hmelo-Silver, 2016; Resta & Laferrière, 2007). CSCL can be implemented in diverse ways, such as collaborative writing, project-based learning, and technology-mediated discourse. A common implementation of technology-mediated discourse is the threaded discussion board, known as a forum. Forums allow learners to interact asynchronously, collaborate, and construct new knowledge (Garrison, Anderson, & Archer, 1999; Hew et al., 2009; Janssen & Bodemer, 2013). Discussion boards permit permanent access to interactions, enabling learners to reflect and learn from each other in a transparent way (B. Barron, 2003; Janssen & Bodemer, 2013; Slof et al., 2010). Forums have become a standard feature of most major learning management systems and continue to constitute among the most commonly used modules by educators (Dahlstrom, Brooks, & Bichsel, 2014).

9.4 Analysis of collaborative learning
Although CSCL may represent an effective medium for collaboration, research has shown that different mechanisms are required to motivate learners, facilitate discourse, manage conflicts, and most importantly monitor the collaboration itself. Four main methods are commonly used for the analysis of collaborative learning: interaction analysis, content analysis, educational data mining (content space), and social network analysis (relational space) (Brindley et al., 2009; Dado & Bodemer, 2017; De Wever, Schellens, Valcke, & Van Keer, 2006a; Jeong & Hmelo-Silver, 2016).

9.4.1 Interaction analysis
Interaction analysis presents statistics regarding forum activities and the contributions of participants such as the frequency of posts by a student, participation statistics, and the number of replies in a discussion. Although these statistics are informative about the level
of activity in a forum, they fall short in terms of relational aspects and can hardly be used to judge the quality or dynamics of the interactions thread (Dimitracopoulou, 2009; Rodríguez-Triana et al., 2013).

9.4.2 Content analysis
Content analysis is concerned with the qualitative analysis of transcripts in order to reveal the process of knowledge construction and the social dynamics of interactions (De Wever et al., 2006a; Stemler, 2015). Content analysis requires an analysis instrument and manual coding. Content analysis instruments have been criticized for lack of clear synergy between theory and design with less than optimal validation procedures (Calvani, Fini, Molino, & Ranieri, 2010; De Wever, Schellens, Valcke, & Van Keer, 2006b; Mayring, 2014). Manual coding is time-consuming and labor-intensive and may produce contradictory results. Given that no unequivocal standard exists by which the reliability of analysis can be judged, the practice is only suitable for experienced researchers. Although manual content analysis can be useful in a research setting, it is an impractical choice for teachers or administrators wishing to monitor students’ interactions in real-time for intervention.

9.4.3 Educational data mining
Educational data mining (EDM) tools may offer an automated analysis of large amounts of text (Al-Razgan, Al-Khalifa, & Al-Khalifa, 2014; Peña-Ayala, 2014; Romero & Ventura, 2013). However, existing tools are very technical for teachers to use or interpret (Al-Razgan et al., 2014; Papamitsiou & Economides, 2014; Romero & Ventura, 2013), are regularly misinterpreted (Papamitsiou & Economides, 2014; Peña-Ayala, 2014), and the implementations often produce contradictory judgments (Papamitsiou & Economides, 2014).

9.4.3 Social network analysis
Social network analysis may be particularly useful for the analysis of the relational space of CSCL as it offers the detailed mapping of
relations and interactions. The technique was discussed in the analysis in Section 6.3.4.
The five studies of this thesis were based on 13 different courses that share several features. I will discuss the general context of education in the university as well as the specific context of the studies. This will be followed by the approach used for data capture, preprocessing and analysis.

10.1 The general context

Data collection for the studies took place at the College of Medicine, Qassim University, Saudi Arabia, (Studies 1-3), and the College of Dentistry (Studies 4 and 5). Both colleges have curricula based on problem-based learning (PBL), which uses problem scenarios (for instance, clinical cases) as triggers to facilitate discourse and interaction among students. Discussions occur in small groups and are facilitated by teachers (Neville, 2009). The PBL process is structured in a way to help students to elaborate and activate previous information (Neville, 2009; Schmidt, Rotgans, & Yew, 2011). Several implementations of PBL can be identified; in particular, the university adopted the Maastricht seven jumps approach (Davis & Harden, 1999). A schematic representation of the classic seven jumps approach is shown in Figure 7.

In the seven jumps approach, students are expected to attend two sessions physically, one at the beginning of the week and another at the end. During the first session, students clarify new vocabulary to find its meaning, then identify the problems that need to be addressed. They subsequently brainstorm possible solutions to iden-
tify the learning objectives they need to study. In the second ses-
sion, students share the knowledge and wrap up what they have
learned (Davis & Harden, 1999).

![Diagram of the classic seven jumps approach]

Figure 7: A schematic representation of the classic seven jumps
approach.

Qassim University has introduced a blended problem-based learn-
ing approach by introducing an online forum so that students can
continue to interact asynchronously for the whole week until a new
problem is introduced. In fact, most interactions are taking place
online. As such, students discuss learning issues, share informa-
tion, collaboratively construct the required knowledge, and
work towards achieving the goals and objectives of the PBL prob-
lem on the online forum (Alamro & Schofield, 2012). A schematic
representation of the approach adopted by Qassim University is shown in Figure 8.

Problem-based learning is the primary learning strategy in the basic science phase (years 1-3). In the clinical phase, the curriculum is based on the traditional lecture and clinical sessions approach. Studies 1, 3 and 4 were conducted in problem-based courses, and Study 2 was conducted in a traditional course. E-learning has been used extensively at Qassim University since its introduction in 2008 by the author of this thesis, where he served as the supervisor of the e-learning unit. All courses rely on e-learning for announcements, lecture delivery, formative assessments, and collaborative computer-supported discussions. Students are expected (but not obligated) to use the e-learning portal for a wide range of learning activities, rendering it key to their educational progress. As such, the activities of students in e-learning are reflective of students’ engagement, self-regulation, and self-evaluation. The e-learning portal is based on the open source learning management system, Moodle, available at www.moodle.org. Moodle is a
free modular software built to enhance interactivity through a constructivist-centered design (Dougiamas & Taylor, 2003). Moodle has built-in modules for course delivery, interactivity, assessments, announcements, and feedback, as well as very vibrant, community-developed plugins that allow for the extension of functionality. Although Moodle has a very mature logging system, it lacks the features necessary for analytics. Nevertheless, Moodle’s developers recently launched a new project, Inspire, which can be deemed Moodle’s answer to learning analytics. Inspire is still in the early stages of its development, with limited functionality. Therefore, researchers interested in applying learning analytics research must use external plugins and manual analysis of the Moodle logs to extract the required information (Falakmasir & Habibi, 2010; HQ, 2017).

10.2 Research methodology

The methodological outline of learning analytics research follows the learning analytics process discussed earlier, depicted in Figure 1 (H. Chen et al., 2012; Clow, 2012b; Siemens, 2013). The process starts with data capture, followed by preprocessing and preparation, and later analysis, interpretation, and action. The first three stages are analogous to the standard data mining procedures common to analytics research (Gandomi & Haider, 2014; Wolff, Zdrahal, Herrmannova, Kuzilek, & Hlosta, 2014). The general structure of the methods process is outlined, followed by specific information concerning each step.

10.2.1 Data capture

This stage involved the recording and gathering of data from the learning management system as well as other sources such as performance data and attendance data. The methods used for data acquisition in this thesis included external Moodle plugins and manual extraction of data from the Moodle logs using structured language queries. While the process was similar in all studies, Study
1 involved slightly different indicators. The details of data collection of each study are as follows:

**Study 1:** Three types of indicators were used in this study: generic indicators, calculated engagement sub-indicators, and access to specific resources (orientation, evaluation, course objectives, etc.) (Saqr et al., 2017). The data were acquired using external Moodle plugins and database queries. The data were extracted each week separately to account for temporal activities and the regularity of using the online portal. The plugin Attendance Register was used to calculate the total duration of students being online each week. The configurable reports plugin was used to extract page views, logins, forum posts, forum edits and forum reads. The plugin Analytics Graphs was used to calculate the number of unique days of course and resource access. These plugins are available at the Moodle plugins directory at www.moodle.org/plugins. Lastly, NodeXL, a social network analysis software, was used to calculate the Betweenness centrality (Smith et al., 2009). Data for views of course orientation (course objectives, requirements, and assessment), and self-evaluation (formative assessment trials and grades obtained) were also collected using the configurable reports plugin queries.

While the external plugins helped gather the indicators of students’ activities, the collected indicators comprised the counts of clicks and views, or “generic indicators” that have in the past been criticized for being less representative of learning and for having only a weak connection to learning theory. Therefore, I calculated the engagement sub-indicators. A detailed explanation of the concept and method of the engagement indicators will be discussed in the data preprocessing and preparation section.

**Studies 2, 3, 4 and 5:** The four studies used social network analysis methods, hence data acquisition and processing was similar. Given that Moodle does not have a plugin to handle or export social interaction data, I used custom structured language queries to extract raw interaction data with the attributes of the users. The attributes collected were: the course ID, the group ID, the source of the interaction, the target of the interaction, the subject of the interaction, the content of the interaction, the forum ID and the sub-forums’
IDs, and the time-stamp) (Saqr, Fors, & Nouri, 2018; Saqr, Fors, & Tedre, 2018; Saqr, Fors, Tedre, & Nouri, 2018; Saqr, Nouri, & Fors, 2018).

10.2.2 Data preprocessing and preparation
In this stage, data extracted through plugins were combined, inspected, and verified for quality. Mislabeled data were corrected, and corrupted records were removed. For Study 1, the engagement sub-indicators were calculated to reflect the regular usage of the online portal over the course duration, comparing each student to the average activity of his or her peers (Saqr et al., 2017). In other words, a student was considered engaged if his or her online activity was more than a 1-z score of the activity of peers. The engagement sub-indicators were calculated for each week, and the total score of the six weeks for each student was computed. The engagement sub-indicators were computed for course views, forum posts, online time, and forum reads. For Studies 2, 3, 4 and 5, the data were assembled, verified, and converted to a file format ready for analysis by the social network analysis software (Saqr, Fors, & Nouri, 2018; Saqr, Fors, & Tedre, 2018; Saqr, Fors, Tedre, & Nouri, 2018; Saqr, Nouri, & Fors, 2018).

10.2.3 Data analysis, interpretation, and reporting
During this stage, I performed summarization of data through visualization and descriptive statistics to detect patterns and the distribution of variables and to attain a general overview of the data set. This was followed by exploratory statistical analysis, a step towards the identification of the outliers and significant predictors, and to test the possible analytical methods (Velleman & Hoaglin, 2012). According to the insights from the visualization and the descriptive and exploratory statistical analysis of the data, the appropriate statistical models were identified.
10.2.3.1 Statistical analysis
In order to answer the question of which variable correlates with better performance, a correlation test was performed in Studies 1, 2 and 4. The correlation test was computed using permutation methods in cases where the data set constituted social network data to avoid the problem of interdependence. Interdependence is a concern as SNA data are relational by design and each connection may be dependent on other connections. In such situations, permutation methods might be considered robust and credible when inferring SNA data (Borgatti, Everett, & Johnson, 2018; Hothorn, Hornik, Van De Wiel, & Zeileis, 2008). In permutation, the software generates plentiful resampled random distributions (10,000 in this thesis), which are later compared to the observed data to calculate the significance.

In order to predict students’ performance, I used automatic linear regression (ALM) in Study 1 as well as in Study 2 where the data belonged to a single course (Saqr et al., 2017; Saqr, Fors, & Tedre, 2018). ALM is a machine-learning algorithm that has the capacity to normalize outliers, including all predictors in the analysis, and to identify and remove collinear predictors. Such features are lacking in simple linear regression models. Furthermore, the model offers ensemble capabilities; hence I used the ensemble method bootstrap aggregation (bagging). Bagging is a machine-learning method that enhances the accuracy and stability of the model, reducing the variance and overfitting. Bagging is performed by generating random subsamples (100 subsamples in this thesis) of the data and using these subsamples for training. The resulting model is an average of all iterations in the model (Filippou, Cheong, & Cheong, 2015; Yang, 2013). ALM was chosen because it is an automatic model that demonstrates its feasibility for a learning analytic dashboard that offers insights without human intervention (Filippou et al., 2015; Yang, 2013).

In order to classify achievement levels, I used logistic regression (Saqr et al., 2017; Saqr, Fors, & Nouri, 2018; Saqr, Fors, & Tedre, 2018). Logistic regression is a powerful algorithm for the classifi-
cation of binary outcomes. Logistic regression does not require linearity, normality, or equal variance as assumptions. To evaluate the model performance, I used the -2 log likelihood, Cox & Snell R-squared, the omnibus tests of model coefficients, and Hosmer and Lemeshow goodness-of-fit (Bewick, Cheek, & Ball, 2005; Brooks & Thompson, 2017; Gardner & Brooks, 2018; Peng, Lee, & Ingersoll, 2002).

I subsequently used machine learning methods with k-nearest neighbors algorithm (k-NN) and naive Bayes algorithms (Saqr, Nouri, & Fors, 2018). KNN was selected as it classifies students based on similarities, hence students who might share specific activity features may become easier to group together. Naive Bayes algorithm was also used to evaluate the same data as it proved in our case to perform more effectively when classifying underachievers (Brooks & Thompson, 2017; Kotsiantis, Zaharakis, & Pintelas, 2007; Zhang, 2016). Both algorithms were performed with a ten-fold cross-validation method. The ten-fold cross-validation was performed by rotatory partitioning of the data set into ten random equal subsamples. Model training was undertaken 10 times, and in each, training was completed using nine subsamples, with the remaining subsample used for validation. Thus, each subsample was used for training and validation. The method had the advantage of more generalizable models and less overfitting. To judge the model performance, I used accuracy, precision, F measure, and area under the curve (AUC) (Saqr, Fors, & Nouri, 2018; Saqr, Nouri, & Fors, 2018).

Other statistical tests were also used. Indeed, the Wilcoxon signed ranks test was used to compare the difference between students’ interactivity parameters before and after intervention. The Shapiro-Wilk test of normality proved that most variables in the study followed a non-normal distribution (Saqr, Fors, Tedre, et al., 2018). Model stability, reliability, and validity were stressed in all of the studies. In the Studies 1 and 2, an ensemble technique was used to create 100 subsamples and to calculate the average prediction results (Saqr et al., 2017; Saqr, Fors, & Tedre, 2018). In Study 4, a following year’s data set was used to validate the results and demonstrate the possibility of forecasting future performance.
(Saqr, Fors, & Nouri, 2018). In Study 5, a ten-fold cross-validation technique was used to improve the generalizability of results. Moreover, the data from each course was used to validate the predictions of the subsequent courses (Saqr, Nouri, & Fors, 2018).

10.2.3.2 Social network analysis

A combination of visual and mathematical analysis was used. Gephi 0.9 was used for the visual network analysis and for the calculation of centrality measures (the mathematical analysis) (Bastian, Heymann, & Jacomy, 2009). SNA visualization was used to summarize the interactions and relationships and to visualize the social structure. Visualization was also used to guide the mathematical analysis and identification of collaborators’ roles. Visualization was undertaken using a force directed algorithm, which is an aesthetically plausible and simple-to-interpret algorithm that renders the nodes according to their structural properties. Furthermore, it can be used to create a continuous animated timeline of interaction events (a dynamic network). An animated compilation of all interactions was created in the form of a time-lapse video to demonstrate the timeline of events and better represent the changes of behavior across the time dimension (Bastian et al., 2009; Jacomy, Venturini, Heymann, & Bastian, 2014).

Mathematical analysis

Two levels of mathematical analysis were performed, at the individual level and the group level. At the individual level, a model was developed to capture the interactions, relationships, and attributes of the collaborators. The model can be summarized as follows:

Individual collaborators’ level:

The quantity of interactions:

- Outdegree centrality represents the quantity of outgoing interactions;
• Indegree centrality represents the total count of received interactions;
• Degree centrality is the total of both outdegree and indegree centrality measures.

Role in relay and transfer of information:
• Betweenness centrality represents the number of times a collaborator connected others or mediated their interactions;
• Closeness centrality represents the closeness of a collaborator with others for communication;
• Information centrality measures the role of a collaborator in information flow within his or her group.

Connectedness:
• Eigenvector centrality measures the connectedness of an individual, and the strength of his or her network connectedness;
• Indegree prestige is the number of directly connected collaborators and an estimate of the ego network;
• Proximity prestige is the number of directly or indirectly connected collaborators;
• Rank prestige represents the strength of the network;
• Domain prestige measures the popularity of the collaborator.

Role in collaboration:
Based on the visual and centrality measures analysis, roles were extrapolated for students and teachers guided by the work of Marcos-García, Martínez-Mónes, and Dimitriadis (2015). The following roles were identified:
• Leader: an active collaborator who leads the discussion;
• Coordinator: an active collaborator who coordinates interactions;
• Active collaborator: an active collaborator;
• Peripheral: an isolated participant.
The group level:
The parameters of group size, average degree, average indegree, average outdegree, density, and clustering coefficients were calculated. These parameters capture the different interactivity constructs in a network as well as the community’s cohesiveness. A detailed discussion of the centrality measures and social network analysis was presented in Chapter 6 in Section 6.3.4.

10.3 Research ethics
Navigating ethical approval options was not a straightforward endeavor. In my institution, Qassim University in Saudi Arabia, no precedent existed to guide me to the available options of using students’ online data for research. Given that working with such kinds of data was new, I tried to tackle the problem by first creating an online privacy and data protection policy; second an informed consent covering general data usage purposes such as daily trends of system usage; and third obtained ethical approval from the governing body.

10.3.1 Data protection and privacy policy
My first step was to create a data protection and privacy policy that governs the ethical use of data and protects users’ rights. The data protection policy was inspired by policies from leading institutions, namely the Open University United Kingdom (University, 2014), the Learning Analytics Community Exchange (LACE) (Drachsler & Greller, 2016), and the Centre for Educational Technology and Interoperability and Standards (CETIS) (Sclater & Bailey, 2015). The policy contains ten guarantees for the students (use the data for students’ benefit, transparency, compliance with law, no harm, anonymity, personal data protection, right to correct data, right to withdraw, control over data, and consent for any usage beyond terms). The policy was distributed to all students and teachers in the institution via email and published on the website, and students...
had to acknowledge it on their first login. The policy was later revised, updated, and ratified among the college policies by the Policies unit.
Throughout this research, the data protection policy was followed strictly. All data were anonymized, personal identifying information was masked and no private data or personal information were stored, used in the analysis or published. It is also important to mention that use of the e-learning portal is neither graded nor mandatory, and depends solely on a student’s interest. Moreover, the researcher was not a teacher and did not participate in the grading of any of the courses used in this research.

10.3.2 Consent
The second step was to develop an online consent form that users of the LMS could sign. The consent form granted limited rights for using the anonymized data within the framework of “Data protection and privacy policy” for research purposes. The consent form was also used to communicate information and privacy rights to the students, and detailed cases requiring the student’s explicit approval or approval from a governing ethical body. The consent form was built in compliance with the National Committee of Bioethics (NCBE) on Living Creatures and good research practice (Hermerén, 2011), and guided by international standards (Tsai & Gasevic, 2017a). The consent form also offered students the option to withdraw with no consequences or denial of service, and emphasized their rights to privacy as well as the possible uses of data for research (Hermerén, 2011; Qassim College of Medicine, 2015).

10.3.3 Ethical approval
Given that the research contained data from multiple sources including students’ grades and logged records from the LMS, I applied for and obtained ethical approval from Qassim Regional Research Ethics Committee. The proposal was reviewed and approved by the university research center. It was then reviewed by the Qassim Regional Research Committee, which reviewed the
study proposal, data collection methods, reporting methods, data handling policies, and outline of the five studies. Having reviewed the documents, approval was issued, and the committee stressed that the privacy of students should be protected. A copy of the ethical approval is shown in Figure 9.

Figure 9: A copy of the ethical approval
11. Overview of the results

This thesis is composed of five interrelated studies that build on insights from one another. The first study attempted to use engagement and self-regulation as proxy indicators for learning in the early identification of underachievers (Saqr et al., 2017). Through the first study, I identified the importance of collaborative and networked aspects of learning. The second study then investigated the use of collaborative learning indicators that are contextually aligned with the process of collaborative learning (Saqr, Fors, & Tedre, 2018). The study objectives comprised how to inform teachers and improve the prediction of students’ performance. The third study built on the second, using the insights gained from the monitoring of collaborative learning to create an informed intervention and evaluate its efficacy (Saqr, Fors, Tedre, et al., 2018). The fourth study sought to build on the insights gained from the previous three studies and create a more reliable predictive model of achievement in the online PBL environment (Saqr, Fors, & Nouri, 2018). The fifth study tried to account for the dimension of time and temporality in a collaborative learning environment (Saqr, Nouri, & Fors, 2018). The five studies thus help answer the two research questions. The results will be discussed in detail in the following section.

RQ1: How can learning analytics and social network analysis reliably predict students’ performance in collaborative learning environments using contextual, theory-guided indicators?

The first study was conducted with the aim of investigating the online activity variables that best correlate with student performance in a PBL-based course, as well as to identify the indicators
that can support the early prediction of underachievers (Saqr et al., 2017). The results indicated a weak correlation between the generic indicators (count of page hits, resource views, number of forum postings and duration of being online) and student performance. Nonetheless, the parameters of engagement and self-regulation demonstrated a greater and more consistent correlation with performance. The most significant variables were those reflecting the task orientation, self-monitoring, evaluation, and consistency of using the e-learning portal.

The results of automatic linear modeling demonstrated that the variables that reflect self-regulation (task orientation, engagement sub-indicators, and self-evaluation) and forum postings (interactions) represented the most significant predictors of performance. The adjusted R-squared of the model was 0.64, an indication that the model could explain 64% of the variance of the grades. A binary logistic algorithm was then used to investigate the possibility of classifying underachievers using the significant parameters identified in the previous step. I was able to detect 21 out of 26 underachievers with a sensitivity of 80.8% CI (60.7% to 93.5%) and a specificity of 96.3% CI (93.4% to 99.8%); the AUC value was 0.9 and the F-score was 0.96. Using the data of early participation up to the midterm, the prediction accuracy was 85.7%, the sensitivity of detecting high achievers was 96.2%, and the specificity was 42.3%. The AUC was 0.69 and the F-score was 0.91.

The results of Study 1 indicated that engagement and self-regulation indicators were reliable predictors of performance and could be used to early predict a substantial number of potential underachievers. Given that engagement and self-regulation are modifiable, they are amenable to positive intervention and may be used by educators and stakeholders for data-driven action.

While Study 1 included CSCL forum postings, reads, and views, and investigated how students self-regulated their learning activities, Study 2 examined the relational, structural and interactive side of collaborative learning (Saqr, Fors, & Tedre, 2018), specifically how participation, interactions and network parameters were cor-
related with academic performance. The operationalization of interactivity was achieved using a framework capturing the quantity, role, information transfer and connectedness, along with the group’s structural and interactive properties. Two statistical methods were used – correlation coefficients and automatic linear regression – in order to investigate how social network parameters could predict the final grades or account for grade variance. The results of the correlation coefficients and automatic linear regression showed that the parameters of the quantity of interactions, akin to the generic indicators of counts of activities, were weakly correlated with better performance. However, indegree centrality showed the highest correlation with performance, indicating that the quality of contributions that stimulated replies from peers was most significant. The results also showed that the SNA variables that were indicative of the role in information transfer and connectedness were moderately correlated with performance. Using a classification binary logistic algorithm, I was able to successfully classify 85.7% of the students according to their performance (91.7% of high achievers, 72.7% of underachievers). These results demonstrate the value of using social network analysis indicators as predictors of performance.

Having obtained reasonable predictive accuracy using SNA as indicators for performance, the next step was to investigate the possibility of creating a model that is reproducible from one course to another, via a theory-guided approach. Although a general model may not be feasible in all contexts, a context-specific model might be achievable in some. Online problem-based learning may represent a good candidate for such a trial because online PBL is a structured approach to collaborative learning and is relatively uniform. The teaching model of PBL has three essential factors: the students, the group, and the tutor (Neville, 2009; Schmidt et al., 2011). As such, the fourth study was conceptualized to study the operationalization of the interactions and collaboration among the three factors (Saqr, Fors, & Nouri, 2018). The study accounted for the roles of students, tutors and their groups using SNA in four courses spanning a full year of education.
The results indicated a positive correlation between interactivity parameters and performance in all the courses studied, regardless of the course subject. Students who were strongly connected to significant collaborators in small interactive groups performed best. The results of the correlation test were confirmed by the regression analysis. The regression model was able to classify students’ performance with an accuracy of 93.3% and an F-measure of 90%. To validate the results and to investigate if learning analytics data can be used to forecast future students’ performance, the following year’s data set was used. The results revealed that students’ future performance could be predicted with an accuracy of 83.1% and an F-measure of 87.6% when all of the students’ data were used. The average accuracy improved to 90.9% when each course was used to predict the corresponding next year’s course iteration.

The results of the fourth study offered evidence that learning analytics can use a theory-guided approach to create reliable future predictions. Furthermore, the results provided proof of the power of SNA in operationalizing the activities of students in online collaborative learning environments. Such results present an opportunity to create support strategies that can contribute to better learning and teaching in collaborative learning environments.

Having studied the interactivity constructs in the previous studies, I looked into the time factor in Study 5 (Saqr, Nouri, & Fors, 2018). I hypothesized that temporality might offer a clue regarding students’ self-regulation of their collaborative learning and may help improve the replicability and reproducibility of the predictive models. The study investigated the temporal patterns of interactions in collaborative problem-based learning. The study investigated four patterns of temporality (day, week, course, and year level) over a full academic year, in which there were 2,783 interactions by 185 students in four courses.

In the first part of the study, I used visualization to summarize the interactions in a representative graph for each time period so that any patterns could be observed. Plotting the early weekly interactions demonstrated that high achievers were more likely to participate early and consistently. The same findings were also evident at the course level: early course interactions indicated a pattern of
high achievers, although this was statistically insignificant. However, at the year level, high achievers tended to be active during the first course of the academic year. This finding was reversed by the end of the year and close to the exams, when low achievers participated more often.

Using the temporality variables, I was able to predict high and low achievers with reasonable accuracy. I used data from the first course to predict the second course; the first and second courses to predict the third course; and data from the first three courses to predict achievement in the fourth course. The results showed that temporality data could provide a reasonable predictor of performance, where the class recall of identifying low achievers ranged from 91.7% to 92.9%. The findings of the predictive algorithm confirm an important finding: that timing can be used with a good level of accuracy to predict students’ achievement. Furthermore, given that early participation is the predictor, an early alert indicator may be yielded to stimulate timely intervention.

**RQ2: How can social network analysis be used to analyze online collaborative learning, guide a data-driven intervention, and evaluate it?**

In order to investigate the role of social network analysis in monitoring the efficiency of collaborative learning, visualization was used as a tool to inform students and teachers at a single discussion level, and the whole course level in Study 2 (Saqr, Fors, & Tedre, 2018). The use of social network analysis to map the course interactions enabled the analysis of a large number of interactions among students in a single visualization. The course’s network visualization demonstrated the role of the teacher, which was dominating the discussions. The visualization also enabled the identification of engaged and disengaged students. Such insights offered an obvious opportunity to use the monitoring of collaborative learning in guiding teachers about the status of their students. The study also highlighted the value of visualizing individual discussions as the unit that makes up the full course network, and a possible target for intervention where it becomes dysfunctional or non-participatory. Furthermore, visualizing the information-giving and
receiving networks demonstrated how information exchange could be tracked in a course, as well as how these visualizations can shed light on the key actors of information exchange. Visualization was augmented by mathematical network analysis or centrality measures in four aspects: the quantity of interactivity, the role and position in information transfer, and connectedness as well as group interactivity. Mathematical network analysis offered a precise view of the course network, each user’s position, level of connectedness, and interactions. Such insights are not plausible using traditional interaction analysis methods, which only count the hits or replies while ignoring the importance of structure, relationships and interactions.

Given that the previous study demonstrated that SNA could effectively enable the monitoring of collaborative interactions, the next logical step was to apply the findings. Thus, Study 3 investigated the possibility of using social network analysis to monitor the efficiency of online collaboration among students and to identify the gaps and shortcomings that may need to be addressed. The second aim of the study was to use the insights from the monitoring to create insightful actions that would help improve and evaluate collaborative learning (Saqr, Fors, Tedre, et al., 2018). The studied courses make use of online clinical case discussions to motivate students to engage in collaborative discussion regarding patients’ management, clinical reasoning, and co-construction of knowledge.

The study was divided into three stages. The aim of Stage 1 was to implement a monitoring mechanism, whereby I monitored the three courses using SNA visual and mathematical analysis. Monitoring included the centrality measures relevant to collaborative knowledge construction, namely level of activity, role, and position in information exchange and the role played in the group. The group was also monitored for interactivity and cohesion. In Stage 2, the data were analyzed to identify gaps and areas in need of improvement, before designing an applicable, data-driven intervention. In Stage 3, I evaluated the efficacy of the intervention by comparing pre-to-post intervention and assessing the value of the procedures.
Monitoring revealed a non-participatory pattern of interactions, where the teacher dominated the interactions and few student-to-student interactions were seen. It also revealed limited information exchange or negotiation on the part of the students. This information was used to design an intervention to address the shortcomings identified in the monitoring stage. The intervention included raising awareness, orientation about the goals of collaboration, and training of teachers and students. The clinical case discussions were improved using a collaborative script to stimulate the argumentation and co-construction of knowledge. The evaluation of intervention revealed a significant improvement in collaborative interactions among students, greater number of active students participating in discussions, and increased group cohesion. As such, this study offered evidence of an impactful intervention using learning analytics methods.
12. Discussion

The discussion chapter begins with a general introduction, followed by a discussion of the approach and findings of each research question. Examination of the significance of the findings of each research question will be introduced thereafter. I will subsequently discuss why SNA was chosen over other methods. The final section discusses the difficulties and the limitations of the thesis.

Although learning analytics for the early identification of underachievers can offer an obvious opportunity for educators to implement supportive remedial strategies, the process has to provide insights into the corrective strategy that should be implemented. Moreover, learning analytics research has to contribute to the understanding of learning and how the learning process unfolds. The prime focus of this research was to use learning analytics to understand the collaborative learning process through pedagogically and contextually aligned indicators of learning. It also sought to produce insights to help guide an informed, data-driven intervention with real impact for educators and students alike. I additionally sought to include the different dimensions where learning occurs and evaluate the ways in which they affect learning and learners.

**RQ1:** How can learning analytics and social network analysis reliably predict students’ performance in collaborative learning environments using contextual, theory-guided indicators?

The first study built upon previous research findings (Saqr et al., 2017) and sought to extend data collection by including indicators regarding access to resources, course views, time on task, online interactions, social network parameters, and formative assessments
across different points in time (Gašević et al., 2016; Macfadyen & Dawson, 2010; Tempelaar et al., 2015; Wolff et al., 2014). In order to improve inference, I calculated the engagement sub-indicators (computed indicators that reflect sustained engagement with the online course materials and normalize extreme clicking behavior) as well as access to resources that reflect students’ self-regulation (course objectives and orientation materials, engagement with the tasks, formative assessments and self-evaluation). As such, I used indicators that are more pedagogically aligned with learning and avoided the one-shot approach that merely accounts for the final counts of clicks and views of course pages and resources. This approach highlighted the importance of time, sustainable effort, and regularity. Building the inference of students’ learning on the proxy indicators of engagement and self-regulation has the advantage of being based on a valid pedagogical model, which teachers can use to scaffold, support, and proactively apply a meaningful intervention (Cicchinelli et al., 2018; Reschly & Christenson, 2012).

The subsequent studies sought to increase the resolution of data collection and account for the interactions, relations, connectedness, and the role collaborators in the collaborative learning environment. In Study 2, the results from the exploratory statistical analysis with correlation tests as well as the predictive models across different points of time demonstrated that students’ indegree, role in information exchange and strength of connectedness were significant indicators of performance. However, indicators of the quantity of interactions were not correlated with performance. Such findings demonstrate the importance of the quality rather than quantity of students’ contributions and underscore the role of SNA over traditional interaction analysis, which offers the simple quantification of students’ number of postings (Saqr, Fors, & Tedre, 2018).

In the next study, I used a theory-driven model that accounted for the different factors influencing the learning process, in order to produce a model that may have the advantage of generalizability and that could augment understanding of the learning environment. In order to achieve this goal, the problem-based learning context
was chosen because it is a fairly uniform teaching model (Neville, 2009; Schmidt et al., 2011). Study 4 offered evidence that models can be generalizable when guided by theory and context (Saqr, Fors, & Nouri, 2018). In contrast to studies that face the replicability problem of using SNA in collaborative settings (Dowell et al., 2015; Hernández-García et al., 2015; Joksimović et al., 2016; Jiang et al., 2014), the results of this thesis were reasonably replicated from year to year and from course to course. The reason might be the carefully chosen variables, the operationalization of the context, and the similarity of the teaching methods in the four courses. Due to the potential significance of the time construct and seeking to further explore the dimension of temporality in collaborative learning, I investigated the role of time as a factor that may explain students’ self-regulation (Saqr, Nouri, & Fors, 2018). The results showed that time plays a pivotal role in students’ self-regulation of their learning and can be used as a valid metric to differentiate students according to achievement. A predictive model based on temporality indicators was reliable from one course to the next, which is an important step towards more generalizable and reproducible models.

The online learning environment is rather demanding and requires autonomy and support. This renders the promotion of self-regulation skills of utmost importance to performance (Barnard et al., 2009; Sonnenberg & Bannert, 2015). Research has shown that strategies that target the disengagement of learners can efficiently improve learning and students’ achievement (Ballard & Butler, 2016; Cruz-Benito, Therón, García-Peñalvo, & Lucas, 2015; Tempelaar et al., 2015). Accordingly, interventions that target the promotion of self-regulation can have a significant and positive impact on student academic achievement, as reported in strategies implemented in classrooms, online and in the workplace (de Boer, Donker-Bergstra, Kostons, Korpershoek, & van der Werf, 2013; Kistner et al., 2010; Sitzmann & Ely, 2011; Zumbrunn, Tadlock, & Roberts, 2011). Therefore, it is reasonable to conclude that using learning analytics to enhance self-regulation and to support students is timely and in the best interests of the students and the teachers.
A positive association between performance and social network metrics has been described by different researchers. For instance, degree and closeness centrality have been reported by Cho et al. (2007) and Gašević et al. (2013). Similarly, degree centrality was reported by Hommes et al. (2012), weighted degree by Joksimović et al. (2016), and out-degree and clustering coefficient by Reychav, Raban, and McHaney (2017). Nevertheless, there is a lack of conclusive evidence regarding which social network constructs correlate best with performance (Dowell et al., 2015; Hernández-García et al., 2015; Joksimović et al., 2016). In my thesis, I have attempted to create a broader model that captures the different interactivity constructs in contrast to previous research that has used different centrality measures that only partially account for students’ interactions (Cho et al., 2007; Joksimović et al., 2015; Romero & Ventura, 2013). The value of the current model is that it combines the significant constructs of interactions: quantity of interactions, information transfer, connectedness, and social capital at the individual level as well as in terms of group interactivity and cohesion variables. As such, it lays the groundwork for a structured approach to using social network analysis for a greater understanding of collaboration and the factors behind improved performance. Indeed, the model helped create a predictive model that has been replicated from year to year.

**RQ2: How can social network analysis be used to analyze online collaborative learning, guide a data-driven intervention, and evaluate it?**

Monitoring the interaction using visualization offered an easy-to-interpret general view of the course status of collaboration, and mapped the roles played by teachers and students. In Study 2, SNA visualizations were used as a tool to inform teachers, while the mathematical analysis was used to offer a precise estimation of the properties of the actors and their networks (Saqr, Fors, & Tedre, 2018). Given that the second study offered insights regarding course monitoring, it was important to investigate how SNA can be used to monitor online collaboration, and how the monitoring data can be used to design, apply and evaluate a data-driven intervention.
using social network visual and quantitative analysis (Saqr, Fors, & Nouri, 2018). The monitoring relied on an improved version of the framework developed in the previous study, which included the three constructs for each learner (level of interactivity, information exchange, and role in collaboration) as well as the group parameters. The monitoring of the three courses during the first term of the study revealed a seemingly active (in terms of quantity of interactions) yet non-interactive or participatory learning environment. Moreover, few students played an effective role in information exchange, whereas teachers assumed a central dominating role. The intervention was designed to address such gaps identified in the monitoring stage of the study. The evaluation revealed a significant increase in student-student and teacher-student interactions, improved group cohesion, as well as a participatory pattern of interactions among course participants.

A considerable volume of research exists in the field of collaborative learning using social network analysis. Most studies have addressed the potential to study patterns of interactions in online learning, knowledge construction and the identification of prominent actors in collaborative groups (Cela et al., 2014; Dado & Bodemer, 2017). However, my study is probably one of the first empirical reports regarding the potential of SNA-based interventions in a real-life application (Cela et al., 2014; Dado & Bodemer, 2017). The findings of my study underscore the role of SNA-based indicators as valid and reliable metrics that can guide a real impact in collaborative contexts.

This research project can be considered a rather complete representation of a complete learning analytics cycle. A learning analytics cycle is made up of five interrelated and consecutive stages, as highlighted in Figure 1 and detailed in a separate chapter. The stages of the cycle comprise data capture, preprocessing, analysis, action, and feedback (Clow, 2012b). Data capture, preprocessing and analysis were performed in all of the studies. The analysis of the first study data led to the improvement and refinement of Study 5 (Saqr et al., 2017; Saqr, Fors, & Tedre, 2018). The insights of the monitoring in the first study were further used as feedback to create an intervention in Study 3 and to conceptualize Study 4 (Saqr, Fors,
As such, this thesis is a complete realization of the whole learning analytics research process. The breadth, speed, and feasibility of SNA, as well as the rich information it offers, renders it more practical than traditional, time- and labor-intensive content analysis methods that necessitate lengthy manual coding and subjective analysis (Mayring, 2014; Stemler, 2015). Given that the study included more than 18,000 interactions from around 950 students and educators, manual content coding and analysis of interactions would have been very lengthy and demanding, and also laden with issues such as problematic validity, reliability, trustworthiness, and credibility. Furthermore, it would contradict the spirit of this study, which attempted to present an automated, cost-effective and easy-to-implement method that can help support learners and educators without placing additional burdens on their time or energy. While qualitative content analysis methods are useful as research tools and bring rigor and richness to the analysis of discourse, they remain far from practical to implement in the classroom or in terms of the real-time monitoring of interactions. The difficulty of applying manual content analysis methods on the overwhelming abundance of interactions data is of particular concern and stimulates the search for automated content analysis methods such as EDM (Mayring, 2014; Stemler, 2015).

The inclusion of automated content analysis methods or text analytics in tandem with social network analysis might have added valuable insights to the interaction analysis. However, like manual content analysis, the current status of automated content analytics is not optimal for producing reliable insights into students’ performance or grades. Furthermore, EDM tools are difficult to standardize, sophisticated for non-experts and difficult to interpret by ordinary educators (Al-Razgan et al., 2014; Dutt et al., 2017; Papatitsiou & Economides, 2014; Romero & Ventura, 2013; Slater et al., 2017). I have led some early trials that use automated content analysis tools, but the results were less than satisfactory for performance prediction and difficult to recommend beyond the research setting (AlGhasham et al., 2013; Al-Gadaa et al., 2014;
Saqr, AlGhasham, et al., 2013; Saqr, Kamal, et al., 2014). The most troubling issue I found was the scarcity of tools designed for educational use, especially in problem-based settings. Given that reliable “student modeling commands the focus of more than half the approaches”, new tools that are specific to education must be developed and validated before being used reliably (Peña-Ayala, 2014, pp. 1457).

Although LA has enabled the early prediction of students’ performance as well as the introduction of remedial strategies, a wide array of challenges and issues behind the successful application of the methods remain. A previous section was dedicated to discussing the general challenges. However, it is critical to discuss the challenges identified through this research. I will discuss these in tandem with the learning analytics process framework.

**Data capture**: although it might be expected that computer-recorded data are structured, this was rarely the case. Data recorded by learning management systems demonstrated errors such as corrupted records, incorrect information, missing fields and erroneous encodings of information, as well as an absence of important fields due to software or configuration errors. Therefore, the data were carefully inspected and verified, and the corrupted records or erroneous information were excluded. Furthermore, the process of data extraction was arduous because the tools required for the extraction of raw data were not always readily available. The available tools also necessitated strenuous efforts to keep them current and compatible with the learning environments. For instance, when the thesis project was conceptualized, the intention was to use a group of plugins called “engagement analytics” to extract the Moodle data, but by the time the first study was performed, the Moodle logging system had been completely overhauled and the “engagement analytics” plugins were no longer maintained.

Similarly, SNA data extraction was planned to be performed using a software application called Meerkat-ed, but this became outdated one year before the studies were conducted. Another issue that was discussed in the ethics chapter comprised the ethical and privacy issue of collecting, storing and handling students’ data. At the time when the thesis project was conceptualized, no policy existed to
handle students’ LMS data for research purposes, necessitating the creation of a new policy.

**Data preprocessing:** given the complexity and time required to extract information, the extracted data were not easy to compile, label, and verify. Moreover, as different methods were used to extract the data, the resulting extracts were in different formats and structure. Compiling, as well as verifying and labeling the data, required a serious effort. The quality assurance of data integrity at this stage is of paramount importance; otherwise the results may be flawed.

**Analysis and interpretation:** students’ data were plentiful and perhaps even overwhelming. A large amount of the data recorded by the learning management systems were not relevant to students’ learning and added unnecessary noise to the analysis. For instance, the LMS records every click a student makes, such as clicks on a profile page, editing online personal data or looking at other participants’ information. Moreover, the length of time a student was logged in to the online system is also recorded, regardless of the fact that the student may be idle or multi-tasking. Thus, the simple counts of clicks or views or the total duration of being online are problematic metrics. Therefore, thoughtful identification of the relevant information to be included in the analysis was conducted. This may, however, be easier said than done, because a clear guide does not exist and researchers have to use multiple exploratory and experimental methods before arriving at a useful answer. As data, methods and software evolve, this challenge will continue to be of serious concern. Another serious problem is how algorithms and artificial intelligence work in general. In most of the studies where machine-learning algorithms were used, they proved effective in predicting the future performance of students. However, they offered no indication about the predictors. As such, I used two predictive algorithms to identify the significant predictors.

**Insightful action:** using student data to stimulate action constitutes the basis of learning analytics. However, action can involve numerous parties, including students, teachers, administrators, as well as stakeholders. It requires a considerable effort to change attitudes, practices, and policies. While this thesis project attempted to bring
tangible changes to students’ learning, the change was not easy because it is required at different levels, such as how students’ approach their learning, how teachers deal with students, and how curricula are organized. Data insights may be fast, and so the changes required may not be as responsive as desired. Furthermore, the availability of data regarding students does not always translate into institutional change; in our case, we encountered various levels of resistance at every level during the process of implementing learning analytics.

### 12.1 Limitations

Although the visual and mathematical analysis of collaborative learning provides rich information covering a significant side of collaboration and interaction, it is far from perfect, and considerable improvement is required in order to help streamline students’ online interactions by building better forums that record messages and replies with greater accuracy. Existing forums are confusing: for instance, a reply might be recorded to the thread initiator even if it was directed to another collaborator. In my experience, this issue requires manual correction and adjustment. Moreover, the current mathematical analysis of social networks builds upon interaction models created outside the educational field. Educational researchers may need to review these concepts, and improve and create novel ways of quantifying interactivity at the individual and group levels.

Throughout the five studies, emphasis was placed on the generalizability of the approach so that they may be applied to a broad range of contexts. Nevertheless, the results of these studies are contextually limited, and the generalization of the findings remains a question for further research.

Another limitation of this research project was the relatively small number of students in Study 2, who were selected to demonstrate an easy-to-understand unit of the collaborative structure with a reductionist approach. A general limitation of the five studies may be the exploratory nature of the research approach. As a new research
endeavor, exploratory methods continue to be the main impetus of research and this will possibly continue in future. The lack of content analysis represents another limitation. Indeed, a valid and reliable content analysis method would have added to the insights. However, as argued in the discussion section, such a method was not readily available. Moreover, the methods that are currently available are reused from other domains. Given that the studies were conducted in blended learning environments, the data captured were limited to the online environment. As such, the face-to-face dimension was not represented in this research.
13. Methodological and theoretical contributions

The methodological contribution of this thesis lies in the operationalization of social network analysis in the context of collaborative learning through a structured framework for both mathematical and visual analysis. The framework accounted for the individual learner, connections, interactions, role in information exchange, and the influence of the social structure. The centrality measures outlined by the framework were used to predict a substantial number of potential underachievers with reasonable accuracy, at a time when they can still benefit from early supportive measures. Furthermore, social network analysis was used to monitor, guide, and test an informed intervention that helped produce a significantly more collaborative pattern of learning, offering a proof of concept of the real potential of the technique.

Using a compilation of all interactions to depict dynamic networks represents a new approach that can facilitate understanding of the timeline of events in online collaborative environments. In Study 2, this approach enabled identification of how the online community was built, the initiators of the interactions, and how the interactions were driven. In Study 3, the time-lapse video of interactions demonstrated the differences in course status before and after the intervention.

A further methodological contribution of this thesis is its approach to the study of the time dimension in learning analytics. I used time in the first study as a dimension of self-regulation by stressing regularity. In the fifth study, the patterns of interactions across different periods were structured into a time-based temporality model. The inclusion of the temporality dimension has proven fruitful in
the characterization of the patterns of self-regulation, improving the accuracy of predictive models and bringing important insights regarding a critical aspect of learning that represents a fertile area of inquiry. Furthermore, the methods used in this study have laid a basic framework on which future research can build, improve, and potentially innovate.

Another notable characteristic of this thesis was its focus on the reproducibility and validation of results. In each of the five cases studied, the validation of results was central to the method of analysis, including resampling, permutation, 10-fold cross-validation, and data from different courses to replicate the results. Although the replication of results in learning analytics remains a challenging task, it was achieved with reasonable reliability using a combination of temporality and interaction metrics. The course data were used to reliably predict future underachievers in the subsequent courses. This finding offers evidence of the possibility of creating more reliable models that can enhance the portability of models, add to the understanding of theory, and embolden trust in the field.

The research presented in this thesis contributes to evidence regarding learning analytics in three main areas: the creation of learner models, support for teaching, and impact on teaching (Ferguson & Clow, 2017). The results of research concerning learners proved that learning analytics could be used to create informative learner models regarding self-regulation, collaboration, and time strategies. The results have also proven that collaboration analysis can be used to monitor, support and improve collaborative learning.

The research articles presented in this thesis are among the first empirical research studies in the field of medical education in terms of learning analytics and social network analysis. They are presented with the hope that they might stimulate dialogue about opportunities for a data-driven approach to learning, as well as the challenges that the medical education community needs to discuss (Isba et al., 2017; Saqr, 2018).
14. Conclusions

The findings of these studies have demonstrated that a significant number of underachievers can be early predicted using learning analytics methods. Metrics of self-regulation, engagement, interactivity, and temporality were the most significant indicators of better performance. Educators can use these indicators (which are essentially modifiable and malleable) to provide interventions and early remedial strategies. A theory-grounded approach to data collection, analysis and interpretation has been proved both feasible and fruitful.

Time and temporality are essential dimensions of how students self-regulate their learning. Students’ early and proactive engagement with their tasks, sustained effort and consistency proved to be an important potential predictors of performance. Moreover, using time-based social network analysis has highlighted the social dynamics of collaborative learning, especially regarding how students engage over time and how the timeline of interactions evolves.

Using social network visualization to summarize the interactions, the relationships and structure of collaborative groups have presented a feasible and understandable perspective of the status of collaborators and groups. Through combining visualization with mathematical network analysis and centrality measures, an accurate view of collaboration was derived. Using social network analysis to monitor and guide a data-driven intervention proved feasible, cost-effective and efficient. SNA-based interventions helped create more cohesive and interactive groups with fewer isolated collaborators.
A framework that captures the gamut of interactivity and relationships and the position of collaborators has helped create a reasonably accurate picture of the collaborative environment. In the process, this has facilitated the prediction of underachievers, as well as the monitoring and correcting of inefficient collaborative practices.
16. Future work

This research might stimulate future research in numerous ways. An analysis of self-regulation could be performed using finer-grained, high-resolution data points. The inclusion of fine-grained activities such as highlighting concepts, taking notes, or creating learning artifacts may contribute to better contextual inferences (Cicchinelli et al., 2018; Molenaar & Järvelä, 2014; Winne et al., 2017). The abundance of such thorough information could be used to create an accurate contextual timeline of students’ activities, strategies for learning and self-regulation (Sonnenberg & Bannert, 2015; Winne, 2017). Furthermore, it might facilitate understanding of the different factors behind self-regulation that may need to be targeted via intervention or support (Molenaar & Järvelä, 2014; Shum & Crick, 2012; Tempelaar et al., 2018).

The focus of this research was on online learners’ interactions. However, students also interacted with learning resources. Using social network analysis to map the interactions with different resources, concepts, and competencies could offer a detailed map of the interconnected networks of learning competencies that enable us to understand the full picture of learning. Using network analysis along with self-regulation and the temporality of students might provide further insights into the learning process.

Another venue for future research may be the introduction of centrality measures that are specific to collaborative learning and information exchange in educational settings. Although centrality measures are numerous and often relevant to learning, situations occur where present centrality measures lack the necessary specificity. An example is the lack of a centrality measure that correlates with argumentation in discussions. While betweenness centrality
can serve as a measure here, it does not represent an accurate measure of such property. Many unanswered questions remain, including the breadth of tracing, the temporality of self-regulation in the long term beyond time-on-task, the optimum delivery of online resources that do not place excessive stress on learners, and the online strategies that could be used effectively to support learners (Winne, 2017). Additional research venues that might be explored in the future are listed below:

- The development and validation of a content analysis method specific to an online problem-based learning context;
- The further development of the social network analysis framework with the inclusion of dynamic and temporal parameters;
- The study of signed and weighted networks in the context of online collaboration;
- Using temporal generative network analysis models to understand why and how ties are formed and maintained or dissolved over time.
References


PBL Based Medical Course. Paper presented at the AMEE2013.
Avella, J. T., Kebritchi, M., Nunn, S. G., & Kanai, T. (2016). Learning analytics methods, benefits, and challenges in
higher education: A systematic literature review. *Online Learning, 20*(2), 13-29.


systematic review. *Educational Psychology Review, 27*(1), 219-246. doi:10.1007/s10648-014-9276-0


international conference on learning analytics and knowledge.


Proceedings of the seventh international learning analytics & knowledge conference.


conferencing in higher education. *The Internet and Higher Education, 2*(2), 87-105.


Alberta, Canada: Society for Learning Analytics Research (SoLAR). DOI: 10.18608/hla17.007


Ma, J., Han, X., Yang, J., & Cheng, J. (2015). Examining the necessary condition for engagement in an online learning environment based on learning analytics approach: The role of the instructor. The Internet and Higher Education, 24, 26-34.


collaboration analysis support adapted to the participants in CSCL situations. *Computers & Education*, 82, 335-353. doi:10.1016/j.compedu.2014.10.027


Educational research, 81(1), 4-28.
doi:10.3102/0034654310393361
International Conference on Learning Analytics and Knowledge.


Studies included in the thesis