Discovering Interaction Patterns in Online Learning Environments: A Learning Analytics Research

THÈSE N^{O} 8238 (2018)

PRÉSENTÉE LE 13 AVRIL 2018

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

LABORATOIRE D'ERGONOMIE ÉDUCATIVE

PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Mina SHIRVANI BOROUJENI

acceptée sur proposition du jury:

Dr D. Gillet, président du jury
Prof. P. Dillenbourg, directeur de thèse
Prof. H. U. Hoppe, rapporteur
Prof. A. Wise, rapporteuse
Prof. R. West, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2018

To my parents,

for their endless love, support, and encouragement

And to Sina, with all my love . . .

Acknowledgements

This thesis would not have been completed without the help and support of many people around me to whom I am extremely grateful. First and foremost, I would like to express my gratitude to my thesis supervisor **Prof. Pierre Dillenbourg**, who continuously and consistently provided me with inspiration, encouragement, and advice, whilst giving me the freedom to develop and realize my own ideas. His unrelenting support throughout this process has made this Ph.D an enjoyable and valuable experience.

I would like to extend my gratitude to the members of my review committee: **Prof. Alyssa Friend Wise**, **Prof. Ulrich Hoppe**, and **Prof. Robert West**, for the time and effort invested in the evaluation of this thesis. Their insightful comments helped me improve the final manuscript.

I am sincerely thankful to **Kshitij** for his valuable contribution in the development of this research and its final form. The quality of this thesis was improved thanks to his careful review and constructive feedback. I am grateful for our discussions and appreciate his readiness to offer counsel and support in varied subjects. I further thank **Łukasz Kidzinski**, working with whom, even though for a short period of time, was a great opportunity for me. I am also thankful to **Francisco Pinto** and **Patrik Jermann** for making this research possible by providing me access to the MOOC datasets.

Special thanks to **Lorenzo Lucignano**, for always being available to discuss my research and results, and for being a great source of creative and inspiring ideas. It was a pleasure for me to share all the different steps in this Ph.D with him, from the beginning to the end. Thank you for your friendship and support over the past four years.

I have been lucky to be surrounded by wonderful colleagues and friends in CHILI laboratory who made my time in the office and outside of it more joyful. The times we had together and their support, make them invaluable to me whether they were directly involved in my research or not. I am thankful to **Ayberk Özgür**, **Kshitij**, **Lorenzo Lucignano**, **Łukasz Kidzinski**, **Mirko Raca**, **Himanshu Verma**, **Luis Prieto**, **Wafa Johal**, **Elmira Yadollahi**, **Khalil Mrini**, **Stian Håklev**, **Hamed Alavi**, **Nan Li**, **Kevin Gonyop**, **Catharine Oertel**, **Thibault Asselborn**, **Alexis Jacq**, **Arzu Göneysu Özgür**, **Konrad ŻoŁna**, **Jennifer Olsen**, **Teresa Yeo**, **Fu-Yin Cherng**, **Louis Faucon**, **Shruti Chandra**, **Julia Fink**, **Sébastien Cuendet**, **Jessica Dehler Zuf-**

ferey, **Daniela Caballero**, **Sophie Dandache**, **Séverin Lemaignan**, **Maria Jesus Rodriguez-Triana**, **Konrad Żołna**, **Sophia Schwar**, **Tuhina Dargan**, **Khuyen Duong**, **Alexander Nebel**, and **Thanasis Hadzilacos**, for all the funny conversations shared over lunch, coffee, concerts, and much more. I am sincerely grateful to **Florence Colomb**, the "backbone" of the lab, for being so supportive and making all the administrative processes as smooth as possible.

I was fortunate to meet many wonderful people during my stay in Lausanne and make lifelong friends. Many thanks to **Samaneh**, **Mehran**, **Masoumeh**, **Elham**, **Mohammad**, **Elahe**, **Reza**, and **Mostafa** for being my extended family away from home, and making my days in Lausanne memorable and unforgettably happy days of my life. I am further thankful to my especial friends whose support and friendship have been with me for long years despite the long distance between us. Thank you **Maryam**, **Elham**, **Zahra**, **Hadiseh** for your warmth, sincerity, and presence.

I would never find the right words to express my gratitude to my parents, **Shahin** and **Esmaeil**. I remain thankful to you for your everlasting love, dedication, sacrifice, and infinite support that always gave me the strength to keep moving forward. Thank you for believing in me more than I ever did and for giving me the courage to purse my dreams. Without your guidance, support, and compassion I would never be able to succeed in this path and be the person who I am today. I am deeply thankful to my brother, **Ali**, my sister, **Nargess** and my brother-in-law, **Farsad**. Thank you for being always there for me, even when hundreds of kilometers away. I have always been proud of you and I hope with this work I make you feel the same.

To my extended family, specially to my wonderful parents-in-law, **Behnaz** and **Razi**. I am deeply grateful for your love and support and I truly appreciate your kindness and encouragement in the last stages.

Last but not the least, my heartfelt and forever thanks goes to my beloved husband and my best friend, **Sina**. Thanks for your unfailing love and kindness, your exceptional understanding, and great patience. Going together with you through the Ph.D adventure was the best part of this journey, which was made possible with your constant encouragement and support.
I dedicate this thesis to Sina, my family, and the memory of my grandfather.

Lausanne, 2018-02-09 Mina Shirvani Boroujeni

Abstract

The increasing amount of data collected in online learning environments provides unique opportunities to better understand the learning processes in different educational settings. Learning analytics research aims at understanding and optimizing learning and the environments in which it occurs. A crucial step towards this goal is to adapt and develop adequate computational methods to process and analyze the learning-related data and to present information in intelligible ways to educational stakeholders.

In this thesis we investigate interaction patterns of learners in two different online learning environments: Massive Open Online Courses (MOOCs) and Realto, an online platform for integrated Vocational Education and Training (VET). We analyze interaction patterns across three principal dimensions: time, activity, and social. To obtain a better understanding of the complex learning behaviours, it is essential to consider these different aspects of the educational data. We develop novel methods and use existing techniques from sequential pattern mining, content analysis, and social network analysis to model and track interaction patterns of learners across the three mentioned dimensions.

As regards the **time dimension**, we present methods to model temporal patterns of learners' participation. We introduce novel techniques to discover and quantify online regularity in terms of following a certain daily or weekly time schedule. We investigate the relation between students' regularity level and their performance in a MOOC course.

Concerning the **activity dimension**, we analyze learners' activity sequences in order to identify and track the evolution of their study approaches over time. By clustering study pattern sequences in a MOOC course, we extract different engagement profiles among learners and describe their properties. Furthermore, we propose a complete processing pipeline for the unsupervised discovery of study patterns from sequential interaction logs. This pipeline is applicable at different levels of actions granularity and time resolution and enables to perform temporal analysis of learners' interaction patterns throughout the course duration.

For the **social dimension**, we explore the attributes of social interactions among learners. In the MOOC context, we combine content and social network analyses to study dynamics of forum discussions and the evolution of students' roles over time. In the context of Realto, we employ social network analysis to model the social interactions among learners and to study the structure of Realto-mediated communication among different stakeholders in the VET system.

Using the presented analytic methods, we provide novel insights into the interaction patterns of learners in MOOCs and Realto. Moreover, we present an implementation of these methods into an analytics dashboard for Realto researchers.

Keywords: Learning analytics, Educational data mining, Temporal analysis, Sequential pattern mining, Social network analysis, Massive Open Online Courses, MOOCs, Educational dashboards, Vocational education and training.

Résumé

La quantité croissante de données collectées dans les environnements d'apprentissage en ligne offre des opportunités uniques pour une meilleure compréhension des processus d'apprentissage dans différents contextes éducatifs. Le domaine de recherche "learning analytics" vise à comprendre et à optimiser l'apprentissage et les environnements dans lesquels il se produit. Une étape cruciale vers cet objectif est d'adapter et de développer des méthodes computationnelles adéquates pour traiter et analyser les données liées à l'apprentissage et pour présenter l'information de manière intelligible aux acteurs de l'éducation.

Dans cette thèse nous étudions les modèles d'interaction des apprenants dans deux environnements d'apprentissage en ligne différents : les cours en ligne ouverts à tous (MOOCs) et Realto, une plateforme en ligne pour la formation professionnelle connectée. Nous analysons les modèles d'interaction à travers trois dimensions principales la temporalité, les activités, et les interactions sociales. Pour obtenir une meilleure compréhension des comportements d'apprentissage complexes, il est essentiel de considérer ces différents aspects des données éducatives. Nous développons de nouvelles méthodes et utilisons les techniques existantes de l'extraction de modèles séquentiels, d'analyse de contenu et d'analyse de réseaux sociaux pour modéliser et suivre les modèles d'interaction des apprenants dans les trois dimensions mentionnées.

En ce qui concerne la **dimension temporelle**, nous présentons des méthodes pour décrire les schémas temporels de la participation des apprenants. Nous introduisons de nouvelles techniques pour découvrir et quantifier la régularité en ligne en termes de suivi d'un certain horaire quotidien ou hebdomadaire. Nous étudions la relation entre le niveau de régularité des étudiants et leur performance dans un cours MOOC.

Concernant la **dimension de l'activité**, nous analysons les séquences d'activités des apprenants pour identifier et suivre l'évolution de leurs approches de leurs études au fil du temps. En groupant les modèles d'étude d'un cours MOOC, nous extrayons différents profils d'engagement parmi les apprenants et décrivons leurs propriétés. De plus, nous proposons un algorithme d'apprentissage non-supervisé qui identifie le stratégies d'apprentissage à partir des traces d'interactions. Ce pipeline est applicable à différents niveaux de granularité des actions et de résolution temporelle et nous permet d'effectuer une analyse temporelle des modèles d'interaction des apprenants tout au long de la durée du cours.

Pour la **dimension sociale**, nous explorons les caractéristiques des interactions sociales parmi

les apprenants. Dans le contexte du MOOC, nous intégrons le contenu et l'analyse de réseaux sociaux pour étudier la dynamique dans les forums de discussion et l'évolution des rôles des étudiants au fil du temps. Dans le contexte de Realto, nous utilisons l'analyse de réseaux sociaux pour modéliser les interactions sociales entre apprenants et pour étudier la structure de la communication via Realto entre les différents acteurs du système de formation professionnelle.
En utilisant les méthodes d'analyse présentées, nous proposons de nouvelles façons de modéliser les interactions des apprenants dans les MOOCs et Realto. De plus, nous présentons une implémentation de ces méthodes dans un tableau de bord analytique pour les chercheurs de Realto.

Mots-clé : Analytique de l'apprentissage, Exploitation de données éducatives, Analyse temporelle, Exploitation de modèles séquentiels, Analyse des réseaux sociaux, Cours en ligne ouvert à tous, Tableaux de bord éducatifs, Enseignement et formation professionnels.

Contents

Acknowledgements i

Abstract iii

Résumé v

Contents vii

List of figures xi

List of tables xv

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Contexts . . . 2
- 1.3 Objectives . . . 3
- 1.4 Thesis roadmap . . . 4

2 Related Work 7
- 2.1 Learning analytics and educational data mining . . . 7
- 2.2 LA models . . . 8
- 2.3 LA objectives . . . 9
- 2.4 LA methods and applications . . . 12
  - 2.4.1 Prediction . . . 12
  - 2.4.2 Structure discovery . . . 13
  - 2.4.3 Relationship mining . . . 15
  - 2.4.4 Visualization . . . 17
  - 2.4.5 Discovery with models . . . 18
  - 2.4.6 Knowledge tracing . . . 18
- 2.5 Discussion . . . 18

3 Realto: Online Platform for Integrated Vocational Education 21
- 3.1 The Swiss vocational education and training system . . . 21
- 3.2 School and company gap . . . 22
- 3.3 Realto: the online platform for integrated VET . . . 24

3.3.1 Awareness tools in Realto . . . 25
3.4 Discussion . . . 30

4 Temporal Patterns of Online Participation 31
4.1 Context . . . 31
4.1.1 Time factor in educational research . . . 31
4.1.2 Time flexibility in online education . . . 32
4.1.3 Temporal analysis in educational research . . . 33
4.2 Problem formulation . . . 34
4.3 Method . . . 35
4.3.1 Regularity patterns . . . 35
4.3.2 Design of measures . . . 35
4.4 Temporal participation patterns in MOOC . . . 43
4.4.1 Dataset . . . 43
4.4.2 Examples of regularity measures . . . 44
4.4.3 Correlation between regularity measures . . . 48
4.4.4 Regularity and performance . . . 48
4.4.5 Other applications . . . 53
4.5 Temporal participation patterns in Realto . . . 53
4.6 Discussion . . . 56

5 Activity Patterns of Online Interactions 59
5.1 Activity patterns in MOOCs . . . 60
5.1.1 Context . . . 60
5.1.2 Problem formulation . . . 61
5.1.3 Dataset . . . 62
5.1.4 Hypothesis-driven approach . . . 63
5.1.5 Data-driven approach . . . 68
5.2 Activity patterns in Realto . . . 78
5.3 Discussion . . . 80

6 Online Social Interactions 83
6.1 Social interactions in MOOC . . . 84
6.1.1 Context . . . 84
6.1.2 Problem formulation . . . 86
6.1.3 Dataset . . . 87
6.1.4 Forum activity over time . . . 87
6.1.5 Discussion content over time . . . 88
6.1.6 Social communication structure . . . 91
6.1.7 Social structure and discussion content . . . 105
6.1.8 Predicting forum activity level . . . 106
6.2 Social interactions in Realto . . . 109
6.2.1 Method . . . 109

6.2.2 Results 110
6.3 Discussion 115

7 General discussion **119**
7.1 Summary 119
7.2 Contributions 120
7.2.1 Methods for interaction pattern analysis 120
7.2.2 Analytic tools for Realto platform 121
7.3 MOOC related findings 122
7.4 Reflections 123
7.5 Limitations and future work 125

A Realto analytics dashboard **127**
A.1 Time analysis module 127
A.2 Activity analysis module 128
A.3 Social analysis module 129

Bibliography **151**

Curriculum Vitae **153**

List of Figures

1.1 Thesis roadmap: Chapter 4 focuses on temporal patterns of learners' participation, Chapters 5 and 6 respectively focus on activity and social interaction patterns and their evolution over time. 5

2.1 Learning analytics reference model proposed by Chatti et al., taken from [49] . 9
2.2 Learning analytics life cycle (right) and constraints (left) proposed by Khalil et al., taken from [118] 10

3.1 The Swiss educational system and possible pathways [197]. 22
3.2 The Erfahrraum model: a pedagogical model to inform the design of technology-enhanced VET learning activities [193] 23
3.3 Realto: The online learning platform for integrated vocational education. . . . 24
3.4 Flow overview in Realto awareness dashboard 27
3.5 Individual view in Realto awareness dashboard 27
3.6 Comparison view in Realto awareness dashboard 29
3.7 Post view in Realto awareness dashboard 29

4.1 Example of a learner's binarized daily time signal F_{24}. Points with value of one represent days on which the learner had online activities in the course platform. 37
4.2 Examples of weekly and daily activity histograms for one learner, respectively representing the distribution of study time over week days and day hours (time zones are not compensated and the start time of the course is considered as the reference point with $t = 0$). 38
4.3 Example of weekly profile matrix for a learner. Rows represents weeks, columns represent days and intensity of colors encodes estimated amount of study time (in hours). 40
4.4 Example of daily activity signal in time (left) and frequency (right) domains for one learner with a regular daily pattern. Dashed red line in the periodogram shows the frequency corresponding to one week period (1/7). 42
4.5 Examples of CDH and CWD measures: Daily (top) and hourly (bottom) histograms of four learners with high and low values. Clearly a high value of CDH and CWD reflects a peak in the corresponding activity histograms. 45

4.6 Example of WSB, WSN and WSR measures: (a) Binary (b) Normalized and (c) Raw weekly study profiles of a learner with high values for the three profile similarity measures. 46
4.7 Example of WSB, WSN and WSR measures: (a) Binary (b) Normalized and (c) Raw weekly study profiles of a learner with low values for the three profile similarity measures. 46
4.8 Example of WSB, WSN and WSR measures: (a) Binary (b) Normalized and (c) Raw weekly study profiles of a learner with different values for the three profile similarity measures. 46
4.9 Examples of PWD measure: Daily activity signal in time (left) and frequency (right) domains for two learners with (a) high and (b) low values of PWD. Dashed red lines in the periodograms show the $(1/week)$ frequency. The periodic daily pattern is clearly reflected by the high value of PWD measure. 47
4.10 Example of PDH and PWH measures: Hourly activity signal in time (left) and frequency (right) domains for a learner with high values of PWH and PDH. Dashed red and solid blue lines in the periodogram respectively show the $(1/week)$ and $(1/Day)$ frequencies. 47
4.11 Scatter plots of grade vs. WSN and PWD measures. Red lines shows the linear smoothed estimations and the blue curves represent the local smoothed estimations. The gray areas show the 0.95 confidence intervals. 49
4.12 Interaction effect between weekly regularity (WSN) and total study time in regression model of final grade. 50
4.13 Clusters of learners based on regularity measures. All values were scaled to [0,1] for visualization purpose. Learners in **Cluster 1** are not-regular but are responsive, **Cluster 2** are weekly regular and responsive, **Cluster 3** are daily regular and responsive, **Cluster 4** are not-regular and not-responsive. 51
4.14 Average grade of clustered learners. Daily/weekly regular and responsive learners (Cluster 2 and 3) achieve significantly higher grades. 52
4.15 Regularity over time for passed and failed students. The gray area shows the 95% confidence interval. 53
4.16 Examples of weekly histograms and CWD measure in Realto. 55
4.17 Examples of daily histograms and CDH measure in Realto. 55
4.18 Examples of weekly profile matrix and WSB measure in Realto. 55
4.19 Example of average weekly and daily histograms for apprentices in a dual-tack apprenticeship. Thursday is the school day in this case. 56
4.20 Example of average weekly and daily histograms for apprentices in a single- tack apprenticeship. 56

5.1 Hypothesis-driven study patterns, (a) Overal frequency and (b) distribution over different assessment periods in the MOOC dataset. 64

5.2 Transition probabilities between different study patterns for learners who change their approach over time. Node size is proportional to the pattern frequency and edge thickness is proportional to the transition probability. . . . 65
5.3 The average activity models for the four resulting clusters in the first simulated scenario (cluster merge). In the transition diagrams, node color intensity is proportional to the action probability and edge thickness is proportional to the transition probability. The resulting clusters, correctly capture the four simulated behaviours. . . . 72
5.4 Sequences of the learners' interaction patterns over 50 sessions in the four simulated scenarios. In the sequence charts, each horizontal line represent the interaction pattern sequence of one learner. Our proposed pipeline correctly captures the changes in cluster count and size and detects clusters merge, split, dissolve, and form. . . . 73
5.5 Transition probabilities between data-driven study patterns in MOOC dataset. Node size is proportional to the pattern frequency and edge thickness is proportional to the transition probability (Transitions with probability smaller than 0.1 are not displayed). . . . 75
5.6 Activity sequence of florist apprentices (top) and teachers (bottom) in Realto. . 79
5.7 Clusters of Realto usage patterns by apprentices (top) and teachers (bottom) in clothing design profession. . . . 80

6.1 Number of messages created per day in $Scala$ and $Reactive$ courses. Vertical lines represent video release days (dashed blue) and assignment deadlines (solid red). . . . 88
6.2 Average number of new messages depending on the proximity to video release (left) and assignment deadline (right) in Scala and Reactive courses . . . 89
6.3 Distribution of content-related posts over time, in Scala and Reactive courses. Vertical lines represent video release days (dashed blue) and assignment deadlines (solid red). . . . 91
6.4 Examples of domain-specific concepts, and their distribution over time in Scala and Reactive courses. Vertical lines represent video release days (dashed blue) and assignment deadlines (solid red). . . . 92
6.5 Example of structural and regular equivalent sets . . . 95
6.6 Blockmodel representation of example network in Figure 6.5, according to structural and regular equivalence based clustering. . . . 96
6.7 Network attributes over time, based on one week network slices using sliding window in Scala and Reactive courses. Vertical lines represent video release days (dashed blue) and assignment deadlines (solid red). . . . 98
6.8 Role structure of information exchange network in bi-weekly slices . . . 100
6.9 Structural role sequences in Scala and Reactive courses. Each horizontal line denotes the role sequence for one learner over four bi-weekly time periods. Label of the vertical axis represent total number of forum participants. . . . 102

6.10 Clusters of role sequences in Scala and Reactive course. Each horizontal line denotes the role sequence for one learner over four bi-weekly time periods. Vertical axis labels denote number of learners in each cluster. 104
6.11 Average grades by participants in role sequence clusters 105
6.12 Average number of keywords in posts by user in structural role clusters for *Scala* course 106
6.13 Average number of keywords in posts by user in structural role clusters for *Reactive* course 106
6.14 Example of Realto communication network for participants in one school . . . 111
6.15 Teacher-apprentice sub-network: connections between apprentices and teachers 113
6.16 Supervisor-apprentice sub-network: connections between apprentices and supervisors 113
6.17 Role structure of Realto network 115

List of Tables

4.1 Regularity patterns in time domain and examples. 36
4.2 Overview of regularity measures and corresponding regularity patterns they reflect. 36
4.3 Overview of regularity measures in the dataset 44
4.4 Linear model for final grade estimated using regularity measures 49
4.5 Linear model of final grade based on the interaction between weekly regularity and amount of study time. 50
4.6 Comparison of weekly and daily regularity measures for Cluster 2 and 3 in Figure 4.13, using one-way Anova test without assuming equal variances. 51

5.1 Clusters of study pattern sequences extracted using hierarchical clustering. Cluster 1 to 3 represent learners with fixed study approach during the course. Cluster 4 to 11 represent categories of role sequences for learners who change their approach over time. Vertical axis in the pattern sequence charts represent students in each cluster and horizontal axis represent assignments. Note that the charts height is **not** proportional to the cluster size. Other columns in this table represent cluster size, average final grade of cluster members, ratio of passed students in each cluster, and description of the study pattern profiles. 67
5.2 Estimated number of clusters and list of identified study patterns in each assessment period. New patterns in each period are highlighted in bold blue font. Study patterns are described in Table 5.3 74
5.3 Data-driven study patterns extracted from MOOC learners interaction logs. For each pattern, transition diagrams (left) show the average activity model (node color intensity is proportional to the state probability and edge thickness is proportional to transition probability). The grid charts (right) show the 20 most frequent daily state sequences for each study pattern. Horizontal axis in sequence charts represent days in the assessment period and rows represent sample sequences (row height is proportional to the sequence frequency). In the patterns description, **N** represents the frequency of each pattern and **AE** is the average error (average distance between activity models and cluster mean vector). 76

6.1 Dataset overview 87
6.2 Examples of indicator phrases used to track discussion topics over time 90

6.3 Examples of classified posts 93
6.4 Overview of overall knowledge exchange network attributes 97
6.5 Attributes of structural roles in information exchange network slices for Scala course 101
6.6 Attributes of structural roles in information exchange network slices for Reactive course 101
6.7 Description of features used in the predictive model 107
6.8 Overview of predictive models and results 108
6.9 Distribution of user roles and centrality measures in Realto social network . . . 111
6.10 Users distribution and reciprocity of Realto sub-networks. 112
6.11 Attributes of structural roles in Realto network 115

1 Introduction

1.1 Motivation

In recent years, there has been a growing interest in Learning Analytics (LA) and Educational Data mining (EDM) among educational researchers and practitioners. LA refers to the process of collecting and analyzing data from learners and their contexts, to understand and optimize the learning processes, environments, and pedagogical scenarios; and EDM is concerned with developing and applying methods to detect patterns in large amounts of educational data [77]. Recent works in LA and EDM research, employ techniques from data-driven fields like Data Mining and Machine Learning to gain insights on learners' behaviors, interactions, and learning paths. Such data-driven insights, enable a better understanding of different learning processes in a variety of educational settings and could provide the basis to support learners, teachers, and educational institutions.

In the recent past of LA research, a typical question would be how to accurately predict students' performance and provide early indicators of whether a learner is likely to successfully complete a course or a study program. As it becomes possible to track the behaviour of learners and teachers in digital learning environments, the landscape of LA research is getting broader and the variety of questions that are being asked of the data are getting richer [131]. With the increasing popularity of online learning environments and emergence of MOOCs, a large volume and variety of educational data is being collected which often include a granular record of learners' interactions with different learning materials. The collected sequential data provide opportunities to gain insights into the learning behaviours and latent traits of students.

In this thesis, we seek to contribute to the growing body of LA research that aims to utilize the student-generated actions data for modeling their behavioral patterns and attributes. Our research is focused on modeling, quantifying, and representing the interaction patterns in online learning environments. The outcome of this research could provide opportunities towards the design of personalized and adaptive learning environments which could tailor the learning activities to the needs and attributes of the students. Moreover, by identifying the

interaction patterns of successful students it could be possible to provide recommendations to the struggling learners to optimize their learning pathways.

In order to model and analyze learners' interaction patterns, we consider different aspects of the educational data. In particular we study the temporal patterns in learners' online participation and timing of their study sessions (time dimension), the type and sequence of performed actions (activity dimension), and the social interactions among learners (social dimension). Most learning analytics research focus only on one (or sometimes two) of these dimensions to describe learners' activities. Taking into account different aspects of the educational data in this thesis, could enable us to gain a better understanding of the learning behaviours and the modeled processes.

In our analysis we go beyond static analysis of certain situations or fixed snapshots of the data and address the temporal and sequential dynamics of learners' interactions. Consideration of the temporal aspects is fundamental for better understanding the learning processes. However, this aspect of educational data is yet under-explored in learning analytics research [50]. As an step towards filling this gap, in all the different studies in this thesis we particularly focus on the temporal attributes of learning behaviours. More specifically, we study how learners' participation and activity patterns, their social interactions, and also the roles they undertake in the social communications change over time. This continuous attention on the time dimension could be a valuable contribution to the learning analytics research.

1.2 Contexts

We investigate learners' interaction patterns in two different learning environments: MOOCs and Realto.

- **MOOCs**: MOOCs deliver online learning on a wide variety of topics to a large number of participants across the world. With minimal cost (often zero) and no entry barriers (e.g. prerequisites, or skill requirements) a large number of learners with different backgrounds can engage in learning opportunities that are often curated by leading universities. MOOC courses offered through platforms such as edX[1], Coursera[2], and Udacity[3], comprise different learning materials such as video lectures, reading materials, in-video quizzes, and assignments. In addition, discussion forums are a key part of MOOC platforms to support social learning and peer-to-peer information exchange. According to self-report surveys, learners have different motivations for enrolling and participating in a MOOC. Personal interest in learning certain topics, learning topics relevant to the current job, and earning a certificate for future career opportunities or university credits are among the main MOOC enrollment reasons[138].

[1] `https://www.edx.org/` (last accessed 10 December 2017)
[2] `https://www.coursera.org` (last accessed 10 December 2017)
[3] `https://www.udacity.com/` (last accessed 10 December 2017)

- **Realto**: Realto[4] is an online platform developed in Dual-T[5] project, the research program in which this the thesis is framed. In a nutshell, the Dual-T project aims at bridging the dual contexts of vocational education, school and workplace, through learning technologies which allow apprentices to connect workplace experiences to classroom instructions. Realto platform is implemented towards this goal to bridge between different contexts and stakeholders in vocational education. This platform provides a digital space for sharing experiences captured in different learning locations and consequently, feeding them into reflective activities.

In this thesis we aim to provide analytic methods that are applicable in both MOOC and Realto contexts. Considering the different pedagogical settings, objectives, and target users in MOOCs and Realto, the findings in one context might not be directly transferable to the other. On the other hand, it could be possible to apply similar analytic methods to investigate interaction patterns of learners in both settings. Our approach in this thesis is to use MOOC data to develop such analytic methods and demonstrate their application and value to MOOCs context. We will then adapt these methods to the context of Realto in order to provide tools for modeling and analyzing interaction patterns of participants in this platform. One important advantage of this approach is that the large volume and variety of interaction data available in MOOCs, enables us to examine different methods and find the most suitable ones for capturing learners' behavioral patterns from interaction sequences. This is of particular importance given that at the current stage, Realto is not populated with large number of participants who actively use it in their learning practices, and therefore not enough data is available to develop the analytics tools for this platform.

1.3 Objectives

The main objectives of this thesis could be summarized as follows. These objectives will be translated into concrete research questions in the following chapters.

- **Providing analytic methods to model learners' interaction patterns and analyze temporal dynamics of learning behaviours**: Being able to collect larger volumes and varieties of educational data is only one of the necessary steps for understanding the learning behaviours. In order to transform the low-level interaction logs into interpretable indicators and models, it is essential to adapt computational methods or develop new techniques to process and analyze the collected data. In this thesis we aim to provide analytic methods and processing pipelines, generalizable to different context and platforms, to model and investigate learners' interaction patterns over time. We present methods and introduce novel metrics of learners' behaviours, which integrate the time, activity, and social dimensions. Towards this goal, we adapt, expand, and combine established techniques such as clustering, optimal sequence matching, and

[4] `https://www.realto.ch/` (last accessed 10 December 2017)
[5] `https://dualt.epfl.ch/` (last accessed 10 December 2017)

social network analysis, but also develop entirely new approaches, such as methods for modeling temporal patterns and assessing online regularity.

- **Providing analytic tools to monitor and analyze participation patterns in Realto**: In order to enable monitoring, modeling, and analyzing users' activities and interactions in Realto, this platform needs to be equipped with adequate analytic tools. Providing such tools is the other objective of this thesis. We intend to integrate into an analytics dashboard, methods and indicators that enable researchers to understand how Realto is being used by different user groups. This in turn could support data-informed planning and decision making towards improving the features of this platform. In addition, we will implement awareness tools for teachers, in the form of an interactive dashboard that integrates indicators of students' activities in Realto.

The role of MOOC scenarios and Realto environments in this thesis are indeed different and in a way complementary. The large number of learners with variety of learning styles in MOOCs and continuous records of their interaction traces, provides us with the data required to develop analytic methods for modeling the interaction patterns. We use MOOC data to introduce and assess new methods and demonstrate their application and value in providing novel insights on temporal dynamics of learners' behaviour. On the other hand, transferring these methods to Realto for developing analytics tools for this platform is more of a system engineering and application challenge in this work.

1.4 Thesis roadmap

The remainder of this thesis is organized as follows. In the next chapter (**Chapter 2**) we provide an overview of the previous research in the field of learning analytics, focusing on the main objectives, methods, and applications.

In **Chapter 3** we present the Swiss vocational eduction system and describe the dual approach of school-workplace training. We then portray the vision of Dual-T research project for bridging the gaps between different learning locations in this system, and introduce the online platform, Realto, designed towards this goal. We describe different features of Realto and present our contribution to the development of awareness tools for teachers in this platform.

Chapters 4, 5, and **6** present our methods for analyzing learners' interaction patterns across time, activity, and social dimensions. As mentioned before and depicted in Figure 1.1, what all these studies have in common is the focus on the temporality and patterns of change in learners' activities over time. This consistent focus on temporal dimensions is a unique contribution and a unifying thread throughout this work. In each chapter, we start by introducing our analytic methods and present findings from their application in MOOC context. We will then describe how we transfer the proposed methods to Realto and integrate them into an analytics dashboard for this platform. This will be followed by examples of patterns and relations that could be extracted from the developed dashboard.

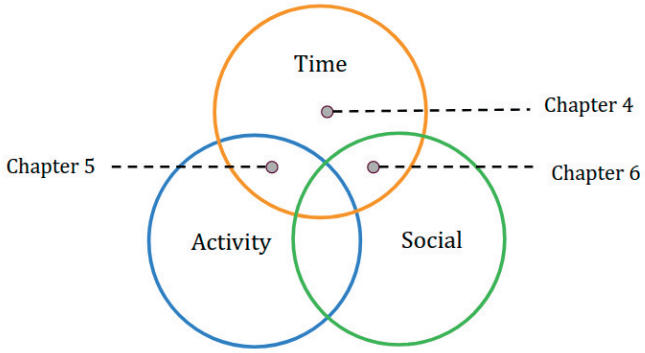


Figure 1.1 *– Thesis roadmap: Chapter 4 focuses on temporal patterns of learners' participation, Chapters 5 and 6 respectively focus on activity and social interaction patterns and their evolution over time.*

In **Chapter 4** we introduce methods to model temporal patterns of learners' online participation and propose metrics to measure different weekly or daily regularity patterns. We demonstrate the application of the introduced methods in a MOOC course and then show how they could enable to investigate the temporal patterns of activities in Realto.

In **Chapter 5** we present methods to analyze types and sequences of the activities performed by learners. We present two different approaches for extracting study patterns of MOOCs learners and present results on how learners' study approaches change over time. We then describe activity analysis tools for Realto and present examples of different platform usage patterns among Realto participants.

In **Chapter 6** we integrate social network and content analyses to study the interaction among learners in MOOC discussion forums. In this context, we explore the evolution of forum activity level, discussion topics, social network structure, and learners' roles over time. We then present how network analysis techniques could enable to model and investigate the structure of communication among different groups of participants in Realto.

Finally, in **Chapter 7** we summarize the main contributions and findings of this thesis, mention its limitations, and highlight potential future research directions.

2 Related Work

Learning Analytics (LA) and Educational Data Mining (EDM) are closely related and growing fields of research which focus on the analysis of learning-related data, in order to understand and improve the learning and teaching processes. In this chapter, we present an overview of the previous research in the field of learning analytics and describe the overall picture of this research area. We do not aim to provide an exhaustive literature review. Instead we refer to few studies to exemplify the applications of different analytics method in LA research. In each of the following chapters, we will then review the most relevant works to the particular topic of the chapter.

The remaining of this chapter is organized as follows. We start by the description of LA and EDM research in Section 2.1. We then present an overview of the models describing the LA process and the main research objectives in this domain, respectively in Sections 2.2 and 2.3. In Section 2.4 we review the most common methods and techniques applied in LA research. We finally conclude the chapter in Section 2.5.

2.1 Learning analytics and educational data mining

LA and EDM significantly overlap in objectives and methods, despite having slightly different perspectives. Enhancing the educational practice by analyzing large-scale data, extracting useful information, and providing data-driven insights for supporting the stakeholders, is the common goal of both research areas. The Society for Learning Analytics Research (SoLAR)[1] defined Learning Analytics as:

> *"The measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environment in which it occurs."*

In a similar definition, EDM is described by the Educational Data Mining Society [2] as:

[1] `https://solaresearch.org/about/`, (last accessed 10 December 2017)
[2] `http://educationaldatamining.org/`, (last accessed 10 December 2017)

> *"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."*

In the literature, few distinctions are made between LA and EDM fields. For instance, EDM is sometimes seen as rather focusing on methods and techniques for extracting information from learning-related data; On the other hand, LA is seen to have relatively higher focus on the applications of the derived information on the learning process [17]. Moreover, it has been argued that EDM has a greater focus on automated methods for the discovery of trends and also for applications in automated adaptation and personalization; while LA is considered as a more holistic approach, where leveraging human judgment is a key factor and automated discovery is considered as a tool towards this goal [205]. However, as both fields evolve over time, their differences get less and less noticeable [137, 205]. Several review papers conjointly describe LA and EDM concepts and methods [168, 207, 210, 17] and even the two terms are sometimes used interchangeably [210]. In this thesis we use the term LA to refer to the wider research area and process of LA and EDM.

2.2 LA models

LA is usually defined as a cyclic process [54, 49]. For instance, Chatti et al. [49] describe the LA process as an iterative cycle which is generally carried out in three major steps: **(1)** data collection and pre-processing, **(2)** analytics and action, and **(3)** post-processing. Data collection and pre-processing refers to the gathering of educational data from different learning environments and preparing it for the next step. The analytics and action phase refers to the actual application of analytic methods to discover meaningful patterns and extract useful information from the data. Post processing involves refining the data set, determining new required data sources or attributes, identifying new indicators, or refining the analytic methods. Post processing is considered as a fundamental step for continually improving the analytics practice.

Chatti et al. [49] proposed a reference model for LA, depicted in Figure 2.1. This model provides a classification schema of LA solutions based on four principal dimensions: **(1)** data and environments (what kind of data does the system gather, manage, and use for the analysis), **(2)** stakeholders (who is target user of the analysis?), **(3)** objectives (why does the system analyze the collected data?), and **(4)** methods (how does the system perform the analysis of the collected data?). This reference model also includes several challenges in relation to each of the four dimensions, which need to be addressed in LA practices. Some examples include handling big data volume from heterogeneous sources, finding meaningful indicators, ethics and data privacy, performance, scalability, and integration of LA into everyday practice.

Later on, Khalil et al. [118] proposed another model for describing the learning analytics life cycle and constraints. As shown in Figure 2.2, their proposed model consists of four main parts: **(1)** learning environment, where stakeholders such as learners and instructors produce

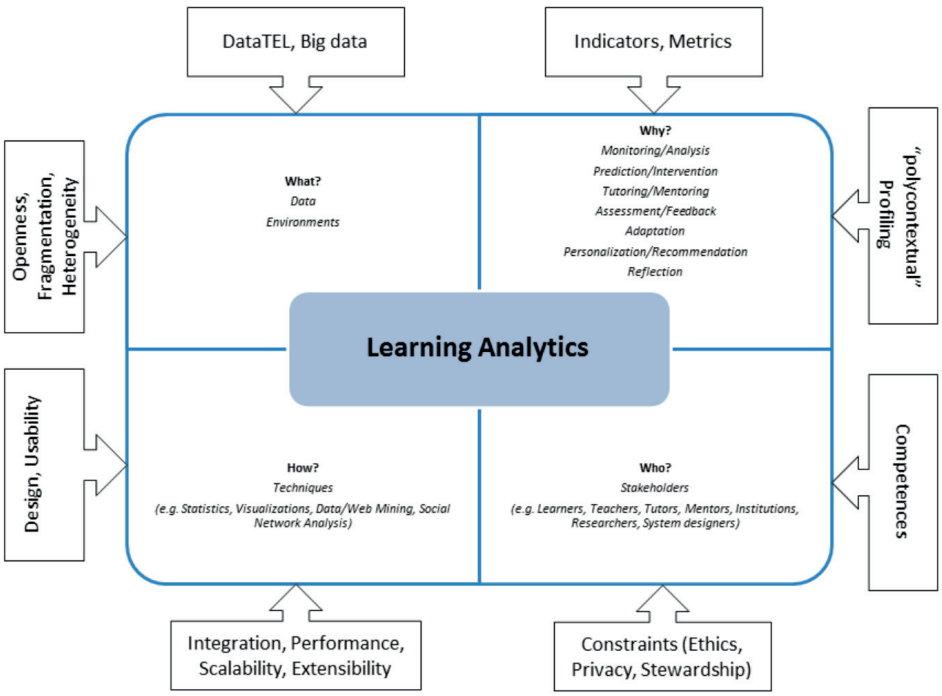


***Figure 2.1** – Learning analytics reference model proposed by Chatti et al., taken from [49]*

data, **(2)** big data, which consists of large-scale datasets, including interaction traces and other informations about the learners, **(3)** analytics, which refers to the different techniques used for analyzing the collected data, and **(4)** act, where the outcome of the analysis is interpreted to achieve the objectives of learning analytics and optimize the learning environment. The challenges that encompass the LA cycle (privacy, access, transparency, policy, security, accuracy, and restrictions) were also integrated in this model.

Despite their differences, both of the described models consider similar dimensions (stakeholders, dataset, objectives, and methods) to portray the LA cycle and mention similar constraints and challenges (such as security, privacy, policy, and ethics issues) which affect the LA process and need to be considered in the research in this domain.

2.3 LA objectives

The overall objective of LA is to support different stakeholders in learning and education. This includes not only learners and instructors, but also educational administrators, and researchers. Each group of stakeholders, could have different objectives and perspectives towards the analytics process and its outcome. Therefore Romero and Ventura [186] classified LA objectives depending on the viewpoint of stakeholders:

- **Learners**: LA could be advantageous for learners by personalization of learning environ-

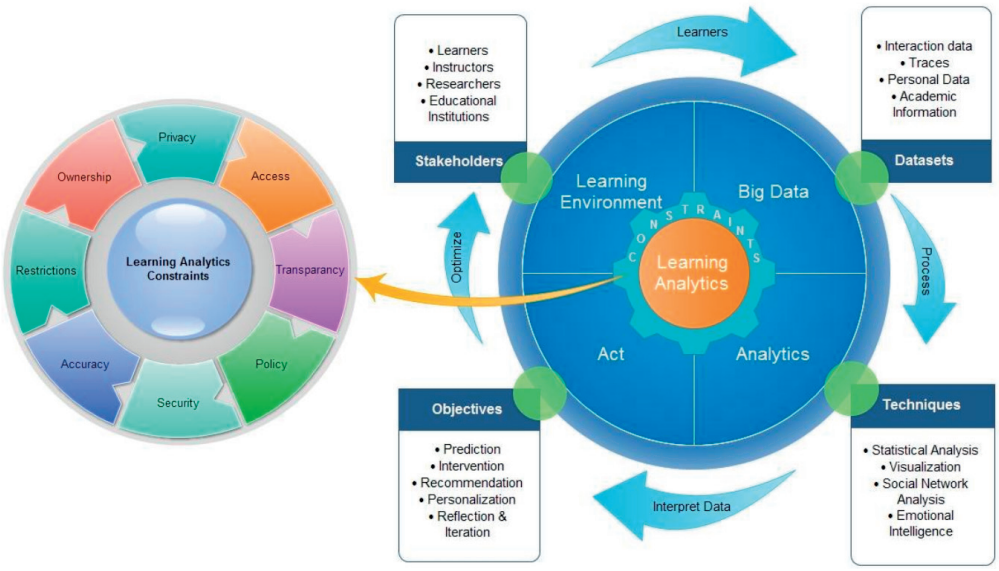


***Figure 2.2** – Learning analytics life cycle (right) and constraints (left) proposed by Khalil et al., taken from [118]*

ments, supporting their reflection on the learning process and providing adaptive feedback or recommendation to improve their learning performance

- **Instructors**: LA could inform instructors about their students' learning processes (e.g. performance, social, cognitive and behavioral aspects) to enable them to reflect on their teaching methods and improve the teaching process.
- **Administrators**: Data-driven insights from LA could assist educational administrators to identify the best way to organize institutional resources (human and material) and support their decision process for achieving higher educational goals.
- **Researchers**: Educational researchers could benefit from LA methods to assess the effectiveness of learning in different settings, recommend the most suitable method for each specific educational tasks, and evaluate and improve the course models and information delivery methods.

Although the described viewpoint clearly shows the benefits of LA research to the main actors in education, some LA applications especially those addressing multiple user groups, could not be easily classified according to this scheme. Chatti et al. [49], mentioned a different set of objectives to reflect the research goal of LA applications. This includes monitoring and analysis, prediction and intervention, assessment and feedback, adaptation, personalization and recommendation, reflection, tutoring and mentoring. Learner modeling, and detecting undesirable behaviours are the other goals which can be added to this list [185, 168]. The listed objectives certainly have overlaps and a specific LA application often addresses several of them. We briefly describe the mentioned goals in the following.

- **Monitoring and analysis**: Monitoring and analyzing learners' activities, aim to provide the

basis for supporting decision making by teachers and/or educational institutions. Through the monitoring process, students with difficulties in following the course and the challenging course units could be identified. This way, more objective feedbacks could be provided to the instructors, enabling them to evaluate the learning processes and consequently improve the course structure and design of learning activities.

- **Prediction and intervention**: The goal of prediction is to develop a model based on learners' current activities and accomplishments to estimate their future performance or knowledge. Predictions of students performance and identifying early indicators of their success or failure, could enable the instructors to offer proactive interventions and support the learners who need further assistance.
- **Assessment and feedback**: This objective is concerned with supporting the (self-)assessment of improved efficiency and effectiveness of the learning processes. This in turn could enable to provide more meaningful feedback to the learners and instructors.
- **Adaptation**: The goal of adaptation is to tailor the learning resources, instructions, and sequence of activities to learners' requirements. Adjustments of the learning processes according to the needs of individual learners is of particular importance in intelligent tutoring systems and adaptive learning environments.
- **Personalization and recommendation**: The aim of LA in personalization is to help learners shape their own learning and learning environment towards achieving their objectives. To foster self-directed learning, LA could provide learners with recommendations based on their preferences and the activity profiles of other learners with similar preferences. Unlike adaptation which is triggered by the teacher or the learning environment, personalization is highly learner-centric and the control of the learning process is left to the learner [49].
- **Reflection**: Promoting (self-)reflection on the teaching and learning practices is another aspect in which LA could be a valuable tool. LA could support students and teachers to reflect on the effectiveness of their approach and progress, by providing comparisons with past performance or comparison between learners, across courses or institutes.
- **Tutoring and mentoring**: Tutoring and mentoring aim at supporting students in the learning process. Tutoring is mainly focused on helping students in domain-specific learning tasks limited to the context of a particular course. In contrast, mentoring has a broader focus and is concerned with providing advices and guidances to learners, through the whole learning process.
- **Learner modeling**: The aim of learner modeling is to develop cognitive models of students to represent their skills, knowledge, learning styles, and behaviour. Such models are constructed based on learners' previous activities and preferences and could describe other characteristics such as motivation, satisfaction, affective or meta-cognitive states. Learner profiling is a crucial step to achieve several of the described learner-centered LA objective, for instance to provide effective intervention and recommendations, and support adaptation and personalization in the learning environments. This is a highly challenging

task considering the complexity of learners' behaviours, in addition to the diversity and increasing complexity of the learning environments.

- **Detecting undesirable student behaviours**: This objective is concerned with detecting students who are facing some type of problem in their learning process (e.g those with learning difficulties, low motivation level, in risk of dropout or academic failure) or learners who show undesirable behaviour (e.g. erroneous actions, cheating, misusing or gaming the system). Early detection of undesirable behaviours could enable educators or learning systems to provide appropriate and timely support.

Apart from the described categories, several other objectives such as constructing courseware, providing reports, developing concept maps, planning and scheduling have also been mentioned in the LA literature [185, 207]. As mentioned, the listed objectives have overlaps and measuring them usually requires to define tailored metrics and performance indicators. As shown by literature reviews in [168] and [134], the most common goals in LA literature include performance prediction, behaviour modeling, reflection, and monitoring.

2.4 LA methods and applications

The methods used in LA research to extract patterns from educational data, originate from different fields including data mining, psychometrics and educational measurement, statistics, and information visualization [210, 17]. The choice of LA techniques is affected by the analytics task and objectives, and also the nature of the collected data [210]. Baker et al. [17] classify LA methods into five main categories: prediction, structure discovery, relationship mining, visualization, and discovery with models. In the following we describe these categories and provide examples of their application in LA research.

2.4.1 Prediction

Prediction is one of the most common categories of methods used in LA research. Predictive models are developed to estimate a certain variable (e.g. grade or knowledge level) based on a combination of the other indicators from the educational data set. To train and validate the predictive models, a set of labeled data for the output variable is required. The two general types of prediction methods used in LA research include classification and regression.

Classification

Classification models are used to predict binary or categorical variables (e.g. pass/fail, or achievement level). Popular classification methods in this field include decision trees, random forest, support vector machines (SVM), logistic regression, and K-nearest neighbor (KNN). As an example, Pardos et al. [172] applied different classification methods to detect students' affective state or behaviours in a tutoring system, including boredom, concentration, confu-

sion, frustration, off-task or gaming behavior. For this purpose they used learners' interaction features such as number of attempts, hint request, or incorrect actions. Classification methods have also been widely used to predict dropouts in traditional education systems [150, 87] or online learning environments [144, 243]. In few recent studies deep neural networks were used for dropout prediction [228, 76]. Other examples include predicting correct or incorrect answers to a question in intelligent tutoring systems [171] and in MOOC assessment tasks [36], predicting students' performance level [107, 167, 162, 239], or their satisfaction and the main factors influencing it[63].

Regression

Regression models, are used for predicting continuous variables. Predicting final grade, assignment grades or test scores is the most common application of regression methods in LA research [148, 188, 72, 182, 117]. Linear regression, neural networks, and SVM are among the common methods used for this purpose.

2.4.2 Structure discovery

In this category of methods, the aim is to detect structure in educational data. Structure discovery methods attempt to extract the naturally emerging structures from the data, without strong a priori assumptions of what should be found [210]. This is in contrast to the prediction methods which require a set of labeled data to model a specific variable. Common structure discovery methods include clustering, factor analysis, domain structure discovery, and social network analysis [17].

Clustering

Clustering methods aim to split the data into a set of categories based on the similarity of the data points with respect to certain features. In LA research, clustering methods have frequently been used to identify groups of similar learners [120, 220, 26, 126]. For instance Khribi et al [120] clustered learners based on the similarity of their access patterns to the learning objects in e-learning environments, in order to provide personalized resource recommendations. To improve test score prediction in a tutoring system, Trivedi et al. [220] clustered students into subgroup based on their activity features (e.g. number of problems solved, percent of correct answers, average time spent per question). Next, they trained a grade predictor separately for each of the resulting learner groups. Learner modeling and discovering different learning behaviours from interaction logs is another application of clustering methods in LA research [199, 9, 129, 128, 97].

Factor analysis

Factor analysis is a technique for studying the interrelationships among variables and finding dimensions of variables which group together[157]. This technique could be used for both confirmatory and exploratory analysis. Confirmatory factor analysis could be applied to test a theoretically proposed factor model and is commonly used to validate scales in psychometrics personality theories. This method has been extensively used in educational psychology before the educational data was considered to be "big". Some studies in educational research, have employed factor analysis to study the association between existing scales and outcome measures. For instance to study the relation between motivational measures and course completion in MOOCs [229], or the impact of students' achievement emotions (enjoyment, anxiety, boredom, hopelessness [175]) on their learning choices between face-to-face and online instruction modes [213].

On the other hand, exploratory factor analysis could be applied to determine the latent factors from data, without strong theoretical assumptions about the factor model [157]. This method could also be used for dimensionality reduction by collapsing a large number of observed variables into a few interpretable underlying factors (unobserved variables). Each factor in this case, consists of interrelated variables and explains a portion of variability within the dataset [216, 223]. As an example, Deane et al. [61] performed exploratory factor analysis on features extracted from learners' writing process in an automatic essay grading system, to infer latent factors which reflect students writing strategies and literacy skills.

Domain structure discovery

This category of methods aims to find the structure of knowledge in educational domain [210]. This could consist of identifying the relation between different knowledge components, or the mapping between course content and the knowledge components[17]. A well-know example is mapping question items to skills in intelligent tutoring systems, and Q-matrix [212] is a standard means to model this mapping. Items to skills mapping plays a pivotal role in tutoring systems towards effective grouping of the problems, monitoring learner's progress, providing personalized hints, and adopting the problems difficulty level and learning pace to individual students [45]. The algorithms used for domain structure discovery in educational research, range from fully automatic methods [67, 65, 23, 215] to methods which incorporate human judgment within the process of model discovery [45].

Social network analysis

Social network analysis (SNA) focuses on the analysis of relationships among learners. SNA techniques could be used to investigate the attributes and structure of networks composed of individual learners and relations among them. In LA research, SNA has been applied to analyze patterns of interactions among learners in knowledge exchange communities [59],

collaborative learning activities [151, 52], and discussion forum communications [163, 101, 246, 237]. SNA methods were also used to study students' access patterns to learning resources [100].

Some studies have used SNA tools to measure the degree of interactions and quantify learners' attributes in the learning networks (e.g. using centrality or density measures). Other studies have used such quantitative analyses to identify important or isolated learners [60, 94], to study the link between learning performance and network attributes [173, 111], and to identify learners at risk of dropping out [25, 241]. In several studies, SNA techniques are complemented with content analysis methods [102, 99, 248, 106]. Combining content and network analysis tools could provide a deeper understanding of the nature and type of interactions among learners [44]. As an example, Hecking et al. [102] analyzed social and semantic structure of learners community in MOOC discussion forums. They determined the thematic areas in which learners seek or provide information and modeled the socio-semantic roles of learners in the communication.

2.4.3 Relationship mining

Relationship mining methods aim to discover relations among variables. This entails measuring the strength of relations between pairs of variables in the dataset, or determining the most strongly associated variables to a particular variable of interest. Three most common relationship mining approaches in LA research are association rule mining, correlation mining, and sequential pattern mining [17].

Association rule mining

This group of methods describe the relations between variables in the form of a if-then rules [5]. Association rule mining is among the most popular methods for discovering and representing strong interesting relations among frequent items in a database [119]. The interestingness of an association rules is evaluated based on two factors: support and confidence, which respectively reflect the usefulness and certainty of the discovered rule. Support measures the frequency of a rule in the entire dataset, and confidence measures its strength according to the number of times the if-then statement has been found to be true. Most algorithms for extracting association rules, such as Apriori algorithm [4], require the user to define threshold values for support and confidence. The algorithm then identifies the set of rules which satisfy the minimum support and confidence restrictions [81]. In LA research, association rule mining is frequently used to evaluate students' performance and identify the factors that affect their academic achievements [130, 32, 119, 7], and also to provide recommendations to students or teachers [6, 81]. For instance Garcia et al. [81], used the association rules extracted from students' usage data as the basis of a collaborative recommender system to support teachers in maintaining and continuously improving e-learning courses.

Correlation mining

Correlation mining methods are mainly focused on identifying positive or negative linear correlations between different variables in a dataset. In educational research, correlation analysis has been widely used to study the relation between performance and different factors such as students demographic information [92, 8, 114], personality factors and learning approaches [47, 2], affective states [169], and forum participation [173]. Other examples include analysis of the relation between students' learning attitudes and help-seeking behaviors in a tutoring system [15], and relation between students' learning strategies and interaction with different course materials [125].

Sequential pattern mining

This category of methods aims to detect temporal associations between variables. Sequential pattern mining is concerned with discovering interesting subsequences in sequential data, where interestingness of a subsequence could be measured in terms of its occurrence frequency, length or other domain specific criteria [78]. In LA research, sequential pattern mining has been used to analyze students' navigation and access patterns to learning resources in order to develop personalized learning scenarios and activities [16, 154, 227], to adapt learning resource sequencing [115], and to provide recommendations on learning content based on students' preferred learning styles [250, 187, 165]. Another example is to indicate characterizing behaviors of successful and unsuccessful groups in online collaborative environments in order to support learner groups by early recognition of problems [177]. Sequential pattern mining has also been applied for constructing student models in intelligent tutoring systems [13], identifying meaningful characteristics and updating the learner models to reflect newly gained knowledge [12].

Causal data mining

In causal data mining, the goal is to discover causal relationship between variables, which in turn could provide a basis for action. Research in LA domain frequently relies on post-hoc analysis and provides descriptive, correlational, and predictive findings which do not necessarily imply causality [235]. However several studies in this field have investigated the causal relation among variables through controlled experimental studies. For instance, Sonnenberg et al [209], conducted an experiment to investigate the effects of metacognitive prompts in a computer-based learning task on learning performance and regulation activities by learners (e.g. planning, monitoring, and evaluation). Using HeuristicsMiner algorithm [231] which searches for causal dependencies between activities and indicates the certainty of a relation between two activities, they found that receiving meta-cognitive prompts increased the occurrence of regulation activities (especially monitoring), which in turn enhanced the performance. Similarly, Rau et al. [181] conducted experiments in a cognitive tutor for fractions, to compare the effect of single and multiple graphical representations on learning.

Using TETRAD package [192] for causal data mining, they found that multiple representations increase error rate, which in turn inhibits learning.

2.4.4 Visualization

The results of LA analysis are usually either fed into recommendation and adaptation mechanism, or reported back to the learners, teachers, or other stakeholders to increase awareness and support the teaching and learning processes [210]. Inspecting and interpreting the analysis outcome and fine-grained statistics available in LA, is sometimes cumbersome and time consuming to perform. Visualization techniques could facilitate this process by making big sets of learning-related data and results more easily accessible and understandable for the end users [185]. Baker et al. [17] refer to this category as "distillation of data for human judgment". The aim of this process is to depict the data in intelligible ways, using summarization, visualization and interactive interfaces, which allow educational stakeholders to easily recognize patterns that may otherwise be difficult to interpret [186]. Suitable visualization play an essential role in gaining insights into the teaching and learning processes and their interrelations [210]. This is in turn a prerequisite for empowering data-driven decision making and pedagogical actions to improve the teaching and learning processes [42].

Visualization of learning traces are commonly integrated in learning dashboards [222, 70, 155]. A learning dashboard aggregates different indicators about learners, learning processes, or learning contexts into one or multiple visualizations in a single display [194]. Research on learning dashboards aims to identify what data is meaningful to different stakeholders and how data can be presented to support sense-making processes. Dashboards can provide teachers and learners with an overview of their activities and how they perform in comparison to others [217, 222]. Some dashboards, represent visualizations of different aspects of the information side-by-side in one single view. Steiner et al. [210] refer to this approach as "all-at-one-time" dashboards. Other dashboards, start with a single overview visualization and allow the user to access further information and details from there [210, 42].

In our previous work we performed a systematic literature review of the research on learning dashboards [195], and tools for monitoring, awareness, and reflection [184]. We analyzed the state-of-the-art research with respect to the contexts in which dashboards were being applied (including educational settings, target users, and learning activities), characteristics of dashboard solutions (including purpose, data sources, indicators, visualization types, underlying technologies), and their maturity regarding evaluation. In general, potentials of learning dashboards for fostering awareness, reflection, sense making, and in the end, improving learning are well-known [222]. However, evaluation of the dashboards regarding their adoption and learning impact is a challenging task, and probably the main yet under-explored aspect of research in this domain.

2.4.5 Discovery with models

Discovery with models, refers to the general approach of using the outcome of one analytics method within another analysis, and so it does not denote a particular category of analytic methods [210]. Most commonly, the model of a phenomenon obtained using prediction or clustering is used as a component in further prediction or relationship mining analyses [34]. As an example Bouchet et al. [33] used clustering to categorize learners into different groups based on their prior knowledge, learning performance, and strategies. In the next step, they employed differential sequence mining to identify differentiating activity patterns among the identified student groups, and interpret these patterns in terms of relevant learning behaviors.

2.4.6 Knowledge tracing

Knowledge tracing is the task of modeling students' knowledge over time and estimating their mastery of skills. It has been widely used in intelligent tutoring systems for predicting how students will perform on future interactions and adopting the learning content accordingly. Knowledge tracing could enable more effective resource suggestion and tailoring learning activities sequence, for instance by skipping or delaying activities which are predicted to be too easy or difficult for the learner. Bayesian Knowledge Tracing (BKT) [55] is the most popular approach for building temporal models of students' knowledge. This method models learners' latent knowledge as a set of binary variable, where each variable denotes whether the student has mastered a particular concept. As the learner proceeds through the learning tasks, a Hidden Markov Model (HMM) is used to update the latent knowledge states based on correct or incorrect answers to exercises of a given concept.

One assumption in BKT in its original formulation is that a skill is never forgotten once it is learned. Several extensions to the original BKT could be found in the literature which include contextual estimation of guessing and slipping probability [57], introducing students' prior knowledge and learning speed parameters [245], and including problem difficulty in the knowledge model [170]. Deep knowledge tracing (DKT) [178] is another recent method in this category which applies recurrent neural networks (RNNs) to the problem of knowledge tracing and has shown improved performance over BKT.

2.5 Discussion

In this chapter we presented an overview of the broad field of learning analytics. Research in this domain is rapidly growing and has great potential to empower the educational processes. As described, a wide range of computational methods are employed in LA research to analyze the educational data. In this thesis, we introduce novel methods, which in combination with the other well-established analytic methods, could provide valuable insights on learners' interactions and behaviours. To fully exploit the potential of LA research for addressing the outlined objectives, the high-level indicators and findings from the analyses need to be

incorporated into the workflow of educational practice. This could enable the LA research to have impact on optimization of teaching and learning experiences, providing support for educational stakeholders, and refinement of the educational structures. Moreover, the critical issues of ethics, privacy, and transparency, in relation to the data collection and use, need to be taken into account to establish proper implementations of learning analytics methods.

3 Realto: Online Platform for Integrated Vocational Education

3.1 The Swiss vocational education and training system

Vocational Education and Training (VET) is the most popular form of upper-secondary level education and training in Switzerland. About two-thirds of Swiss youngsters, after finishing their ninth year of compulsory education, enroll in a VET program which provides them with their first exposure to working life and opens a wide range of career prospects [197]. The Swiss VET system offers training in about 230 occupations, including electrician, IT technician, logistician, florist, clothing designer, hairdresser, cook, carpenter, and car mechanic. Upon successful completion of their vocational training, apprentices receive a federal certificate and are considered as qualified workers in their field. After the upper-secondary VET program, which is three to four years long, apprentices have the possibility to continue further with professional training or get into higher education schools (Figure 3.1).

Two different types of programs are offered in Swiss VET system: single-track (or school-based) and dual-track. The single-track training approach, involves full-time classroom instruction and is more common in trade or commercial schools [197]. On the other hand, the dual-track approach, which is the most common one, consists of part-time classroom instruction at a vocational school (one or two days per week) combined with part-time apprenticeship at a host company (three to four days per week). Vocational classrooms provide instruction in general subject matters (e.g. language, mathematics, communication, and history), as well as the profession specific theoretical aspects. The practical skills and know-how knowledge are acquired during the remaining days spent in the host company. In the company, apprentices are assigned to a supervisor who is usually a senior worker in the same field with several years of experience and a license for training young employees. With the support of their supervisor, apprentices acquire the practical competencies in authentic settings and actively take part in the host company production processes.

Workplace experiences might differ among apprentices. Apprentices may get limited exposure to some important aspects of their profession, depending on the size or geographical location of the company they work for, and the roles they are assigned to. For instance, a carpenter

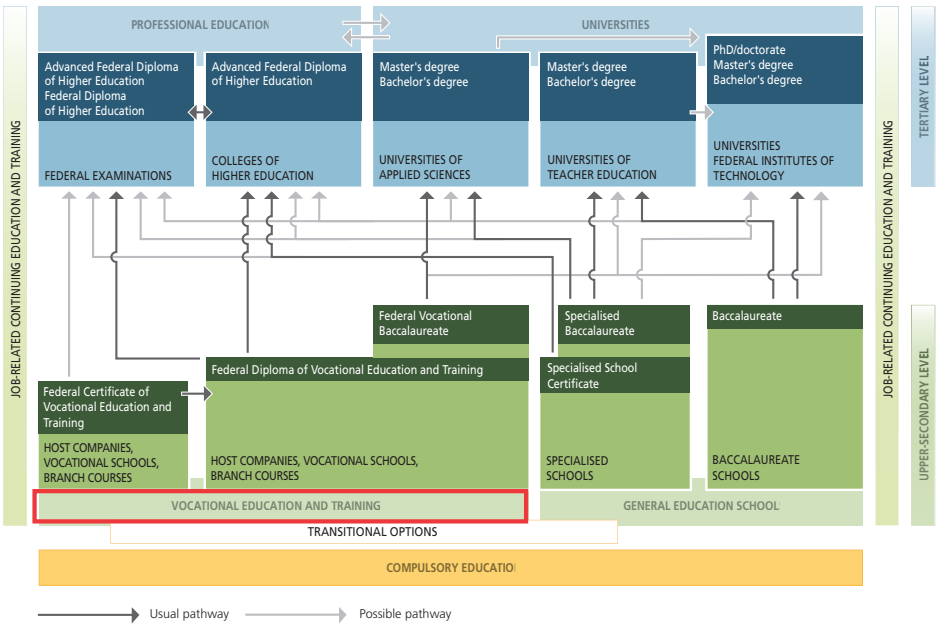


Figure 3.1 *– The Swiss educational system and possible pathways [197].*

apprentice working in a small company might only get to work on a certain types of roof structures, or an apprentices working in a company which uses computerized numerical control (CNC), might not get exposure to other wood cutting machines. The role of school in dual-track settings is to ensure that all apprentices obtain a certain level of theoretical knowledge in all aspects of their profession.

3.2 School and company gap

In dual-track VET system, apprentice acquire different forms of knowledge by alternating between the company and school contexts. In this setting, school and companies are supposed to work together to support aggregation of the information obtained in different locations into a coherent body of knowledge. However, as Gurtner et al. [93] have mentioned, these two contexts often have separate aims, content and sociological organization. Consequently, putting together the skills, knowledge, and attitudes acquired from the two contexts is incumbent on the apprentice her/himself [93]. In workplace, apprentices often obtain knowledge in implicit and concrete form, contextualized in the specific practice. In contrast, the knowledge from school is mainly abstract, explicit and theoretical, situated outside of the target context [73]. Transferring the knowledge and skills from one context to the other is not obvious and does not take place spontaneously. Without adequate support, the obtained knowledge often remains encapsulated in its original context. Consequently, apprentices often perceive gaps between their learning locations and do not see adequate links between what they learn at school, and what they face and do in practice [193].

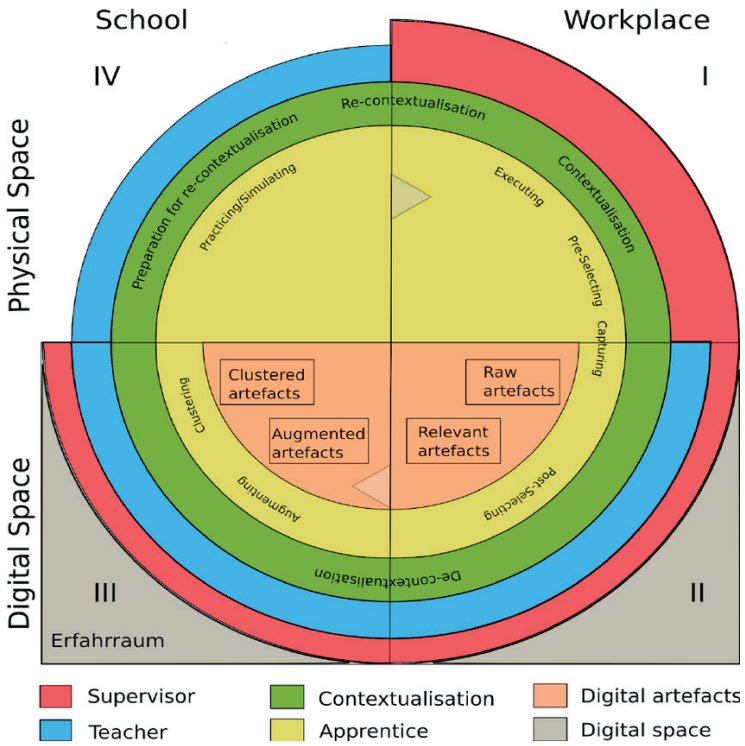


Figure 3.2 – *The Erfahrraum model: a pedagogical model to inform the design of technology-enhanced VET learning activities [193]*

.

The Dual-T research project aims to bridge the gaps between different learning locations and connect the actors of these locations in dual-track VET programs. Taking into account the importance of differences between workplace and school for the success of dual model, the project considers the necessity of connecting these two context without suppressing their specificities. A primary hypothesis of Dual-T project is that learning technologies have the potential to connect classroom activities and workplace experience by providing a space within which knowledge can be reflected upon, communicated back and forth from one context to another, and shared with all actors. This hypothesis is translated into the concept of *Erfahrraum*, a pedagogical model that informs the design and implementation of learning technologies, activities, and scenarios to connect school and work contexts in VET programs [193]. The term *Erfahrraum* is a portmanteau consisting of the German words *Erfahrung* (reflected experience) and *Raum* (space). *Erfahrraum* consists of technology-enhanced spaces that facilitate integrating practical and theoretical knowledge through scaffolded reflection activities. *Erfahrraum* describes boundary-crossing spaces to capture, share and process experiences from different learning locations and facilitates conversations between work and school contexts. As shown in Figure 3.2, the *Erfahrraum* model includes physical and digital learning spaces (vertical axis) which can be found in school or workplace contexts (horizontal axis). This model distinguishes between the role of supervisors (red ring) and teachers (blue ring) in supporting apprentices' actions (yellow ring) for the contextualisation of vocational

knowledge (green ring).

3.3 Realto: the online platform for integrated VET

Realto (Figure 3.3), is an online learning platform for VET, which implements the *Erfahrraum* principles[1]. This platform provides a digital space where apprentices, teachers, and supervisors can upload and share digital artefacts, including photo, text, audio, video, or other digital files. Realto is accessible through multiple devices and platforms, including smart-phones, which enables apprentices to quickly capture experiences and use them for later reflection activities at school or at workplace. For instance apprentices can take a photo of an interesting experiences at their workplace, upload it into Realto, add additional information such as description or annotations to it, and share it with their teacher, supervisor, and peers. Teachers can use the uploaded materials for classroom lessons to illustrate abstract and theoretical concepts with concrete examples from the apprentices workplace experiences.

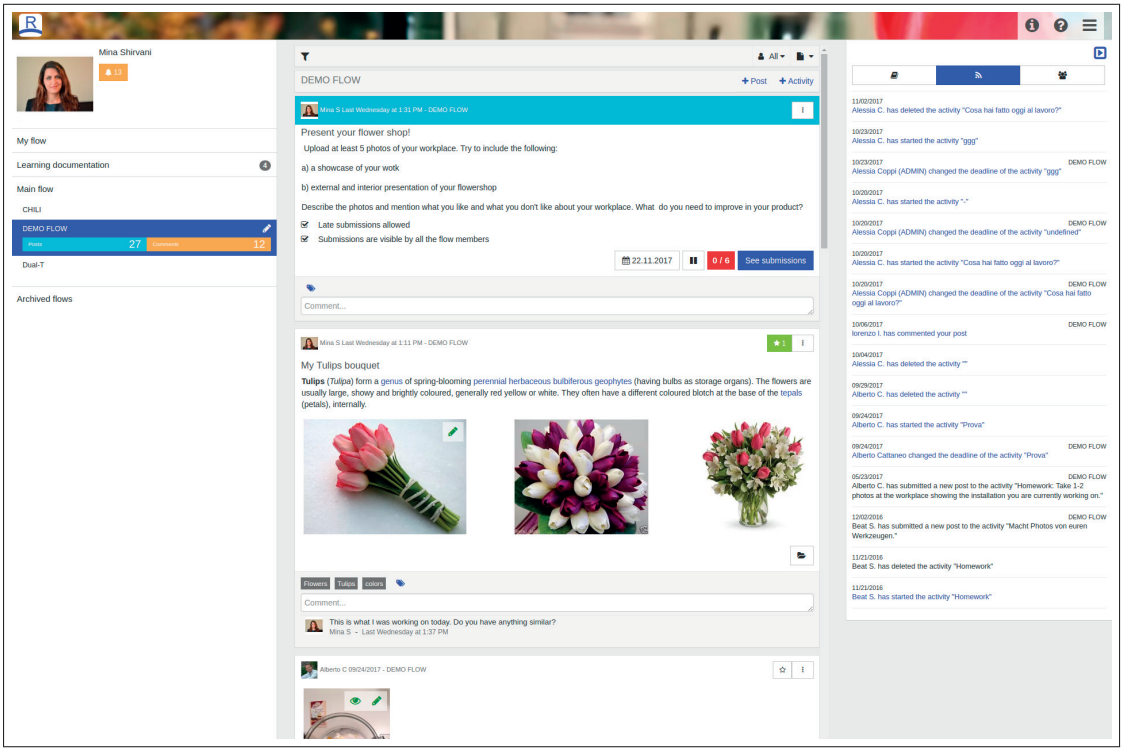


Figure 3.3 – *Realto: The online learning platform for integrated vocational education.*

In Realto, teachers can create a *flow* (group) for a class (or a certain topic), which provides a space for the members to share pictures or other digital resources. As in a social platform, flow members can comment on and rate each others contributions. Moreover, the teachers may define activities and assign tasks to the apprentices, for instance ask them to upload

[1] `http://dualt.epfl.ch/page-121584-en.html`, (last accessed 10 December 2017)

photos of specific tools, or workplace experiences relevant to current school topics. Using features such as picture comparison and annotation in Realto, teachers can then exploit apprentices' submissions during the next session. It is also possible to perform classroom activities using Realto. For instance, teachers can create a collaborative image annotation activity, where apprentices can annotate a particular picture and the system then superposes the created annotations on the original image. Alternatively, the teacher can select one or multiple apprentices to overlay only the annotations made by them and project the resulting image on the classroom screen to initiate discussion about certain mistakes or different solutions made by the learners.

Besides creating posts and sharing experiences in Realto, apprentices can also prepare learning documents (LD) to document their workplace activities. Throughout their training, apprentices are required to regularly create LDs to document and reflect on their workplace training procedures and their professional development. LDs can also serve as personal records of experiences made during the vocational training. Apprentices' LDs in Realto are made available to their workplace supervisors. Supervisors have their Realto profiles linked to the ones of their apprentices to facilitate their communication. Supervisors are responsible for controlling and validating their apprentices' LDs. They can provide feedback to apprentices and suggest modifications for improvement. They can also ensure that the materials which are protected by company restrictions are not included in the LDs.

3.3.1 Awareness tools in Realto

In online learning platforms, it is often challenging for the teachers to be aware of what their students are doing and how they are performing, to provide adequate guidance and intervene when necessary. This task is not trivial for both small and large classes. Adequate tools should be provided to the teacher to enable understanding of learners' activities and improve awareness [226]. In this section, we describe our contribution to the development of learning analytics tools to support awareness of teachers in Realto platform.

The literature in learning analytics research provides several examples of tools developed for facilitating awareness of teachers in a classroom [222]. Learning dashboards, with real-time analysis and visualization of important information about learners' online interactions are considered to be an adequate mean for this purpose [217, 222]. The existing learning dashboard solutions, could be divided into two broad categories with respect to the way they are delivered to the user [225]. The first group includes external or standalone solutions, which are not integrated into the learning platform, but use the data recorded by it. Examples in this category include general web analytics services such as Google Analytics[2] and Woopra[3]. The second group consists of dashboards integrated into the learning platform, such as the analytics dashboard in Graasp, a social media platform for learning and knowledge sharing

[2] `http://www.google.com/analytics/` (last accessed 10 December 2017)
[3] `http://www.woopra.com/` (last accessed 10 December 2017)

[226]. Integrating analytics tool in the target platform facilitates users' access to the analytics without requiring them to switch between the platform and the analytics tools.

To provide awareness tools for teachers in Realto, we chose the integrated approach and developed an interactive learning dashboard which aggregates various indicators of apprentices' activities in Realto. Our proposed dashboard solution is similar to Graasp analytics tools [226] in the sense that it is integrated in the target platform (Realto), is embedded in the interaction context within the platform, and presents contextual information. In Realto, the teacher can access the dashboard directly from each flow, and the content of the dashboard is adopted to include only the information about the activities made by learners in that particular flow. Consequently the teacher does not need to leave the flow to observe and understand the interactions happening there. The dashboard is accessible directly in the interaction context and includes relevant information scoped by this context.

We could not directly involve the vocational teachers in the design process of Realto awareness dashboard. The main limitation was that at the time when we started development of the dashboard, the platform was not yet in a stable state and the teachers were not familiar with its functionalities. Hence we decided to prototype an initial functioning version based the guidelines from the literature on learning dashboards [48, 226, 152], and the internal brainstorming sessions with the researchers involved in the Dual-T project. We implemented an interactive awareness dashboard which is integrated into Realto and is accessible by the teachers registered on the platform. This dashboard represents timeline of students' activities and their overall activity indicators such as sum of posts, comments, and responses to teacher-defined activities. It represents indicators both at the class level and individual level, and provides the possibility to compare students' activities and identify at-risk learners. Our proposed dashboard comprises four components which we describe in the following. This includes: flow overview, individual view, comparison view, and post view.

- **Flow overview**: (Figure 3.4) provides an aggregated view of activities made by all flow members. As shown in Figure 3.4, the top chart in this view presents the overall activity by all flow members over time, and the bottom chart shows the total activity by each individual. Learners' activity level is measured by the number of produced artefacts including posts, submissions to teacher-defined activities, LDs, and comments. The teacher can exclude/include any of these metrics either by clicking on the chart area or using the control buttons on top. It is also possible to change the visualization type and customize the time period of the data included in the visualizations. The information provided in this view, could enable the teacher to track changes in learners' engagement over time, identify most or least active time periods and learners, and investigate the influence of interventions or certain events on learners' overall activity level.

- **Individual view**: (Figure 3.5) provides detailed information about activities of individual learners and how they compare to the other flow members. In addition to the number and type of produced entries, this view includes other indicators such as the level of details in learners' posts, their activity profile over time, and their last access time. The level of

posts details is determined based on the resources and descriptions included in the posts. Well-elaborated posts which include pictures and descriptions, are considered to have *high*

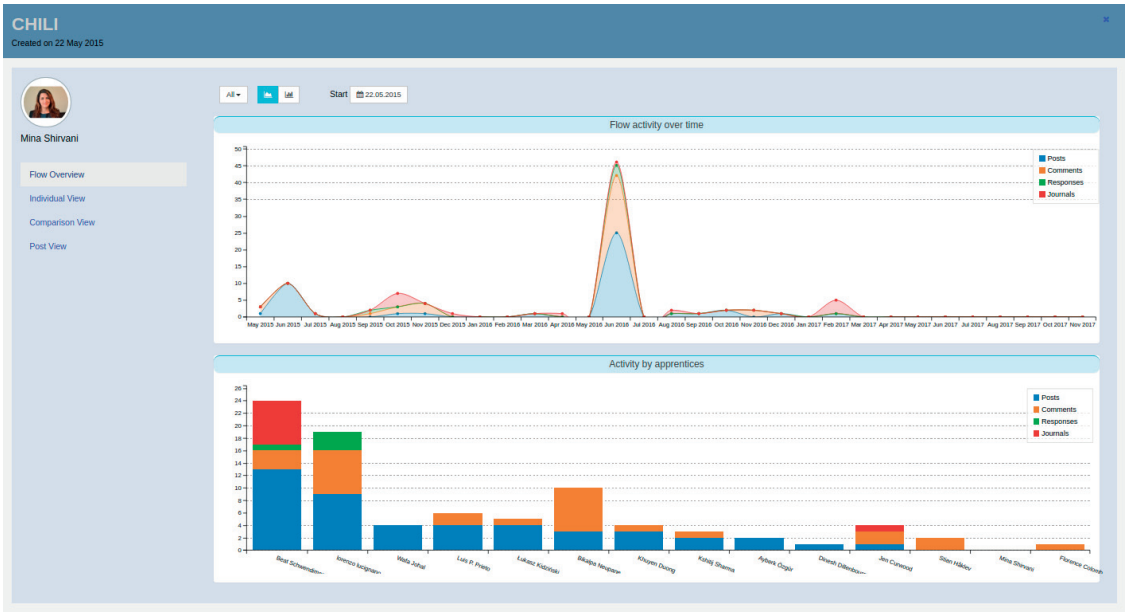


Figure 3.4 *– Flow overview in Realto awareness dashboard*

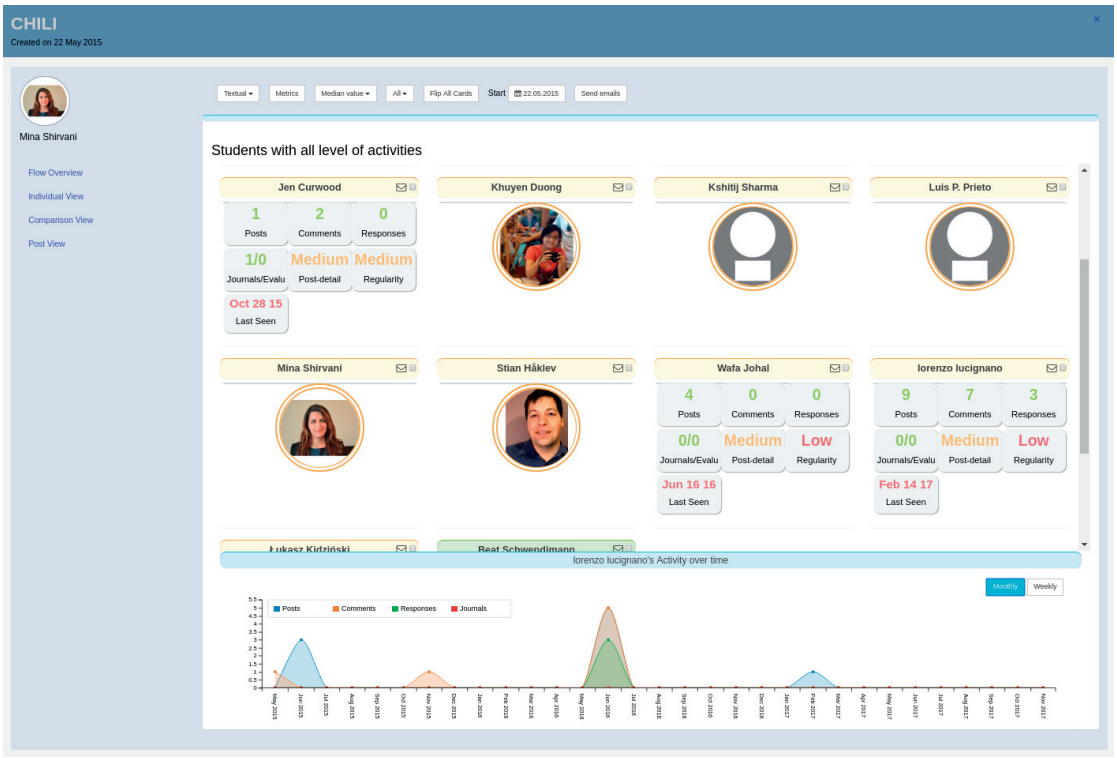


Figure 3.5 *– Individual view in Realto awareness dashboard*

level of details; Posts with only pictures or textual descriptions are considered as *medium* detail level, and posts with no resource and no description are considered as *low* detail level. Similar to the previous case, the control buttons on the top enable the teacher to select the metrics to be included in the visualizations, the visualization type, and the time period for the data being used. When hovering the mouse over the area for a particular learner, the bottom chart in Figure 3.5 gets updated to show activities of the selected learner over time. The time granularity (weekly/monthly) for this chart can be selected from the interface. By clicking on the area for each learner, the teacher can show/hide the detailed view of the learner's activity indicators. The color of the region representing each learner encodes learner's status in comparison with the other flow members (green, yellow, or red to respectively show high, medium, or low activity status). By default, learners' activities are compared to the most active learner in the flow. However this comparison criteria (flow average or maximum) can be changed through the dashboard controls. In this view, the teacher may also filter learners based on their activity status.

The individual view in the dashboard provides the possibility for the teacher to send direct feedback to the learners who might need further attention. In order to do that, the teacher needs to select the learners using the checkbox next to their names and then click on the *send email* button on the dashboard control bar. This opens up a modal page for the teacher to write down a message, which will then be sent to the selected learners. Therefore, without requiring to leave the dashboard context, the teacher can provide personalized feedback to the learners based on their activity status. This feature could therefore facilitates the process of taking actions based on the insights from the provided analytics and visualizations.

- **Comparison view**: (Figure 3.6) includes similar information about individual learners as in the previous view. The main difference between the two components is that in the comparison view, the teacher can select any two learners to view their activity indicators side-by-side. The information provided in the individual view could be overwhelming for the teacher specially when dealing with large classes. The comparison view provides a less cluttered view of the selected learners and facilitates their comparison.

- **Post view**: (Figure 3.7) represent the information about the most popular posts in the flow during the time period indicated by the dashboard user (by default during last three months/weeks). The popularity of a posts is measured by the number of comments, votes, and views it has received from the flow members. This module could bring to the attention of the teacher, the posts and resources which might be interesting for future classroom activities or discussions.

The design of dashboard is a very subtle trade-off: too much information could overload the teachers, too poor information would make the dashboard useless. The optimal solution can only be approximated through an iterative participatory design process. Within the time constraints of this thesis, we could not perform further design iterations. The usability of our dashboard solution and its impact on the practice need to be evaluated and teachers' feedback should be taken into account for the future iterations.

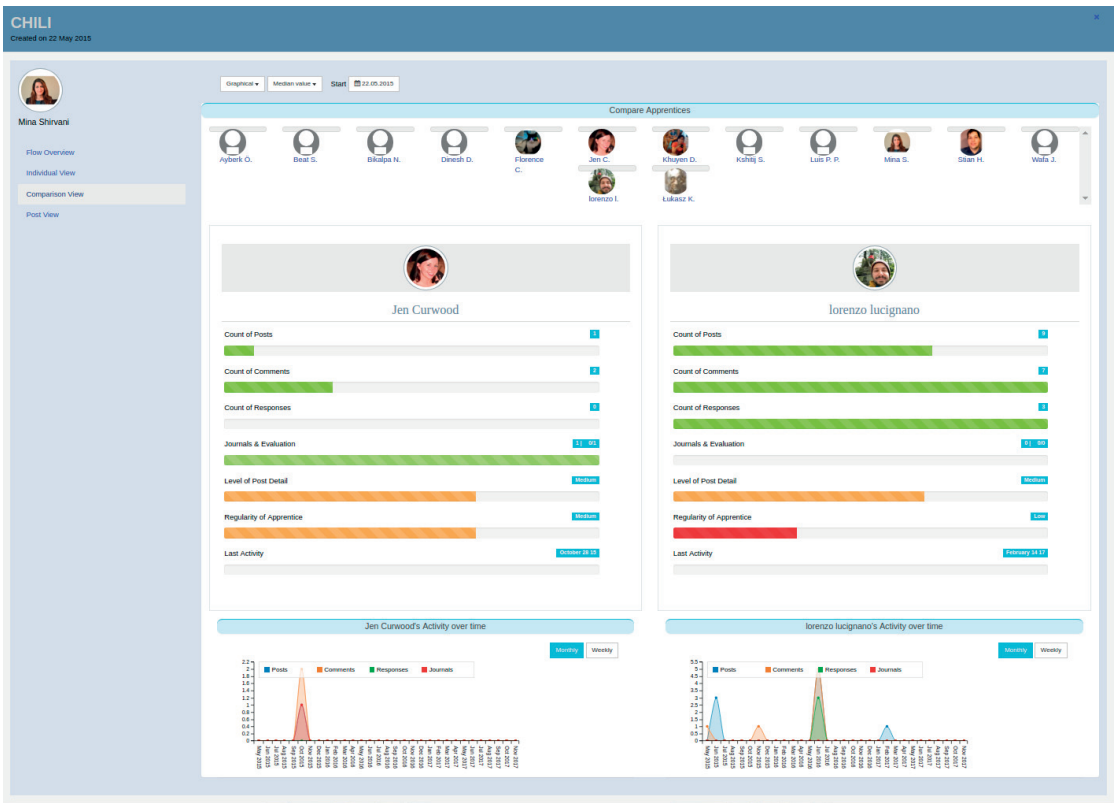


Figure 3.6 *– Comparison view in Realto awareness dashboard*

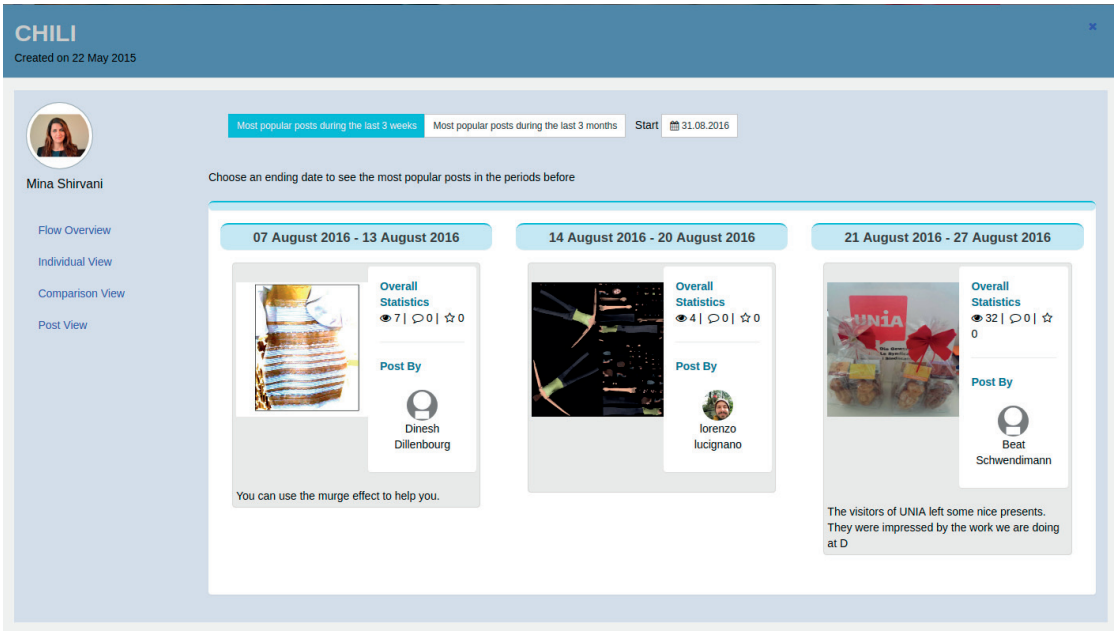


Figure 3.7 *– Post view in Realto awareness dashboard*

3.4 Discussion

In this chapter we outlined the structure of Swiss VET system and mentioned the characteristics and limitations of dual-track vocational training programs. We described the *Erfahrraum* model, which illustrates the vision of Dual-T research project to bridge the gaps between school and workplace practices through technology-enhanced spaces that facilitate conversations between the two contexts. We then described Realto, the online platform developed in Dual-T project to implement the principles of the *Erfahrraum* model by providing a digital space for reflection on experiences and facilitating communication among different actors in the vocational system. We presented our contribution to the development of awareness tools for teachers in this platform. We implemented a contextual learning dashboard which aggregates different indicators of students' activities in Realto. Our proposed dashboard allows for customizing the dashboard content, and enables the teachers to send direct feedback to the learners who need attention.

It should be mentioned that this thesis is not dedicated to educational dashboards. The focus of this thesis is upstream. We are looking for new metrics of learners' behaviors, namely metrics that integrate the time, activity, and social dimensions. Elaborating these metrics is our scientific endeavor (which will be presented in details in the following chapters). We produced some basic visualizations of these metrics in a teacher dashboard but did not consider the design of these visualizations as our research question.

4 Temporal Patterns of Online Participation

The primary focus of this chapter is on the analysis of temporal patterns of learners' participation in online learning environments. Effective use of time has been recognized as a crucial factor for success in many different fields, including education and learning. In educational research, the concept of time has not always been considered in the same way, however it has been constantly regarded as a crucial explicative factor due to its important role in key aspects of teaching and learning process [19, 21, 89]. Consideration of temporal aspects is therefore essential for understanding these processes. However in educational research which use aggregated features or fixed snapshots for describing learners' activities, the temporal dynamics are often overlooked [50].

In this chapter, we start by an overview of the role of time factor in traditional and online educational research in Section 4.1 and formulate the research questions in Section 4.2. In Section 4.3 we introduce quantitative methods facilitating the analysis of temporal activity patterns. In particular we focus on methods for determining repeating patterns in timing of learners' online participation. After detailed description of our proposed methods and metrics, in Sections 4.4 and 4.5 we demonstrate their application on two different educational contexts, MOOCs and Realto. We conclude the chapter in Section 4.6.

4.1 Context

4.1.1 Time factor in educational research

Empirical educational research incorporating time aspect mainly focus on measuring learners' time management behaviours and its relation with learning outcomes. Time management competencies are most commonly considered as: time analysis, planning, goal setting, prioritizing, scheduling, organizing, and establishing new and improved time habits [105]. The common methods used for obtaining individuals' time regulation attitudes and behaviours are self-report measures captured through questionnaires or interviews [53, 71, 79, 147]. The most

Parts of this chapter have been previously published in [203].

commonly used questionnaires for this purpose include *time management behaviour scale (TMBS)* [145], *time structure questionnaire (TSQ)* [30] and *time management questionnaire (TMQ)* [38]. *TMBS* is based on a list of popularized concepts of time management behaviours such as setting goals and priorities, mechanics of time management (e.g. making to-do lists), preference for organization, and perceived control of time. *TSQ* on the other hand consists of items referring to the extent to which time is used in a structured and purposeful way. Finally *TMQ* includes items on attitudes towards time management (e.g. "Do you feel you are in charge of your own time?") and planning the allocation of time. Items related to planning behaviours are a common feature among the available time management questionnaires [53]. Alternatives to these questionnaires include personal diaries [71, 79] and self-report time usage questions that measure how individuals manages their time in a typical weekday [10].

Management of time and its associated advantages in academic performance have been highlighted in numerous empirical studies. Effective time management has shown to be associated with greater academic achievement [24, 145, 147, 153] whereas, poor time management practices, such as failure in proper time allocation or meeting deadlines are frequently cited as major engagement obstacles and to be associated with poor academic performance [141, 145]. Similarly, procrastination, defined as the tendency to delay of the task completion [132] has been found to be negatively correlated with the learning outcomes [86, 109, 127]. Time management shares a strong empirical relationship with conscientiousness [139, 146]. Conscientiousness as a personality trait has also been found to be associated with attainment in higher education [166, 218, 179]. Recent studies suggest that the relationship between conscientiousness and academic performance is mediated by time management [56, 147]. In particular, highly conscientious students tend to regulate their own learning through time management strategies, particularly those related to organization, planning and self-discipline, which in turn lead to academic success. Time management therefore appears as a behavioral expression of high conscientiousness [147].

In summary, in empirical educational research, time-management behaviours were commonly measured using self-report questionnaires. In this context, time-management has been found to be associated with academic performance and learners' conscientiousness.

4.1.2 Time flexibility in online education

With the emergence of open online education and MOOCs, the nature of educational processes and in particular the temporal dimension has been highly transformed [98]. Flexibility in relation to time, place, and pace of study is one of the main proposed benefits of online education [143, 159, 19]. Online learners are required to self-regulate their learning and determine when, how, and with what content they engage [22]. For instance, in MOOCs, course materials including videos and assignments are usually made accessible on a weekly basis, or at the beginning of the course and remain accessible during the course period. Although some MOOCs have specific due dates for assessment tasks and learners are encouraged to

follow the course regularly, they still have the possibility to adopt a study plan which suits their lifestyle and may follow the course out of step with other participants [143].

The flexibility offered by online learning environment, may be a challenge for learners, as much as a benefit. In this context, self-regulation is considered as a critical factor for success. However, not all learners might be equipped with the self-regulation skills (including time management [22]) to manage the flexibility and openness provided [82, 214, 143]. According to recent studies in online educational settings, difficulties with time management are the main obstacle for engaging in a MOOC. About 60% of the respondents to an online survey in [122], indicated time-related reasons as influencing their decision to stop participating. Similarly poor time organization and loosing the rhythm of the course were among the principal MOOC dropout reasons for survey participants in [160].

4.1.3 Temporal analysis in educational research

In online education settings, availability of digital traces of online interactions enables detailed analysis of the temporal aspect of learners' activities. However, time factor in this context is yet under-explored [89]. According to reviews of online educational research literature [20, 19], time is explicitly taken into account only by few articles (23 in [20] and 40 in [19]). Analysis of temporal patterns of learning activities could provide insights about learners' study habits, rhythms, and possible challenges. This aspect has been considered in few recent studies. As an example, Brooks et al. [41] analyzed lecture view patterns in a lecture capture environment by tracking in which weeks of the course learners had accessed the videos. This analysis revealed five types of lecture viewing habits among learners: *high activity* (learners who habitually watch lectures throughout the semester), *just-in-time* (learners who observed lectures the week before the midterm exam), *early* (learners with consistent viewership in the first half of the course), *deferred* (learners with consistent viewership in the second half of the course), and *minimal activity* (learners who habitually did not watch lectures).

In a similar approach Goda et al. [86] tracked participants weekly completion rates of the learning materials in a mandatory online course. They identified seven learning behavioral types through manual categorization of weekly behaviours. These patterns include: *procrastination* (tendency to procrastinate), *learning habit* (appropriate learning pace throughout the course), *random* (learning behavior without a definite tendency), *diminished drive* (starting with a high learning pace, but gradually slowing down), *early bird* (completing the assigned task far before the deadline), *chevron* (increment of learning activity towards the middle of course and slowing down afterwards), and *catch-up* (slow learning pace at the beginning and then catching up to the appropriate pace). Students exhibiting learning habits in this study scored significantly higher compared to procrastinating students. In MOOCs context, Loya et al. [143] considered participants who engaged with the course at roughly the same day every week as being conscientious and self-disciplined, who in turn showed to have lower dropout probability.

Considering essential differences between traditional and online educational settings with respect to the time factor, further empirical studies on time and its relation with learning in the online education context are needed [19]. Furthermore, self-report measures in this context, although effective for reflecting learners' attitudes, intended time plans, and self-perception of their performance, are insufficient for capturing actual learning behaviours. The literature in this domain lacks adequate computational methods for investigating the temporal aspect of learning and indicators to translate temporal analysis into actionable insights [50].

4.2 Problem formulation

The importance of time factor in education and learning, besides the absence of adequate computational methods for studying temporal patterns of learning activities are the main motivations for the presented work in this chapter. A temporal pattern refers to a structure appearing periodically within a given temporal rhythm which enables the understanding of past and anticipating future trends. In this chapter, we propose a quantitative approach for analyzing the temporal patterns of learners' online sessions.

Some learners following an online course, such as in MOOCs, might follow an adaptive approach in which they regulate their learning time according to their daily work or other personal activities and constraints. On the other hand, some learners plan their learning activities and dedicate fixed time slots during the week to their online participation. The timing of online sessions for such learners would represent a regular pattern, whereas this is not necessarily the case for the learners affirming to the first strategy.

The primary focus of this chapter is to introduce methods to assess whether a learner follows specific time schedule and in particular to measure his(her) regularity level in time domain. Investigating the link between time regularity and performance in MOOCs context is our other objective. The research questions we address in this chapter can be summarized as the following:

Question 1. How can we quantify regularity in time domain?
Question 2. Is regularity related to performance?

Concerning the first question, we propose quantitative methods, inspired by time series analysis techniques, to measure time regularity based on the digital traces of learners' actions. Regarding the second question, considering that planning and scheduling are important constructs of time management, in addition to the evidence of positive links between time management and learning outcomes in traditional education context, we hypothesize that regular learners in MOOCs would have higher achievement level. However, considering the time structure of MOOCs which permits participants to flexibly access learning resources at any time during the course period, this hypothesis needs to be empirically validated.

4.3 Method

Time regularity, observed as a repeating pattern in participation time, can be considered as a seasonal component of a time series. In classic time series analysis, researchers often remove this seasonal pattern and focus on modeling the remaining behaviour of the process. However, in the case of students time regularity, the pattern varies for each individual and becomes a characteristic of interest for discriminating students. Therefore, to quantify regularity we can get inspiration from seasonality detection methods in time series analysis [39, 113].

We study regularity in two main domains: time and frequency. For the time domain methods we slice the time series into segments of the length of interest and compare repeatability of the slices. Frequency domain methods are based on the fact that inner product of a signal with a periodic function is large if the signal has the same period as the function [176, 206].

4.3.1 Regularity patterns

Regularity can be assessed at different levels of time such as the day, the week, and longer periods such as the duration of a learning activity. Without loss of generality, we consider weekly and daily regularity, corresponding to periodic participation patterns in a weekly or daily basis. Regularity may emerge as many different temporal patterns. Loya et al. [143] considered only the first access of learners to videos of a week and defined a regular learner as one who accessed videos on the same day every week. This definition, although intuitive, but fails to cover many other forms of regularity, such as studying at multiple fixed week days (e.g Mondays and Fridays) or studying at a certain time of the day (e.g at 8-10).

In this work, we consider six regularity patterns with varying levels of temporal granularity, summarized in Table 4.1. This set of regularity patterns, is not inclusive but covers a wide range of possible regular time patterns. In the next section, we introduce metrics for detecting the described regularity patterns. The first five patterns in Table 4.1 ($P1$ to $P5$), correspond to weekly and daily regularity. However, some learners could be considered as regular not due to a timing routine, but because they follow the schedule of the course. An example could be a learner who is responsive to the course related events and does not postpone watching videos or submitting to the assignments. We refer to this type of regularity as course-based ($P6$) in Table 4.1.

4.3.2 Design of measures

In the following, we propose nine measures in time and frequency domains to quantify the regularity patterns listed in the previous section. Table 4.2 provides an overview of the measures and patterns they reflect and in the remaining of this section we present the detailed description of each measure.

Table 4.1 – *Regularity patterns in time domain and examples.*

Pattern	Type	Description	Example
P1	Weekly	Activity on a certain day(s) of the week	Studying on Mondays, or Mondays and Fridays every week but not necessarily at the same hour
P2	Daily	Activity on certain hour(s) of the day	Studying at 8-10, but not necessarily on all or fixed week days
P3	Weekly	Certain distribution of participation time among week days	Studying for 4 hours on Mondays and 2 hours on Fridays
P4	Daily	Periodic hourly pattern across days	Studying at 8-10 on every day.
P5	Weekly	Periodic hourly pattern across weeks	Studying at 8-10 every Monday, or 8-10 on Mondays and 16-18 on Fridays
P6	Course-based	Following the course schedule	Accessing course materials after they are released, without postponing

Table 4.2 – *Overview of regularity measures and corresponding regularity patterns they reflect.*

Measure	Description	Method	Pattern
CWD	Certain Week Day	Time domain	*P1*
CDH	Certain Day Hour	Time domain	*P2*
WSB	Weekly Similarity Binary	Time domain	*P1*
WSN	Weekly Similarity Normalized	Time domain	*P3*
WSR	Weekly Similarity Raw	Time domain	*P3*
PWD	Periodicity of Week Day	Frequency domain	*P1*
PDH	Periodicity of Day Hour	Frequency domain	*P4*
PWH	Periodicity of Week Hour	Frequency domain	*P5*
RSI	Responsiveness Index	Time domain	*P6*

Binarization of action sequences

In order to study the temporal patterns of online sessions, we start by reconstructing learners' action sequences using the timestamped interaction records from the platform data logs. We then transform the action sequences into binary signals representing activity intervals during the course timeline for each learner. This binarization procedure is explained in the following.

Let n be the number of actions (of any type) performed by a learner and $T = \{t_1, t_2, ..., t_n\}$ be the set of actions timestamps. We set the course start time to $t = 0$. Let L_h, L_d and L_w be the course duration length in hours, days and weeks respectively. We convert learner's action

sequence into a binary signal defined as:

$$F_W(x) = \begin{cases} 1 & \text{if } \exists t_i \in T : x = \lfloor \frac{t_i}{W} \rfloor \\ 0 & \text{otherwise} \end{cases}, \text{ where } x \in \{1, 2, ..., L_h/W\} \tag{4.1}$$

where W is the length of a time window in hours.

With a time window of one hour ($W = 1$), we obtain hourly time signal F_1, where $F_1(x) = 1$ implies that learner had at least one action at hour x after the course start. Similarly, considering a one day time window ($W = 24$) results in the daily time signal F_{24} where $F_{24}(x) = 1$ indicates that at least one action was performed by learner at day x of the course. Figure 4.1 depicts an example of daily time signal for a leaner following a 10 weeks long MOOC. The hourly and daily time signals are the basis of regularity measures we introduce in the remaining of this section.

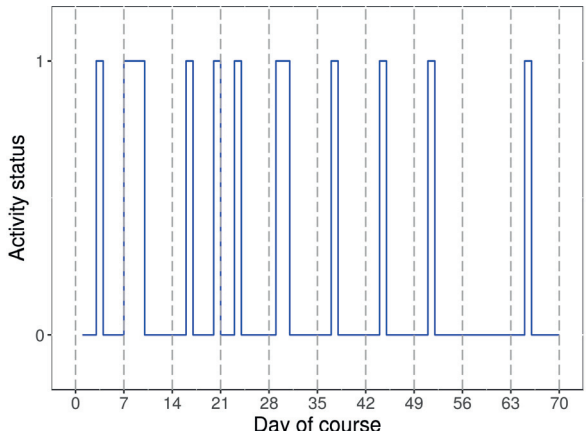


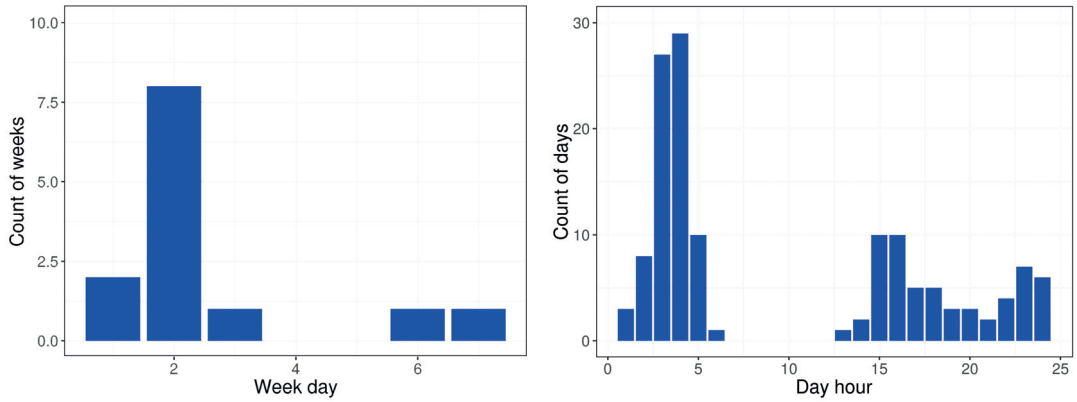
Figure 4.1 – *Example of a learner's binarized daily time signal F_{24}. Points with value of one represent days on which the learner had online activities in the course platform.*

Time domain: Entropy based measures

The first two regularity patterns we consider are studying on certain day of the week (*P1*) and on certain hour of the day (*P2*). In order to detect the first pattern, we build the weekly histogram which represents the distribution of learners' activities across different week days. We define weekly histogram as a function $W(d)$ on every day d of a week:

$$W(d) = \sum_{i=0}^{L_w - 1} F_{24}(7i + d), \text{ where } d \in \{1, 2, ..., 7\}. \tag{4.2}$$

Following a similar approach we construct the daily histogram to represent distribution of activities over different day hours. Daily histogram is defined as a function $D(h)$ on every hour



(a) *Weekly histogram* ***(b)*** *Daily histogram*

Figure 4.2 – *Examples of weekly and daily activity histograms for one learner, respectively representing the distribution of study time over week days and day hours (time zones are not compensated and the start time of the course is considered as the reference point with $t = 0$).*

of the day:

$$D(h) = \sum_{i=0}^{L_d - 1} F_1(24i + h), \text{ where } h \in \{1, 2, ..., 24\}. \tag{4.3}$$

Therefore $W(d)$ represents the number of weeks in which learner was active at day d, and $D(h)$ corresponds to the number of days on which learner was active at hour h. Figure 4.2 depicts examples of weekly and daily histograms. If study time for a learner is concentrated around a particular day (hour) it would appear as a peak at the corresponding point in the weekly (daily) histogram. Therefore to reveal such pattern, we focus on detecting spikes in the histograms. A simple approach would be to check if the largest value is above a certain threshold. This however would require to manually define the threshold value. A popular measure to identify if a given distribution is uniform or has a spike, is entropy. Based on its definition, we suggest weekly and daily entropy as:

$$\begin{aligned} E_W &= -\sum_{d=1}^{7} \hat{W}(d) \ln(\hat{W}(d)), \\ E_D &= -\sum_{h=1}^{24} \hat{D}(h) \ln(\hat{D}(h)) \end{aligned} \tag{4.4}$$

where $\hat{D}$ and $\hat{W}$ are normalized histograms.

The value of E_D is bounded in $[0, ln(24)]$ and similarly E_W is bounded in $[0, ln(7)]$. A small entropy value encodes presence of spikes in the distributions. However, since entropy is computed on the normalized histogram, it does not reflect the magnitude of the spike in the original histogram. To overcome this limitation, we define our first regularity measure, ***C****ertain*

Week Day (***CWD***) as:

$$CWD = (ln(7) - E_W)\max(W(d)), \text{ for } d \in \{1,2,...,7\} \tag{4.5}$$

In this formulation, a high value of the first component ($ln(7) - E_W$) encodes the presence of a spike in the weekly histogram and the second component ($\max(W(d))$) reflects its magnitude. A high value of CWD therefore reflects concentration of activities on a particular day. CWD is bounded in $[0, ln(7).L_w]$, where L_w is the course duration in weeks.

Similarly, we define ***Certain Day*** **H**our (***CDH***) measure as:

$$CDH = (ln(24) - E_D)\max(D(h)), \text{ for } h \in \{1,2,...,24\}. \tag{4.6}$$

CWD is bounded in $[0, ln(24).L_d]$, where L_d is the course duration in days. A high value for this measure denotes that learner's activities are concentrated around a particular time of the day.

Time domain: Profile similarity measures

The second set of regularity measures we introduce in time domain, are based on the comparison of weekly study profiles. We define study profile ($\vec{P}_k$) of a learner during week k as a vector encoding distribution of study time over week days:

$$\vec{P}_k = [p_{k1}, p_{k2}, ..., p_{k7}] \in \mathbb{R}^7 \tag{4.7}$$

where p_{kd} represents amount of study time (in hour) by the learner in day d of week k and is computed using the hourly activity signal (F_1 in equation 4.1) as:

$$p_{kd} = \sum_{i=1}^{24} F_1(24(7(k-1) + (d-1)) + i), \text{ where } d \in \{1,2,...,7\},\ k \in \{1,2,...,L_w\}. \tag{4.8}$$

Figure 4.3 provides an example of weekly profile matrix for a learner following a 10 weeks long MOOC. In the profile matrix, rows correspond to course weeks, columns represent week days and color intensities show learner's estimated study time (in hours) on a particular day. Learners who follow a certain weekly plan ($P3$ in Table 4.1) would have similar rows in their profile matrix. In order to compare the weekly profiles, we define three similarity functions:

- Binary profile similarity (Sim_B): compares the binary version of the learner's weekly profiles and measures if the learner works on the same week days.
- Normalized profile similarity (Sim_N): compares the normalized profiles and measures if learner has a similar distribution of workload among week days.
- Raw profile similarity (Sim_R): compares the raw profiles and reflects if the time spent on each day of the week is similar in different weeks of the course.

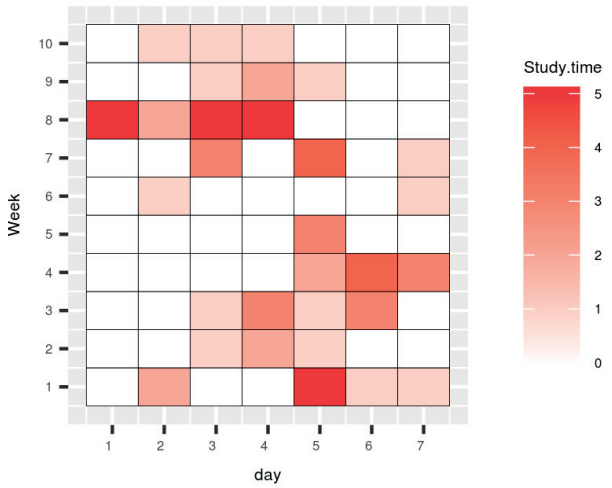


Figure 4.3 – *Example of weekly profile matrix for a learner. Rows represents weeks, columns represent days and intensity of colors encodes estimated amount of study time (in hours).*

Binary profile similarity: Let A_k be the set of days in week k, on which the learner had some activity. We define the binary profile similarity function as the ratio of common active days in the weeks of comparison (Jaccard similarity coefficient):

$$Sim_B(\vec{P}_i, \vec{P}_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|} \tag{4.9}$$

The value of Sim_B is bounded in $[0,1]$ and for two weeks in which the learner is active on the exact same days, irrespective of the count of such days, this similarity function has the maximum value.

Normalized profile similarity: To compare the normalized profiles ($\hat{P}_k$) of two weeks, we use Jensen-Shannon Divergence (JSD) [136], which is a symmetric and bounded metric for measuring the similarity between probability distributions:

$$Sim_N(\hat{P}_i, \hat{P}_j) = 1 - \frac{JSD(\hat{P}_i, \hat{P}_j)}{ln(2)} \tag{4.10}$$

where $\hat{P}_k$ is the normalized profile ($\hat{P}_k = \vec{P}_k / \sum_d p_{kd}$) and JSD based on its definitions is:

$$JSD(\hat{P}_1, .., \hat{P}_n) = H\Big(\sum_{i=1}^{n} \pi_i \hat{P}_i\Big) - \sum_{i=1}^{n} \pi_i H(\hat{P}_i), \tag{4.11}$$

where π_i is the selected weight for the probability distributions $\hat{P}_i$ and H is the entropy function. We consider uniform weights for all weeks, hence $\pi_i = 1/n$. The value of Sim_N is bounded in $[0,1]$ and high value of this measure reflects similar shapes of study profiles in the weeks of comparison. An example could be profiles with relatively more study time on Monday compared to Tuesday and Wednesday.

Raw profile similarity: In order to capture the similarity in shape and magnitude of weekly profiles, we define the third similarity function, based on χ^2 divergence [46] as:

$$Sim_R(\vec{P}_i, \vec{P}_j) = 1 - \frac{1}{|A_i \cup A_j|} \sum_{d=1}^{7} \left(\frac{P_{id} - P_{jd}}{P_{id} + P_{jd}} \right)^2 \tag{4.12}$$

The value of Sim_R is also bounded in $[0,1]$ and it highest value is achieved if the two weekly profiles of comparison are identical.

Finally based on the presented similarity functions, we define three regularity measures ***W**eekly **S**imilarity **B**inary* (***WSB***), ***W**eekly **S**imilarity **N**ormalized* (***WSN***) and ***W**eekly **S**imilarity **R**aw* (***WSR***) as the average pairwise similarity of weekly profiles according to Sim_B, Sim_N and Sim_R respectively.

Frequency domain measures

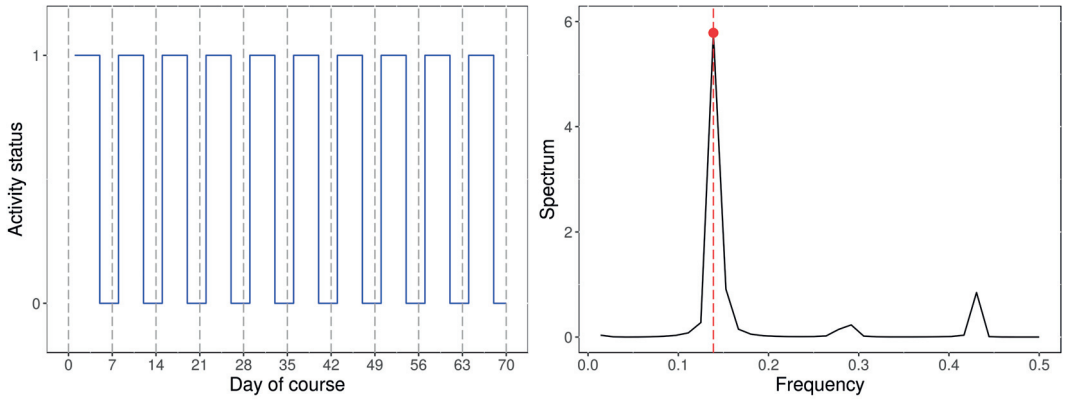
The third category of regularity measures we introduce are calculated in frequency domain, aiming to detect seasonality in learners' activity time signals which were defined in Equation 4.1. In signal processing and time series analysis, one common approach to detect seasonal components of a signal is to convert it from the original domain (often time) to a representation in the frequency domain by applying Fast Fourier Transform (FFT) algorithm [35]. Frequency domain representation of a time signal $X(t)$ is computed as:

$$\mathscr{F}(\theta) = \sum_{t=0}^{N-1} X(t) e^{(-i2\pi t\theta/N)} \tag{4.13}$$

where N is the sequence length.

The resulting signal $\mathscr{F}(\theta)$ is referred to as spectral density or periodogram and is used to detect periodicity in the data, by observing peaks at the frequencies corresponding to these periodicities. The upper bound of $F(\theta)$ is the coherent summation of all samples of the original signal: $\mathscr{F}(\theta) \leq \sum_{t=0}^{N-1} |X(t)| \leq N.max(|X(t)|)$ and therefore if $\forall t |X(t)| \leq 1$ this upper bound would be smaller than N.

For the purpose of detecting weekly or daily regularity, we compute spectral density of learner's time signals, F_1 and F_{24}, to obtain hourly and daily periodograms $\mathscr{F}_h(\theta)$ and $\mathscr{F}_d(\theta)$. From the resulting periodograms we then extract values corresponding to daily and weekly periods. If there is a daily or hourly repeating pattern in learner's activities time signal, we expect the extracted values to be relatively large. This is illustrated by an example in Figure 4.4 which presents the time and frequency domain representation of the activity signal for a learner who follows a regular daily pattern. As it can be seen, this regular pattern is reflected by a spike in the frequency signal at the point corresponding to one week period. Following this approach, we define three regularity measures, ***P**eriodicity of **W**eek **D**ay* (***PWD***), ***P**eriodicity of **D**ay **H**our*



(a) *daily time signal (F_{24})* ***(b)*** *daily periodogram ($\mathscr{F}_d(\theta)$)*

Figure 4.4 – *Example of daily activity signal in time (left) and frequency (right) domains for one learner with a regular daily pattern. Dashed red line in the periodogram shows the frequency corresponding to one week period (1/7).*

(***PDH***) and ***P****eriodicity of* ***W****eek* ***H****our* (***PWH***) as:

$$\begin{aligned} PWD &= \mathscr{F}_d(1/week) = \mathscr{F}_d(1/7) \\ PDH &= \mathscr{F}_h(1/day) = \mathscr{F}_h(1/24) \\ PWH &= \mathscr{F}_h(1/week) = \mathscr{F}_h(1/(24\times 7)) \end{aligned} \tag{4.14}$$

PWD captures if the daily pattern of activities is repeating over weeks (e.g. the learner is active on Monday and Tuesday in every week). PDH measures the extent to which the hourly pattern of learner's activities is repeating over days (e.g. the learner is active at 8h-10h and 12h-17h on every day). PWH identifies if the hourly pattern of activities is repeating over weeks (e.g. in every week, the learner is active at 8h-10h on Monday, 12h-17h on Tuesdays, etc.). The upper bound of PWD is L_d (course duration in days) and the upper bound of PDH and PWH is L_h (course duration in hours).

It is worth noting that since our aim was to assess daily and weekly regularities, we focused on the corresponding value of the periodograms at these intervals. However, other intrinsic seasonalities in the activity signals could be detected by identifying the peaks at the frequency domain signals.

Adherence to course schedule

The last regularity measure we define, reflects student's responsiveness to course events (pattern *P*6 in Table 4.1). Some students watch the lecture right after it is released whereas others postpone watching lectures or submitting to assignments. Therefore some learners are regular not due to time routine, but as they follow the course schedule and are responsive

to course-related events. This constitute the basis for another approach to the analysis of regularity. Since not all MOOCs consist of formative assessment during the semester, we focus on video watching behaviours to capture adherence to the course schedule. We define ***Res**ponsiveness **I**ndex* (***RSI***) measure as:

$$RSI = 1 - \frac{1/n \sum_{i=0}^{n} (V_i - R_i)}{L_h} \tag{4.15}$$

where V_i is the time of learner's first view of the video i, R_i is the release time of that video, n is the number of watched videos, and L_h, as mentioned before, is the course duration in hours. RSI is therefore opposite to the average delay in watching the course videos. Its value is bounded to $[0,1]$ and a learner who does not postpone watching video lectures, achieves a high value for this measure.

4.4 Temporal participation patterns in MOOC

In this section, we present the results of applying the described methods on a MOOC dataset. We provide examples for each of the regularity measures and investigate the link between regularity and performance.

4.4.1 Dataset

Our analysis in this section is based on an undergraduate engineering MOOC, Produced by EPFL university and offered in Coursera entitled *"Functional Programming Principles in Scala"*[1]. This course covers the principles of functional programming paradigm including the use of functions as values, recursion, immutability and several other topics, using Scala programming language. Total duration of the course was 10 weeks and lectures were released on a weekly basis during the first seven weeks. The final grade was calculated based on six graded assignments and passing grade was 60 out of 100. For the analysis of regularity patterns, we considered learners with at least three weeks of activity on the platform (14,900 learners). Learners who did not submit any assignments and therefore scored zero in the course (4,644 learners) were removed from the dataset. Some participants, never watched a video on the platform, instead they downloaded the lectures and probably watched them off-line (225 learners). Since activity traces for such learners is not available, we removed them from the dataset as well. Therefore, in our analysis we considered all events by remaining 10,031 participants. Their average grade was 55.7 and 51% scored higher than the passing threshold (60).

[1] https://www.coursera.org/learn/progfun1

Table 4.3 – *Overview of regularity measures in the dataset*

Measure	Mean	SD	Max
CWD	1.12	1.08	13.62
CDH	4.65	3.65	49.92
WSB	0.14	0.13	0.90
WSN	0.17	0.15	0.88
WSR	0.11	0.10	0.74
PWD	0.36	0.35	4.64
PDH	0.34	0.64	14.65
PWH	0.17	0.25	4.2
RSI	0.86	0.11	1

4.4.2 Examples of regularity measures

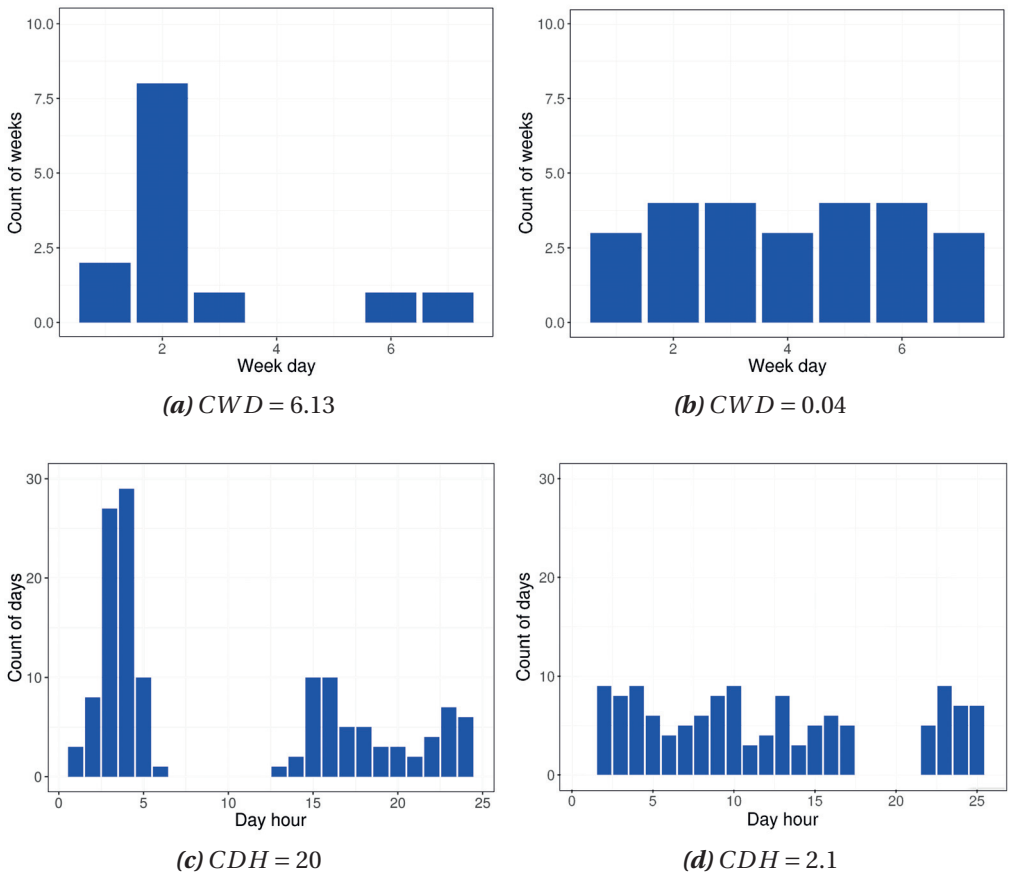
Table 4.3 provides an overview of the computed regularity values for the 10,031 participants in our MOOC dataset. In the following we present examples of proposed features to verify if they capture the regularity patterns as expected.

Entropy based measures: CWD and CDH

These two measures, were defined respectively based on the entropy of the weekly and daily activity histograms (Equations 4.5 and 4.6), to detect strong peaks in the histograms and determine whether learner's activities are concentrated around a particular day of the week, or at a certain time of the day. Figure 4.5 illustrates examples of weekly and daily histograms for learners with high and low values for CWD and CDH measures. The learner in Figure 4.5a, is mostly active on the second day of the week and as expected achieves a relatively high value (6.13) for CWD. Learner in Figure 4.5b on the other hand has no modal day in the weekly histogram as his(her) activities are quite uniformly distributed among different days. As expected, this learner achieves a low value (0.04) for CWD measure. Regarding the CDH measure, a similar trend is observable in Figure 4.5c and 4.5d where a high value (20) of CDH represents peak of activity at particular day time. These examples therefore show how CWD and CDH measures capture regularity patterns $P1$ and $P2$, introduces in Table 4.1.

Profile similarity measures: WSB, WSN and WSR

These three measures were defined based on the similarity between weekly study profiles of the learner during the course duration. WSB, WSN, and WSR respectively results from the comparison of binary, normalized, and raw version of the learner's weekly study profiles (Equations 4.9, 4.10, 4.12). Figure 4.6, 4.7, and 4.8 provide examples of study profile matrix and the corresponding profile similarity measures for three different learners. Activities of the learner in Figure 4.6 are clearly concentrated on the second half of the week and the amount

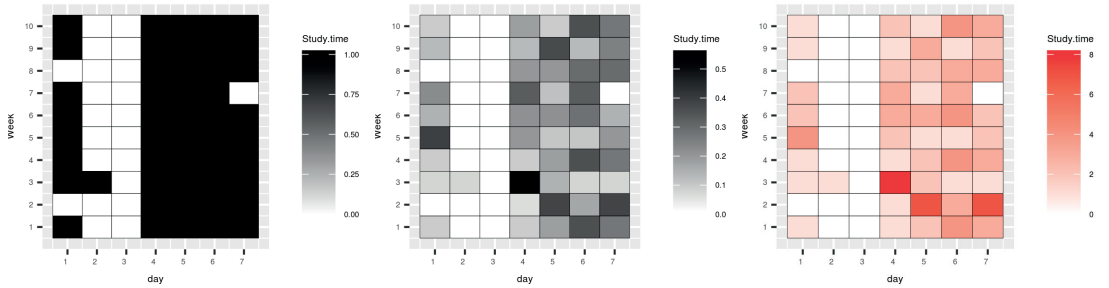


(a) $CWD = 6.13$ ***(b)*** $CWD = 0.04$

(c) $CDH = 20$ ***(d)*** $CDH = 2.1$

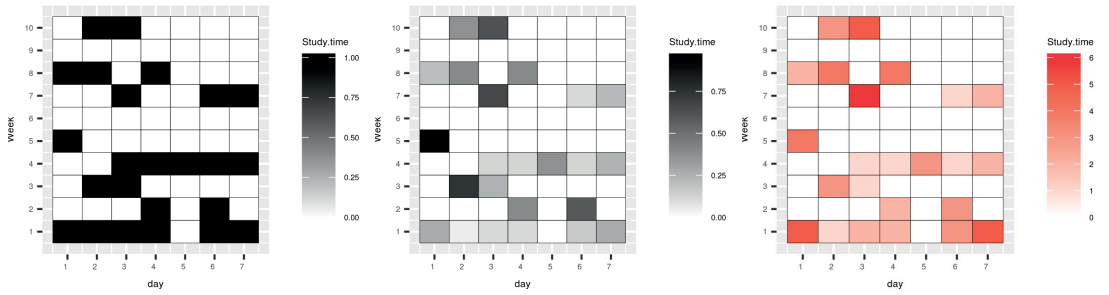
Figure 4.5 – *Examples of CDH and CWD measures: Daily (top) and hourly (bottom) histograms of four learners with high and low values. Clearly a high value of CDH and CWD reflects a peak in the corresponding activity histograms.*

of study time does not vary largely. This learner obtains high values (0.86, 0.84 and 0.74) for all the three regularity measures. For the learner in Figure 4.7 on the other hand, no regular pattern is evident in weekly study profiles and all three profile similarity measures return a low value (0.17, 0.2 and 0.13) in this case. Figure 4.8 provides an example highlighting the difference between these three measures. The learner in this figure, is mainly active during the forth and fifth week days and his(her) study time on day five is almost two times more than day four in all weeks. This regular time distribution pattern results in similar rows in the normalized weekly study matrix in Figure 4.8b which is also reflected by the high value (0.91) of WSN measure. However, since in some weeks the learner has slight amount of activity in few other days, the WSB returns a smaller value (0.69). Furthermore, according to Figure 4.8c the absolute amount of study time for this learner varies between different course weeks, which in turn results in a smaller value (0.55) for WSR measure. These examples therefore confirm that WSB, WSN and WSR measures capture the corresponding regularity patterns as expected.



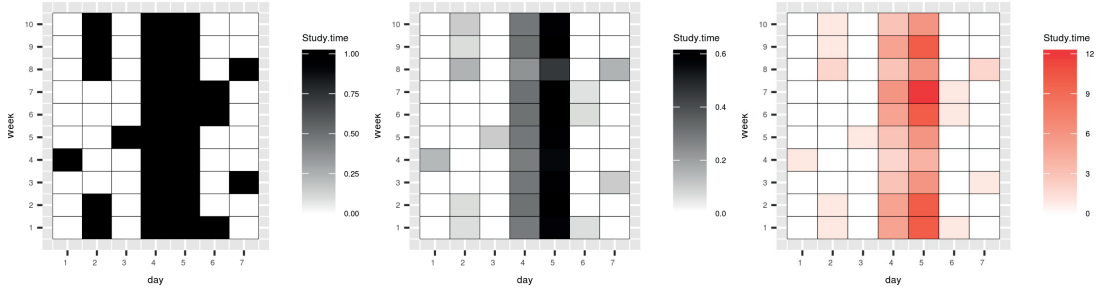
(a) *WSB=0.86* ***(b)*** *WSN=0.84* ***(c)*** *WSR=0.74*

Figure 4.6 – *Example of WSB, WSN and WSR measures: (a) Binary (b) Normalized and (c) Raw weekly study profiles of a learner with high values for the three profile similarity measures.*



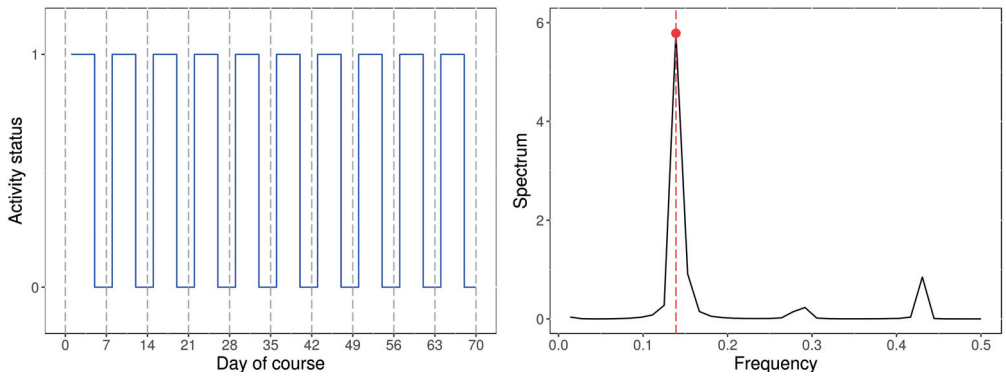
(a) *WSB=0.17* ***(b)*** *WSN=0.2* ***(c)*** *WSR=0.13*

Figure 4.7 – *Example of WSB, WSN and WSR measures: (a) Binary (b) Normalized and (c) Raw weekly study profiles of a learner with low values for the three profile similarity measures.*

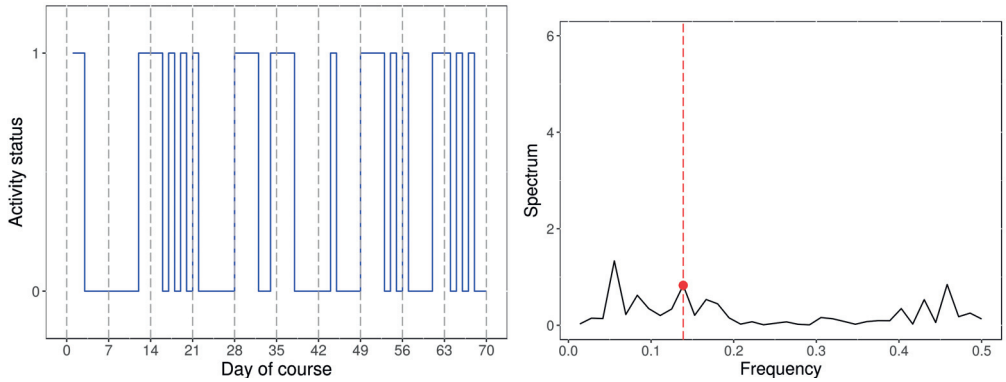


(a) *WSB=0.69* ***(b)*** *WSN=0.91* ***(c)*** *WSR=0.55*

Figure 4.8 – *Example of WSB, WSN and WSR measures: (a) Binary (b) Normalized and (c) Raw weekly study profiles of a learner with different values for the three profile similarity measures.*

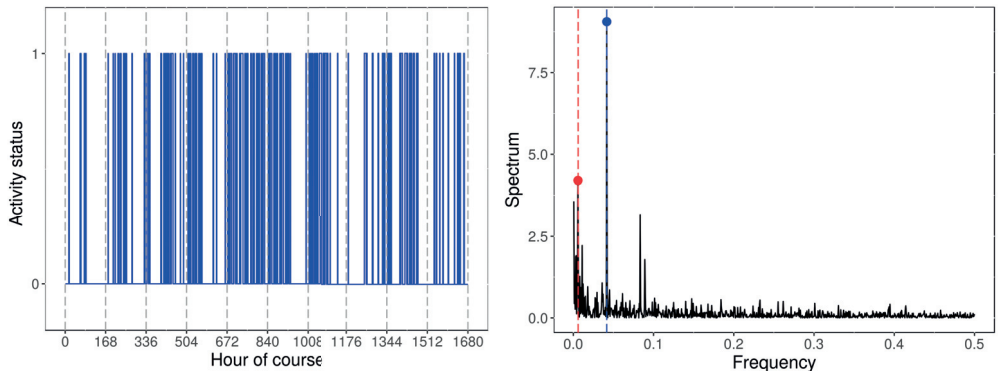


(a) *PWD=5.79*



(b) *PWD=0.83*

Figure 4.9 – *Examples of PWD measure: Daily activity signal in time (left) and frequency (right) domains for two learners with (a) high and (b) low values of PWD. Dashed red lines in the periodograms show the $(1/week)$ frequency. The periodic daily pattern is clearly reflected by the high value of PWD measure.*



(a) *PDH=9.6, PWH=4.19*

Figure 4.10 – *Example of PDH and PWH measures: Hourly activity signal in time (left) and frequency (right) domains for a learner with high values of PWH and PDH. Dashed red and solid blue lines in the periodogram respectively show the $(1/week)$ and $(1/Day)$ frequencies.*

Frequency domain measures: PWD, PDH and PWH

These three measures were defined based on the frequency domain representations of learner's time signals (Equation 4.14) in order to capture periodic daily patterns over weeks (PWD), periodic hourly patterns over days (PDH) and periodic hourly patterns over weeks (PWH). Figure 4.9 illustrates examples of two users with high and low value of PWD measure. The daily time signal in Figure 4.9a clearly shows a repeating weekly pattern. This periodic pattern is captured by the large value (5.79) of the PWD measure, which is the magnitude of the periodogram at the frequency corresponding to one-week period. On the contrary, no seasonal pattern is evident in user's time signal in Figure 4.9b and consequently PWD obtains a small value (0.04). PDH and PWH measure follow the same principle and Figure 4.10 shows an example of these two measures.

4.4.3 Correlation between regularity measures

Pairwise Pearson's correlations between regularity measures shows that the profile similarity measures, WSB, WSN and WSR, although sensitive to different study profiles (as shown in Figure 4.8), result to have strong correlation ($r(10029) = 0.9$, $p < .01$) in the MOOC dataset of our study. PWD measure is also moderately correlated with profile similarity measures ($r(10029) = 0.57$, $p < .001$). The remaining set of measures are not strongly correlated with each other, inferring that they capture orthogonal patterns of regularity.

4.4.4 Regularity and performance

This section addresses our second research question on the link between regularity and performance (**Question 2**). We start by correlation analysis between regularity measures and final grade. We then use clustering methods to extract different categories of learners with respect to their regularity behaviours and investigate their performance level. Finally we track the regularity level of learners over time, and explore differences between students who passed or failed the course.

Predictive power of regularity measures

Correlation analysis reveals positive links between final grade and regularity measures. Pearson's correlation between regularity measures and final grade shows that final grade is strongly correlated[2] with WSN ($r = 0.71$) and moderately correlated with PWD ($r(10029) = 0.47$). Figure 4.11 show the relation between grade and the two mentioned regularity measures. Final grade also shows a moderate correlation with PDH ($r(10029) = 0.32$) and PWH ($r(10029) = 0.37$), slight correlation with CDH ($r(10029) = 0.26$) and RSI ($r(10029) = 0.25$). The high correlation between final grade and weekly regularity measures (WSN and PWD) suggests

[2] $p < .001$ for all reported correlations

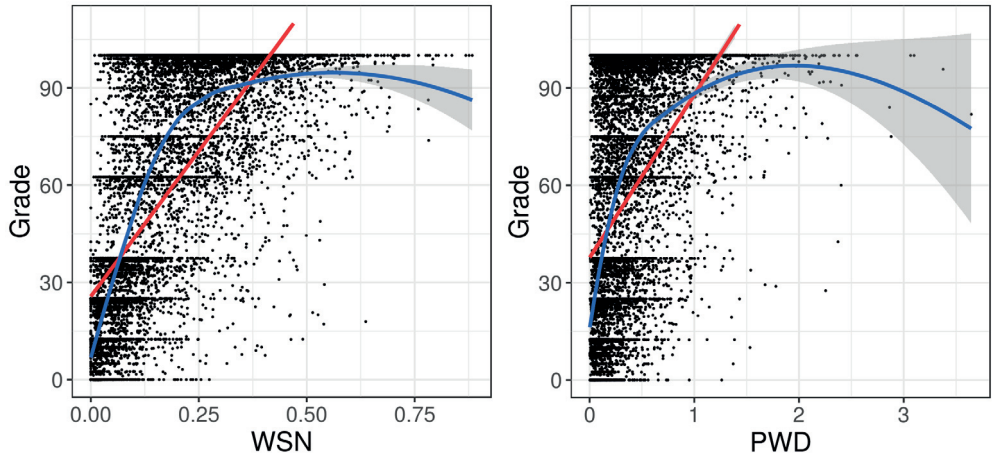


Figure 4.11 – *Scatter plots of grade vs. WSN and PWD measures. Red lines shows the linear smoothed estimations and the blue curves represent the local smoothed estimations. The gray areas show the 0.95 confidence intervals.*

that learners who follow a certain weekly plan, also achieve higher grades in the course, whereas following a certain hourly plan (daily regularity) is less strongly associated with the performance.

In order to assess the predictive power of the regularity measures we build a linear model of the final grade including all of the described regularity measures and use penalized regression to improve the model by removing features of low importance. In our dataset, linear model with variables WSN, PWD, PWH and RSI, captures 52% of the variability of the final grade ($R^2 = 0.52$). This indicates the predictive potential of the designed metrics which makes them promising for being integrated in performance prediction models. Table 4.4 shows the estimated model.

Table 4.4 – *Linear model for final grade estimated using regularity measures*

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	-9.6	2.0	-4.7	2.6e-06
WSN	166.8	2.3	72.3	<2e-16
RSI	40.4	2.4	16.9	<2e-16
PWD	11.6	0.9	11.7	<2e-16
PWH	-9.9	1.3	-7.3	2.5e-13

Model: $lm(Grade \sim WSN + PWD + PWH + RSI)$

But do the regularity measures provide an added value to simple aggregated features such as time on task? To assess this we do the following test. We extract a new feature to represent the total amount of study time (in hours) during the course duration for each student. We then build two different linear models of final grade, one including only this new feature, and the

other including also the regularity measures. According to the results, integrating regularity measures significantly improves the model ($F[1,9] = 602.5, p < 0.001$). The total study time feature captures only 27% of the variability of the final grade ($R^2 = 0.27$ for the first model) whereas the combined model reflects 53% of grade variance ($R^2 = 0.53$ for the second model). This results show that regularity measures provide a novel view of learners participations which is different from and complementary to the total study time.

To explore in more details the effects of time regularity on final grade and its interaction with the amount of study time, we build a regression model of final grade based on the interaction of total study time and WSN weekly regularity features. The resulting model which is represented in Table 4.5 has an R^2 of 0.56 which is considerably large for a model with only two parameters. Figure 4.12 depicts the interaction plot between total study time and weekly regularity in the resulting model and provides interesting insights on the interaction effect of these two variables. According to this figure, learners with large amount of study time (e.g. 200 hours) obtain high final grades regardless of their regularity value. On the other hand for learners with limited study time (e.g. less than 50 hours), regularity is a factor which could positively influence the final grade. That is, learners with few hours of study who follow a weekly regular pattern (upper tercile of WSN) achieve high final grades, whereas those with low amount of

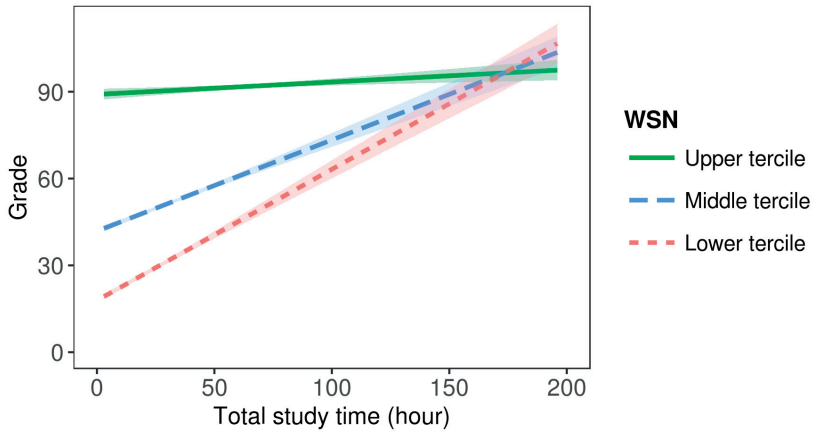


Figure 4.12 – *Interaction effect between weekly regularity (WSN) and total study time in regression model of final grade.*

Table 4.5 – *Linear model of final grade based on the interaction between weekly regularity and amount of study time.*

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	12.3	0.6	21.6	<2e-16
WSN	222.4	2.8	78.8	<2e-16
Total study time	0.48	0.02	23.9	<2e-16
WSN : Total study time	-1.3	0.03	-36.4	<2e-16

Model: $lm(Grade \sim WSN * Total\_Study\_Time)$

study time and not-regular study patterns (lower tercile of WSN) receive low final grades. In summary, this results suggest that regularity is important for performance specially when the dedicated study time is limited.

Learners categories

To investigate behavioral categories among learners with respect to time regularity, we applied hierarchical agglomerative clustering in combination with euclidean distance on the calculated regularity measures. Prior to clustering, all measures were scaled and centered for comparability. We used Silhouette method [191] to estimate the optimal number of clusters, and set the number of clusters to four, as it resulted in the maximum average clustering coefficient. The obtained categories are represented in Figure 4.13. Learners in Cluster 1 and 4 have

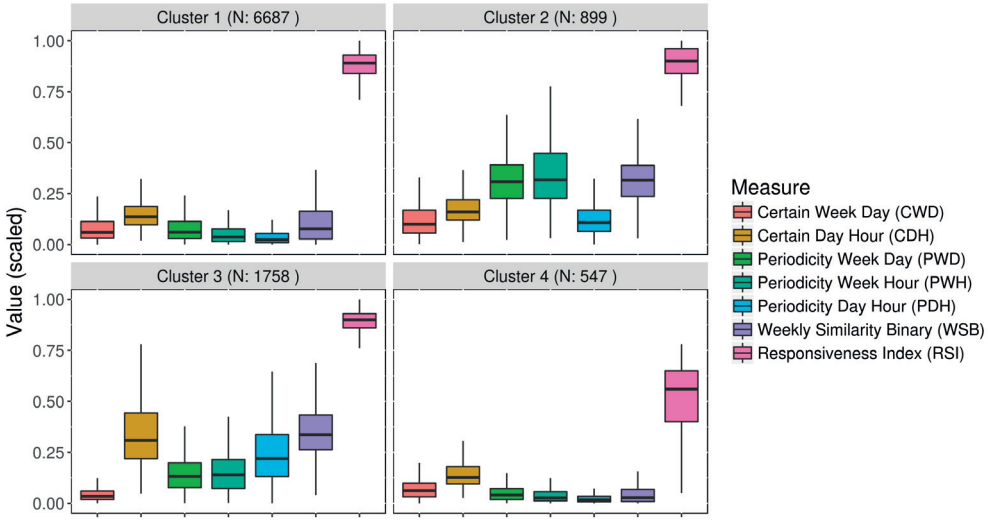


Figure 4.13 – *Clusters of learners based on regularity measures. All values were scaled to [0,1] for visualization purpose. Learners in* ***Cluster 1*** *are not-regular but are responsive,* ***Cluster 2*** *are weekly regular and responsive,* ***Cluster 3*** *are daily regular and responsive,* ***Cluster 4*** *are not-regular and not-responsive.*

Table 4.6 – *Comparison of weekly and daily regularity measures for Cluster 2 and 3 in Figure 4.13, using one-way Anova test without assuming equal variances.*

Measure	Cluster 2	Cluster 3	F statistics
CWD	**1.7**	0.6	F[1, 1059.8]= 639.8, p<.001
PWD	**0.99**	0.46	F[1, 1334.7]= 1096.8, p<.001
PWH	**0.56**	0.26	F[1, 1298.8]= 755.7, p<.001
CDH	4.27	**8.3**	F[1, 2649.5]= 1087.2, p<.001
PDH	0.46	**0.91**	F[1, 2649.1]= 576.1, p<.001

relatively low values for the time regularity measures. Regarding the responsiveness index, according to one-way Anova test, this measure is relatively high for the first three clusters and is significantly lower for learners in Cluster 4 (0.5 vs. 0.9, $F[1,555.5] = 2549, p < .001$). Considering Cluster 2 and 3, as summarized in Table 4.6, learners in Cluster 2 are attributed with significantly higher weekly regularity (CWD, PWD, and PWH) and lower daily regularity (CDH and PDH).

Based on the described trends, Cluster 1 to 4, could respectively be labeled as not-regular and responsive, weekly regular and responsive, daily regular and responsive, not-regular and not-responsive learners. Comparison of average final grade by learners in each cluster (Figure 4.14) shows that weekly/daily regular and responsive learners (Cluster 2 and 3) achieve significantly higher grade compared to not-regular and responsive (Cluster 1) learners (86 vs. 46, $F[1,8036.8] = 5482.6, p < .001$). The lowest average grade (17) is associated to not-regular and not-responsive learners.

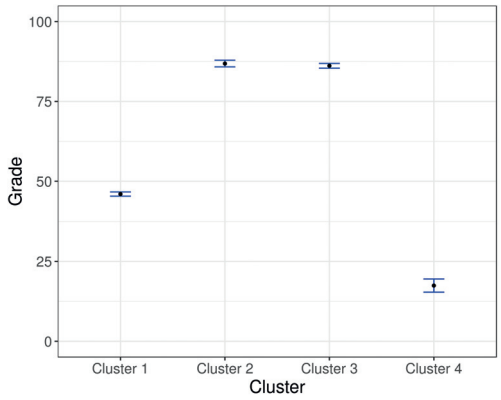


Figure 4.14 – *Average grade of clustered learners. Daily/weekly regular and responsive learners (Cluster 2 and 3) achieve significantly higher grades.*

Regularity over time

The results presented so far, were based on the overall regularity level estimated at the end of the course. The proposed measures, in particular the profile similarity measures (equations 4.9, 4.10 and 4.12), could also be computed throughout the course duration, making it possible to track learners' regularity over time. As an example, Figure 4.15 presents the weekly regularity for passed ($N = 5096$) and failed ($N = 4935$) students over the course duration. In this example, regularity in each week is computed based on the similarity between learners' binary weekly profiles over the past two weeks (equation 4.9, the pattern of studying on the same week days). According to this figure, both groups establish a more regular study pattern after about three weeks from the course start. Passing students have relatively higher regularity level throughout the course; whereas there is an evident drop of regularity around the fourth week for the failed students. The final set of lectures and assignments were released on week seven which could

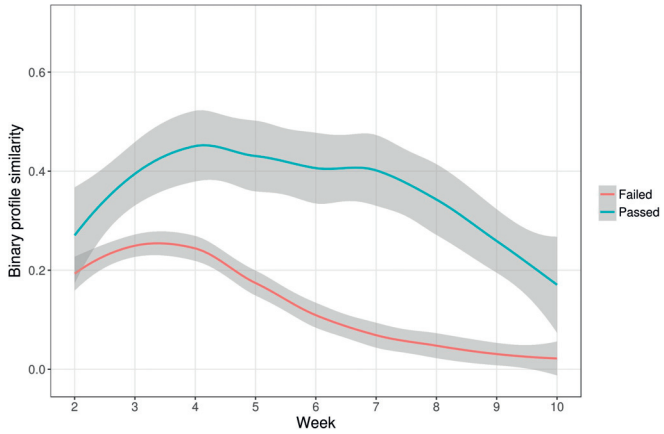


Figure 4.15 – *Regularity over time for passed and failed students. The gray area shows the* 95% *confidence interval.*

explain the decrement of passing students' regularity level around this week.

4.4.5 Other applications

The regularity features proposed in this chapter could also be used for measuring the effect of external factors (e.g. platform features, instructor's interventions, or employment status) on participation patterns. As an example, here we investigate the link between regularity and employment status. The MOOC database used for this study contains employment information for about 9.6% of the participants. Based on these information we extract two categories of learners: *full-time employed* and *full-time students* (559 v.s. 113 learners). We assume that participants in both categories have a daily or weekly routine imposed by their occupation or school schedule. One-way Anova test (without assuming equal variances) with regularity measure as dependent and employment status as independent variable shows that employed participants have higher time regularity both in weekly and daily basis. This is reflected by significantly higher values of WSN measure ($F[1,170.7] = 5.43$, $p = .02$), PWD measure ($F[1,187.2] = 5.51$, $p = .02$) and CDH measure ($F[1,295.9] = 19.83$, $p = 1e-5$) for employed participants.

4.5 Temporal participation patterns in Realto

In the context of Realto, investigation of temporal patterns of users' activities is of particular interest for the researchers involved in the Dual-T project. Information about the creation time of entries in Realto (weekdays or weekends, during working hour or in the eventing and etc.) could help in understanding the importance given by teachers, supervisors and apprentices to the use of Realto. Furthermore identifying the time patterns adopted by different user categories could be insightful for customizing the platform features. For instance, the timing of notification and reminder messages could be adopted based on users' activity time; Or the

required steps for creating a new post or uploading materials in Realto could be simplified to facilitate this process for the apprentices who mainly use the platform during working hours, and reflective prompts could be sent out to them during other day times.

The methods presented in Section 4.3 are transferable to different contexts. As the only input for the proposed methods is the set of actions timestamps, they can be applied to any context and platform where interaction logs and their associated time information are being recorded. We used the described methods to provide the time analysis infrastructure in Realto and implemented the *time analysis* module in Realto analytics dashboard (Appendix A.1). Figure 4.16, 4.17 and 4.18 provide examples of the information included in the dashboard for the members of a particular flow, selectable from the dashboard interface. In Figure 4.16, each chart corresponds to a learner's weekly histogram (Equation 4.2), depicting number of weeks on which he/she was active on a particular weekday. Similarly the charts in Figure 4.17 represent the daily histograms (Equation 4.3), illustrating the number of days with activity at each hour. Similar to the MOOC context, in Realto, the regularity measures could be applied as indicators at the individual level to assess the presence of a periodic time schedule in user's activities, or can be utilized at the group level for comparing different user categories. The chart headers in aforementioned figures include the value of the CWD and CDH regularity measures and the boxplots on the right, show the distribution of the corresponding regularity measure for all the apprentices in the selected flow. The dashboard also includes visualization of the users' weekly activity matrices (Equation 4.8), modeling the distribution of activity time among week days. Considering the weekly similarity measure proposed in Section 4.3.2, we included WSB it in the dashboard as shown in Figure 4.18.

Apart from the individual level information, the dashboard further includes group level time histograms. Figure 4.19 and 4.20 show examples of average weekly and daily histograms for the apprentices in two different flows. Figure 4.19 corresponds to a group of 10 apprentices whose apprenticeship is based on a dual-track model where they spend four days at a workplace and only one day at school (Thursday in this case). As shown by Figure 4.19, apprentices in this class mainly use Realto on Wednesday evening (6-7 pm) just before the school day. Further investigation of the activity types reveals that in this flow, the teacher defines frequent activities (assignments) requesting apprentices to upload photos of their workplace activities, describe their productions, and write down their comments on possible improvements. Therefore the main entries produced by apprentices in this flow are submissions to the teacher-defined activities. Figure 4.20 on the other hand, corresponds to a flow of 15 apprentices (same flow as Figure 4.16 to 4.18), following a single track apprenticeship model, where practical training is performed at in-school workshops instead of a workplace. Therefore these apprentices spend all the five weekdays at school and according to their time histograms, their Realto usage is more uniformly distributed among different weekdays (Monday to Friday, mostly between 7am to 3pm). The main use of Realto by apprentices in this case is creating learning documents on the practical training procedures. These two examples clearly show contrasting Realto usage patterns within dual-track and single-track apprenticeship models and exemplify possible information, that could be extracted from the time module in Realto analytics dashboard.

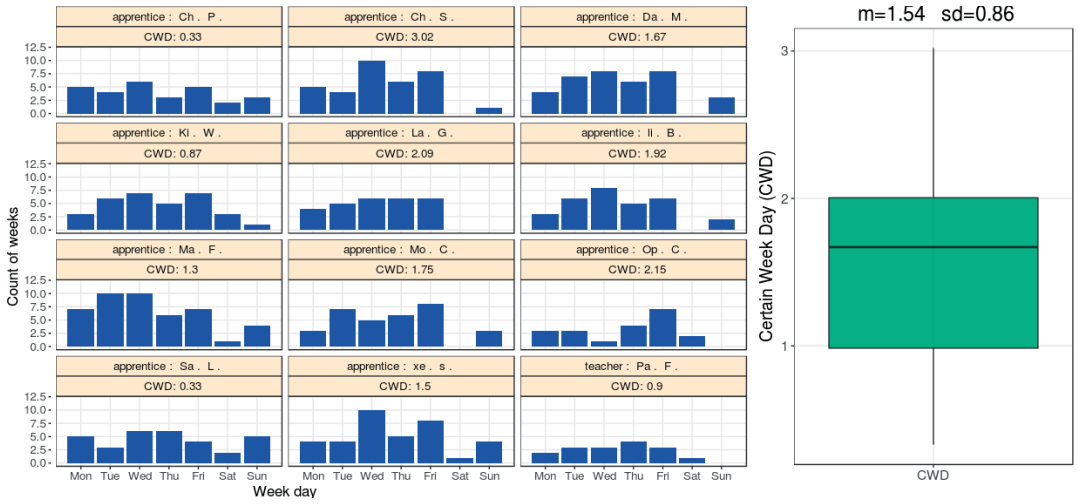


Figure 4.16 – *Examples of weekly histograms and CWD measure in Realto.*

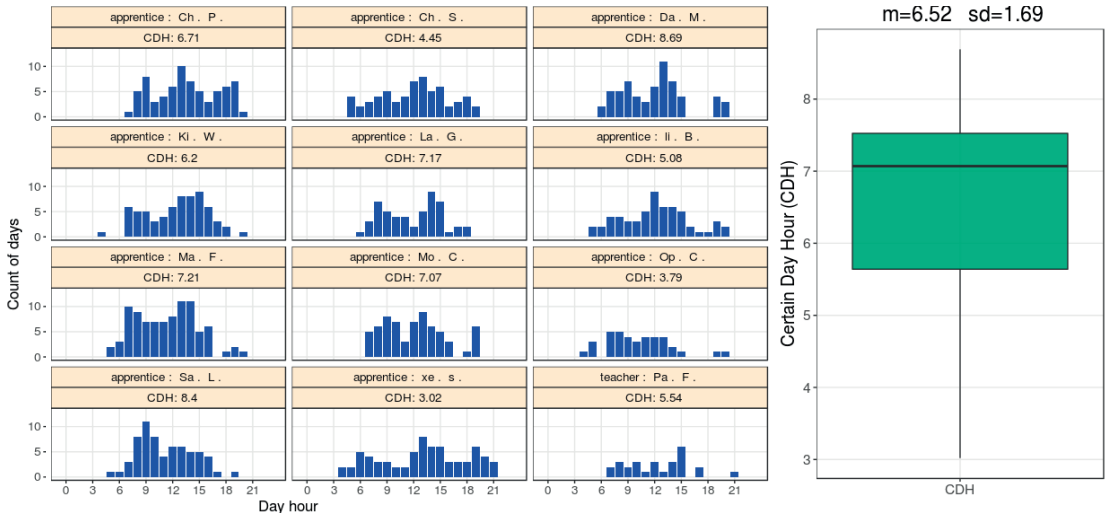


Figure 4.17 – *Examples of daily histograms and CDH measure in Realto.*

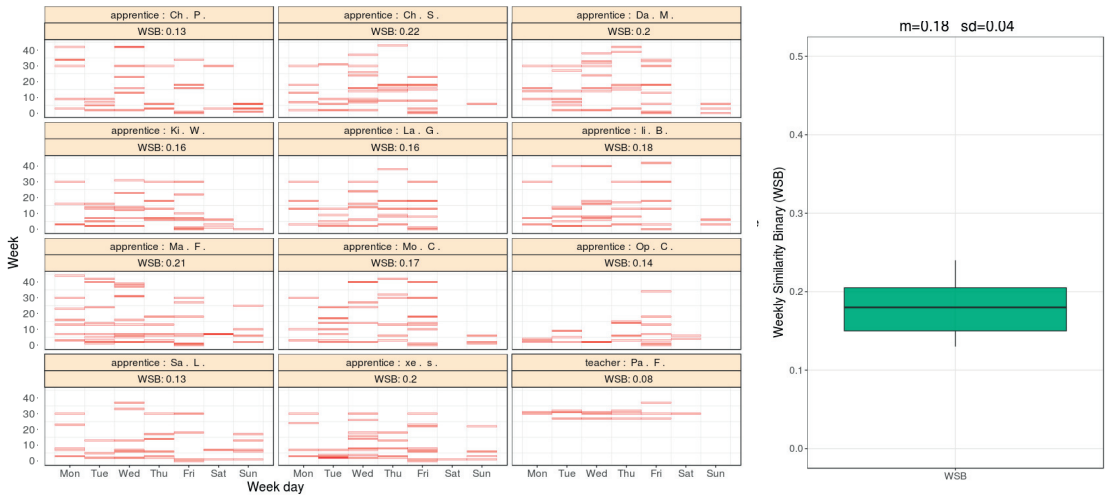


Figure 4.18 – *Examples of weekly profile matrix and WSB measure in Realto.*

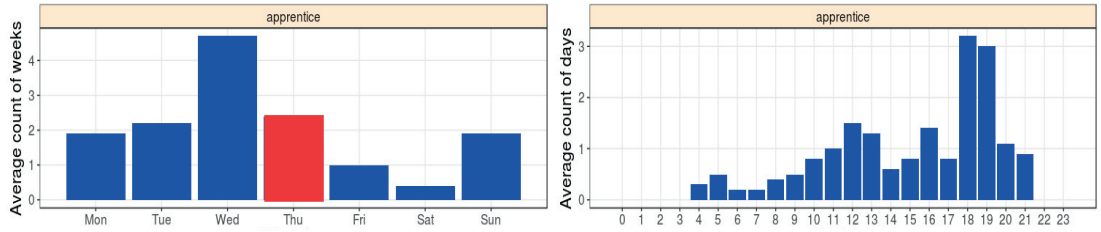


Figure 4.19 – *Example of average weekly and daily histograms for apprentices in a dual-tack apprenticeship. Thursday is the school day in this case.*

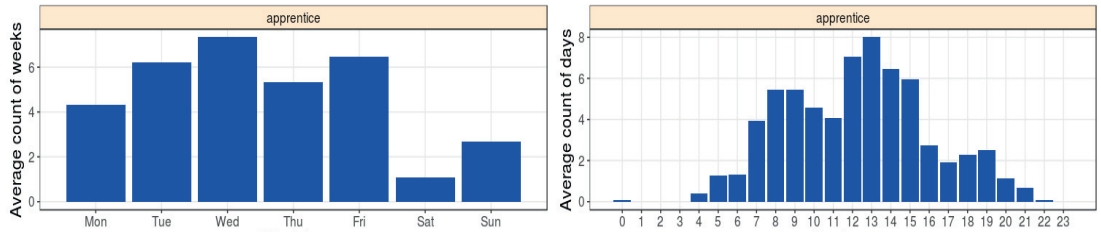


Figure 4.20 – *Example of average weekly and daily histograms for apprentices in a single- tack apprenticeship.*

Since apprentices' activities in Realto is not being graded, we could not investigate the link between regularity and performance in this context.

4.6 Discussion

In this chapter, we provided quantitative methods for analyzing temporal patterns, described their properties and demonstrated their application in MOOC and Realto.

Concerning our first research question on quantifying time regularity (**Question 1**) we proposed nine measures, in time and frequency domains, to capture three overall types of regularity: daily, weekly, and course-based. The value of these measures could reflect if a learner follows a particular time schedule and serve as indicators of planning and time management behaviours. We showed that a subset of the introduced measures are not strongly correlated with each other and capture different regularity patterns.

Concerning our second question on the link between regularity and performance (**Question 2**), we found positive correlations between regularity and final grade in a MOOC course. The regularity measures explained over 50% of the grade variability which reflects their predictive potential making them promising to be included in the existing performance models. We showed that our proposed measures are complementary to features such as total study time and provide additional information on learners' participation. Interaction analysis between weekly regularity and total study time showed that regularity is particularly important for performance when the amount of study time is limited. Furthermore, through clustering of learners based on their regularity values, we identified four categories of learners and

showed that weekly/daily regular participants who did not postpone reviewing the course materials, achieved higher grades. We further showed that the students who passed the course, displayed higher regularity level during the course duration as opposed to the failed students. The positive link between regularity measures and students' performance, supports the hypothesis that students who plan their learning activities in a regular manner have better chances of succeeding in MOOCs [68, 160]. There are however two plausible explanations for this observation. First, regular students follow the structure of the course and therefore attains higher achievement. Second, time regularity is related to certain factors internal to the students such as motivation, conscientiousness, commitment or learning strategies [29, 253]. Further investigations is therefore required to identify the factors influencing time regularity.

The methods and regularity measures presented in this chapter are general and applicable to different contexts and platforms. They can be employed in different domains where time pattern analysis and detecting regularity is of interest. The described methods could be utilized for several applications, besides what presented in this chapter. For instance, regularity measures can serve as indicators for quantifying the extent to which certain features of a course or platform influence regularity and engagement of participants. They can also be used to compare the courses and platforms regarding their habit inducing properties, or to measure the effect of interventions and external factors on users' temporal rhythms. For instance, the regularity measures enabled us to confirm the impact of employment status on learning time patterns in MOOCs [202], revealing that employed learners had higher regularity level both on weekly and daily basis compared to not-employed learners.

In this chapter, we compared students' activity against their own previous behaviour to determine their regularity. However, the proposed metrics and in particular profile similarity measures (Equations 4.9, 4.10 and 4.12) could also be employed to compare the activity profiles of different learners and identify students with similar temporal patterns. The time similarity criteria could in turn be integrated into automatic team formation modules in online learning platforms [208, 232]. Teams composed of learners with similar time patterns, have higher chances for synchronous communication which could be beneficial for particular collaborative tasks.

Personalized notification or reminder systems are another potential application domain for the proposed measures. For learners with high weekly regularity, the typical activity day(s) could be identified based on their recent weekly profiles and the platform notifications could be adopted accordingly. For instance, a reminder could be sent out to the learners in case of no activity on the expected day, or the notifications time (hour of the day) could be adopted based on the learners' daily time patterns. Although our analysis was a posteriori, the described features can be estimated over time, making it possible to track learners' engagement patterns and identify potential problems.

One limitation of the regularity measure we proposed is that, using our measures one cannot distinguish between the different strategies used by those students who adaptively plan

their learning activities based on their daily activities. Furthermore, the set of time regularity patterns we addressed in this chapter, despite covering a broad range of possibilities, is not inclusive. As any projections, our measures can only discriminate patterns that they were designed for and should be combined for accurate assessment of regularity. Our MOOC analyses in this chapter were based on a single structured course in engineering field. Therefore the generalizability of our findings on the relation between regularity and performance needs to be further assessed on a broader range of courses. Considering the time analysis module in Realto dashboard, we showed examples of information which can be obtained from this module. A systematic evaluation is still required to assess its usability and effectiveness for gaining actionable insights into users' time habits.

Finally, in this chapter, we did not discriminate between different action types and treated learners' activity sequence as a binary signal and considered equal weights for all the performed actions. However some learning activities demand more effort compared to others, or learners could be interacting with the course platform without really getting involved in deep learning activities. Therefore complementary methods which take into account the types and details of actions performed by learners, are essential to gain better understanding of the interaction patterns and learning strategies adopted by different learners. We will consider this aspect in the next chapter.

5 Activity Patterns of Online Interactions

In the previous chapter, we studied the temporal patterns of online participation. In this chapter we investigate activity sequences in online learning environments and present methods for detecting patterns in participants' interaction sequences. Similar to the previous chapter, we study activity patterns in MOOC and in Realto. Individual differences among MOOCs participants with different backgrounds, motivations and learning styles, in combination with the flexibility offered by MOOC platforms for navigating through the learning materials, could result in different engagement patterns among learners. Similarly, different participation styles could be observed in Realto, as this platform provides a set of functionalities to participants from different professions and with different roles. Analysis of learners' activity patterns in both contexts, could provide insights on learners' engagement and their preferred learning styles.

In this chapter, we describe methods for extracting and temporal analysis of learners' study patterns in MOOCs in Section 5.1. We provide an overview of the context in section 5.1.1, formulate the research question in Section 5.1.2, and describe the dataset in Section 5.1.3. In Section 5.1.4 we describe a hypothesis-driven approach to capture predefined learning styles in MOOCs and present the obtained results. In Section 5.1.5 we present an unsupervised approach for automatic discovery of learning behaviours from interaction sequences. We present experiments with synthetic data to demonstrate the properties of our proposed method and further employ it for detecting and tracking learners' study patterns in a MOOC course. In Section 5.2 we present our activity analysis approach in Realto and provide examples of the detected usage patterns among Realto participants. Finally, we conclude the chapter in Section 5.3

Parts of this chapter will be published in proceedings of LAK18 [201].

5.1 Activity patterns in MOOCs

5.1.1 Context

Mining sequential interaction logs to identify behavioral patterns of learners has gained great deal of interest in educational data mining and learning analytics communities over the recent years. Leveraging computational methods could reveal interesting hidden patterns in students' activity traces and provide insights about their learning strategies [185]. This in turn could open up possibilities for improving adaptivity and personalization within educational environments [129, 128].

In general, two main approaches could be considered for identifying learners' study patterns: *hypothesis-driven* and *data-driven*. Hypothesis-driven (or pattern-driven) methods aim to detect predefined learning styles from interaction sequences, and require the set of possible learning approaches to be defined by human operator. As an example, a hypothesis-driven approach was employed in [123] for classifying learners' interaction sequences based on submission time to MOOCs assignments. Four engagement types were considered in this work: *on track* (on-time submission to the assignment), *behind* (late submission to the assignment), *auditing* (watching the videos, without submitting to the assignment), and *out* (no participation in the course at all). The hypothesis-driven approach could be utilized for mining theoretically grounded and interpretable learning styles. However, due to the complexity of students' behaviour, it is not often feasible to accurately define a priori, the set of possible learning patterns.

To overcome this limitation, data-driven methods are used for unsupervised discovery of concrete behavioral patterns from learners' interaction data. In this approach, human intervention in the process is being reduced to the assessment of validity and utility of system findings [129, 83]. Clustering methods have received growing attention in this domain, as they allow semi-automatic or open ended behavioral style detection. In particular clustering of sequential data is commonly applied for discovering learners' study patterns. In some studies, learners' activity sequences are being compared in their original format using sequence similarity measures [199, 66, 174, 28], whereas some other works make use of a summarized form or an aggregated representation of the fine-grained activity sequences [128, 83, 200].

Common techniques applied for modeling and analyzing learners' activity sequences include sequential pattern mining [161, 149, 121], Markov chain [97, 128, 75, 129], Hidden Markov Models (HMM) [83, 200, 110], and process mining [221]. In the following we presents examples of previous studies using these methods.

Sequential pattern mining methods [3] seek for the most frequent patterns across a set of action sequences. Nesbit et al. [161] applied this method to study learners' self-regulation behaviours in a multimedia learning environment, and Maldonado et al. [149] used it to identify frequent interaction sequences which differentiate high and low achieving groups in a collaborative tabletop activity. Similarly, Kinnebrew et al. [121] employed sequential pattern

mining techniques in combination with time-series segmentation to identify and compare segments of students' productive and unproductive learning behaviors.

Markov chain representation aggregates sequences of users' actions into memory-less state transition models, encoding the probability of performing one action type after the other. In [75] Markov chains were used to model learners' interaction logs as transition probabilities between different learning activities, and Expectation-Maximisation (EM) algorithm [64] was employed to identify behaviour profiles that characterize groups of similar students. Similarly, students' activities were modeled and clustered based on different similarity measures such as euclidean distance [129] or Jensen-Shannon Divergence (JSD) [128] defined on the transitional probabilities in the Markov chain models.

HMMs have also been broadly applied to model students' learning processes in online learning environments. For instance the use of HMM-based clustering techniques for automatic discovery of students' learning strategies in a tutoring system was investigated in [200]. In [133] HMM were employed to extract stable groups from temporal data by joint optimization of the model parameters and the cluster count. In [83] a two-layer HMM was proposed to discover students' behavioral patterns and transition between them over time. Following this approach authors identified four behavioral patterns (states) for MOOC students: *low activity*, *active*, *forum browsing*, and *passive*. By contrasting state transition of high and low performing students, authors showed that high performing students show longer concentration on quizzes and forum participation.

Process mining [221] is another technique that has been applied on educational data to analyze students' learning processes [156, 219, 18, 27]. This technique originates from business community. It aims to extract process models from activity logs and provide insights into the underlying processes to improve their efficiency. Process mining methods could be adopted to compare students' interaction patterns with predefined models (conformance checking), or to discover the underlying process model from the activity sequences (process discovery). As an example compliance between students MOOC video watching behaviours over the duration of a MOOC course and the predefined sequential video viewing model was assessed in [156], revealing that successful students are more likely to study sequentially than unsuccessful ones. However, while dealing with large scale unstructured data such as interaction logs from thousand of students in online courses, the discovered process models are often "spaghetti-like" showing all details and failing to distinguish the important trends [91]. This makes process discovery methods in their original format inefficient to identify study patterns in MOOC context.

5.1.2 Problem formulation

In this chapter, we investigate learners' study patterns in MOOCs and perform temporal analysis of their longitudinal behaviour. Previous works of mining students' activity sequences in MOOCs, often focus on characterizing relatively short interaction sessions as a composition of

learners' interaction with different course materials [75, 83, 233]. In this work, we aim to identify learners' study patterns during assessment periods, that is their learning sequences from the time when an assignment is made available until the submission deadline. Furthermore, similar to [83], we consider that learners might change their study approach over the course duration depending on the context of the new assignment or for instance, if they find their previous approach inefficient. Through temporal analysis of interaction patterns, we explore the evolution of learning approaches over time. Temporal dynamics of student' behaviour is overlooked in many of the previous studies such as [75, 97, 200] which assume students to exhibit a fixed behavioral patterns in a course. The question we aim to answer in this work could be summarizes as:

Question 1. What are the different study patterns exhibited by learners during MOOCs assessment periods and how do learners' study patterns evolve over time?

We employ two different methods to answer this question. In the first method, following a hypothesis-driven approach, we label students' activity sequences according to predefined patterns and perform clustering to identify prototypical engagement trajectories over the course duration. In the second method we propose a data-driven approach to automatically capture study patterns from learners' interaction sequences. We introduce a complete processing pipeline which starts by modeling learners' activity sequences, applies clustering to identify common study patterns based on the modeled sequences, and performs cluster matching to enable tracking learning approaches over time. We present detailed description of both methods and the obtained results in the following sections.

5.1.3 Dataset

The dataset used for this study consists of the interaction logs of participants in *"Functional Programming Principles in Scala"*[1], an undergraduate engineering MOOC, Produced by EPFL university and offered in *Coursera*. The course is composed of seven sets of video lectures and six graded assignments. Course materials (videos and assignments) are released on a weekly basis and no assignment is anticipated for the sixth week. Submissions to each assignment are accepted before the (hard) deadline and duration of the assessment periods (assignment release day to hard deadline) varies between 11 to 18 days. The final grade is computed as a weighted average of individual assignment grades with a passing threshold of 60 out of 100. The dataset includes three categories of events, describing learners' interaction with video lectures (play, pause, download, seek, change speed), assignments (submit) and discussion forums (read, write a post or comment, vote).

In order to analyze learners' study patterns during assessment periods, we split the full sequence of interaction logs into subsequences corresponding to each assessment period. As the assessment periods of assignments might overlap, we refine the resulting subsequences to contain learners' interactions only with the materials of the corresponding week. In our

[1] https://www.coursera.org/learn/progfun1

analysis we consider learners who were active in at least three assessment periods. Following these data preprocessing steps, the final dataset used in this study contains interaction subsequences of 7527 learners during six assessment periods.

5.1.4 Hypothesis-driven approach

Method

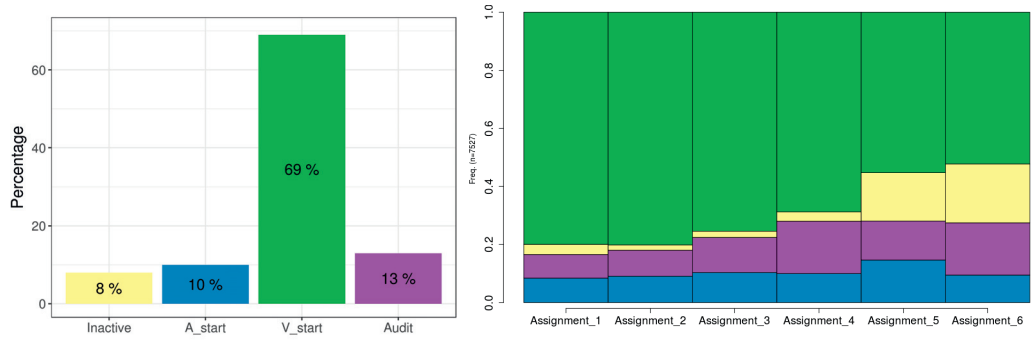
A typical study pattern that could be assumed for MOOC participants, is to watch video lectures, read or write in the discussion forum concerning the difficult concepts and then solve and submit the assignment. However, learners can freely navigate through the course content once it is made available. It is therefore presumable that not everyone would follow the mentioned approach. Some learners might prefer to directly attempt to solve the assignments using trial and error approach. Other learners might decide to skip the videos as they are already familiar with the topic, or since they prefer to attempt the exercises first, and then refer to selected parts of videos if needed. Learners might also skip submitting to the assignment or even watching the videos in some periods.

To identify such study patterns in learners' interaction logs, we examine activity sub-sequences for each assessment period according to two criteria: **(1)** whether the learner starts his(her) learning sequence by watching a video or by submitting to the assignment, **(2)** whether the learner submits to the assignment before the hard deadline. Considering these criteria, we label activity sub-sequences with one of the following study patterns:

- **V_start**: learner watched the video(s) before submitting to the assignment.
- **A_start**: learner submitted to the assignment without having watched the corresponding video(s).
- **Audit**: learner watched the video(s) but did not submit to the assignment.
- **Inactive**: learner did not watch the video(s) and did not submit to the assignment.

Based on the labels that a learner is assigned to for each assessment period, we construct study pattern sequence, describing his(her) engagement over the course duration.

Once we have the study pattern sequences for all learner in the course, we apply hierarchical agglomerative clustering on the sequences to extract categories of learners with similar study profiles and identify prototypical study pattern sequences over the course duration. To determine the optimum number of clusters, we use Calinski-Harabasz (CH) index [43] which is a well-known method for cluster count estimation. To assess pairwise distance between study pattern sequences we use *optimal matching (OM)*, a common distance measure for sequence alignment [1]. In optimal matching, the degree of dissimilarity between two sequences is determined as the least number of required edit operation to turn one sequence into the other (i.e. to match the two sequences). Three kinds of edit operations are generally used: insertion, deletion, and substitution. However, in the case of equal length sequences, similar to study



***(a)** Overall frequency* ***(b)** Distribution over time*

***Figure 5.1** – Hypothesis-driven study patterns, (a) Overal frequency and (b) distribution over different assessment periods in the MOOC dataset.*

pattern sequences in our case, substitutions are the only relevant edit operation.

Results

Study pattern distribution and attributes

Following the described approach, study patterns of 7527 learners during each of the six assessment periods were identified (a total of 45162 study sessions). The overall frequency of study pattern types and their distribution over time is represented in Figure 5.1. In the most common case (69% of all study sessions), learners watch videos before submitting to an assignment (*V_start*). However, in all assessment periods, around 10% of the learners skip video lectures and directly submit to the assignments (*A_start*). Proportion of learners who watch the videos but do not submit the assignments (*Audit*), gradually increases towards the course end (8% vs. 18% in the first and last assignment respectively). Proportion of *Inactive* students also considerably increases in the last two assessment periods (4% vs. 20% respectively in the first and last assignment).

Comparison of the identified study patterns shows that learners who start by watching videos, start their learning sequence earlier than those starting by assignment. This is reflected by significantly longer time between the activity sequence start time and the assignment deadline in *V_start* and *Audit* approaches, compared to *A_start* approach (8 vs. 4.2 days, $F[1, 41575] = 3031, p < .001$). Considering the assignment resubmission behaviours in *V_start* and *A_start* sessions, χ^2 test shows a significant relation between number of submissions and the study approach ($\chi^2 = 254, df = 1, p < 0.001$). According to the test residuals, learners in *A_start* sessions are less likely than the other group to have multiple attempts for solving an assignment (65% of *A_start* sessions, include only one submission to the assignment). However, learners in both approaches perform equally well and get an average score of 8.8 out of 10 ($sd = 2$) in their first attempt. Further investigation of learners' activity sequences in *A_start* session reveals that only in 6% of such cases, learners access the lectures after

submitting to the assignment. These results suggest that learners in *A_start* approach, are likely to have prior knowledge about the assignment topic, since they achieve a high grade in their first (and often only) attempt, without viewing the course lectures.

Fixed study pattern

Analysis of individuals' study pattern sequences shows that 53% of learners continue with their initial study approach during the course duration. These learners can be clustered into three categories, represented as Cluster 1 to 3 in Table 5.1. Learners who follow the *V_start* approach in all the assessment tasks, form the largest cluster, **Cluster 1**, comprising 44% of participants. This group represent typical MOOC learners who rely on lectures to attain the knowledge required for solving the assignments. On the other hand, 2% of participants, represented by **Cluster 2**, do not spend time on watching the videos before submitting to any of the assignments. Their high performance level (average grade of 90 out of 100), reflects their proficiency in the course topics. Earning the completion certificate could therefore be the main participation motivation for these learners. On the contrary, 7% of learners, **Cluster 3**, do not submit to any of the assignments, but they follow most of the video lectures during the course period. This group of learners watch the videos as a source of knowledge without having the intention of receiving a certificate. This group can be referred to as *auditing* students, similar to the categorization scheme in [123].

Changing study pattern

Unlike the described groups with fixed approaches, 47% of the learners change their study approach at least once during the course duration. Transition probabilities among study pattern types for this group of learners are represented in Figure 5.2. Several interesting

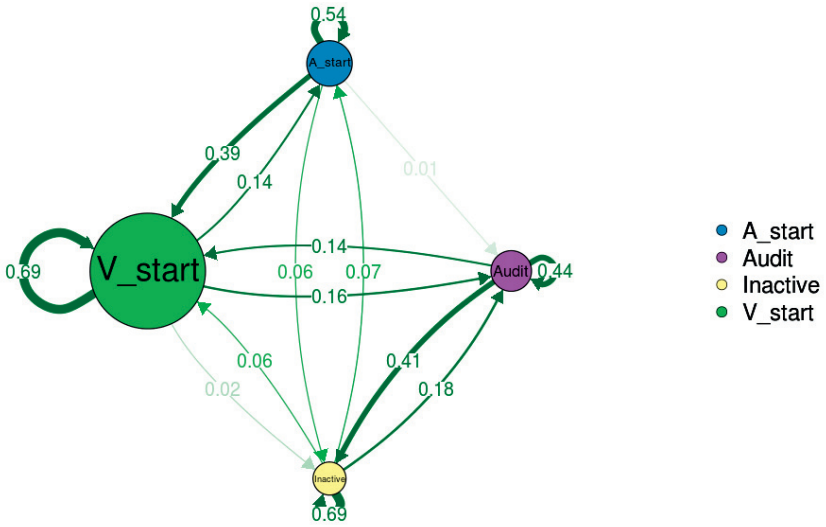


Figure 5.2 *– Transition probabilities between different study patterns for learners who change their approach over time. Node size is proportional to the pattern frequency and edge thickness is proportional to the transition probability.*

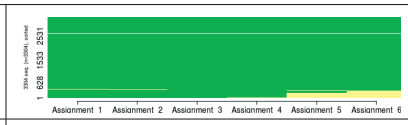
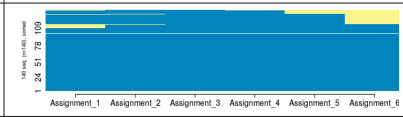
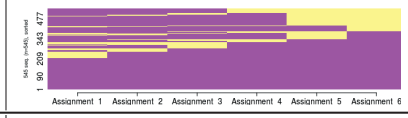
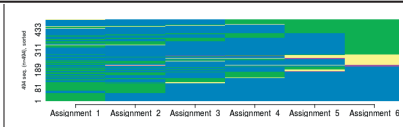
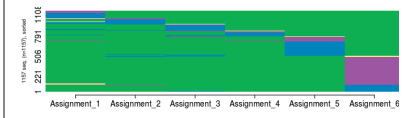
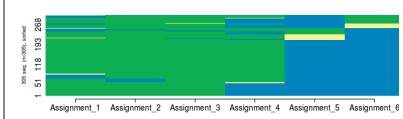
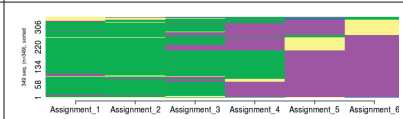
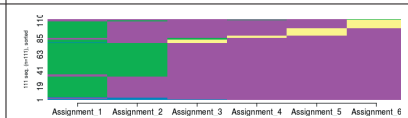
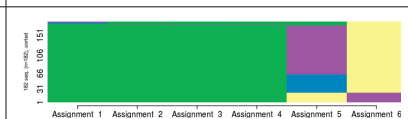
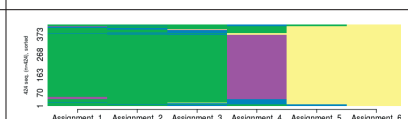
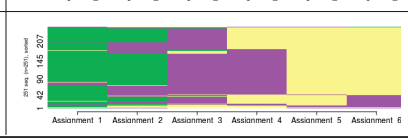
observations can be made from this diagram. In general, high probabilities associated with the self-loops suggest that in each assessment period, learners are more likely to continue with their previous study approach, specially for the *V_start* and *Inactive* states which have a stay probability of 0.69. Learners who start by watching videos, have a low probability of skipping videos in the next period (transition probability from *V_start* to *A_start* = 0.14); whereas learners with *A_start* approach show a relatively high probability (0.39) of watching the videos before submitting the next assignment. Students who *audit* the course in one assessment period, will most likely continue auditing (probability = 0.44) or go *Inactive* (probability= 0.41) in the next period. Once entering the *Inactive* state, participants are not very likely to get engaged in solving the next assignments, but they might continue watching the videos in the next period (transition probability from *Inactive* to *Audit* = 0.18).

To identify the common study pattern trajectories for learners with change of approach during the course duration, we applied clustering on their study pattern sequences (hierarchical clustering in combination with OM distance) and obtained eight clusters, represented as Cluster 4 to 11 in Table 5.1. Description of the resulting clusters and their attributes, including cluster size, average final grade and ratio of passed students, in addition to the visualizations of study pattern sequences in each cluster are provided in the same table.

Learners in **Clusters 4, 5, and 6**, despite having different study pattern profiles, complete the course by submitting to almost all the assignments and more than 94% of them pass the course with high average grades (above 84). Learners in **Cluster 4**, mainly follow *A_start* approach, but in few assignments, mostly the first or last ones, they watch the videos prior to submitting. **Cluster 5** on the contrary comprises learners whose main approach is *V_start*, but they skip the videos in one or two assignments during the course. The start time of the learning sequence for these learners is closer to the assignment deadline in the *A_start* sessions, in comparison with the previous assessment period (4.2 vs. 8.4 days left for the deadline respectively in *A_start* session and the preceding *V_start* session, $t[1,977] = -14.5, p < 0.001$). One possible explanation could be that during such periods the learners procrastinate their activities and consequently, the proximity of deadline makes them to temporally change their study approach and submit to the assignments without watching the videos. Learners in **Cluster 6**, also prefer to watch the videos first, in most of the course duration. But in the last two assessment periods, they submit to the assignments without watching the videos. These learners achieve nearly complete grade in the first four assignments (average grade 9.7 out of 10 $sd = 0.8$). Considering that the final grade is calculated based on the assignment grades, such learners are likely to have a high final grade even without receiving the complete score in the remaining assignments. This might be one factor influencing their decision to skip videos in the last periods and directly submit to the assignments. However, more information about learners' experience and conditions is required to precisely determine the factors triggering changes in learners' study approaches.

The last five clusters (Clusters 7 to 11), show learners who start the course with an active approach as they get engaged both in watching the videos and submitting to the assignments

Table 5.1 *– Clusters of study pattern sequences extracted using hierarchical clustering. Cluster 1 to 3 represent learners with fixed study approach during the course. Cluster 4 to 11 represent categories of role sequences for learners who change their approach over time. Vertical axis in the pattern sequence charts represent students in each cluster and horizontal axis represent assignments. Note that the charts height is* ***not*** *proportional to the cluster size. Other columns in this table represent cluster size, average final grade of cluster members, ratio of passed students in each cluster, and description of the study pattern profiles.*

		Study Pattern sequences	Size (%)	Final grade	Pass %	Description
		A_start Inactive Audit V_start				
Fixed approach	Cluster 1		3304 (44%)	92 (sd:13)	96%	**Submit all, *V_start***: watch videos before submitting to all the assignments
	Cluster 2		140 (2%)	90 (sd:15)	94%	**Submit all, *A_start***: submit to all the assignments without having watched the videos
	Cluster 3		545 (7%)	0 (sd:0)	0%	**Auditing**: watch most videos without submitting to any assignment
Changing approach	Cluster 4		494 (7%)	88 (sd:15)	94%	**Submit all, mainly *A_start***: submit to all the assignments, occasionally watch videos before submission
	Cluster 5		1157 (15%)	84 (sd:14)	95%	**Submit all, mainly *V_start***: submit to all the assignments, skip videos before only one/two assignments
	Cluster 6		305 (4%)	89 (sd:12)	99%	**Submit all, *V_start* then *A_start***: watch videos before submission at the beginning, skip them at the final assignments
	Cluster 7		349 (5%)	44 (sd:15)	20%	**Complete, *V_start* then *Audit***: stop submitting to the assignments after the first half of the course, continue watching videos
	Cluster 8		111 (1%)	19 (sd:10)	0%	**Complete, *V_start* then *Audit***: submit only to the first one/two assignments, continue watching videos without submitting
	Cluster 9		182 (2%)	63 (sd:11)	66%	**Disengage at the end** : start by *V_start* approach, in the last two weeks switch to audit and then drop out
	Cluster 10		424 (6%)	47 (sd:11)	12%	**Disengage in the middle**: start by *V_start* approach, switch to audit in the second half of the course and eventually drop out
	Cluster 11		251 (3%)	23 (sd:10)	0%	**Disengage at the beginning**: start by *V_start* approach, switch to audit after only one/two weeks and then drop out

(*V_start*), but their engagement level decreases over the course duration. Learners in **Cluster 7 and 8** remain engaged until the course end. However, over time they loose motivation for submitting to the assignments and continue watching the course lectures without making any submission. learners in Cluster 7 submit to nearly half of the assignments, whereas those in Cluster 8 submit only to the first one or two, before switching to the auditing state. **Cluster 9, 10, and 11**, demonstrate profiles of disengaging learners or dropouts. The dominant pattern in learners' study profiles in these clusters is to start by *V_start* approach, change to *Audit* state (stop submitting to the assignments) and finally stop watching the videos and drop out. The three clusters differ in the point at which learners' engagement level decreases. Participants in Cluster 9 submit to the first four assignments and 66% of them acquire enough points to pass the course. Whereas those in Cluster 10 and 11 stop doing assignments after one to three weeks, and eventually drop out about a week after.

The identified engagement profiles, could inform the design of intervention mechanisms to support and improve the engagement level of learners who might be facing problems in completing the assignments (e.g. learners in Cluster 7 to 11). An example could be providing supplementary learning materials or connecting them to the well performing learners, in the discussion forum.

5.1.5 Data-driven approach

In the previous section we presented a hypothesis-driven approach for analysis of MOOC study patterns. In this section, we present a data-driven approach for unsupervised discovery of behavioral patterns in different online learning environments.

Method

We introduce an unsupervised processing pipeline to discover and track latent study patterns from students' interaction sequences. The proposed pipeline consists of four steps: **(1)** Activity sequence modeling, **(2)** distance computation, **(3)** clustering, and **(4)** cluster matching. The method receives as input the action sequences extracted from learners' log data, transforms them into probability distributions which model transitions between different action types, computes pairwise dissimilarities between the modeled sequences, estimates the optimal number of clusters and performs clustering to identify groups of learners with similar study patterns in each time period. At each time step t, matching clusters with those at times $t-1$ and earlier are identified. This enables us to track learners' study patterns over time and capture changes in their study approaches, which is an advantage of our proposed method to a recent clustering method proposed in [128]. As the only input to our method is the sequence of learners' activities, it can be used to model and track learners' interaction patterns at different levels of actions granularity or time resolutions. Moreover, our clustering pipeline is able to automatically capture changes in the number and size of clusters, and can be used to detect cluster evolution events such as cluster forming, dissolving, splitting and merging [40].

Activity sequence modeling

Let $A = \{a_1, a_2, ..., a_k\}$ be the set of possible actions in a platform and $S^t = (s_1, s_2, ..., s_n)$, $s_i \in A$ be the sequence of actions performed by a learner during time period t. We model learner's action sequence as a matrix $F_{k\times k}$ where f_{ij} represents the frequency of observing action a_i right before a_j in S^t. We then transform F into a normalized vector P, by normalizing frequencies to represent proportions. Since the entries in P sum up to one, we can consider P as a probability distribution. P provides an aggregated view of the original sequence, encoding probabilities of transitions between different action types. Unlike Markov chain models (as used in [128] and [129]) our representation can directly reflect frequent transitions in learners' action sequences. Hereafter we refer to P as learner's activity model.

Distance computation

To perform clustering on the modeled sequences, a dissimilarity measure needs to be defined to compare learners' activity models. Since the introduced models are in the form of probability vectors, we can use Jensen-Shanon Divergence (JSD) [136], a distance metric designed for comparison of probability distributions. The JSD for two probability distributions is bounded in $[0, 1]$ and the value of zero denotes identical distributions.

Clustering

Based on the pairwise dissimilarity matrix between learners' activity models, any standard clustering method can be used to identify learners with similar activity patterns. We use hierarchical agglomerative clustering for this purpose. In cluster analysis, determining the optimal number of clusters is a major challenge. Several methods have been proposed in the literature to automatically estimate the number of clusters based on the information intrinsic to the data (see a review in [95]). Calinski-Harabasz index [43] and Silhouette Coefficient [191] are among the most well-know methods for this purpose. Such methods in general, measure compactness of clusters (similarity between points in same cluster) and separateness between different clusters (how far points in different clusters are). In our pipeline, we use Calinski-Harabasz index for estimating the number of clusters based on the distance matrix between learners' activity models. Since learners might change their behaviour, some clusters might disappear or new clusters might emerge over time. Therefore we separately compute the number of clusters in every time period.

Cluster matching

After extracting clusters of activity models in each time period, cluster matching is required to identify the correspondence between clusters in the most recent time step and those of previous steps. In social network analysis, cluster matching is often employed for group evolution discovery [40] or tracking dynamic communities over time [88]. In this context, the overlap between cluster members is a criteria considered for computation of clusters similarity. However, in our processing pipeline, this step aims to identify corresponding study patterns in the clustering results of different time periods. Therefore, the similarity of activity pattern models should be taken into account for assessing clusters similarity. We apply a method, similar to Ward method [158], for computing the similarity of activity pattern clusters.

Ward method is used in hierarchical agglomerative clustering for selecting the clusters to be merged in each step[2]. According to Ward method, the most similar clusters are the ones which minimize the increase in the sum of squared errors (euclidean distance) once being merged. Inspired by this approach, we define the distance, d, between two clusters C_i and C_j, as the amount of increment in the sum of errors (JSD distances in our case) when they are combined:

$$d(C_i, C_j) = SE_{C_{ij}} - (SE_{C_i} + SE_{C_j}) \tag{5.1}$$

where C_{ij} is the union of the two clusters, C_i and C_j, and SE_C is the sum of errors for cluster C defined as:

$$SE_C = \sum_{x_i \in C} JSD(x_i, m_C) \tag{5.2}$$

where m_C represents the centroid of cluster C, defined as the mean vector, and JSD refers to the Jensen Shanon Divergence.

Based on the defined cluster distance measure, for each cluster C_i at each time step, we identify the closest one to it, C_j, from the set of clusters obtained in previous time steps. In case of multiple candidates for the closet cluster, we choose the most recent one. If the distance between C_i and C_j is smaller than a threshold (95% quantile of the set of distances between candidate matching clusters), we consider the two clusters to be matching and assign the same labels to them. Otherwise we consider C_i as a new cluster and associate a new label to it.

Results

In this section, using a synthetic dataset, we first demonstrate the application of our proposed clustering pipeline for modeling learners' activity sequences and tracking their behaviour over time. For this purpose we use the synthetic dataset presented in [128], which simulates students' action sequences in an intelligent tutoring system. This synthetic dataset consists of different scenarios including changes in learners' interaction patterns and can be used to validate the ability of our method in capturing behavioral changes and detecting cluster evolution events. Next, we employ the proposed method to capture and analyze learners' study patterns during the assessment periods in the previously described MOOC dataset (Section 5.1.3).

[2] In hierarchical agglomerative clustering, each point is initially assigned to its own cluster, and at each step, the two clusters with the smallest distance are merged into a new one.

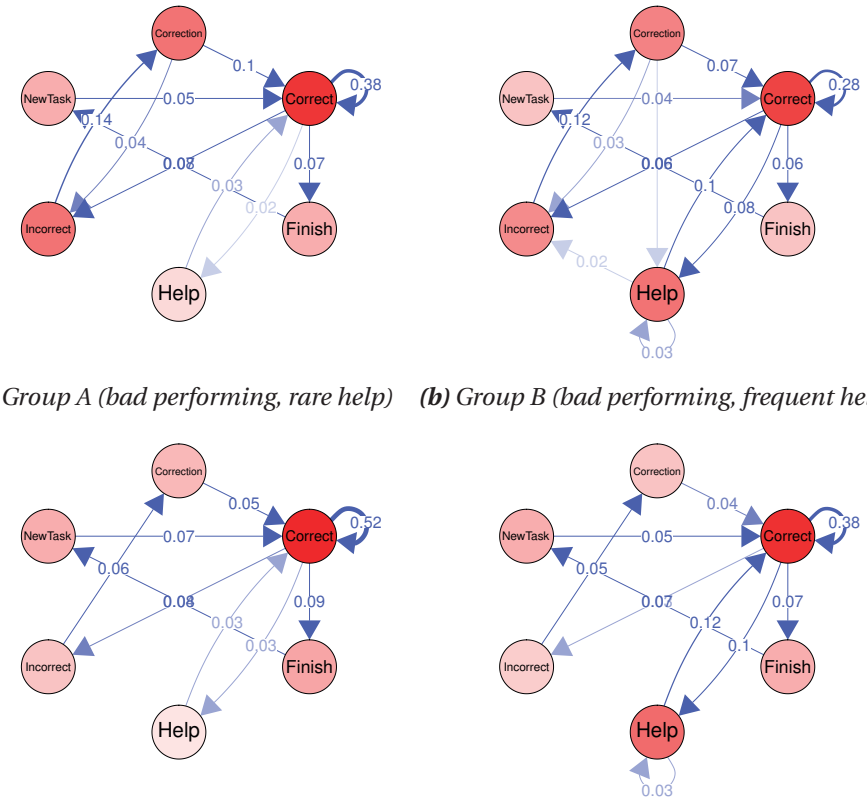
Simulated study

The synthetic data in [128] simulates action sequences of 80 learners over 50 training sessions in a tutoring system. In each session, students needed to complete 20 tasks. Each task was composed of eight steps and students had to correctly solve all the steps in order to finish the task. Learners' abilities θ were sampled from a normal distribution with mean $\mu = 0$ and variance $\sigma = 1$, and task difficulties d were sampled uniformly from the $[-3,3]$ range. The probability of correctly solving a task for each student was given as $p(y) = (1 + e^{-(\theta - d)})^{-1}$. Students could request for help at any point during the training session, with a probability of p_H. Six types of actions were considered for the learners: $S = \{$*New task*, *Correct*, *Incorrect*, *Correction*, *Help*, *Finish*$\}$. In the simulated data, good performing students were modeled by setting $\theta = 1$ and poor performing learners were simulated by setting $\theta = -1$. Moreover, normal help seeking behaviour was modeled by a small probability of help request $p_H = 0.05$, whereas frequent help seeking behaviour (help abuse) was simulated by a large probability of asking for help $p_H = 0.2$. Following this approach, four groups of learners with different behaviours were simulated, including:

- **Group A**: bad performing learners with rare help requests
- **Group B**: bad performing learners with frequent help requests
- **Group C**: good performing learners with rare help requests
- **Group D**: good performing learners with frequent help requests

In the synthetic dataset in [128], four artificial scenarios were considered, simulating different cluster evolution events including cluster merge, split, dissolve, and form. The **first** scenario, simulates merging clusters. In this scenario, after about 20 sessions, bad performing learners with rare help requests (groups A) start abusing the help, and eventually group A completely merges into group B. The simulation in the **second** scenario, starts with three groups, B, C, and D. Over time, some of the bad performing students with frequent help calls (group B), stop abusing the help and consequently group B splits into group A and B. In the **third** scenario which simulates a dissolving cluster, learners in group B switch to the other approaches and eventually group B completely dissolves into the other three groups. Finally, forming cluster event is simulated in the **fourth** scenario. In this case, the simulation starts with three groups, A, C, and D. Over time a fourth group, B, is formed which gradually absorbs students from the other groups until all the four groups have equal sizes.

For all the described scenarios, we used our processing pipeline to model and cluster learners' action sequences in each session, and identified corresponding clusters in different sessions using the cluster matching step. Using this approach, four clusters of interaction patterns were identified in all the scenarios. Figure 5.3 represents the average activity models for the resulting clusters in the first scenario (similar results were obtained for the other scenarios). According to the transition probabilities between different action types, the resulting clusters clearly correspond to the four simulated learner groups. The two clusters in Figures 5.3a and 5.3b, depict bad performing learners as reflected by the relatively high transition probabilities



(a) *Group A (bad performing, rare help)* ***(b)*** *Group B (bad performing, frequent help)*

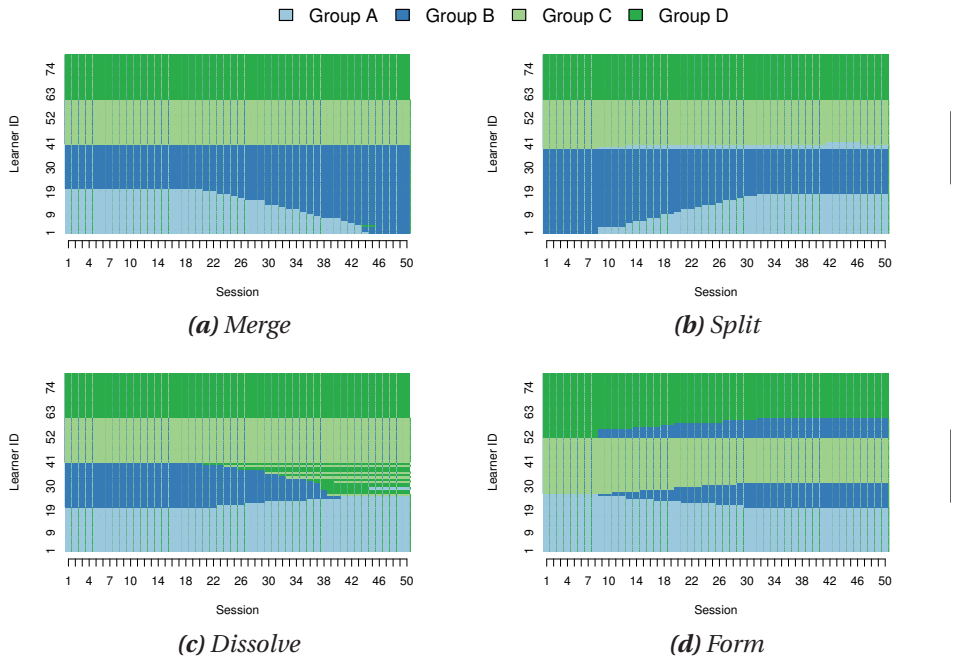
(c) *Group C (good performing, rare help)* ***(d)*** *Group D (good performing, frequent help)*

Figure 5.3 – *The average activity models for the four resulting clusters in the first simulated scenario (cluster merge). In the transition diagrams, node color intensity is proportional to the action probability and edge thickness is proportional to the transition probability. The resulting clusters, correctly capture the four simulated behaviours.*

between *Incorrect* and *Correction* actions. Such learners therefore make more mistakes compared to the good performing learners in Figures 5.3c and 5.3d. Regarding the help seeking patterns, the help abusing behaviour is reflected by the frequent transitions between *Help* and *Correct* actions in Figures 5.3b and 5.3d, whereas such transitions are quite rare for learners with normal help seeking behaviour in Figures 5.3a and 5.3c.

Based on the clusters that students were assigned to in each session, we build their interaction pattern sequences in the four simulated scenarios. According to the resulting sequences, illustrated in Figure 5.4, our method successfully captures the described cluster evolution events in all the scenarios. Furthermore, comparison of our clustering results with the ground truth, confirms the high accuracy of our method in labeling learners' study sessions. Our proposed method, achieves 95% accuracy in first and third scenarios (clusters merging and forming) and 93% accuracy in second and forth scenarios (clusters split and dissolve).

Overall, the presented experiments with simulated data, demonstrate that our processing



(a) Merge *(b) Split*

(c) Dissolve *(d) Form*

***Figure 5.4** – Sequences of the learners' interaction patterns over 50 sessions in the four simulated scenarios. In the sequence charts, each horizontal line represent the interaction pattern sequence of one learner. Our proposed pipeline correctly captures the changes in cluster count and size and detects clusters merge, split, dissolve, and form.*

pipeline is able to detect different interaction patterns among learners, and provides models which are easy to interpret. The validity of the clustering and cluster matching steps is also confirmed by these results, showing that our method correctly captures changes in the number and size of clusters and is able to detect changes in learners' behaviours over time.

MOOC study patterns

In order to employ the described data-driven approach to model study patterns in MOOC, we choose daily granularity of actions. We label each day according to the type of activities (regardless of their order) performed by the learner (***V***ideo access, ***F***orum access and ***A***ssignment submission), with one of the following states: {*A*, *F*, *V*, *AF*, *AV*, *FV*, *AFV*, *Inactive*}. We then describe individuals' daily state sequences as the list of daily states between their first and last activity day during each assessment period. As an example, if a learner starts the learning sequence by watching the videos on two successive days, does not perform any action on the next three days, accesses the forum (read/write) and submits to the assignment the day after, his(her) daily state sequence would be as: {*V*, *V*, *Inactive*, *Inactive*, *Inactive*, *AF*}.

Following this data preparation procedure, we construct the daily state sequences for all the six assessment periods for a randomly selected sample of 2000 learners in our dataset[3]. The

[3]Sampling was done due to high memory requirement for pairwise distance computation on the full dataset.

set of state sequences is then provided as input to the described clustering pipeline. For each assessment period, learners' activity models, which in this case represent transition probabilities between different daily states, are constructed and clustered according to the estimated number of clusters. Cluster labels are then refined based on the cluster matching results. The center of each cluster (average vector) is considered as the representative study pattern for the learners in each cluster.

Following this approach, we identified 13 different study patterns (clusters) from learners' interaction logs. Table 5.3 provides a summary of the results, including visualization of the study patterns and the most frequent daily state sequences in each cluster. Description of the study patterns and their attributes (size and average cluster errors) are also provided in Table 5.3. The low average cluster errors (0 to 0.1) reflect the accuracy of the clustering results.

The resulting clusters, capture meaningful patterns in learners study sequences. According to the state transition diagrams and their descriptions in Table 5.3, the extracted study patterns differ in the duration of study sequences and also the performed daily activity types. In most cases, learners work on materials of a week during one or multiple consecutive days. For instance, learners with *Patterns 8*, *10*, and *12* have a single activity day. In *Pattern 10*, learners directly submit to the assignments without accessing any other course materials, whereas in *Pattern 8*, they watch the videos and submit to the assignment, and in *Pattern 12*, they also access the discussion forum. Learners in *Patterns 4*, *5*, *6*, and *9* study during two or more successive days, whereas, in *Patterns 3*, *7*, and *13*, learners have multiple inactive days during their learning sequence.

Table 5.2 provides an overview of the estimated number of clusters and the list of detected study patterns at each assessment period. As reflected by the cluster counts, a higher variability is observed in learners' study approaches at the beginning of the course. Most of the patterns detected in the first assignment, remain present in learners' interaction sequences over the course duration (*Patterns 1*, *2*, *4*, *7*, *8*, *10*, and *11*), whereas some other patterns such as *Patterns 5* and *9* disappear over time. During the second and fourth assessment periods, two new study patterns are formed (*Patterns 12* and *13*), both of which dissolve into other patterns after only

Table 5.2 *– Estimated number of clusters and list of identified study patterns in each assessment period. New patterns in each period are highlighted in bold blue font. Study patterns are described in Table 5.3*

Assignment	Estimated cluster count	Clusters list (study patterns)
1	11	Pattern **1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, **9**, **10**, **11**
2	12	Pattern 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, **12**
3	10	Pattern 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
4	10	Pattern 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, **13**
5	8	Pattern 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
6	8	Pattern 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

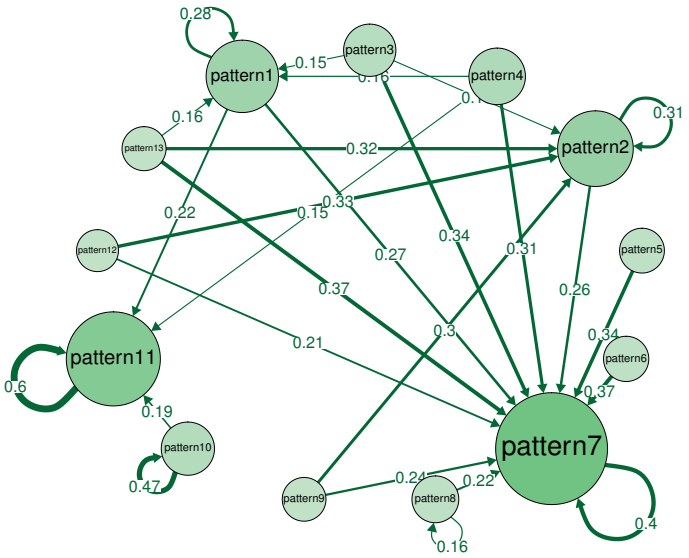
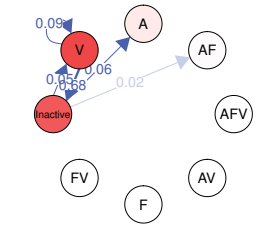
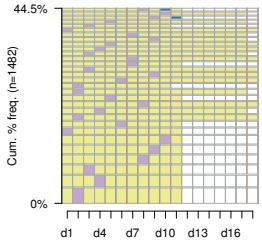
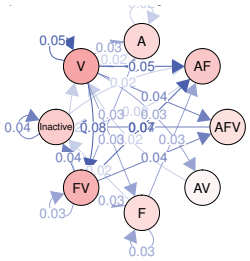
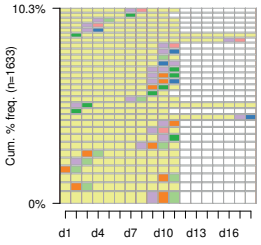
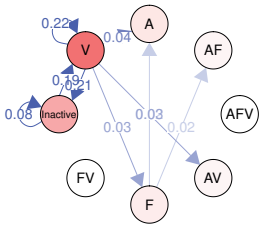
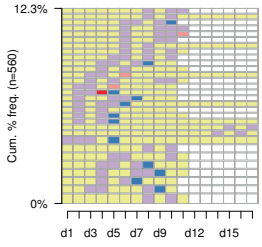
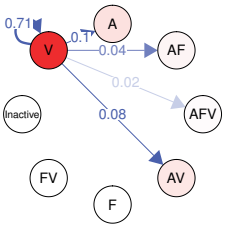
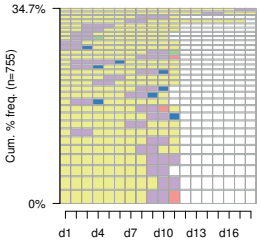
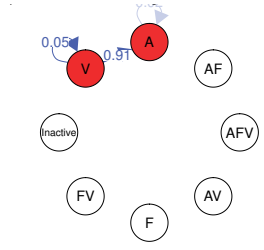
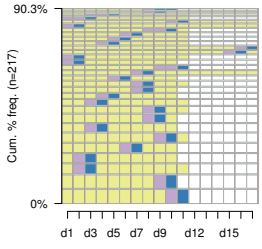
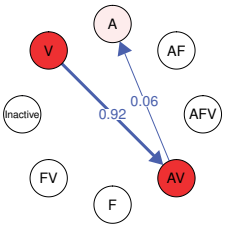
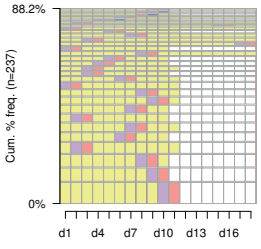


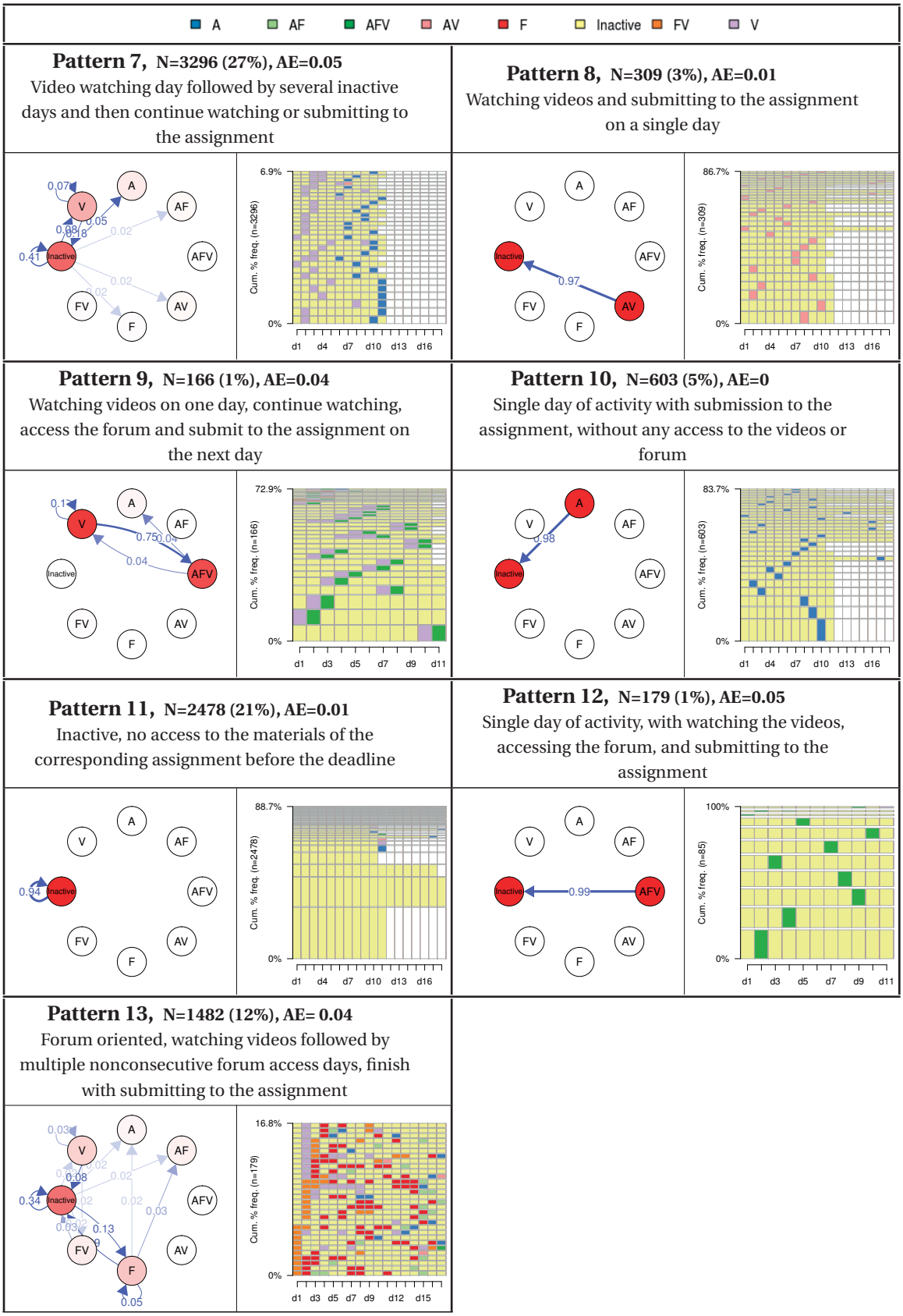
Figure 5.5 – *Transition probabilities between data-driven study patterns in MOOC dataset. Node size is proportional to the pattern frequency and edge thickness is proportional to the transition probability (Transitions with probability smaller than 0.1 are not displayed).*

one or two periods.

Figure 5.5 depicts the transition probabilities between different study patterns, extracted from learners' study profiles over the course duration. According to the self-loop probabilities, *Patterns 10* and *7* are the most stable study patterns, in the sense that learners following these approaches, are likely to continue with the same approach in the next assessment period. *Pattern 11*, which represents inactive learners during a period, is also associated with high stay probability, suggesting that inactive learners in one period would remain inactive in the next period, with a probability of 0.6. *Pattern 10* represents learners with only one activity day on which they submit to the assignment and do not access any videos or the discussion forum. This pattern could represent a similar approach as *A_start*, described in Section 5.1.4. *Pattern 7*, which is the most frequent study pattern, represents students who watch the lectures, and after few inactive days they either continue watching the videos or submit to the assignment. This pattern receives relatively strong connections form the other nodes (except *Pattern 10*), suggesting that learners with other approaches, might adopt this study pattern for the next assessment period, with probabilities between 0.2 to 0.4.

Table 5.3 – *Data-driven study patterns extracted from MOOC learners interaction logs. For each pattern, transition diagrams (left) show the average activity model (node color intensity is proportional to the state probability and edge thickness is proportional to transition probability). The grid charts (right) show the 20 most frequent daily state sequences for each study pattern. Horizontal axis in sequence charts represent days in the assessment period and rows represent sample sequences (row height is proportional to the sequence frequency). In the patterns description,* ***N*** *represents the frequency of each pattern and* ***AE*** *is the average error (average distance between activity models and cluster mean vector).*

Study pattern	Sample state sequences	Study pattern	Sample state sequences
A AF AFV AV F Inactive FV V			
Pattern 1, N=1482 (12%), AE=0.04 Watching videos on a single day, with low probability (0.06) of submitting to the assignment after one inactive day		**Pattern 2,** N=1633 (14%), AE=0.13 No dominant pattern, similar probability of transition between different daily states	
			
Pattern 3, N=560 (5%), AE=0.06 Multiple nonconsecutive days of video watching, submitting to the assignment after the last video watching day		**Pattern 4,** N=755 (6%), AE=0.04 Successive days of video watching, followed by a submission day on which learners are also likely to access forum or videos	
			
Pattern 5, N=217 (2%), AE=0.01 Single day of video watching, followed by a submission day (without any access to forum or videos)		**Pattern 6,** N=237 (2%), AE=0.01 Video watching day, followed by watching videos and submitting to the assignment on the next day	
			



Pattern 7, N=3296 (27%), AE=0.05
Video watching day followed by several inactive days and then continue watching or submitting to the assignment

Pattern 8, N=309 (3%), AE=0.01
Watching videos and submitting to the assignment on a single day

Pattern 9, N=166 (1%), AE=0.04
Watching videos on one day, continue watching, access the forum and submit to the assignment on the next day

Pattern 10, N=603 (5%), AE=0
Single day of activity with submission to the assignment, without any access to the videos or forum

Pattern 11, N=2478 (21%), AE=0.01
Inactive, no access to the materials of the corresponding assignment before the deadline

Pattern 12, N=179 (1%), AE=0.05
Single day of activity, with watching the videos, accessing the forum, and submitting to the assignment

Pattern 13, N=1482 (12%), AE= 0.04
Forum oriented, watching videos followed by multiple nonconsecutive forum access days, finish with submitting to the assignment

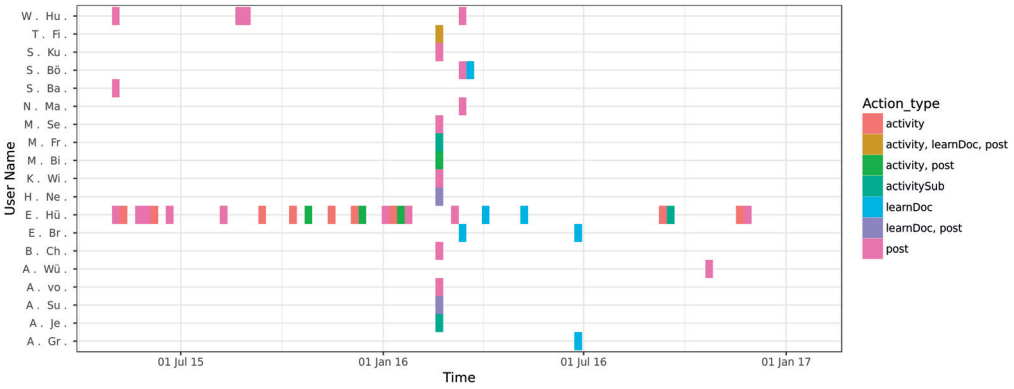
5.2 Activity patterns in Realto

As described in Chapter 3, Realto is designed as a generic platform accessible by different professions in vocational education and training. It provides a range of functionalities to users with different roles, i.e. apprentices, teachers and supervisors. Analysis of platform usage patterns by Realto users, could provide insights on its employment in vocational schools. Understanding how the platform is being used by different user categories, which features are of higher importance for particular sub-populations, and which aspects of the platform are under-explored by the users, could inform the design of new features and enhancement of the existing functionalities in order to better address users' requirements in Realto.

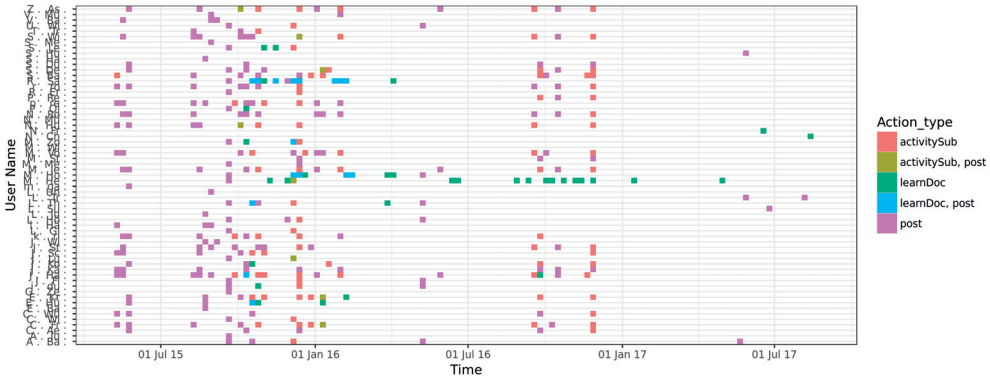
To enable monitoring and analysis of users' activity patterns, we developed *activity analysis* module in Realto analytics dashboard (Appendix A.2). The dashboard allows for selecting specific user groups based on their profession, role, school, and language, and provides two types of information for the selected users: **(1)** visualization of their activities over time, and **(2)** clusters of participants with similar usage patterns. Five type of activities are considered in the activity analysis module: creating a post, comment, or learning documentation entry, creating a classroom activity (specific to teachers and supervisors), and submission to teacher-defined activities.

An example of the activity sequence charts is provided in Figure 5.6, representing the weekly activity types by teachers and apprentices in florist profession, during a period of 28 months (May 2015 to August 2017). The time period and time window size (weekly or monthly) for the activity sequences can be selected from the interface. According to Figure 5.6a, one of the three initially subscribed florist teachers in Ralto, remains engaged in the platform till the end of 2016. This teacher creates several classroom activities and posts, at different points of time, and also explores the learning documentation functionality. In the beginning of 2016, 16 other florist teachers are registered in Realto, however they do not continue using this platform in their classes. According to Figure 5.6b, florist apprentices mainly used Realto for creating standard posts before November 2015, and after this period, they also submit to teacher-defined activities. Few apprentices in this profession have created learning document entries in Realto. The decreasing engagement level of florist participants in 2016 and afterwards is evident from Figure 5.6. Such information could help the research team to plan follow-up meetings and training workshops for user groups who might need further support.

Although the data-driven method described in Section 5.1.5, could directly be applied to determine usage patterns in Realto, due to the sparsity of platform usage data at the current stage, we decided to follow a different approach. To identify groups of participants with similar engagement pattern in Realto, we cluster users based on their total number of posts, comments, learning documents, activity creation and activity submissions. The number of clusters can be selected in the dashboard interface and Clara algorithm is used as the clustering method. Clara is an extension to K-medoids clustering method, adapted for fast clustering of large datasets. This method perform K-medoids on a randomly selected subsample of the



***(a)** Teachers*

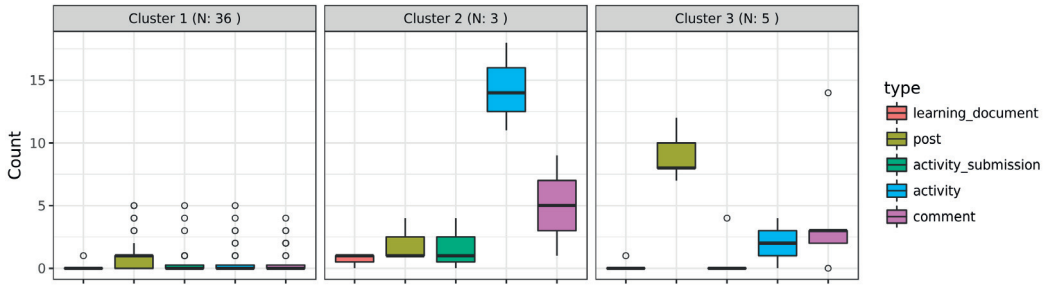


***(b)** Apprentices*

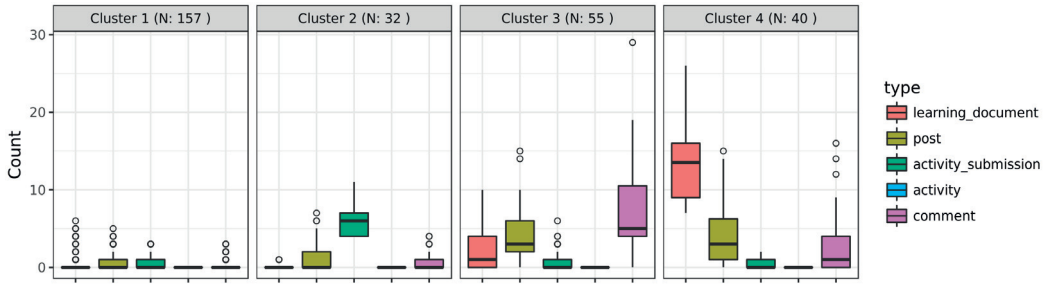
***Figure 5.6 –** Activity sequence of florist apprentices (top) and teachers (bottom) in Realto.*

data, finds the optimal set of medoids, and assigns the objects in the entire dataset to the closest medoid. To alleviate sampling bias, this process is repeated for a predefined number of times and the clustering result with the minimal cost (average dissimilarity between objects and cluster medoids) is returned as the output.

Figure 5.7 provides an example of Realto usage clusters by teachers and apprentices in clothing design profession. As depicted in Figure 5.7a, apart from inactive teachers in Cluster 1, three teachers in Cluster 2, frequently create classroom activities, and also write comments on the posts in Realto. Creating standard posts is the main activity of the other five teachers in Cluster 3. Considering the apprentices in this profession, according to Figure 5.7b, the first cluster ($N = 157$) represents inactive users and the second cluster ($N = 32$) represents apprentices with relatively low level of activity, who mainly respond to teacher-defined activities. Apprentices in the third cluster ($N = 55$) mostly use the social features of the platform, as reflected by their larger number of comments and posts. On the contrary, creating learning documentations is the primary use of Realto for apprentices in the forth cluster ($N = 40$). This example shows how activity analysis module in Realto analytics dashboard could enable investigation of engagement patterns of different user groups in Realto.



(a) Teachers



(b) Apprentices

Figure 5.7 – *Clusters of Realto usage patterns by apprentices (top) and teachers (bottom) in clothing design profession.*

5.3 Discussion

In this chapter, we investigated engagement patterns of participants in online learning environments. We described methods for analyzing interaction sequences in order to extract behavioral patterns and inspect learners' engagement over time. We applied the described methods to analyze learners' participation patterns in MOOCs. To investigate users' engagement patterns in Realto, we introduced the activity analysis module in Realto analytics dashboard and provided examples of different usage profiles among participants in this platform.

To answer our research question on learners' study patterns in MOOC context (**Question 1**), we presented two methods for detecting and temporal analysis of study patterns. In the first method, we employed a hypothesis-driven approach to label learners' activity sequences based on interactions with lectures and assignments. According to the results, about 44% of the learners in our MOOC dataset, watch lecture videos prior to submitting to each assignment. On the other hand about 2% of learners, skip the videos in all assessment periods. Moreover, through unsupervised categorization of study pattern sequences, we identified different longitudinal engagement profiles among learners. We showed that some learners temporally change their study approach during few periods, whereas some other learners permanently switch to a new approach. Detecting changes in study approaches could be used for providing personalized support to learners who face difficulties during their learning process.

In the second method, we proposed a processing pipeline for unsupervised discovery and temporal analysis of interaction patterns from sequential activity logs. The proposed method is general and requires only the collections of action sequences as input. It can therefore be employed for modeling and analyzing interaction patterns in various online learning environments, including MOOCs and intelligent tutoring system. Moreover, the presented pipeline allows for analysis of interaction patterns at different levels of granularity and time resolutions. Through experiments with simulated data, we showed that our pipeline enables to detect learners' behavioral patterns, provides interpretable models describing them, and is able to capture temporal dynamics of learning behaviours. We further applied our pipeline to a MOOC dataset to explore learners' study patterns in this context. Using this approach, 13 different study patterns were identified. We presented a description of the obtained patterns and investigated transitions among them.

Generalizability of the detected study patterns in this work, needs to be further explored in different MOOC courses. Additionally, it would be interesting to explore the relation between the adopted strategies and learning performance and also the influence of participants' background, educational context, and demographics on their study approaches. Unlike most existing works with post-hoc analysis of learners' activity traces, our proposed pipeline can be employed for analysis of learners' engagement during the course period. A possible extension of the presented work could be to integrate an overview of the captured behaviours and study pattern sequences in analytics dashboards. This information provided to the teaching team, could help them plan for improving the course design, for instance by identifying the factors which trigger course-wide drifts in learners' engagement patterns. Moreover, the extracted learning behaviours could be employed for improving personalization of online learning platforms and intelligent tutoring systems.

6 Online Social Interactions

In the previous two chapters we studied learners' online participation patterns in time and activity domains. Focusing on individuals' behaviour, we incorporated computational methods to model learners' study habits and identified groups with similar engagement styles. In this chapter we explore the social dimension of online learning through the investigation of participants' online communication. Following a similar structure as in the previous two chapters, we study the social dimension of online interactions in MOOC and in Realto platform.

Social learning is considered as an important element of scalable education in MOOCs [37]. In this context, with the absence of face-to-face communication and lack of individual support by tutors, discussion forums often serve as the only channel for peer-to-peer information exchange. Considering the large number of participants and diversity of their background knowledge and learning styles, social communication can take very complex forms. To better support information exchange between peers, it is essential to understand the current situation and learners' interactions within the discussion forums.

In the case of Realto, as mentioned before (Chapter 3), this platform has been implemented as a social platform connecting different VET stakeholders including apprentices, teachers, and supervisors. Realto provides a digital space for sharing information and experiences captured from different learning locations (e.g. school or workplaces). Social interactions among platform users is enabled through commenting, rating, and providing feedback on the shared materials. Analysis of the emerging communication patterns among participants could therefore provide insights on the use of social features in Realto, shedding light into the connections between different stakeholders.

In this chapter, we present an exploratory study on dynamics of MOOC discussion forums in Section 6.1. By incorporating different analytic methods we study three main aspects of MOOC forum communication including time, discussion content, and learners' social

Parts of this chapter were done in collaboration with Tobias Hecking and have been previously published in [204].

interactions. In Section 6.2, we employ social network analysis measures and methods to uncover the attributes and structure of communication among Realto participants. We provide a discussion on the findings and conclude the chapter in Section 6.3.

6.1 Social interactions in MOOC

6.1.1 Context

Analysis of MOOC discussion forums have received much attention in recent years. A substantial body of research has been developed aiming to understand the use of MOOCs discussion forums as a prerequisite for the development of improved collaboration mechanisms tailored to the specific conditions of collaboration on massive scale [189, 251]. Previous studies had investigated MOOC discussion forums from different perspectives. This includes analysis of learners' engagement and activities [11, 124, 164], discussion themes and topics or linguistic properties of written messages [140, 190, 236], structure of the communication network, group formation and social interactions among forum participants [84, 85, 163].

The question about *how engaged different MOOC users are in discussion forums* has been addressed in various studies. Several studies point out the limitations of MOOC discussion forums such as low overall participation [124, 164], and sometimes a lack of responsiveness [242]. The fact that the discussion forums are mainly used only by a small fraction of the course participants [124] is meanwhile commonly known. Furthermore, the fraction of learners who use the forums intensively is even smaller [164]. On the other hand, there is evidence for a relation between engagement in discussion forums and different levels of engagement with respect to other course activities [11] and that forum activity goes along with completion rates [11, 84]. Forum participation features have also been employed for modeling learners' engagement in the course and predicting dropouts [180, 241]. Discussion volume often represents a continuous decline over the duration of the course. The high decline rate of forum activity besides the behavioral factors, which contribute to maintenance of a robust participation rate are investigated in [37]. In general, several studies point out the discrepancy between the goal of establishing a learning community and the actual implementation of collaboration mechanisms in MOOCs [198]. Considering the asynchronous communication and heterogeneous population of participants in this context, it has been argued that adequate support mechanism are required to engage participants in active collaborative knowledge exchange [189, 251]. Personalization, support in finding peers for information exchange, and formation of learning groups are among the examples of such support methods [189].

Apart from questions about the activity of course participants in the discussion forum the actual content of the discussions is of interest as well. This typically requires natural language processing to analyze the textual contributions of forum users. The types and themes of discussion forums can be diverse and not necessarily related to the actual course subject [164]. Messages like personal introductions, search for learning groups, or requests of technical

and organizational support, are among the examples of non-content related discussions. Identifying discussions relevant to the course content is crucial for the analysis of collaborative knowledge building and information exchange in discussion forums. Towards this goal, Wise and Cui [236] proposed content-based indicators in combination with machine learning methods to determine content-related discussion threads. Similarly, Rossi et al. [190] built supervised models to classify discussion threads into different categories such as social talk, open ended topics, close ended problems, and course logistics. Another strand in content-based analysis is concerned with the nature of forum posts. Classification of speech-acts in MOOC discussion forums such as questions, answers or issue resolution [14, 140], provides insights into the composition of discussion forum from the perspective of contribution types. Apart from speech acts, contributions can also be classified according to constructs of conceptual and operational learning levels according to the Anderson and Krathwohl's taxonomy [240].

Considering the aspect of social forum interactions, network analysis methods have been widely applied to analyze communication structure emerging from participants interaction in discussion forums. Structural patterns and the underlying relational organization of a course community has been studied in several related research. Gillani et al. [84, 85] analyzed networks of forum users connected by co-contribution to the same discussion threads. They argue that the coherence of the social structure mainly depends on a small set of central users and the forum users can be considered as a loosely connected crowd rather than a strongly connected learning community. These difference between regular forum users and occasional posters was explicitly taken into account by Poquet and Dawson [163] showing that regular users shape a denser and more centralized communication network since they have more opportunities to establish connections. In the context of structural analysis of forum communication networks, different studies have used exponential random graph models (ERGMs) [183] or related statistical network analysis models to identify factors that influence the emergence of the observed network characteristics [112, 116, 163, 249]. In general these results reveal an effect of reciprocated ties and a lack of centralization of the networks to few influential users. On the level of individuals, social network analysis is further applied to identify different roles of users based on their social connections and thematic affiliations [102, 99]. This will be explicitly taken up in this chapter later on in Section 6.1.6.

Despite the amount of work described above, there are still many open questions (see Section 6.1.2) and comprehensive studies that integrate different aspects of peer exchange in MOOC forums do not exist to a large extent. Furthermore, although the important aspect of time in online discussions is being recognized in recent studies such as [103] and [104], most of the existing research overlook the temporal dynamics of forum communication by considering aggregated variables over time or a static forum snapshot to describe users interactions. Our work constitutes a step towards filling this gap by providing results of adapting mixed methods to investigate different aspects of users' participation in MOOC discussion forums.

6.1.2 Problem formulation

In this chapter, we aim to extend the body of research on MOOC discussion forums by providing an integrated study on the interplay of temporal patterns, discussion content, and the social structure emerging from learners' forum communication. As mentioned, most of the existing research focus on only one of these dimensions. However, to attain a more complete picture of the discussion forum communication, a combined analysis of all the aforementioned aspects is necessary. A special focus of our analyses is on the yet under-explored aspect of temporal dynamics and influence of the course structure on forum participation. In particular, we cover three main dimensions of discussion forums: time, content, and social. In the **time dimension**, we reconstruct the daily timeline of the course by considering the main course related events, namely video release time and assignment soft deadlines[1], and track the evolution of forum activity with respect to the course timeline. In the **content dimension**, we apply text analysis methods to investigate discussion themes and topics. In the **social dimension**, we study the underlying social structure emerging from the communication (global level) and learners' roles in the communication network (individual level). Contrasting the results of these analyses could shed light on the interrelation between forum activity, discussion content, and social communication structure. Concerning the aforementioned aspects we aim to answer the following research questions:

Question 1. How does the overall activity in discussion forums evolve over time and is it related to the course structure? [*time dimension*]

Question 2. How do the discussion topics evolve over time and is it related to the course structure? [*content* + *time dimension*]

Question 3. Does the course structure influence the structure of information exchange network? [*social* + *time dimension*]

Question 4. How do the learners' roles in discussion forum evolve over time? [*social* + *time dimension*]

Question 5. How are the learners' roles in the communication network related to discussion content? [*content* + *social dimension*]

Question 6. Is the overall forum activity predictable?

In general, our hypothesis is that the course structure has influence on the forum communication and therefore structural features of the course could be used for predicting forum activity level, in order to enable the teaching team to prepare for supporting learners during intense discussion periods. In particular, we expect increased discussion volume at the proximity of course deadlines, along with the increment of content related discussions and learners' connections in the information exchange network. Considering the discussion content, investigation of discussion topics over time, allows for determining the challenging topics and thematic areas during different periods of the course. Furthermore, analysis of learners' roles in the discussion forum serves as a prerequisite for development of improved collaboration mechanisms adaptive to needs of the different user groups.

[1] Submissions made up to three days after the soft deadline are still graded but penalized for a late submission.

In the following, after presenting the dataset in Section 6.1.3, we investigate overall forum activity across course timeline (*Question 1*) in Section 6.1.4. In Section 6.1.5 we explore discussion topics over time (*Question 2*). In section 6.1.6 we present analysis of social communication structure at global (*Question 3*) and individual level (*Question 4*). We further provide analysis of the relation between social and content aspects of discussions (*Question 5*) in Section 6.1.7. In Section 6.1.8, we integrate extracted features from several of previous sections into a machine learning model to predict the forum activity level (*Question 6*).

6.1.3 Dataset

The dataset used in this study consists of two engineering MOOCs, produced by EPFL and offered on Coursera entitled: "*Functional Programming Principles in Scala*"[2] and "*Principles of Reactive Programming*", hereafter referred to as $Scala$ and $Reactive$ respectively. Both courses are taught by the same instructor and comprise seven sets of video lectures released in a weekly basis and six graded assignments corresponding to the different course topics. The final grade is computed as a weighted average of individual assignment grades with a passing threshold of 60 out of 100. Discussion forums in both courses were structured into several sub-forums such as general discussions, search for learning group, questions and clarifications about course lectures and assignments. Learners were instructed to choose the relevant sub-forum while posting new questions. Discussion forums followed a hierarchical structure, similar to most existing forums, where learners could either create a thread to initiate discussion around a new topic, write a post in an existing thread or add a comment to an existing post. We restrict our analysis to the lectures and assignments sub-forums as our focus is in tracking the evolution of discussions related to the course content. According to the statistics reported in Table 6.1, this resulted in 7,699 messages created by 1,175 participants in $Scala$ course and 12,283 messages by 1,902 participants $Reactive$ course.

Table 6.1 – *Dataset overview*

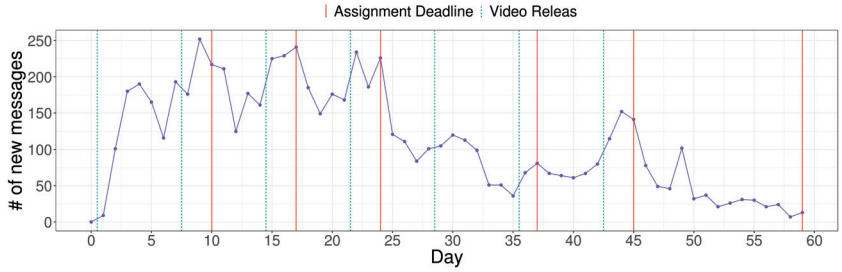
Course	Forum participants	Forum contributors*	Message count	Thread Count
Scala	10,081	1,175	7,699	939
React	12,065	1,902	12,283	1,702

*participants who wrote a message (post or comment)

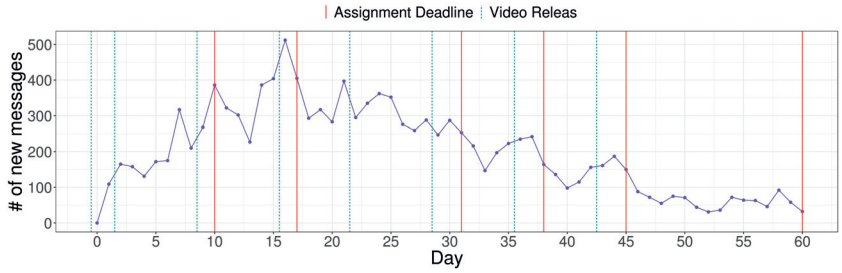
6.1.4 Forum activity over time

To investigate the temporal dynamics of forum activity and its relation with the course structure (**Question 1**), we extracted number of messages (posts or comments), number of forum contributors (participants who wrote a message), and number of new threads added to the content-related sub-forums on each day of the course. Figure 6.1 represents daily count

[2]https://www.coursera.org/learn/progfun1



(a) *Scala*



(b) *Reactive*

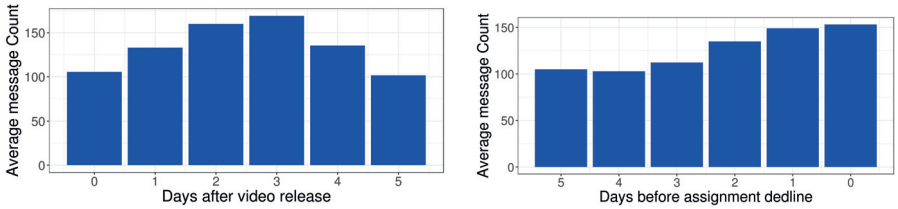
Figure 6.1 – *Number of messages created per day in Scala and Reactive courses. Vertical lines represent video release days (dashed blue) and assignment deadlines (solid red).*

of messages in *Scala* and *Reactive* MOOCs with respect to the main course events: video release (dashed blue lines) and assignment deadlines (solid red lines). Since daily count of new threads and forum contributors were highly correlated with posts count (Pearson linear correlation, $r > 0.9$, $p < .01$ for both courses), here we only represent posts charts.

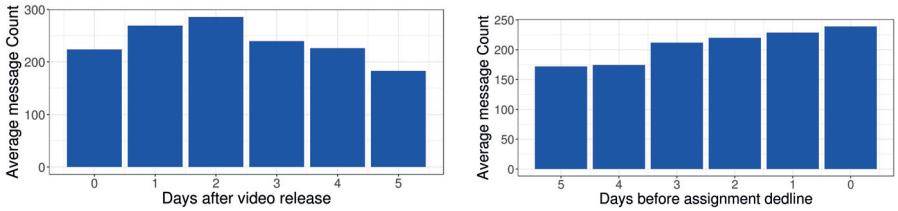
As it can be perceived from Figure 6.1, despite the decline of forum activity over time, at several points close to the video release or assignment deadlines, there is an increment of messages in the discussion forum. This is better perceived from Figure 6.2 which depicts the average count of new messages with respect to the proximity of video release day and assignment deadlines. In *Scala* MOOC, the highest level of forum activity is associated with two to three days after the video release and the forum activity peak in *Reactive* course is on one to two days after video release. Considering the proximity of assignment deadlines, as perceived from the right charts in Figure 6.2, in both courses, the forum activity level increases as the deadline approaches. These observations further confirm the dependency between course structure and forum activity level.

6.1.5 Discussion content over time

With respect to our research question about the evolution of discussion content over the course duration (**Question 2**), in the following we present analysis of the posts' content over time.



(a) *Scala*



(b) *Reactive*

Figure 6.2 – *Average number of new messages depending on the proximity to video release (left) and assignment deadline (right) in Scala and Reactive courses*

Method

To investigate the evolution of discussion content, we extracted certain indicator phrases from the written messages and tracked their distribution over time. As summarized in Table 6.2, we considered two categories of indicator phrases.

The first category, **domain-specific concepts**, contains the set of keywords related to the main concepts of the course subject and were specifically created for each course. For constructing this set, the most frequent concepts in the discussion threads were first determined using Open CalaisAPI[3]. Based on the course outline and detailed knowledge on the course topics (e.g, common tools and concepts in functional and reactive programming), this initial set was then manually refined and different spelling and synonyms were explicitly taken into account (e.g. "lambda function" and "anonymous function" were mapped to the same concept). Following this procedure, 25 domain concepts for $Scala$ and 19 for $Reactive$ were extracted (examples are provided in Table 6.2).

The second category of indicator phrases, **content-related keywords** is comprised of terms and phrases which indicate content-related discourse. Unlike the previous category that could directly be mapped to a specific course topic, these terms are more general in nature and distinguish content-related discussions (e.g. discussions about course topics and materials) from non-content-related discourse (e.g. social talks, search for study groups, questions about deadlines or certificates). They can either appear in combination with domain specific terms (e.g. *"Is there a* ***difference between*** *a lambda and an anonymous function?"*) or without mentions of domain concepts (e.g. *"I have no idea how to approach this problem. Can someone*

[3]http://www.opencalais.com/opencalais-api/

Table 6.2 – *Examples of indicator phrases used to track discussion topics over time*

Category	Example
Domain-specific concepts	*Scala*: subroutine, recursion, concurrent, immutable, anonymous function, patetrn matching, huffman coding *Reactive*: akka, promise, await, replicator, heap, concurrent
Content-related keywords	understand, difference, solution, answer, feedback, clarify, clarification, question, example, explain, mean

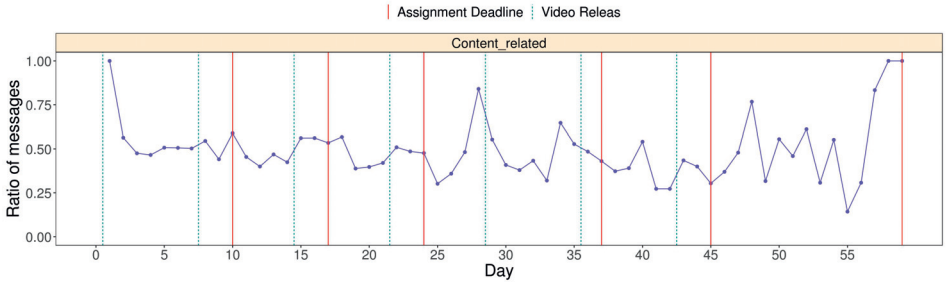
provide a better ***explanation****?"*). Such keywords as "difference_between" and "explanation" have been characterized as "signal concepts" by Daems et al. [58] and are considered as indicators of potential information needs and problems in understanding domain related concepts. Our selection of general content-related keywords was also inspired by Wise and Cui's [236] findings on the identification of content-related threads which suggests that general terms such as "understand", "example", and "difference" are among the predictors of content-related contributions.

Results

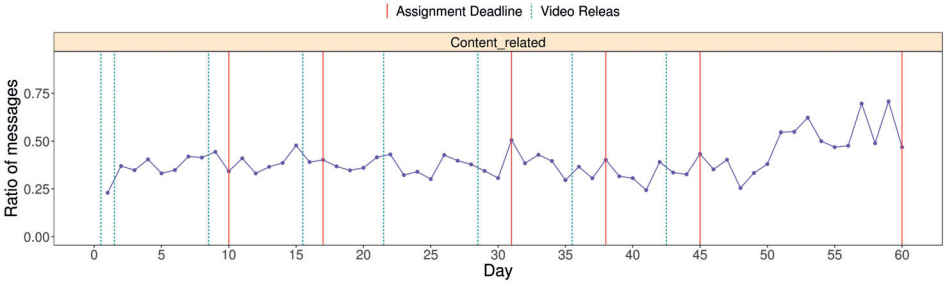
Figure 6.3 represents the ratio of messages including content-related keywords during the course timeline for *Scala* and *Reactive* MOOCs. In general, content-related discourses are present throughout the course duration and, unlike our hypothesis, their distribution is not strongly influenced by video release or assignment deadlines. In *Scala* course, on average 49% (sd = 17%) of the messages on each day contain content-related keyword and this ratio for reactive course is 41% (sd = 12%).

Considering the domain-specific concepts, Figure 6.4 provides concrete examples of distributions of terms related to the course subject. Some domain concepts such as *recursion* and *anonymous function* in *Scala*, are present in the discussions from the beginning and can be related to participants with certain background knowledge, who are able to discuss the specific course concepts independently from the conveyed knowledge in the lectures. On the other hand, some other domain concepts such as *pattern matching* and *huffman* in *Scala* and *heap*, *promise*, *akka*[4] and *replicator* in *Reactive*, are clearly introduced to the discussions after a specific video release and could represent the challenging or unclear concepts for the learners in the corresponding lecture. Interestingly, most of such lecture introduced concepts remain in the discussion until the end of the course. This indicates that discussion forums are to some extent useful for further discussion on lecture introduced knowledge which could be connected with the following course sections.

[4] Akka is a toolkit used for reactive programming.



(a) Scala



(b) Reactive

***Figure 6.3** – Distribution of content-related posts over time, in Scala and Reactive courses. Vertical lines represent video release days (dashed blue) and assignment deadlines (solid red).*

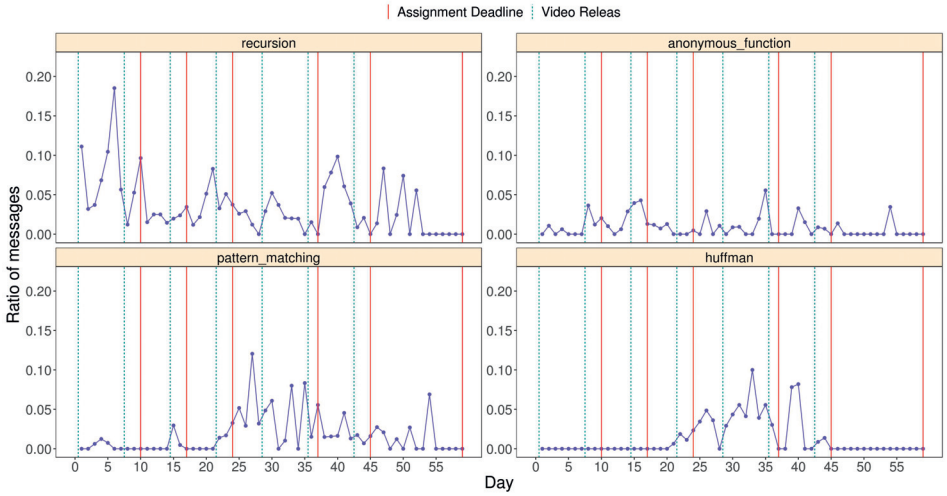
6.1.6 Social communication structure

In this section, we explore the social aspect of discussion forum and investigate the network of information exchange among forum contributors. In particular, we study the information exchange network at two levels: global and individual. At the global level we explore the evolution of network over time (Section **Evolution of network structure over time**), whereas at the individual level we focus on students' roles in the network (Section **Structural roles over time**).

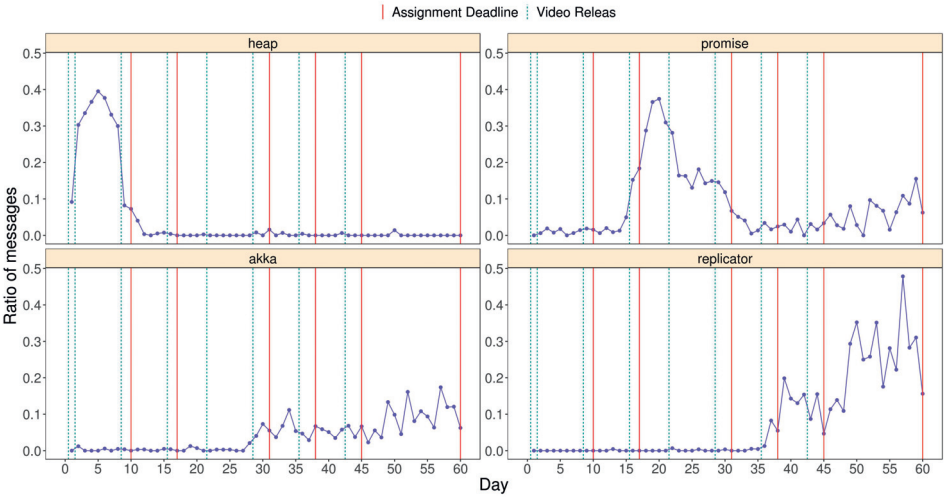
Methods

Network extraction

Several approaches have been used in the literature to model learners forum communication as a social network. One common approach is to consider links between all participants contributing to the same discussion thread [85, 90], or to link contributors in a thread only to the thread initiator [116, 252]. Another commonly used approach is to extract reply relationship between forum participants according to the hierarchical structure of messages in the forum. In this method, participants writing a post in thread are considered to be tied to the thread starter, whereas those writing a comment are linked to the author of the corresponding post [112, 116]. In [237], Wise and Cui have examined the effect of different tie definitions on social



(a) *Scala*



(b) *Reactive*

Figure 6.4 – *Examples of domain-specific concepts, and their distribution over time in Scala and Reactive courses. Vertical lines represent video release days (dashed blue) and assignment deadlines (solid red).*

network structure and interpretation. Their study showed that network properties were not very sensitive to differences in tie definition, with an exception of the first approach (linking all contributors in the same thread, also known as total co-presence), which might result in dramatically distinct network structures.

In this study, we are interested in information exchange relations among participants and their information giving and information seeking behaviours over time. The aforementioned methods do not take into account the semantic attributes of the written messages and hence do not differentiate between message types (e.g. questions, answers, clarifications). Therefore,

Table 6.3 – *Examples of classified posts*

Message	Category
Would it be OK to unfold source list to a string (like 'aaabbc') and then build combinations from this string? I still can't think in terms of Lists.	Information seeking
I'm having a similar problem. Mine is in a worksheet called *patternmatching*. Sorry I'm not understanding this.	Information seeking
Not sure about my solution: I am using recursion, but I found that I must sort the coins list first. Is this the right approach, or am I missing something else?	Information seeking
You're fine, what he means is that there is no need to use an extra stack, since the passed parameter is already kind of like a stack.	Information giving
If you want to use those constructor parameters, just give them names.	Information giving
The point is to implement it in a way that will not need *isEmpty* or *isInstanceOf* but will work implicitly instead.	Information giving
I think it is starting to make sense. Thanks.	Other
Of course, I will implement it myself then. Thanks!	Other
Good to learn. Never used it before.	Other

to extract concrete information exchange network from forum discussions, we adopt the method introduced in [102]. This method incorporates three steps: **(1)** post classification, **(2)** posts relation extraction, **(3)** transforming network of posts into network of participants. We briefly describe these steps in the following and refer to [102] for more details.

In the first step, messages are classified into three classes *information giving*, *information seeking*, and *other* by applying a supervised classification method (random forests) trained on a set of 300 manually labeled posts. All messages that request information, for example, concrete questions on course topics or requests for advice were coded as *information seeking*. Posts that provide any kind of information to information seekers were subsumed as *information giving*, and posts which cannot be associated to any of the other classes were labeled as *other*. The feature set used by the classifier includes a combination of structural features (e.g. absolute and relative position of the message in the thread, number of votes) and content related features (e.g. message length, occurrences of questions words, question and exclamation marks, and specific phrases such as "need help" or "helps you"). Using this method, a classification accuracy of $F1 = 0.77$ was achieved in our dataset, which is close to the values reported in [102]. Table 6.3 provides examples of the classified messages. As shown by the examples, an information seeking message is not necessarily in the form of a concrete question, but it can also be an implicit clarification request or reporting a similar problem already mentioned by another learner in the discussion forum.

In the second step, relation between the posts is extracted. In order to do so, first the posts

labeled as *other* are removed. The remaining messages in the discussion thread (or a sub-thread comprising comments to a parent post), can then be decomposed into alternating sequences of *information seeking* and *information giving* messages. In the most usual case, messages in a information giving sequence refer to the most recent preceding information seeking sequence. This allows to extract the network of posts by connecting information giving messages to their corresponding information seeking message(s) with outgoing links. Furthermore, an edge between two messages carries a timestamp indicating when it was created which is extractable based the creation time of the messages.

In the final step, the final information exchange network among forum participants is derived by collapsing all nodes with the same author in the posts network, into a single node. Therefore, in the resulting network, there exist a directed edge between two nodes (representing forum participants) if the first participant provided some information to the second one.

Based on the timestamps of the edges, the resulting network can be divided into a sequence of network slices corresponding to certain time intervals. Each network slice contains all the nodes (forum participants) but only the edges that where created in the corresponding time window. This allows to study the dynamics of the social communication structure in detail, as it will be explained in the following.

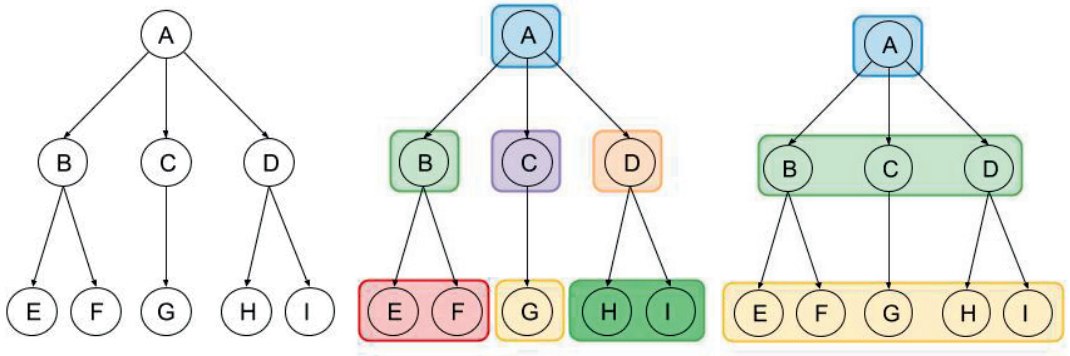
Network structure over time

In order to study the temporal dynamics of information exchange network, we consider overlapping network slices over one week time windows using a sliding window approach. This results in one network slice for every day (d where $d > 6$) of the course modeling participants connections based on their forum activity during the past seven days ($[d-6:d]$). For each network slice, we then extract a set of classic structural attributes, including number of nodes and edges, average total degree[5], network density[6], average path length[7], and global clustering coefficient[8]. Tracking the structural attributes of these overlapping network slices would enable us to investigate how the overall structure of the information exchange network evolves over time.

Role modeling

In order to determine the individual's role in the information exchange network, we use role modeling techniques (also known as positional analysis) in social network analysis. The general idea of role modeling methods such as blockmodeling [69] is to decompose the set of nodes in a network into clusters with equivalent connection patterns to the other nodes. In social network science those clusters are interpreted as users who have similar role or position

[5] Average number of connections that a node has to/from other nodes
[6] Ratio of existing edges in the network to the possible number of edges
[7] Average number of steps along the shortest paths for all possible pairs of connected nodes
[8] Fraction of closed triangles (cliques of three) to the possible triangles



(a) Original network *(b) Structural equivalence sets* *(c) Regular equivalence sets*

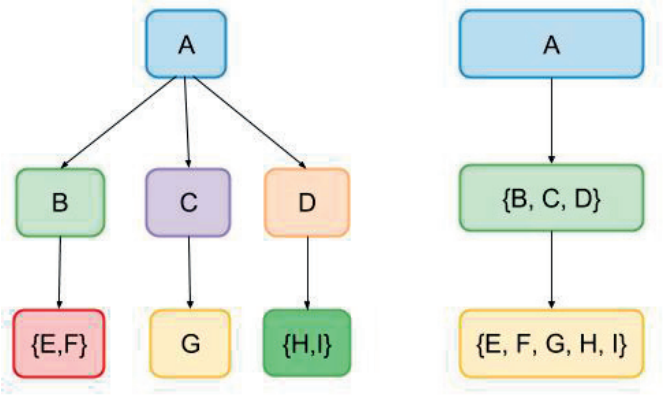
Figure 6.5 – *Example of structural and regular equivalent sets*

in the community, hereby referred to as a **structural role** [196].

The notion of equivalence can be defined in different ways. Structural equivalence [142] and regular equivalence [234] are the most commonly used definitions. Structural equivalence requires two equivalent nodes to have the same type of connections to same set of nodes, which implies having the exact same neighbors[9]. Regular equivalence relaxes this strict criterion such that equivalent nodes should have similar relations to nodes that are equivalent themselves. That is, regular equivalence sets are composed of nodes which have similar types of relations to members of other regular equivalence sets, but do not necessarily share the same set of neighbors. Mathematical details for the computation of structural and regular similarity can be found in [69].

To clarify the concept of structural and regular equivalence with an example, consider a network representing family relations. In such a network, two mothers don't have the same children, husband, or in-laws, so they are not structurally equivalent. However, they have a similar pattern of connections with a husband, children, and in-laws. Therefore the two mothers can be regarded as regularly equivalent due to the similarity of their connection patterns with at least one member in each of the other sets of actors (who are themselves regarded as equivalent due to the similarity of their ties to a member of the set mother) [96]. Another example of structural and regular equivalent sets is presented in Figure 6.5. Considering the structural equivalence in this sample network, as there is no other node with the exact same set of contentions as A, this node forms a cluster by itself. This is also the case for B, C, D and G. However, E and F have the exact same connections as they both have a single incoming link from node B. Therefore this two nodes form a structural equivalence cluster. The same is true for nodes H and I. On the other hand, with respect to regular equivalence, only three equivalent clusters can be identified in this example. The first cluster includes a single node, A. The second cluster is composed of B, C and D. The third

[9] Two nodes in a network are neighbors if they are connected to each other.



(a) *Structural equivalence* ***(b)*** *Regular equivalence*

Figure 6.6 *– Blockmodel representation of example network in Figure 6.5, according to structural and regular equivalence based clustering.*

cluster consists of nodes E, F, G, H and I. The three nodes in the second cluster are regularly equivalent, as each of the nodes has an incoming connection from the first cluster, and has at least one outgoing connection(s) to the third cluster. Similarly, the third set of nodes are regularly equivalent as they have no connection with any node in the first cluster (node A), and each has an incoming connection from a node in the second cluster (B, C or D).

Structural equivalence is a very restrictive form of similarity and in many real and particularly large networks, exact structural equivalence may be rare. On the contrary, the concept of regular equivalence quite closely corresponds to the sociological concept of a role, and therefore it is the most commonly used definition in different applications [96]. Similarly, in the case of discussion forum communication networks, considering the sparsity and usually large size of such networks, clustering learners based on the regular equivalence is more reasonable. Therefore, we use regular similarity in conjunction with hierarchical agglomerative clustering to model learners structural roles in the discussion forums. For the computation of regular similarity we use REGE algorithm described in [31]. The resulting clusters and the relational patterns between them can be represented using a blockmodel. Figure 6.6 depicts the block model representation of the example network presented in Figure 6.5. Each node in the blockmodel, represents one structural role, that is a cluster of nodes in the original network. In the blockmodel resulting from structural equivalence clustering, relations between each pair of connected roles (e.g. $r1$ and $r2$) are either complete (all nodes in $r1$ are connected to all nodes in $r2$) or non-existent (there are no connections between nodes in $r1$ and $r2$). Whereas for two connected roles extracted according to regular equivalence, all nodes in $r1$ point to at least one node in $r2$ and all nodes in $r2$ receive a link from at least one node in $r1$. A blockmodel thus presents the interpretable macro structure of a possibly very complex social network and allows to uncover the inherent organization.

Later on, in Section **Structural roles over time**, we extract the blockmodel from each time

slice of the evolving information exchange network representing the role structure in different periods of the course. The resulting blockmodels are then used to investigate role changes of course participants over time.

Results

Following the described network extraction method, we extracted the information exchange network from the discussion forum communication in *Scala* and *Reactive* MOOCs. Table 6.4 summarizes the resulting networks' attributes. On average each forum participant provides/receives information to/from four and five other forum participants respectively in *Scala* and *Reactive* course. Despite the relatively larger network size in *Reactive*, the low density values reflect the sparsity of the connections. Furthermore, comparison of the forum contributors count (Table 6.1) and nodes count in the resulting networks suggests that a subset of forum participants (272 in *Scala* and 411 in *Reactive*) either had no information seeking/giving messages or their messages did not get replied by any other forum participant. In the following we present the results on network structure and learners structural roles over time.

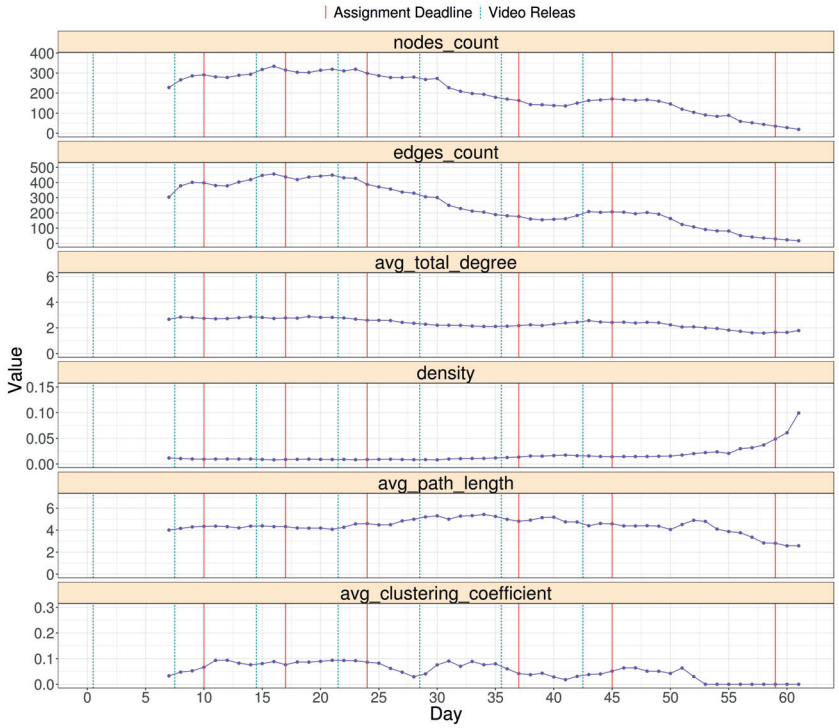
Table 6.4 – *Overview of overall knowledge exchange network attributes*

Database	Nodes count	Edges count	Total degree average (SD)	Density
Scala	903	1806	4 (9)	$4.4e-3$
Reactive	1491	3740	5 (9.2)	$3.2e-3$

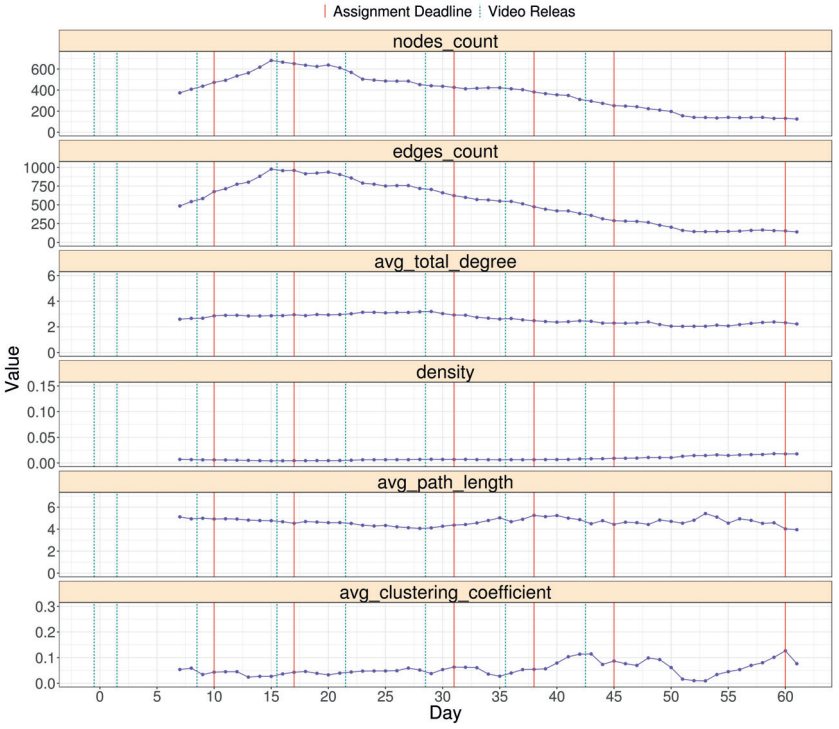
Evolution of network structure over time

Considering our research question about the evolution of information exchange network structure (**Question 3**), based on the trends observed in Section 6.1.4, we hypothesize that course schedule would influence the network structure. For instance with the increment of contributors and messages in the discussion forum close to the video release or assignment deadlines, new nodes or edges could appear in the network which in turn could influence network attributes such as size, average degree, or density.

Figure 6.7 represents the evolution of network attributes over weekly network slices (extracted using sliding window as mentioned before) for both the courses. The overall decrease of forum activity towards the end of the course is also reflected by network size metrics (nodes and edges count). The networks are very sparse since the low average degree in relation to the large network size results in a low density (< 0.02). Despite the comparatively larger network size in *Reactive*, in both courses average path length are relatively small (< 6) throughout the course. Short path length though sparsity of the connections are a typical property of small-world networks [230]. Small-world networks tend to contain cliques, and near-cliques, which



(a) Scala



(b) Reactive

Figure 6.7 *– Network attributes over time, based on one week network slices using sliding window in Scala and Reactive courses. Vertical lines represent video release days (dashed blue) and assignment deadlines (solid red).*

are sub-networks that have connections between almost any two nodes within them. Many technological, biological, social, and information networks fall into this broad category as they consist of tightly interconnected clusters of nodes [108]. However, in contrast to classical small-world network models, the clustering coefficient for the network slices in both courses is low (< 0.1). Clustering coefficient (bounded in $[0,1]$), is a measure of the degree to which nodes in a graph tend to cluster together. Therefore, the low clustering coefficient indicates that the communication structure does not evolve into densely connected communities, but rather into sparse parts interconnected via a few highly connected nodes.

On contrary to our hypothesis, course events do not show any direct influence on the network structure. One plausible explanation could be the structural limitations of the communication network such as the absence of persistent discussion groups throughout the course duration, which is also pointed out in previous studies such as [84] and [163]. Moreover, the increase of messages in the discussion forum in a particular period of time, could be resulting from a sequence of messages between few students, which would not add new edges or nodes to the network. Furthermore, since an edge in the networks aggregates possibly multiple communication events between a pair of users, such message sequences would not be reflected in the network structure.

Structural roles

To model individuals' structural role in the discussion forum, we consider successive bi-weekly slices (hereafter referred to as time phases) of the knowledge exchange network and extract the macro structure of each network slice, using the described role modeling method. According to [247], to obtain a clear image of the online communities, the time slice size needs to be long enough to cover the typical production cycle of the community and therefore should be adjusted according to the latent inherent time and productivity speed of the community under study. The choice of two weeks time windows is therefore reasonable considering that in both courses, similar to most other MOOCs, video lectures and assignments are structured as one or two weeks thematic blocks. It is therefore presumable that the forum discussions would follow a similar pace [104]. The resulting role models for all four time phases in both courses follow the structure represented in Figure 6.8. The derived blockmodels consist of four structural roles (clusters of users) which we label as **core**, **help givers (HG)**, **help seekers (HS)**, and **inactive/isolated**, according to their connection patterns and other attributes summarized in Table 6.5 and 6.6. In the following we describe these roles in more details.

The **core** participants form a cohesive subgroup in the sense that they have communication relations within their cluster (self-loop in the blockmodel), and also with other two clusters (outgoing connection to *HS* and incoming connections from *HG*). Therefore, participants with the *core* role, contribute to the communication both by receiving and providing information from/to other forum participants. This is further confirmed by their relatively high average of

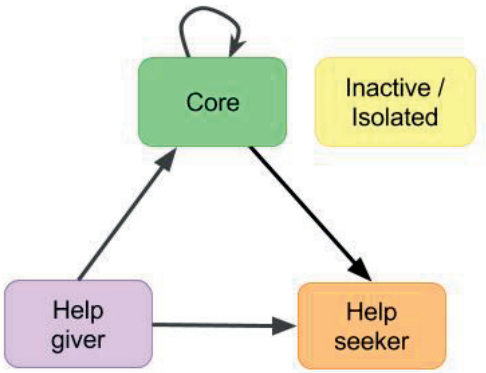


Figure 6.8 – *Role structure of information exchange network in bi-weekly slices*

in-degree[10] and out-degree[11] (> 2.9) in all phases, according to Tables 6.5 and 6.6. Furthermore, despite relatively smaller size (number of participants) of *core* cluster, this category of learners are associated with the highest average number of messages in all time phases. This further reflects their high engagement level in the discussions. The two other roles, HS and HG in Figure 6.8 are not cohesive but are connected to other roles. These two peripheral roles can be characterized as **help givers** and **help seekers** respectively since they have only outgoing or incoming relations. Learners with HG role serve as information providers in forum communication without ever receiving help from others (in-degree of zero). Learners with HS role, show an opposite behavior by only receiving help from other forum members (out-degree of zero). Finally, the fourth role, **inactive/isolated**, includes all participants who do not have any connections to others in a particular time slice. The absence of connections for these learners could have two reasons, either they were not active in the discussion forum during the time span for which the model was created (inactive) or their posts could not be linked to the other posts (isolated). An example could be a help-seeking post without a reply or a post not related to information exchange (Section 6.1.6). According to Tables 6.5 and 6.6, majority of forum participants in all time phases, fall within this category.

Structural roles over time

To track the evolution of individuals' structural role in the discussion forum (**Question 4**) we construct learners' role sequences based on the blockmodels described above. To differentiate late-comers or drop-outs from students who follow the course without participating in the discussions, we consider a fifth role referred to as **course inactive**. This role consists of learners who do not perform any type of activity in the course platform during a time phase. On the other hand, *inactive/isolated* role includes learners who are engaged in other course activities (following lectures or submitting to assignments), but do not contribute to the discussions.

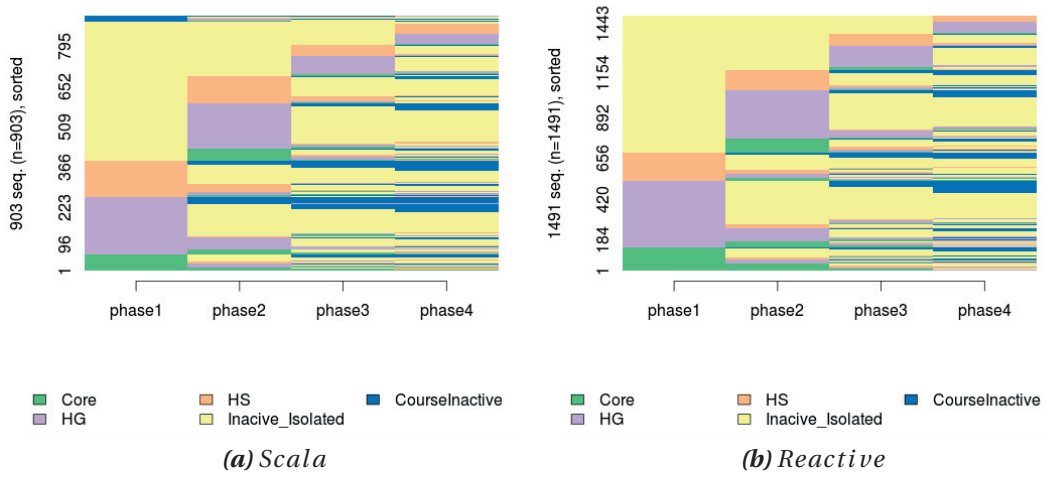
[10] Number of incoming connections to a node.
[11] Number of outgoing connections from a node

Table 6.5 – *Attributes of structural roles in information exchange network slices for Scala course*

Attribute	Time phase	Structural role			
		Core	HS	HG	Inactive
Participants count	1	60	129	200	515
	2	89	140	229	449
	3	48	85	116	655
	4	24	75	91	714
Average message count	1	167	46	52	16
	2	176	59	64	27
	3	133	32	49	3
	4	114	45	55	10
Average in-degree	1	4.4	2	0	0
	2	3.6	2.6	0	0
	3	2.9	2.1	0	0
	4	3.2	2	0	0
Average out-degree	1	4.3	0	1.7	0
	2	4	0	1.5	0
	3	3.1	0	1.4	0
	4	2.9	0	1.7	0

Table 6.6 – *Attributes of structural roles in information exchange network slices for Reactive course*

Attribute	Time phase	Structural role			
		Core	HS	HG	Inactive
Participants count	1	138	171	406	814
	2	184	188	402	755
	3	113	155	263	998
	4	43	124	170	1192
Average message count	1	103	45	59	20
	2	207	69	92	38
	3	176	79	86	32
	4	136	54	58	14
Average in-degree	1	4.5	3	0	0
	2	5.1	2.7	0	0
	3	3.7	2.5	0	0
	4	3.3	2.3	0	0
Average out-degree	1	2.9	0	1.8	0
	2	3.7	0	1.9	0
	3	3.2	0	1.6	0
	4	2.9	0	1.8	0



(a) *Scala* ***(b)*** *Reactive*

Figure 6.9 – *Structural role sequences in Scala and Reactive courses. Each horizontal line denotes the role sequence for one learner over four bi-weekly time periods. Label of the vertical axis represent total number of forum participants.*

Figure 6.9 depicts structural role sequences for forum participants in *Scala* and *Reactive* courses. Every learners' role sequence is represented by a horizontal line in the role sequence charts. According to Figure 6.9, instances of active forum participation throughout the course duration are quite rare and most students are often active only in one or two phases. Furthermore, in each phase, a considerable portion of active students in the discussions, are new forum participants (i.e. for the first time have a role different from *inactive/isolated*). This observations in turn could imply that persistent discussion groups in the forum are not very common.

Next, to identify the common role sequence patterns, we clustered learners based on the similarity of their role sequences. To assess pairwise distance between role sequences we use *optimal matching* distance and determine the degree of dissimilarity between two sequences based on the edit operations (substitutions) required to match the two sequences. In optimal matching, different substitutions can be associated with different costs. In a common data driven approach, substitution costs are determined based on state transition matrix [224]. The general idea in this approach is to define substitution costs inversely proportional to transition rates, that is to consider a higher costs for substituting between states when the transitions between them are rare, and a low cost when frequent transitions are observed [211].

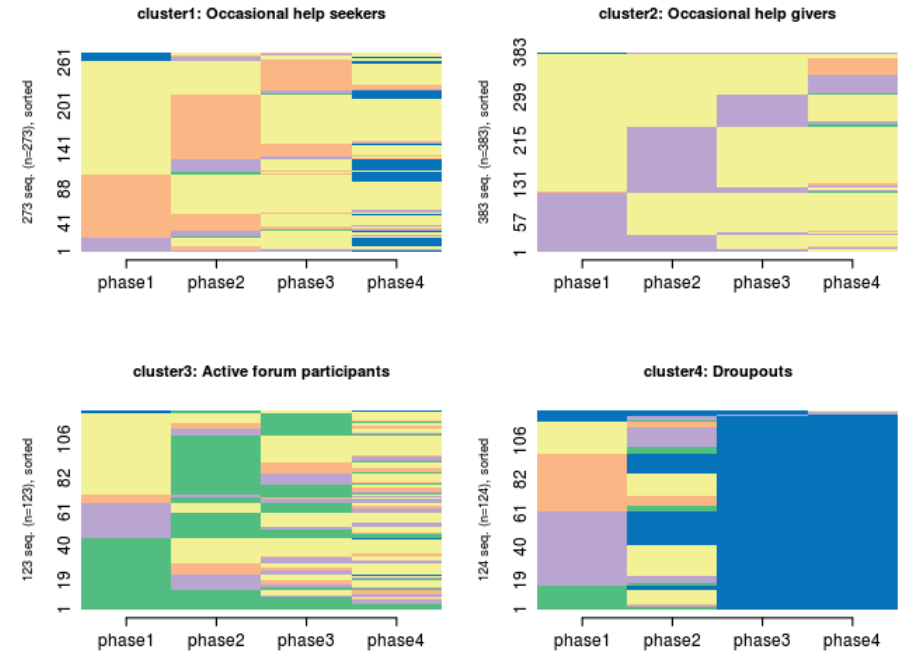
Following a similar approach, along with hierarchical agglomerative clustering method, we extracted four cluster of role sequences in *Scala* and *Reactive* course. Number of clusters was determined using cluster bootstrapping, a method to estimate the optimal number of clusters by minimizing cluster instability, given pairwise points distances and a clustering

function [74].

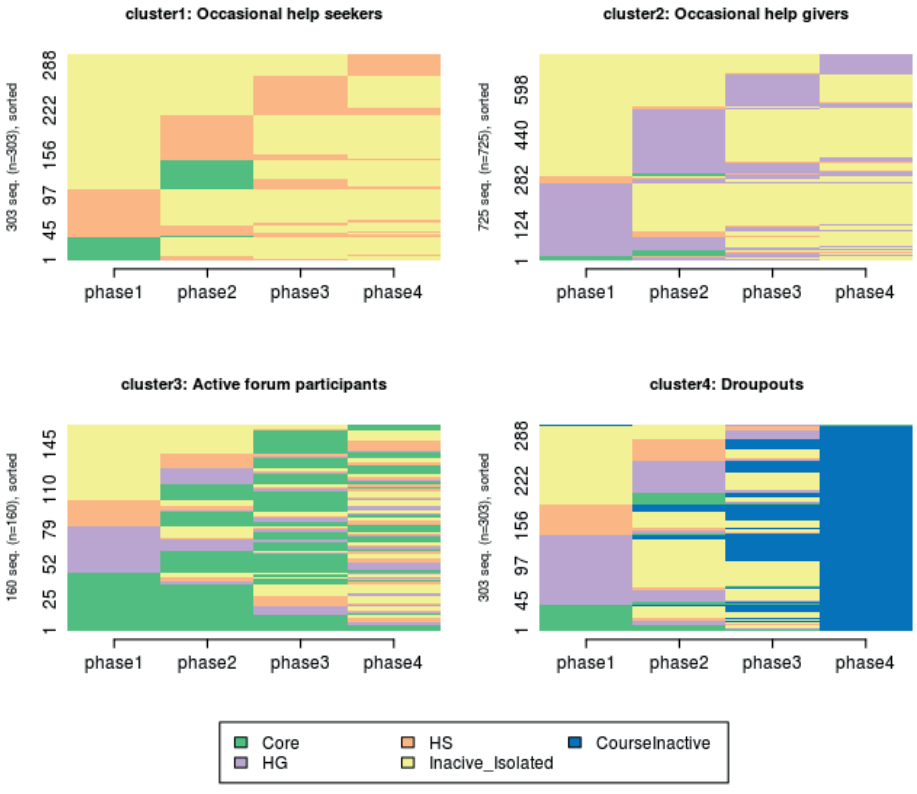
Figure 6.10 represents the resulting clusters of role sequences for both courses. According to the sequence charts, in both courses the first two clusters are composed of occasional forum participants, the learners whose forum participation is limited to one or two time periods during the course duration. As learners in these two clusters are mainly attributed with help seeking and help giving roles, we refer to them respectively as **occasional help seekers** ($N = 273$ in $Scala$ and $N = 303$ in $Reactive$) and **occasional help givers** ($N = 383$ in $Scala$ and $N = 725$ in $Reactive$). On the other hand, the third cluster comprises learners who are active during significantly more time periods in comparison with the previous two categories (2.4 vs. 1.3 in $Scala$; 2.9 vs. 1.4 in $Reactive$, $p < 001$, Mann-Whitney-Wilcoxon test). Furthermore, periods with *core* role are more frequent in this cluster. Overall, during 35% of time phases in $Scala$ and 41% in $Reactive$, learners within this cluster are attributed with *core* role. This cluster can therefore be characterized as **active forum participants** ($N = 123$ in $Scala$ and $N = 160$ in $Reactive$). Finally, the forth cluster clearly includes learners who drop out during the second half of the course and hence we label this cluster as **dropouts** ($N = 124$ in $Scala$ and $N = 303$ in $Reactive$).

Comparison of average grade obtained by each cluster of participants in Figure 6.11, reveals that in both courses, occasional help seekers have significantly lower grades compared to occasional help givers (77 vs. 86, $F[1,1] = 23.3$, $p < .001$ in $Scala$; 86 vs. 92, $F[1,1] = 18.6$, $p < .001$ in $Reactive$). Additionally, despite the fact that occasional help givers have lower forum participation compared to active forum participants, both groups achieve comparably high scores (86 and 88 in $Scala$, 92 in $Reactive$). Interestingly all the three categories of described learners, demonstrate high level of engagement in the other sections of the course. On average they access 75%-80% of the video lectures and submit to 90% of the graded assignments. Therefore one plausible explanation for the observed performance differences could be that active forum participants take advantage of discussion forum to advance their knowledge and resolve difficulties with respect to the course materials, whereas occasional help givers could be students with higher expertise level who during their few periods of forum participation, mainly provide answers to questions asked by other participants. Therefore, it would be promising to foster sustainable knowledge exchange dialogues among learners through development of mechanisms for engaging occasional forum participants in more frequent discussions, connecting peripheral help seekers to proper communication partners and further enabling peripheral help givers to reach information seeking requests.

Intuitively, dropout clusters have the lowest average grade (21 in $Scala$ and 40 in $Reactive$). An interesting observation concerning this category of participants is their active forum roles as help givers, help seeker, or even core during the first weeks of the course. Therefore initial forum roles are not necessarily a predictor of dropping out from the course.

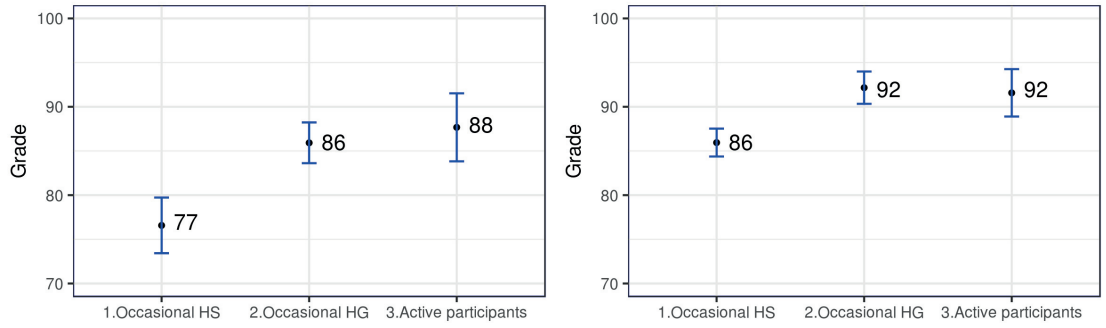


(a) *Scala*



(b) *Reactive*

Figure 6.10 *– Clusters of role sequences in Scala and Reactive course. Each horizontal line denotes the role sequence for one learner over four bi-weekly time periods. Vertical axis labels denote number of learners in each cluster.*



(a) *Scala* ***(b)*** *Reactive*

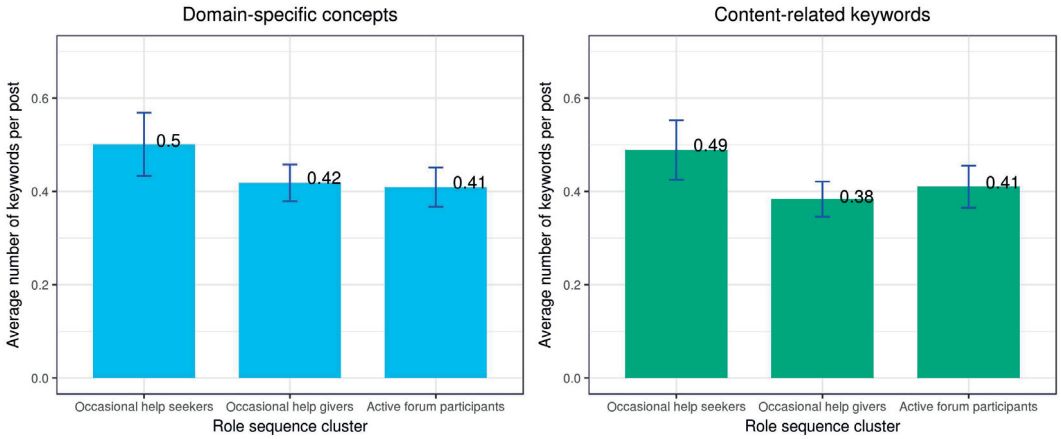
Figure 6.11 – *Average grades by participants in role sequence clusters*

6.1.7 Social structure and discussion content

In this section, we investigate the relation between learners structural roles in the network and the discussion content (**Question 5**), by combining the results of role sequence clusters described in the previous section with the content analysis reported in Section 6.1.5.

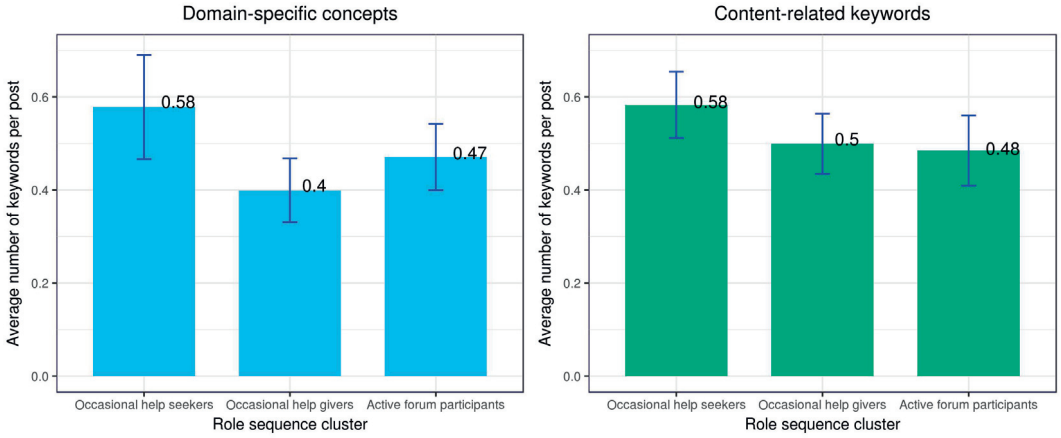
Figures 6.13 and 6.12 present the distribution of indicator phrases (Table 6.2) in the posts made per user in each of the four role sequence clusters (Figure 6.10). Note that one post may contain keywords of different types. Considering the higher engagement of active forum participants (third cluster in Figure 6.10) in knowledge exchange discussions, we expected the content-related and domain-specific discussions to be mainly made by them. However as perceived from Figures 6.13 and 6.12, occasional forum participants have comparable and sometimes more mentions of domain-specific concepts and content related phrases in their posts.

Moreover, it is interesting to note that in both courses, posts by occasional help seekers contain a higher ratio of domain-specific and content-related keywords in comparison to the other groups. This suggests that this group of learners make fewer but important contributions in the information exchange communication. This finding is not obvious and gives interesting insights into the characteristics of forum users who are engaged in discussions in a limited time span. While related works suggest that the structural coherence of the forum communication mainly depends on the small set of very active users [85, 164], the content analysis of the posts shows that the other users could also have an important impact on the discourse by triggering focused discussions on specific subject areas and mentioning concrete problems.



(a) Domain-specific concepts *(b) Content-related keywords*

***Figure 6.12** – Average number of keywords in posts by user in structural role clusters for Scala course*



(a) Domain-specific concepts *(b) Content-related keywords*

***Figure 6.13** – Average number of keywords in posts by user in structural role clusters for Reactive course*

6.1.8 Predicting forum activity level

In the previous sections we explored the temporal dynamics of the discussion forum across time, content and social domain and in Section 6.1.4 we showed that the course structure has some influence on the forum activity level. In this section, we aim to build a model for predicting the forum activity level during the course duration (**Question 6**). This information could enable the teaching team to prepare the logistics to efficiently support students in discussion forums, mainly during the high activity periods, or to plan interventions for boosting forum participation during more silent periods.

Table 6.7 – *Description of features used in the predictive model*

	Structural features
Dv	number of days after the latest video release
Da	number of days after the latest assignment release
Dd	number of days left to the next assignment deadline
Dr	ratio of the current day (d) to the course length encoding what percentage of the course is passed
Na	number of assignments open for submission
	Previous forum activity
M_k	number of new messages created on day $d-k$
TM_k	mean time between successive forum writing events
	Network features
Net_k	network features on day $d-k$, including nodes count, edges count, average degree, average path length, and clustering coefficient
	Initial forum activity (first week)
$W1M$	mean and standard deviation of message count per day, average time between messages

Method

In order to construct the predictive model, we extract four sets of features for each day (d) of the course as summarized in Table 6.7. The first category, **structural features**, describe the characteristics of a day, with respect to the course structure, such as time after video release, time before assignment deadlines and the passed ratio of the course. Consideration of such features is inspired by the trends we observed in Section 6.1.4: the overall decrease of forum activity over time and its increment close to lecture and assignment dates. **Previous forum activity** features encode the volume and intensity of forum activity on previous days. An example is number of messages and the time between massages created on k days before the current day. The third category of features, comprises the structural attributes of the network slice on k days before the current day (See Section **Network structure over time** for details on network partitioning and features description). Finally, the last category of features describes the **initial forum activity** level, during the first week of the course. Such features could act as a normalization factor to compensate the difference in intrinsic popularity of discussion forums in different MOOCs.

Using the described features, we built a regression model for estimating number of new messages in discussion forum on each day of the course. We tested different machine learning methods such as support vector regression model (SVR) with linear and RBF[12] kernel, random forests, and neural networks for building the predictive model. Data from $Scala$ and $Reactive$

[12] radial basis function

Table 6.8 – *Overview of predictive models and results*

Features	R^2 (train)	NRMSE* (train)	NRMSE (test)
All features	**0.89**	**0.27**	**0.23**
All excluding structural features	0.72	0.42	0.38
All excluding previous forum activity	0.86	0.28	0.29
All excluding network features	0.88	0.3	0.25
All excluding initial forum activity	0.77	0.28	0.29
Average baseline	$1e-3$	0.5	0.73

*Normalized RMSE by mean of observed values

courses was randomly partitioned into training (70%) and testing (30%) sets, and 10-fold cross validation on the training set was used to tune the models' parameters. Highly correlated ($r > 0.7$) and linearly dependent features were removed prior to model training. Furthermore, to assess the influence of each feature category on the prediction results, we examined two different setups. In the first setup, all the four categories were included in the model, whereas in the the second setup, four reduced models were trained, each excluding one of the described categories from the features set.

Results

Support vector machine with linear kernel resulted in smallest prediction error, reported in Table 6.8. The full model captures 89% variance of the dependent variable and results in quite accurate predictions as reflected by low value of normalized root-mean-square errors (NRMSE) on the test data (0.23). Considering the reduced models, as reflected by prediction errors, excluding the structural features results in the highest deterioration of the model accuracy (NRME test=0.38) which suggests their high predictive power. On the other hand, excluding network features, has the least effect on prediction accuracy. This could be due to the fact that as shown before (Section **Evolution of network structure over time**) several network attributes such as degree, density, and path length show a low variance during the course duration and therefore their inclusion does not add substantial information to the model.

Regarding the previous forum activity and network features, best prediction results were obtained when the features corresponding to seven days before was used in the model (i.e. $k = 7$ for $k \in [5, 15]$) and hence the introduced model is capable of predicting the forum activity level one week in advance. According to variable importance analysis on the full model, the most influential features in the model are: passed ratio of the course (Dr), count and average time between messages on previous days (M_7, TM_7), time after video release (Dv), number of open assignments (Na).

6.2 Social interactions in Realto

In this section, we apply social network analysis methods to model participants interaction in Realto and investigate how different stakeholders (apprentices, teachers, and supervisors) are connected to each other. In particular we address the following question:

Question 7. What are the attributes of connections between different stakeholders in Realto and what is the macro structure of their communication?

6.2.1 Method

As mentioned in Chapter 3, Realto is composed of two main areas: social area and learning documentation (LD) area. In the social area participants can share different types of resources and react to each others contributions by commenting and rating (likes). Such social features are mainly being used by apprentices and teachers, although they are available to all participants. On the other hand, LD area enables apprentices to document their workplace training procedure and receive supervisor's feedback on it. Once an apprentice creates a LD, it becomes visible to his/her supervisor. The supervisor can then give detailed feedback (comment) on different sections of the document or provide an overall evaluation for the submitted LD.

To construct the network of communication between Realto participants we follow a different approach from the presented MOOC study and extract the directionality of relations between users in this platform. We consider five types of actions: **(1)** commenting and **(2)** rating posts, **(3)** creating LDs, **(4)** providing feedback on LDs, and **(5)** evaluating LDs. The first two action types, describe the social interactions among participants and are modeled as a directed link from the actor to the resource owner, weighted according to the count of such interactions. The other three actions, model LD-related interactions. In this case, creating a LD is modeled as a directed link from the apprentice to his/her supervisor; whereas giving feedback or evaluation of learning documents are represented as links from supervisor to the apprentice. The connection weights in this case represent the number of submitted/evaluated LDs. In case of multiple feedback on a single LD, only one is taken into account in the network.

To analyze the resulting network, in addition to the overall network attributes, we also consider different subnetworks consisting of interactions among particular user categories. We apply role modeling method described in Section 6.1.6 to extract the macro structure of the communication and individuals' structural role in the network. Analysis of the network attributes and the corresponding role model, provides interesting insights about the communication among different stakeholders in Realto.

In order to provide the infrastructure for investigating the social interactions in Realto, we implemented *network analysis* module in the analytics dashboard for this platform (Appendix A.3). This module includes filters to select connection type (comment, like, create, and evaluate a LD), subnetwork type (connections between certain stakeholders, such as apprentices and

teachers) and filters to select a subset of participants (based on profession, school or language). Based on the selected values, the resulting network and its corresponding blockmodel are then constructed and displayed in the dashboard (in addition to the networks attributes). All the results presented in the following are extracted from the network analysis module in Realto dashboard.

6.2.2 Results

To address our research question about the interaction patterns among Realto users**Question 7**), in the following we present results on the attributes of overall communication network in Realto. We then analyze subnetworks of communication among apprentices and their peers, apprentices and their teachers, and apprentices and their supervisors. Finally we investigate the macro structure of communication network in Realto.

Overall network attributes

Interactions among Realto participants over a time period of 26 months (April 2015 to June 2017) form a network comprising 752 nodes (participants) interconnected through 1832 edges with an average weight of 2.1. Social interaction (commenting and rating) account for the 80% of the connections and LD-related interactions (creating, commenting, or evaluating a LD) are represented by the remaining 20% of connections. In total 1372 comments, 1683 likes, and 635 LDs were created in Realto, and 153 of the LDs were evaluated by supervisors or teachers (79 comments and 74 overall evaluation). The relatively large network size in combination with low average degree (4.9), results in a low density (0.01) and reflects the sparsity of connections among Realto users. Furthermore, the network is attributed with small average path length (6.7) and high clustering coefficient (0.43). As mentioned before, this can indicate presence of interconnected clusters of nodes in the network; Realto network can be decomposed into 75 connected component among which 36 consist of only two users. The remaining components have and average size of 17 nodes ($sd = 30$) and average density of 0.5 ($sd = 0.2$) which is relatively high.

Distribution of different stakeholders and their corresponding network centrality measures (total degree and betweenness) are presented in Table 6.9. According to this table, on average each apprentice/teacher is connected to five/six other users, whereas supervisors have connections with only three other users on average. In Realto communication network, teachers have a central role in connecting different user subsets. This is reflected by their high betweenness centrality, which quantifies the number of shortest paths passing through a node. Nodes acting as a bridge between different communities (dense subnetworks) are generally attributed with high betweenness centrality in contrast to the nodes lying inside a community. This measure could therefore reflect the amount of control that a node exerts over the interactions of other nodes in the network [80, 244]. The central role of teacher in connecting communities of learners is evident in the sample network shown in Figure 6.14.

Table 6.9 – *Distribution of user roles and centrality measures in Realto social network*

User role	Number of nodes	Average total degree(SD)	Average betweenness
Total	752	4.9 (5)	146
Apprentice	569	5.1 (4.7)	94
Teacher	104	6.1 (6.1)	477
Supervisor	79	3.3 (5.4)	88

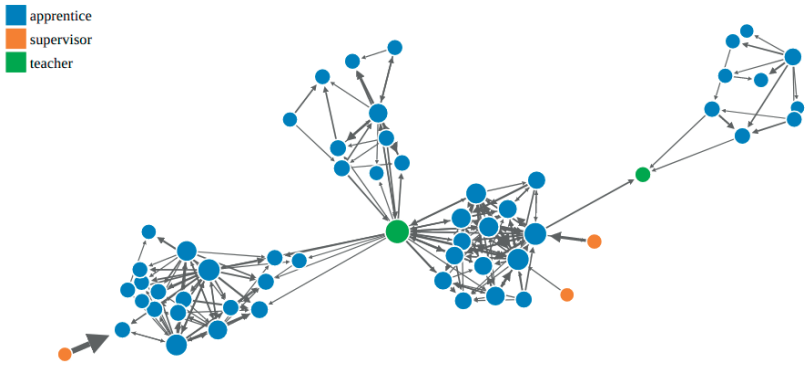


Figure 6.14 – *Example of Realto communication network for participants in one school*

This figure represents the interactions among Realto participants in florist profession in one vocational school.

The presented results reflect the structural constraints and the adopted policies by Realto users. In the current design, apprentices in Realto do not have the possibility to interact with each other unless they are added to the same flow (group) by a teacher or a supervisor, and materials shared within a flow, are only accessible by the members. Furthermore, teachers often prefer to create a separate flow in Realto for each of their classes that are composed of 8-15 apprentices. As a result social interactions among apprentices is limited to relatively small communities interconnected by the teacher (example shown in Figure 6.14). This could explain the high clustering coefficient of the network and high betweenness centrality of the teachers.

One possibility that might be taken into account in the future phases of Realto is to enable communication among participants in a certain profession by inviting them to a meta-flow comprising all the other users within the same profession. This in turn might expand such small communities and enable sharing of experiences among a broader set of users. However, considering the privacy and data confidentiality concerns, the detailed procedure needs to be determined in close collaboration with the stakeholders of vocational training and professional associations.

Table 6.10 – *Users distribution and reciprocity of Realto sub-networks.*

Sub-network type	Apprentices count (%*)	Teachers count (%)	Supervisors count (%)	Binary reciprocity
Apprentices	458 (80%)	0	0	0.44
Teachers-Apprentices	266 (47%)	54 (52%)	0	0.21
Supervisors-Apprentices	147 (26%)	0	68 (86%)	0.49
Teachers-Supervisors-Apprentices	53 (1%)	22 (21%)	15 (19%)	0.28

* proportion of the reported count to the corresponding node count in overall network presented in Table 6.9

Connections among different stakeholders

Next, to explore the connections among different stakeholders in Realto, we consider four sub-networks comprising connections among **(1)** apprentices, **(2)** teachers and apprentices **(3)** supervisors and apprentices, and **(4)** apprentices connected both to a teacher and a supervisor. According to the statistics reported in Table 6.10, from the initial set of apprentices in the network (569), 80% have social interactions with other apprentices in Realto, whereas only 47% are connected to a teacher. The ratio of apprentices in the other two sub-networks is even smaller. Only 26% of apprentices are connected to their supervisors and 1% of them are linked both to their teacher and supervisor in Realto.

Considering the reciprocity of connections, apprentices and supervisors-apprentices sub-networks show relatively high values for this measure. Reciprocity is defined as the ratio of bi-directional links to the total number of links in a directed network and hence reflects the mutuality of the connections. Therefore mutual relations between apprentices and their teacher account for only 21% of the existing connections. In particular 148 of apprentices have only outgoing connections to their teacher, suggesting that they do not receive any feedback from their teacher. On the other hand, reciprocal connections are more frequent among apprentices (44%) and also between apprentices and their supervisor (49%). This might also be related to the fact that teachers as shown in Figure 6.15 are generally associated with more (5-14) apprentices whereas according to Figure 6.16 most supervisors have only one or two apprentices in the platform and hence are more likely to follow their activities. Few supervisors located in the center of Figure 6.16 have a noticeably different connection pattern as they are linked to relatively larger number of apprentices (6 to 30). This difference is due to the fact that these cases represent supervisors in a single-track vocational system. Unlike dual-track system where supervisors generally have few apprentices in a workplace, in single-track system they supervise one or possibly more classes of apprentices at workplace-like workshops and therefore could be connected to a larger number of learners.

The binary reciprocity measure shows the mutuality of interactions but it does not reflect the degree of reciprocity between mutual dyads since it treats the network as unweighted. In order to analyze the strength of connections, we consider weighted in-degree and out-degree (average weight of incoming/outgoing connections) of LD-related interactions in supervisors-

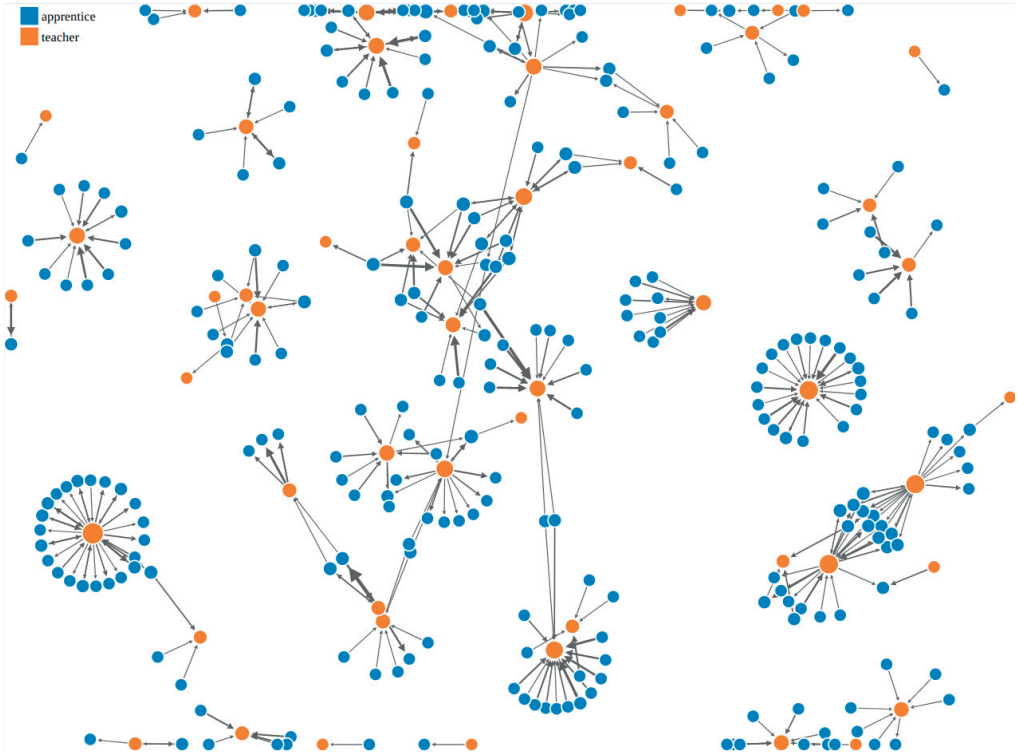


Figure 6.15 – *Teacher-apprentice sub-network: connections between apprentices and teachers*

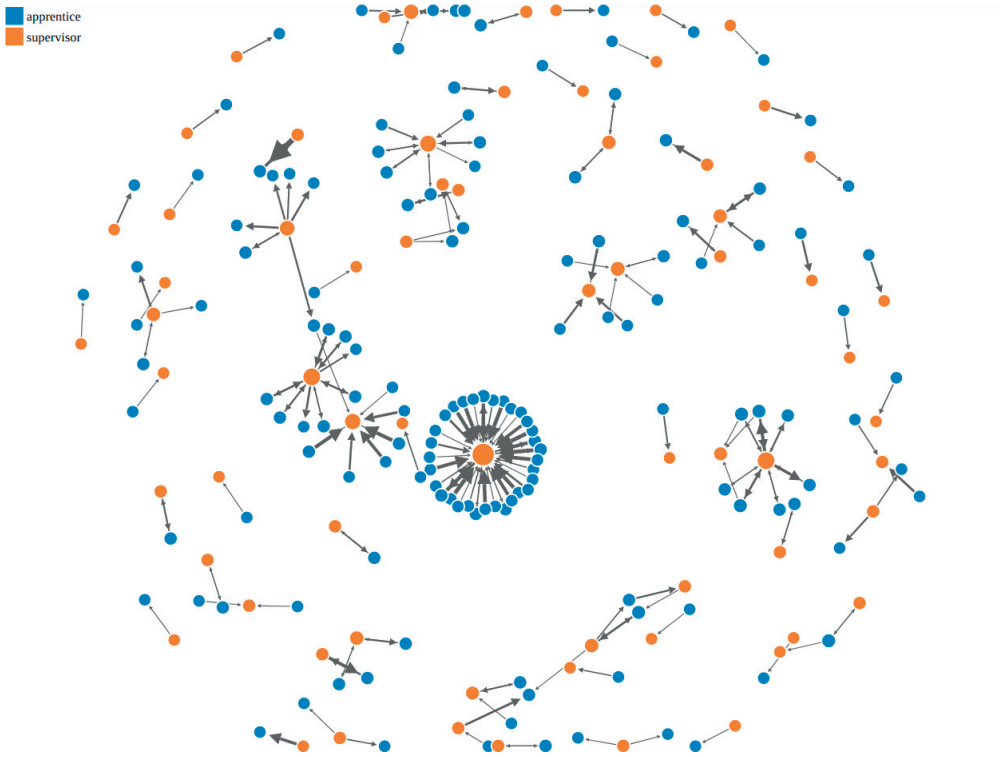


Figure 6.16 – *Supervisor-apprentice sub-network: connections between apprentices and supervisors*

apprentices sub-network. According to the results, supervisors on average provide feedback on only one third of the LDs created by their apprentices. More specifically, average weighted in-degree for supervisors is 7.8 (number of received documents) and their weighted out-degree is 2.5 (number of evaluated documents). Considering the importance of supervisor feedback on the practical training procedure, it is necessary to investigate the underlying reasons and issues faced by supervisors for providing Realto-mediated feedback to their apprentices and design mechanisms to improve this situation.

Structural roles

To investigate the macro structure of participants communication in Realto, we applied role modeling method described in Section 6.1.6. Interestingly, despite essential differences between Realto and MOOC discussion forums, a similar role structure was obtained for Realto communication network. This role model, as depicted in Figure 6.17, consists of an interconnected core group (**R1**) linked to two other peripheral groups which are attributed by only outgoing or incoming connection (**R2** and **R3** respectively). Such core-periphery role models, consisting of a dense cohesive core and sparse, loosely connected periphery, represent a typical structure that could be found in several communication networks.

Participants distribution and degree attributes of the resulting role model are summarized in Table 6.11. About half of Realto users (54%) fall into the core category, **R1**, who take part in the social and LD-related interactions more actively than the other two groups, as perceived by their connection patterns and higher average degrees.

Regarding the second role, **R2**, the absence of incoming connections suggests that apprentices in this group ($N = 160$) do not receive any reaction on their posts from other members or any feedback on their learning journals. Teachers ($N = 33$) and supervisors ($N = 20$) in this role, do not receive any LD from their apprentices and hence their outgoing connections represent their participation in social interactions (writing comments or rating posts).

Concerning the third role, **R3**, as perceived from the absence of outgoing connections, apprentice in this role ($N = 130$) do not create any LD (this action would be modeled as an outgoing connection from apprentice to his/her supervisor) and consequently their connections with the other members is limited to social interactions. This group of apprentices receive comments or likes on their posts from few other participants in R1 and R2 (2.3 on average), but do not react to posts created by others. Teachers ($N = 28$) and supervisors ($N = 18$) in this role do not provide any type of feedback to their apprentices (including comments/likes on their posts or feedback on their LDs).

In summary, analysis of the structural roles of different stakeholders in Realto provided interesting insights about their communication patterns. According to the presented results, the connection between apprentices and their teachers or supervisors needs to be better supported. In addition, the implemented SNA tools in the analytics dashboard of Realto

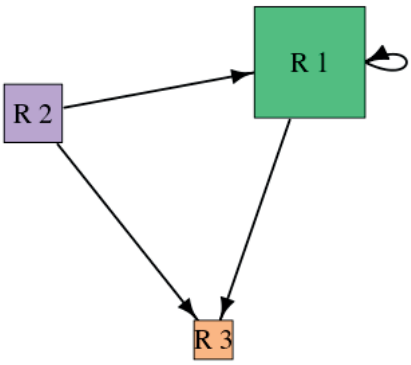


Figure 6.17 – *Role structure of Realto network*

Table 6.11 – *Attributes of structural roles in Realto network*

Attribute	Structural role		
	R 1	**R 2**	**R 3**
Average In-degree	3.8	0	2.3
Average out-degree	3.6	1.7	0
Members count (%*)	409 (54%)	213 (28%)	130 (17%)
Apprentices count(%)	325 (57%)	160 (28%)	84 (15%)
Teachers count (%)	43 (41%)	33 (32%)	28 (27%)
Supervisors count (%)	41 (52%)	20 (25%)	18 (28%)

* proportion of the reported count to the corresponding node count in overall network presented in Table 6.9

enable to identify participants who might require further support and could provide the basis for planning interventions targeting particular user groups.

6.3 Discussion

In this chapter we investigated the social aspect of participants online interactions in MOOC and Realto platforms. Considering MOOC discussion forums, by incorporating different analytic methods we investigated learners knowledge exchange communication in content-related discussions in two Coursera MOOCs. We found similar patterns related to forum participation patterns in both courses, despite their different content, learners population and forum communication volume. In contrast to most existing studies on MOOC discussion forums, in our analyses we explicitly took into account the time dimension and explored the evolution of forum participation and communication structure with respect to the course structure. Our analysis of the interplay of time, content, and social dimensions provided insights regarding the research questions formulated in Section 6.1.2.

In the time dimension, in Section 6.1.4 we investigated how the structure of the course (video release and assignment deadlines) influences the overall forum activity (**Question 1**). We

could observe an increase of the number of posts before deadlines and after video release days, and thus conclude that course events have an impact on the forum communication.

Surprisingly, based on the content analysis of the posts over time (**Question 2**), there was no clear coupling between the course structure and quantity of content-related discussions as reported in Section 6.1.5. However, mentions of some specific concepts regarding the course subject tend to increase after certain video releases indicating that some discussion topics are introduced by the course while others are brought into discussion by the participants themselves. Identification of the main topics of discussion at different periods of the course, could in turn help to identify the challenging concepts in the course lectures that require further clarification or on which support materials should be provided to the learners.

The temporal dynamic of the social structure emerging from the forum communication with respect to course structure (**Question 3**) was analyzed in Section 6.1.6. Here we could show that the global organization (network characteristics) of the communication network is independent of the course structure. One reason could be the absence of a sustainable forum community and the high fluctuation of the active contributors. Consequently, there is no inherent self-organization of the network, which would require coordination and maintenance of social relations. This further supports the claim of Gillani and Eynon [84] that MOOC forums resemble decentralized crowd behaviour rather than a social community.

A more user centric perspective on the integrated analysis of the time and the social dimension was taken in Section 6.1.6 to answer research question about how roles of forum participants evolve over the course duration (**Question 4**). In order to model the structural role of users according to their connection patterns in the communication network, we applied a blockmodeling approach similar to [102] and [116]. Interestingly, the role structure of the information exchange network comprising a small cohesive core of active contributors, and peripheral help-giving or help-seeking users that has also been reported for static snapshots of the network in [102] and [116], persist over several time slices. While the overall structural organization of the networks is stable it could be shown that the association of users to roles changes drastically over time. Only a small subset of the most active (active core) users retain an active role over time, and majority of learners are active in only one or two time slice (occasional forum participants). It could be seen that the fluctuation of active users is so pervasive such that in each time slice even the majority of users are “newcomers” in the sense that they form connections to other users for the first time. This could be considered as a major obstacle for the emergence of a sustainable community and further explains the irregularities in the overall network structure mentioned above.

The research question about the relation between students roles in the communication network and discussion topics (**Question 5**) was addressed in Section 6.1.7. Towards this goal, the previously mentioned results of the structural roles over time (Section 6.1.6) were combined with the content dimension using the coding of discussion posts (Section 6.1.5). Results showed that even if occasional help-givers and help-seekers (peripheral users) are generally

not as important for the structural cohesion of the communication network as the core users, they make fewer but important contributions indicated by their high rates of discussions on domain-specific concepts and content-related posts. Especially the participants in the group of occasional help-seekers show similar or even higher ratio of information requests related to course content or mentions of concrete specific course concepts in their posts. Consequently those users could often be notable as initiators of discussions even if their activity is limited.

The question about the predictability of forum activity (**Question 6**) was answered in Section 6.1.8. It could be shown that the forum activity level is predictable one week in advance given the course structure (video and assignment dates), history of forum activity, and the described network features. Further, the predictive models could be considered as a building block for teaching support tools to forecast periods when increased tutor support for forum discussions is needed.

In summary, the majority of research works focus on single aspects of MOOC discussion forums and point to the conclusion that the current implementation of discussion forums are only used intensively by a small amount of course participants, and further, only a subset of the discussions are relevant for the information exchange. However, our integrated analysis of the interplay between different aspects (time, content and social interactions), provides new insights on the complex structure of forum communication. There are several interdependencies between the progress of the course, the contributed content and participants structural roles in the communication that have to be taken into account to get a clearer picture and to foster the development of future collaboration support, tailored to the characteristics of different user groups. Recommendation mechanisms to find the right information and adequate discussion partners [189, 242] could be one initial step, but in order to transform the loosely connected "crowd" of forum users into a sustainable community in the sense of social learning requires also support for maintaining social contacts. The characteristics of different user roles should be considered in the design of such support mechanisms. Furthermore, combination of predictive models proposed in Section 6.1.8 with content analysis of the forum contributions could potentially support instructors to turn their attention to upcoming important discussions and enable interventions and community management.

Our study on MOOC forum dynamics has certain inherent limitations in that it was based on two MOOC courses. Despite differences in timing of lecture and assignment dates and different forum participation levels, both courses were based on a weekly organization of course materials and included several assessments tasks. Therefore in future work we plan to test the generalizability of our findings on courses with different structures and possibly in different domains. Furthermore, we filtered out discussions irrelevant to the course content based on sub-forum types and considered assignment and lecture sub-forums in our analysis. However as misplaced posts are still possible, selection of content-related threads could be refined using the linguistic models proposed in [238]. One possible extension of our study is to explore engagement of the identified forum user clusters in other aspects of the course such as their access patterns to the lectures and other course materials to identify if

changes in forum roles are coupled with changes in study behaviours. In addition, it would be interesting to explore to what extent forum structural roles adopted by learners is influenced by their internal characteristics, motivation type (e.g. amotivation, extrinsic and intrinsic motivations as described by self-determination theory [62, 51]), or their academic level and cultural characteristics. In our study we explored the effect of course structure on the evolution of forum communication. It would be valuable to identify other external factors which could possibly influence learners' forum participation level or emergence of different user roles in the communication network. Instructor's facilitations, course evaluation criteria and educational setting (formal or open online settings) could be examples of such external factors.

Finally, in Section 6.2 we demonstrated how social network analysis concepts and techniques could be adopted to model and analyze the connections among apprentices, teachers and supervisors in Realto platform (**Question 7**). Application of network analysis methods similar to those used in our MOOC study, enabled us to gain insight on the attributes and structure of Realto communication network both at the global and individual level. In summary, our analysis showed the network of participants in Realto is decomposable into small densely connected communities which in the most common case, represent different classes of apprentices, possibly interconnected by their teacher. Moreover, a considerable proportion of apprentices are not connected to their teacher or supervisor in Realto communication network. Considering that connecting different stakeholders in vocational training is one of the goals in Realto, it is essential to investigate the reasons and obstacles faced by Realto users in this direction. According to the role model analysis, the communication structure in Realto is shaped as a similar core-periphery role model that we observed in MOOC forum communication. Investigation of structural roles in association with participants actual roles (apprentice, teacher, supervisor) revealed different engagement profiles which should be taken into account for planning targeted interventions. Furthermore, several evidences suggest the limited participation of a teachers and supervisors in Realto-mediated communication with their apprentices and also limited feedback provided on their learning journals. Mechanisms to bring the new posts and un-evaluated LDs to the attention of teachers and supervisors, and facilitating the evaluation procedure could possibly improve this situation.

A future extension of our social interaction analyses in Realto, is to incorporate the time aspect to study *how participants' communication and roles evolve over time.* Furthermore, the content of participants' contributions (posts, comments, feedback) is another important aspect which needs to be considered in the future analyses.

7 General discussion

7.1 Summary

In this thesis, we investigated patterns of learners' participation in two different online learning environments, MOOCs and Realto. We presented several techniques to analyze activity sequences and introduced methods to transform low-level interaction logs into interpretable representations. The presented methods, provided indicators and models which could facilitate quantifying, categorizing, and temporal analysis of learning behaviours.

As outlined in the introduction, we analyzed educational data across three principle dimensions: time, activity, and social. This thesis contributed to the learning analytics and educational data mining research by providing analytic methods, novel indicators of learning behaviours, and new insights on learners' interaction patterns from the temporal, activity, and social perspectives. Furthermore, by investigating the dynamics of learners' participation over time, this thesis extended the body of research on temporal analytics, an under-explored aspect of educational data which could provide valuable insights on the learning processes.

In summary, we introduced quantitative methods to study the temporal distribution of actions performed by the learners (**time dimension**). Moreover, we presented hypothesis-driven and data-driven methods to analyze and model the type and sequence of actions performed by learners, in order to discover their study patterns and longitudinal engagement profiles (**activity dimension**). We also studied social interactions among learners and presented an integrated study on the dynamics of MOOC discussion forums, including the evolution of learners' roles in the knowledge-exchange communication (**social dimension**). Finally, we adapted the presented methods to the context of Realto platform and implemented an interactive dashboard to enable investigation of temporal and activity patterns in participants' interactions and the social aspect of their communication in this platform.

7.2 Contributions

This section summarizes the contributions of this dissertation towards our two primary research objectives regarding learning analytics: **(1)** providing methods for analyzing learners' interaction patterns, and **(2)** providing analytic tools for Realto platform.

7.2.1 Methods for interaction pattern analysis

We introduced novel methods to analyze learners' interaction sequences in order to discover their behavioral patterns in online learning environments. An important property of our proposed methods is their generalizability to different contexts and learning platforms, as they require only the time-stamped activity sequences as input. Moreover, the introduced methods provide quantitative metrics and interpretable models of individuals' behaviours, which could support personalization of the learning environments and improve awareness of educational stakeholders about the learning processes. A summary of the proposed methods in time and activity dimensions is presented in the following.

Temporal pattern analysis

We proposed novel methods for modeling and quantifying temporal patterns in online interactions. In particular, we proposed different measures to quantify regularity in terms of detecting repeating patterns in timing of users' activities. These measures can be computed at any temporal granularity (e.g. daily, weekly, or longer intervals). Our proposed measures can be categorized as: **(1)** entropy-based measure, **(2)** profile similarity measures, and **(3)** frequency domain measures. Binarization of user's activity sequences to represent when the user was active/inactive during a certain period is the initial step for all the following computations.

- **Entropy-based measures**: start by constructing the daily(weekly) histogram, which represents for how many days(weeks) the user was active at a certain hour(day). Based on the entropy of the resulting histograms, this category of measures can detect whether user's activities are concentrated around a particular hour(day).
- **Profile similarity measures**: build weekly profiles to represent for how many hours the user was online on a certain day in each week. By comparing activity profiles of different weeks (in their binary, normalized, or raw formats), this category of measures can detect whether the user follows a certain weekly time schedule.
- **Frequency-domain measures**: transform user's activity time signal to its frequency domain representation (periodogram) using Fourier transform. A periodic hourly/daily pattern in user's activity signal, would appear as a spike in the periodogram at the corresponding frequency. Frequency-based measures are designed to capture such patterns.

Activity pattern analysis

We introduced methods to model learners' activity sequences and track the evolution of their study approaches over time. We proposed a processing pipeline which receives as input the sequences of actions performed by the learners and models them as transition probabilities between different action types. The common study patterns among learners at each time period are then extracted by applying a clustering algorithm on the modeled sequences. Next, based on a cluster similarity measure, the matching clusters in different time periods are identified to enable tracking learners' study patterns over time and detecting changes in their approaches. The proposed pipeline does not require any manual parameter tunning and allows for discovering behavioral patterns from users' interaction logs in an unsupervised manner.

7.2.2 Analytic tools for Realto platform

To provide analytic tools for Realto platform, we implemented two different dashboards, briefly described in the following: an integrated awareness dashboard for teachers, and an analytics dashboard for researchers.

Awareness dashboard for teachers

We developed an interactive dashboard to support teachers by raising awareness of their students' activities in Realto. The dashboard aggregates several indicators about learners both at the individual and at class level. This includes type, quantity, and quality of the produced artifacts, and the timeline of learners' activities. The dashboard provides information about learners activity status in comparison with the other class members. Moreover, it provides the possibility for the teacher to send direct feedback to the selected learners who might need further attention.

Analytics dashboard for researchers

Unlike the awareness dashboard which presents to the teachers information about what their students do in Realto, the analytics dashboard provides researchers with higher level indicators to describe platform usage patterns by different Realto user groups (e.g. users in different professions, roles, schools, etc.). The analytics dashboard comprises three main components: **(1)** time analysis, **(2)** activity analysis, and **(3)** network analysis modules. The **time analysis module** includes a subset of our proposed methods in time dimension and enables to study the temporal patterns in users' activities in Realto. The **activity analysis module**, includes visualization of users' activity sequences and provides the possibility to cluster users based on their platform usage patterns. The **network analysis module** provides an interactive visualization of the communication among Realto users, presents the network attributes and its underlying role structure. It also provides functionalities to study the connections among

users with different roles (apprentices, teachers, and supervisors) to assess if Realto has been able to connect different stakeholders in the Swiss VET system.

7.3 MOOC related findings

Using a combination of our proposed methods and other existing analytic methods, we obtained interesting insights into learners' engagement patterns in MOOCs. Our main findings in this context could be summarized as follows:

- **Positive links between time regularity and performance**: We found positive correlations (up to 0.7) between our proposed time regularity measures and students' final grade. We showed that time regularity measures could capture above 50% of the variability in the final grade and that weekly regularity is related to performance mainly for the learners with limited amount of study time. We also found that clusters of learners who are regular in a daily or weekly basis and do not postpone accessing the course materials, achieve significantly higher grade in comparison with the non-regular learners. Moreover, we showed that students who pass the course, display higher regularity level throughout the course duration, compared to the failing students. These findings suggest that despite the time flexibility offered by MOOCs, learners who planned their learning activities in a regular manner had better chance to succeed. The regularity measures could provide a basis for intervention, for instance by providing guidelines to students on how to work more productively or by indicating to teachers when they should intervene with greatest effect to support their students.

- **Different study patterns among learners**: Through modeling and clustering interaction sequences during MOOC assessment periods, we could identify different study patterns among learners. We showed that some learners dedicate only one day to work on the materials of a particular week, for instance, they watch the videos and submit to the assignments on the same day. Some other learners work on multiple consecutive days, for instance they watch videos on one day and submit to the assignment on the day after. Some other learners have several inactive days during their learning sequence. In the most common case, the learners watch videos of a week, and after few inactive days, they submit to the assignment. The extracted patterns provide insights about learners' study strategies and could be useful for improving the adaptivity and personalization of the learning environments.

- **Evolution of study approaches over time**: By clustering sequences of study approaches over time, we identified different longitudinal engagement profiles among learners. We showed that nearly half of the learners follow a fixed approach in all of the assessment periods. The majority of such learners follow the typical approach of watching the videos before submitting the assignments. On the contrary small proportion of learners always skip the videos before submissions, and another group audit the course lectures without making any submission. The other half of the students, change their study

approach over time. We showed that some learners temporally change their study approach during few periods and then continue with their initial approach, whereas others permanently switch to a new study approach. Changes in the study approaches, might be indicator of facing difficulties in following the course and detecting them could enable to provide personalized support to the learners.

- **Discussion forum dynamics**: Our analysis revealed several interdependencies between the structure of the course (video release dates and assignment deadlines), the forum activity level, and the produced content. We showed that the forum activity level could be predicted one week in advance, using a combination of previous forum activity features, and structural features (e.g. passed ratio of the course, or number of days left to assignment deadline, etc.). Considering the discussion topics, our analyses revealed that some domain-specific topics are introduced to the discussions by particular course videos, whereas some other topics are brought into the discussion by the learners who might have the prior knowledge of certain course topics. Identifying the main discussion topics at different course periods and prediction of forum activity level could help to determine the challenging concepts and intense discussion periods when increased tutor support is needed.

- **Evolution of learners' role in discussion forum**: We showed that the structure of knowledge exchange network, modeling learners' forum communication, persists over time and presents a core-periphery structure (a core of active contributors, and two peripheral help-giving or help-seeking roles). However, the association of users to roles changes drastically over time such that in each time period, the majority of users are newcomers. This in turn implies the absence of persistent discussion communities among learners. We identified three cluster of learners based on their role sequences over the course duration. A small group of learners who retain an active role over time, and two groups of occasional forum participants who are active only during limited time periods: occasional help-givers, and occasional help-seekers. Analysis of the content produced by these learner groups, revealed that occasional forum participants, despite having limited forum participation, could have an important impact on the discourse by triggering focused discussions on specific subject areas.

7.4 Reflections

In this thesis we targeted very relevant challenges in the emerging field of Learning Analytics (LA), especially by addressing issues related to time-dependent interaction patterns in different online learning environments. The focus on temporality and patterns of change in learners' activities over time was one of the key contribution of this thesis. In all the presented studies, we explicitly took into account the temporal dimensions of the data and studied temporal and sequential dynamics of learners' activity/interaction patterns and roles. Considering the importance of this rather under-researched aspect of educational data, our sustained

attention on time dimension is a unique, valuable, an much needed contribution to the field of learning analytic.

Exploring three different dimensions of learning analytics has been one of the other key contributions of this thesis. Most of the existing LA research focus only on one aspect of the educational data. Whereas in this thesis we investigated the temporal, activity, and social dimensions and this enabled us to gain a more comprehensive view of the learning behaviours. In this direction, our work built upon and contributed to the advancement of computational methods and approaches used in LA research to analyze the data emerging from online learning environments. The proposed methods in this thesis could be taken up by other researchers in the LA community and could be applied to various problems of practice.

Our analyses across time, activity, and social dimensions showed that learners have different time schedules, exhibit different study approaches and access patterns to the learning material, and undertake different roles in the social interactions. Such individual differences should be taken into consideration for adapting learning environments to the attributes and requirements of learners. Furthermore, by integrating temporal analytics into the activity and social dimensions, we showed different engagement profiles among learners. Our analyses showed that a large ratio of learners change their approaches over time and therefore it is insufficient to represent learning behaviours with static snapshots or aggregated features which neglect the temporal aspects. This in turn confirms the importance of temporal dynamic for understanding the learning processes.

An important point which needs to be further clarified here is the question on causality. In this thesis we presented several results showing evidences of relations between learners' activity patterns (e.g. temporal regularity, study patterns, and social roles) and their performance. However it should be noted that the discovered relations/correlations do not necessarily imply causality. Determining casual relations needs to be addressed methodologically through controlled studies, which was not the focus of this thesis.

One of the main challenges in this thesis was to provide generalizable analytic methods which are in particular relevant for MOOCs and Realto. This was not trivial given the differences and specificities of these two learning environments. This approach however has both advantages and disadvantages. On one hand, it is important to have platform-independent analytics as sciences aims at abstraction and because this allows integrating analytics from multiple platforms. Moreover, some level of abstraction is essential for understanding and describing the underlying learning processes from the interaction data. On the other hand, one could argue that learning is always specific to some contents and platforms which leads to platform-specific analytics. For instance, video lectures are central components of MOOCs and as shown by Li et al. [135] video interaction patterns (e.g. pause, forward/backward seek, and speed change) could tell about learners' challenges and their performance in the course. Similarly, learner-created artefacts are of particular importance in Realto. Analysis of these artefacts and the structure of the actor-artefact networks are among the other important aspects which

should be considered in the analytics module of Realto. The action sequence analyses in Realto could be complemented by considering the attributes of the created artefacts, such as the content of the uploaded images or semantic richness of the textual descriptions. In this thesis, we made one step towards platform-independent analytics, but the future is probably a mix of both.

7.5 Limitations and future work

The presented analytic methods in this thesis enabled us to model and describe interaction traces of thousands of students and shed light on the evolution of students' learning behaviours in MOOCs. However, our analyses were post-hoc in nature and limited to the interaction data available to us. Based on the data-driven insights we could provide certain explanations about the observed learning behaviours. Nevertheless, the lack of data about learners' background, personal constraints or attributes, and their experiences during the course imposed a limitation on our attempt to identify the underlying factors (internal or external to students) which could influence learners' behaviours, such as time regularity, choice of study approaches, or changes in learning behaviours.

The few number of MOOC courses used in our analyses could put limits to the generalizability of our findings in this context. The datasets used in our studies were limited to structured MOOCs (with predefined schedule and deadlines), in computer science. It would be interesting to investigate if similar trends could be found in other courses in different domains and with different structures. An examples could be to study time regularity of learners in self-paced MOOCs (where all the course materials are available upon registration and the assignments do not have predefined start or due dates), or to assess whether similar study patterns and learning trajectories could be observed among learners in different courses. Another possible extension of our analyses could be to see whether study patterns are kept for the same learners studying completely different subjects on the same/other online platforms.

Regarding the presented activity analysis methods, we did not take into account the duration of learning activities and the time elapsed between the performed actions. One reason was that in our MOOC data, although a granular record of video interactions was available, submissions were the only actions recorded with respect to the assignments. Therefore it was not possible to infer the amount of time spent on the assessment tasks. The time invested on different course materials, could be an indicator of the required effort by the learners to perform particular learning tasks. Consideration of this aspect could be a possible extension to our approach. Moreover, certain actions or learning materials in some learning platforms, might be of higher pedagogical importance. Our presented approach could be modified to associate higher weights to certain actions (or transition between actions) to reflect their importance.

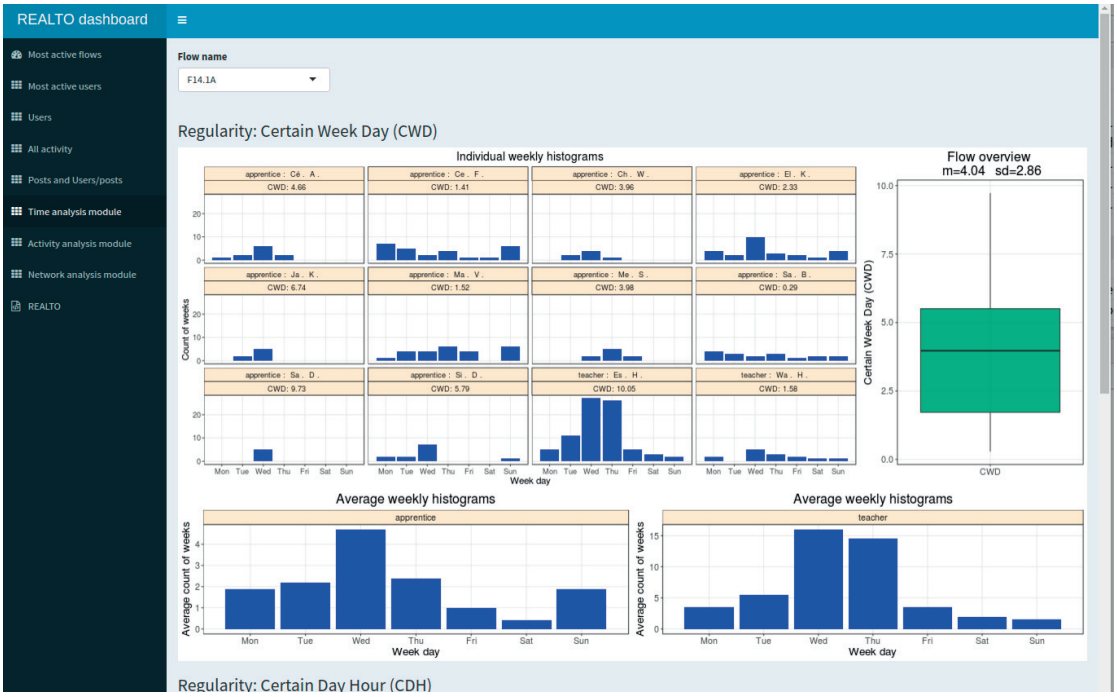
One limitation of our work concerns the analytics tools for Realto platform and in particular the fact that teachers were not involved in the design process of the awareness dashboard for this platform. As mentioned before, it is necessary to perform an evaluation of the presented

solution with the target users and improve the current version based on teachers' feedback to enhance its usability and impact on the practice. An important future direction is to extend this work into analytics for action, by identifying patterns of student activity that should be flagged and brought to the attention of teachers so that they can take timely actions to alert and support the specific students.

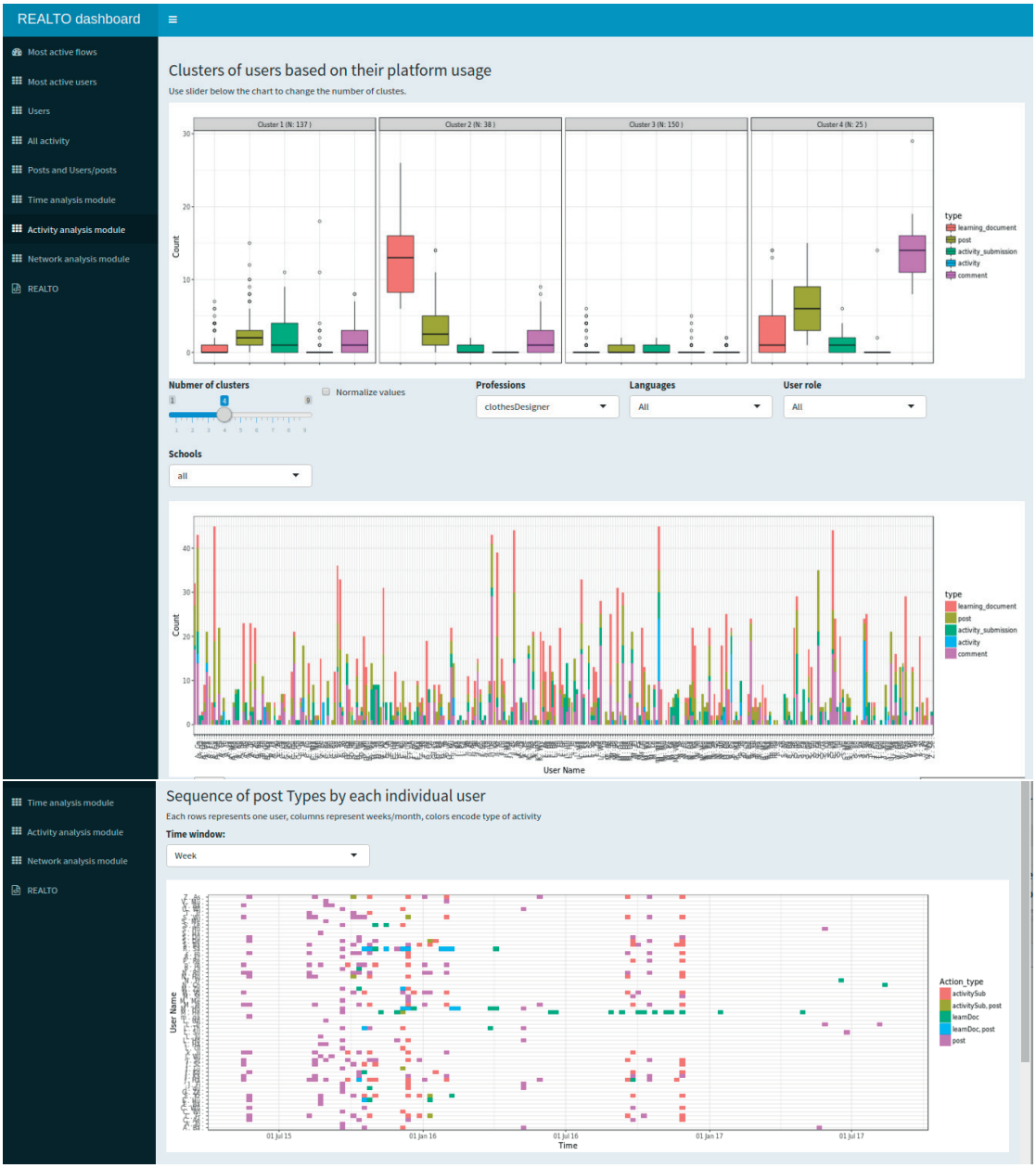
An important future direction for the presented research in this dissertation, is to put into practice the described methods and derived indicators, in order to guide the design of appropriate and timely interventions which could support the learning experiences. This is a necessary step to effectively close the learning analytics loop, as Clow [54] refers to it. The outcome of our described analytic methods could be visualized and integrated into dashboards, to raise awareness of different educational stakeholders about the learning processes. Moreover, our proposed methods generate objective metrics which could be useful for personalization of learning environments. Integration of these methods into the learning platforms, could open up possibilities for providing in-time feedback to the stakeholders and improving the adaptivity and personalization within educational environments.

A Realto analytics dashboard

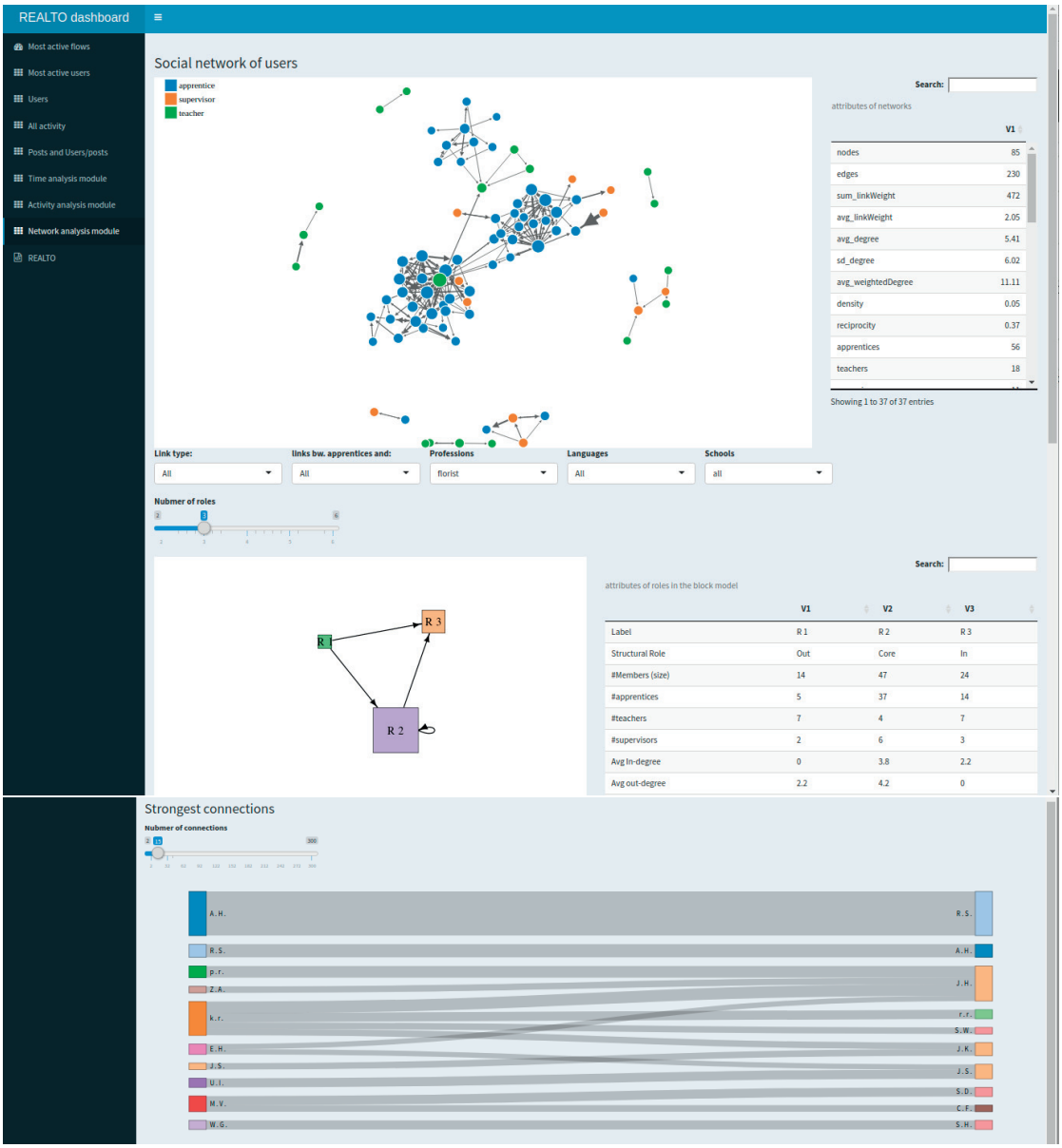
A.1 Time analysis module



A.2 Activity analysis module



A.3 Social analysis module



Bibliography

[1] Andrew Abbott and Angela Tsay. "Sequence analysis and optimal matching methods in sociology: Review and prospect". *Sociological methods & research*, 29(1):3–33, 2000.

[2] Ani Aghababyan, Nicholas Lewkow, and Ryan S Baker. "Exploring the asymmetry of metacognition.". In *LAK*, pages 115–119, 2017.

[3] Rakesh Agrawal and Ramakrishnan Srikant. "Mining sequential patterns". In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 3–14. IEEE, 1995.

[4] Rakesh Agrawal, Ramakrishnan Srikant, et al. "Fast algorithms for mining association rules". In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[5] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, A Inkeri Verkamo, et al. "Fast discovery of association rules.". *Advances in knowledge discovery and data mining*, 12(1):307–328, 1996.

[6] S Aher and LMRJ Lobo. "Mining association rule in classified data for course recommender system in e-learning". *International Journal of Computer Applications*, 39(7): 1–7, 2012.

[7] Shibbir Ahmed, Rajshakhar Paul, and Abu Sayed Md Latiful Hoque. "Knowledge discovery from academic data using association rule mining". In *Computer and Information Technology (ICCIT), 2014 17th International Conference on*, pages 314–319. IEEE, 2014.

[8] Nout M Alhajraf and Aishah M Alasfour. "The impact of demographic and academic characteristics on academic performance". *International Business Research*, 7(4):92–100, 2014.

[9] Saleema Amershi and Cristina Conati. "Combining unsupervised and supervised classification to build user models for exploratory". *JEDM-Journal of Educational Data Mining*, 1(1):18–71, 2009.

[10] Vivek Anand. "A study of time management: The correlation between video game usage and academic performance markers". *CyberPsychology & Behavior*, 10(4):552–559, 2007.

[11] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. "Engaging with Massive Online Courses". In *Proceedings of the 23rd International Conference on World Wide Web*, WWW'14, 2014.

[12] Anton Andrejko, Michal Barla, Mária Bieliková, and Michal Tvarozek. "User Characteristics Acquisition from Logs with Semantics.". *ISIM*, 7:103–110, 2007.

[13] Cláudia Antunes. "Acquiring background knowledge for intelligent tutoring systems". In *Educational Data Mining 2008*, pages 18–27, 2008.

[14] Jaime Arguello and Kyle Shaffer. "Predicting Speech Acts in MOOC Forum Posts". In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, ICWSM'15, 2015.

[15] Ivon Arroyo and Beverly Park Woolf. "Inferring learning and attitudes from a Bayesian Network of log file data.". In *AIED*, pages 33–40, 2005.

[16] Hafidh Ba-Omar, Ilias Petrounias, and Fahad Anwar. "A framework for using web usage mining to personalise e-learning". In *Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on*, pages 937–938. IEEE, 2007.

[17] Ryan Shaun Baker and Paul Salvador Inventado. "Educational data mining and learning analytics". In *Learning analytics*, pages 61–75. Springer, 2014.

[18] Maria Bannert, Peter Reimann, and Christoph Sonnenberg. "Process mining techniques for analysing patterns and strategies in students' self-regulated learning". *Metacognition and learning*, 9(2):161–185, 2014.

[19] E Barbera, B Gros, and PA Kirschner. "Paradox of time in research on educational technology". *Time & Society*, 24(1):96–108, 2015.

[20] Elena Barbera and Marc Clarà. "Time in e-Learning Research: A Qualitative Review of the Empirical Consideration of Time in Research into e-learning". *ISRN Education*, 2012, 2012.

[21] Elena Barberà and Marc Clarà. "The Temporal Dimensions of E-Learning". *E-Learning and Digital Media*, 11(2):105–107, 2014.

[22] Lucy Barnard, William Y Lan, Yen M To, Valerie Osland Paton, and Shu-Ling Lai. "Measuring self-regulation in online and blended learning environments". *The Internet and Higher Education*, 12(1):1–6, 2009.

[23] Tiffany Barnes. "The q-matrix method: Mining student response data for knowledge". In *American Association for Artificial Intelligence 2005 Educational Data Mining Workshop*, pages 1–8, 2005.

[24] Murray R Barrick and Michael K Mount. "The big five personality dimensions and job performance: a meta-analysis". *Personnel psychology*, 44(1):1–26, 1991.

[25] Jaroslav Bayer, Hana Bydzovská, Jan Géryk, Tomás Obsivac, and Lubomir Popelinsky. "Predicting Drop-Out from Social Behaviour of Students.". *International Educational Data Mining Society*, 2012.

[26] Carole R Beal, Lei Qu, and Hyokyeong Lee. "Classifying learner engagement through integration of multiple data sources". In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 151. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[27] Sanam Shirazi Beheshitha, Dragan Gašević, and Marek Hatala. "A process mining approach to linking the study of aptitude and event facets of self-regulated learning". In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 265–269. ACM, 2015.

[28] Yoav Bergner, Zhan Shu, and Alina von Davier. "Visualization and confirmatory clustering of sequence data from a simulation-based assessment task". In *Educational Data Mining 2014*, 2014.

[29] Clancy Blair and Adele Diamond. "Biological processes in prevention and intervention: The promotion of self-regulation as a means of preventing school failure". *Development and psychopathology*, 20(03):899–911, 2008.

[30] Michael J Bond and NT Feather. "Some correlates of structure and purpose in the use of time.". *Journal of Personality and Social Psychology*, 55(2):321–329, 1988.

[31] Stephen P Borgatti and Martin G Everett. "Two algorithms for computing regular equivalence". *Social networks*, 15(4):361–376, 1993.

[32] Suchita Borkar and K Rajeswari. "Predicting students academic performance using education data mining". *International Journal of Computer Science and Mobile Computing (IJCSMC)*, 2(7):273–279, 2013.

[33] François Bouchet, Roger Azevedo, John S Kinnebrew, and Gautam Biswas. "Identifying Students' Characteristic Learning Behaviors in an Intelligent Tutoring System Fostering Self-Regulated Learning.". *International Educational Data Mining Society*, 2012.

[34] Nabila Bousbia and Idriss Belamri. "Which Contribution Does EDM Provide to Computer-Based Learning Environments?". In *Educational data mining*, pages 3–28. Springer, 2014.

[35] E Oran Brigham and RE Morrow. "The fast Fourier transform". *IEEE spectrum*, 4(12): 63–70, 1967.

[36] Christopher G Brinton and Mung Chiang. "Mooc performance prediction via clickstream data and social learning networks". In *Computer Communications (INFOCOM), 2015 IEEE Conference on*, pages 2299–2307. IEEE, 2015.

[37] Christopher G Brinton, Mung Chiang, Shaili Jain, Henry Lam, Zhenming Liu, and Felix Ming Fai Wong. "Learning about social learning in MOOCs: From statistical analysis to generative model". *IEEE transactions on Learning Technologies*, 7(4):346–359, 2014.

[38] Bruce K Britton and Abraham Tesser. "Effects of time-management practices on college grades.". *Journal of educational psychology*, 83(3):405–410, 1991.

[39] Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer Science & Business Media, 2013.

[40] Piotr Bródka, Stanisław Saganowski, and Przemysław Kazienko. "GED: the method for group evolution discovery in social networks". *Social Network Analysis and Mining*, 3(1): 1–14, 2013.

[41] Christopher Brooks, Graham Erickson, Jim Greer, and Carl Gutwin. "Modelling and quantifying the behaviours of students in lecture capture environments". *Computers & Education*, 75:282–292, 2014.

[42] Malcolm Brown. "Learning analytics: Moving from concept to practice". *EDUCAUSE Learning Initiative*, pages 1–5, 2012.

[43] Tadeusz Caliński and Jerzy Harabasz. "A dendrite method for cluster analysis". *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[44] Karina L Cela, Miguel Ángel Sicilia, and Salvador Sánchez. "Social network analysis in e-learning environments: A Preliminary systematic review". *Educational Psychology Review*, 27(1):219–246, 2015.

[45] Hao Cen, Kenneth Koedinger, and Brian Junker. "Learning factors analysis-a general method for cognitive model evaluation and improvement". In *Intelligent tutoring systems*, volume 4053, pages 164–175. Springer, 2006.

[46] Sung-Hyuk Cha. "Comprehensive survey on distance/similarity measures between probability density functions". *International Journal of Mathematical Models and Methods in Applied Science*, 1(4):300–307, 2007.

[47] Tomas Chamorro-Premuzic and Adrian Furnham. "Personality, intelligence and approaches to learning as predictors of academic performance". *Personality and individual differences*, 44(7):1596–1603, 2008.

[48] Sven Charleer, Joris Klerkx, Erik Duval, Tinne De Laet, and Katrien Verbert. "Creating effective learning analytics dashboards: Lessons learnt". In *European Conference on Technology Enhanced Learning*, pages 42–56. Springer, 2016.

[49] Mohamed Amine Chatti, Anna Lea Dyckhoff, Ulrik Schroeder, and Hendrik Thüs. "A reference model for learning analytics". *International Journal of Technology Enhanced Learning*, 4(5-6):318–331, 2012.

[50] Bodong Chen, Alyssa F Wise, Simon Knight, and Britte Haugan Cheng. "Putting temporal analytics into practice: the 5th international workshop on temporality in learning data". In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 488–489. ACM, 2016.

[51] Kuan-Chung Chen and Syh-Jong Jang. "Motivation in online learning: Testing a model of self-determination theory". *Computers in Human Behavior*, 26(4):741–752, 2010.

[52] Zuoliang Chen and Shigeyoshi Watanabe. "A case study of applying SNA to analyze CSCL social network". In *Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on*, pages 18–20. IEEE, 2007.

[53] Brigitte JC Claessens, Wendelien Van Eerde, Christel G Rutte, and Robert A Roe. "A review of the time management literature". *Personnel review*, 36(2):255–276, 2007.

[54] Doug Clow. "The learning analytics cycle: closing the loop effectively". In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 134–138. ACM, 2012.

[55] Albert T Corbett and John R Anderson. "Knowledge tracing: Modeling the acquisition of procedural knowledge". *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[56] Marcus Credé and Nathan R Kuncel. "Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance". *Perspectives on Psychological Science*, 3 (6):425–453, 2008.

[57] Ryan SJ d Baker, Albert T Corbett, and Vincent Aleven. "More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing". In *International Conference on Intelligent Tutoring Systems*, pages 406–415. Springer, 2008.

[58] Oliver Daems, Melanie Erkens, Nils Malzahn, and H. Ulrich Hoppe. "Using Content Analysis and Domain Ontologies to Check Learners' Understanding of Science Concepts". *Journal of Computers in Education*, 1(2):113–131.

[59] Ben K Daniel, Gordon I McCalla, and Richard A Schwier. "Social Network Analysis techniques: implications for information and knowledge sharing in virtual learning communities". *International journal of advanced media and communication*, 2(1): 20–34, 2008.

[60] Shane Dawson. "'Seeing'the learning community: An exploration of the development of a resource for monitoring online student networking". *British Journal of Educational Technology*, 41(5):736–752, 2010.

[61] Paul Deane. "Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks". *ETS Research Report Series*, 2014(1):1–23, 2014.

[62] Edward L Deci and Richard M Ryan. *Handbook of self-determination research*. University Rochester Press, 2002.

[63] Karel Dejaeger, Frank Goethals, Antonio Giangreco, Lapo Mola, and Bart Baesens. "Gaining insight into student satisfaction using comprehensible data mining techniques". *European Journal of Operational Research*, 218(2):548–562, 2012.

[64] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[65] Michel Desmarais. "Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization". In *Educational Data Mining 2011*, 2010.

[66] Michel Desmarais and François Lemieux. "Clustering and visualizing study state sequences". In *Educational Data Mining 2013*, 2013.

[67] Michel Desmarais, Behzad Beheshti, and Rhouma Naceur. "Item to skills mapping: deriving a conjunctive q-matrix from data". In *Intelligent tutoring systems*, pages 454–463. Springer, 2012.

[68] Pierre Dillenbourg, Nan Li, and Łukasz . *From Books to MOOCs? Emerging Models of Learning and Teaching in Higher Education*, chapter The complications of the orchestration clock. Portland Press, 2016.

[69] Patrick Doreian, Vladimir Batagelj, Anuska Ferligoj, and Mark Granovetter. *Generalized Blockmodeling (Structural Analysis in the Social Sciences)*. Cambridge University Press.

[70] Erik Duval, Joris Klerkx, Katrien Verbert, Till Nagel, Sten Govaerts, Gonzalo Alberto Parra Chico, Santos Odriozola, Jose Luis, and Bram Vandeputte. "Learning dashboards & learnscapes". In *Educational Interfaces, Software, and Technology*, pages 1–5, 2012.

[71] I M Van Eekelen, H PA Boshuizen, and Jan D Vermunt. "Self-regulation in higher education teacher learning". *Higher education*, 50(3):447–471, 2005.

[72] Asmaa Elbadrawy, Agoritsa Polyzou, Zhiyun Ren, Mackenzie Sweeney, George Karypis, and Huzefa Rangwala. "Predicting student performance using personalized analytics". *Computer*, 49(4):61–69, 2016.

[73] Michael Eraut. "Non-formal learning and tacit knowledge in professional work". *British journal of educational psychology*, 70(1):113–136, 2000.

[74] Yixin Fang and Junhui Wang. "Selection of the number of clusters via the bootstrap method". *Computational Statistics & Data Analysis*, 56(3):468–477, 2012.

[75] Louis Faucon, Lukasz Kidzinski, and Pierre Dillenbourg. "Semi-Markov model for simulating MOOC students". In *Proceedings of the 9th International Conference on Educational Data Mining*, pages 358–363, 2016.

[76] Mi Fei and Dit-Yan Yeung. "Temporal models for predicting student dropout in massive open online courses". In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, pages 256–263. IEEE, 2015.

[77] Rebecca Ferguson. "Learning analytics: drivers, developments and challenges". *International Journal of Technology Enhanced Learning*, 4(5-6):304–317, 2012.

[78] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage Uday Kiran, Yun Sing Koh, and R Thomas. "A survey of sequential pattern mining". *Data Science and Pattern Recognition*, 1(1):54–77, 2017.

[79] Marga Franco-Casamitjana, Elena Baberà, and Margarida Romero. "A methodological definition for time regulation patterns and learning efficiency in collaborative learning contexts". *eLearn Center Research Paper Series*, pages 52–62, 2013.

[80] Linton C Freeman. "Centrality in social networks conceptual clarification". *Social networks*, 1(3):215–239, 1978.

[81] Enrique García, Cristóbal Romero, Sebastián Ventura, and Carlos De Castro. "A collaborative educational association rule mining tool". *The Internet and Higher Education*, 14 (2):77–88, 2011.

[82] Iolanda Garcia, Begoña Gros, and Ingrid Noguera. "Supporting Learning Self-Regulation through a PLE: Dealing with the Time". *Assessment and Evaluation of Time Factors in Online Teaching and Learning*, pages 127–162, 2013.

[83] Chase Geigle and ChengXiang Zhai. "Modeling MOOC Student Behavior With Two-Layer Hidden Markov Models". In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 205–208. ACM, 2017.

[84] Nabeel Gillani and Rebecca Eynon. "Communication patterns in massively open online courses". *The Internet and Higher Education*, 23:18–26, 2014.

[85] Nabeel Gillani, Taha Yasseri, Rebecca Eynon, and Isis Hjorth. "Structural limitations of learning in a crowd: communication vulnerability and information diffusion in MOOCs". *Scientific Reports*, 4:6447.

[86] Yoshiko Goda, Masanori Yamada, Hiroshi Kato, Takeshi Matsuda, Yutaka Saito, and Hiroyuki Miyagawa. "Procrastination and other learning behavioral types in e-learning and their relationship with learning outcomes". *Learning and Individual Differences*, 37: 72–80, 2015.

[87] Geraldine Gray, Colm McGuinness, and Philip Owende. "An application of classification models to predict learner progression in tertiary education". In *Advance Computing Conference (IACC), 2014 IEEE International*, pages 549–554. IEEE, 2014.

[88] Derek Greene, Donal Doyle, and Padraig Cunningham. "Tracking the evolution of communities in dynamic social networks". In *Advances in social networks analysis and mining (ASONAM), 2010 international conference on*, pages 176–183. IEEE, 2010.

[89] Begoña Gros, Elena Barbera, and Paul Kirshner. "Time factor in e-learning: impact literature review". *eLearn Center Research Paper Series*, (1):16–31, 2010.

[90] Anatoliy Gruzd and Caroline Haythornthwaite. "Automated Discovery and Analysis of Social Networks from Threaded Discussions. Paper presented at". In *the International Network of Social Network Analysts, St. Pete Beach*. Citeseer, 2008.

[91] Christian Günther and Wil van der Aalst. "Fuzzy mining–adaptive process simplification based on multi-perspective metrics". *Business Process Management*, pages 328–343, 2007.

[92] Philip J Guo and Katharina Reinecke. "Demographic differences in how students navigate through MOOCs". In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 21–30. ACM, 2014.

[93] Jean-Luc Gurtner, Alida Gulfi, Philippe A Genoud, Bernardo de Rocha Trindade, and Jérôme Schumacher. "Learning in multiple contexts: are there intra-, cross-and transcontextual effects on the learner's motivation and help seeking?". *European journal of psychology of education*, 27(2):213–225, 2012.

[94] Indira Hamulic and Nina Bijedic. "Social network analysis in virtual learning community at faculty of information technologies (fit), Mostar". *Procedia-Social and Behavioral Sciences*, 1(1):2269–2273, 2009.

[95] Emrah Hancer and Dervis Karaboga. "A comprehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number". *Swarm and Evolutionary Computation*, 32:49–67, 2017.

[96] Robert A Hanneman and Mark Riddle. "Introduction to social network methods", 2005.

[97] Christian Hansen, Casper Hansen, Niklas Hjuler, Stephen Alstrup, and Christina Lioma. "Sequence Modelling For Analysing Student Interaction with Educational Systems". *arXiv preprint arXiv:1708.04164*, 2017.

[98] Linda Harasim. "Shift happens: Online education as a new paradigm in learning". *The Internet and higher education*, 3(1):41–61, 2000.

[99] T. Hecking, I. A. Chounta, and H. U. Hoppe. "Analysis of User Roles and the Emergence of Themes in Discussion Forums". In *Proceedings of the 2nd European Network Intelligence Conference*, ENIC'15, pages 114–121. IEEE, 2015.

[100] Tobias Hecking, Sabrina Ziebarth, and Heinz Ulrich Hoppe. "Analysis of dynamic resource access patterns in online courses". *Journal of Learning Analytics*, 1(3):34–60, 2014.

[101] Tobias Hecking, H Ulrich Hoppe, and Andreas Harrer. "Uncovering the Structure of Knowledge Exchange in a MOOC Discussion Forum". In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1614–1615. ACM, 2015.

[102] Tobias Hecking, Irene-Angelica Chounta, and H. Ulrich Hoppe. "Investigating Social and Semantic User Roles in MOOC Discussion Forums". In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, LAK'16, 2016.

[103] Tobias Hecking, Irene Angelica Chounta, and H Ulrich Hoppe. "Role modelling in MOOC discussion forums". *Journal of Learning Analytics*, 4(1):85–116, 2017.

[104] Tobias Hecking, Andreas Harrer, and H Ulrich Hoppe. "Discovery of Structural and Temporal Patterns in MOOC Discussion Forums". In *Prediction and Inference from Social Networks and Social Media*, pages 171–198. Springer, 2017.

[105] Laurie-Ann M Hellsten. "What do we know about time management? A review of the literature and a psychometric critique of instruments assessing time management". In *Time management*. Intech, 2012.

[106] Heeok Heo, Kyu Yon Lim, and Youngsoo Kim. "Exploratory study on the patterns of online interaction and knowledge co-construction in project-based learning". *Computers & Education*, 55(3):1383–1392, 2010.

[107] Shaobo Huang and Ning Fang. "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models". *Computers & Education*, 61:133–145, 2013.

[108] Mark D Humphries and Kevin Gurney. "Network 'small-world-ness': a quantitative method for determining canonical network equivalence". *PloS one*, 3(4):e0002051, 2008.

[109] Irshad Hussain and Sarwat Sultan. "Analysis of procrastination among university students". *Procedia-Social and Behavioral Sciences*, 5:1897–1904, 2010.

[110] Hogyeong Jeong and Gautam Biswas. "Mining student behavior models in learning-by-teaching environments". In *Educational Data Mining 2008*, pages 127–136, 2008.

[111] Srećko Joksimović, Areti Manataki, Dragan Gašević, Shane Dawson, Vitomir Kovanović, and Inés Friss De Kereki. "Translating network position into performance: importance of centrality in different network configurations". In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 314–323. ACM, 2016.

[112] Srećko Joksimović, Areti Manataki, Dragan Gašević, Shane Dawson, Vitomir Kovanović, and Inés Friss de Kereki. "Translating Network Position into Performance: Importance of Centrality in Different Network Configurations". In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, LAK'16, 2016.

[113] Per Jönsson and Lars Eklundh. "Seasonality extraction by function fitting to time-series of satellite sensor data". *Geoscience and Remote Sensing, IEEE Transactions on*, 40(8): 1824–1832, 2002.

[114] Mehdi Kaighobadi and Marcus T Allen. "Investigating academic success factors for undergraduate business students". *Decision Sciences Journal of Innovative Education*, 6 (2):427–436, 2008.

[115] Pythagoras Karampiperis and Demetrios Sampson. "Adaptive learning resources sequencing in educational hypermedia systems". *Journal of Educational Technology & Society*, 8(4):128–147, 2005.

[116] Shaun Kellogg, Sherry Booth, and Kevin Oliver. "A social network perspective on peer supported learning in MOOCs for educators". *The International Review of Research in Open and Distributed Learning*, 15(5), 2014.

[117] Gregor Kennedy, Carleton Coffrin, Paula De Barba, and Linda Corrin. "Predicting success: how learners' prior knowledge, skills and activities predict MOOC performance". In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 136–140. ACM, 2015.

[118] Mohammad Khalil and Martin Ebner. "Learning analytics: principles and constraints". In *Proceedings of world conference on educational multimedia, hypermedia and telecommunications*, pages 1326–1336, 2015.

[119] Anupam Khan and Soumya K Ghosh. "Analysing the impact of poor teaching on student performance". In *IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pages 169–175. IEEE, 2016.

[120] Mohamed Koutheaïr Khribi, Mohamed Jemni, and Olfa Nasraoui. "Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval". In *Advanced Learning Technologies, 2008. ICALT'08. Eighth IEEE International Conference on*, pages 241–245. IEEE, 2008.

[121] John S Kinnebrew, Kirk M Loretz, and Gautam Biswas. "A contextualized, differential sequence mining method to derive students' learning behavior patterns". *JEDM-Journal of Educational Data Mining*, 5(1):190–219, 2013.

[122] René F Kizilcec and Sherif Halawa. "Attrition and achievement gaps in online learning". In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 57–66. ACM, 2015.

[123] René F Kizilcec, Chris Piech, and Emily Schneider. "Deconstructing disengagement: analyzing learner subpopulations in massive open online courses". In *Proceedings of the third international conference on learning analytics and knowledge*, pages 170–179. ACM, 2013.

[124] René F Kizilcec, Emily Schneider, Geoffrey L Cohen, and Daniel A McFarland. "Encouraging forum participation in online courses with collectivist, individualist and neutral motivational framings". *Experiences and best practices in and around MOOCs*, pages 17–26, 2014.

[125] René F Kizilcec, Mar Pérez-Sanagustín, and Jorge J Maldonado. "Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses". *Computers & Education*, 104:18–33, 2017.

[126] Aleksandra Klašnja-Milićević, Boban Vesin, Mirjana Ivanović, and Zoran Budimac. "E-Learning personalization based on hybrid recommendation strategy and learning style identification". *Computers & Education*, 56(3):885–899, 2011.

[127] Robert M Klassen, Lindsey L Krawchuk, and Sukaina Rajani. "Academic procrastination of undergraduates: Low self-efficacy to self-regulate predicts higher levels of procrastination". *Contemporary Educational Psychology*, 33(4):915–931, 2008.

[128] Severin Klingler, Tanja Käser, Barbara Solenthaler, and Markus H Gross. "Temporally Coherent Clustering of Student Data.". In *Proceedings of the 9th International Conference on Educational Data Mining*, pages 102–109, 2016.

[129] Mirjam Köck and Alexandros Paramythis. "Activity sequence modelling and dynamic clustering for personalized e-learning". *User Modeling and User-Adapted Interaction*, 21 (1):51–97, 2011.

[130] Varun Kumar and Anupama Chadha. "Mining association rules in student's assessment data". *International Journal of Computer Science Issues*, 9(5):211–216, 2012.

[131] Charles Lang, George Siemens, Alyssa Wise, and Dragan Gašević. "Handbook of learning analytics". `https://solaresearch.org/hla-17`, 2017.

[132] Clarry H Lay. "At last, my research article on procrastination". *Journal of research in personality*, 20(4):474–495, 1986.

[133] Cen Li and Gautam Biswas. "A Bayesian Approach to Temporal Data Clustering using Hidden Markov Models.". In *ICML*, pages 543–550, 2000.

[134] Kam Cheong Li, Hoi Kuan Lam, and Sylvia SY Lam. "A Review of Learning Analytics in Educational Research". In *International Conference on Technology in Education*, pages 173–184. Springer, 2015.

[135] Nan Li, Łukasz Kidziński, Patrick Jermann, and Pierre Dillenbourg. "MOOC Video Interaction Patterns: What Do They Tell Us?". In *Design for Teaching and Learning in a Networked World*, pages 197–210. Springer International Publishing, 2015.

[136] Jianhua Lin. "Divergence measures based on the Shannon entropy". *Information Theory, IEEE Transactions on*, 37(1):145–151, 1991.

[137] Laura Calvet Liñán and Ángel Alejandro Juan Pérez. "Educational Data Mining and Learning Analytics: differences, similarities, and time evolution". *International Journal of Educational Technology in Higher Education*, 12(3):98–112, 2015.

[138] M Liu, J Kang, and E McKelroy. "Examining learners' perspective of taking a MOOC: reasons, excitement, and perception of usefulness". *Educational Media International*, 52(2):129–146, 2015.

[139] Ou Lydia Liu, Frank Rijmen, Carolyn MacCann, and Richard Roberts. "The assessment of time management in middle-school students". *Personality and individual differences*, 47(3):174–179, 2009.

[140] Weizhe Liu, Łukasz Kidziński, and Pierre Dillenbourg. "Semiautomatic Annotation of MOOC Forum Posts". In *State-of-the-Art and Future Directions of Smart Learning*, pages 399–408. Springer, 2016.

[141] Debbie Guice Longman and Rhonda Holt Atkinson. *College learning and study skills*. Wadsworth Publishing Company, 1999.

[142] Francois Lorrain and Harrison C White. "Structural equivalence of individuals in social networks". *The Journal of mathematical sociology*, 1(1):49–80, 1971.

[143] A Loya, A Gopal, I Shukla, Patrick Jermann, and Roland Tormey. "Conscientious behaviour, flexibility and learning in massive open on-line courses". *Procedia-Social and Behavioral Sciences*, 191:519–525, 2015.

[144] Ioanna Lykourentzou, Ioannis Giannoukos, Vassilis Nikolopoulos, George Mpardis, and Vassili Loumos. "Dropout prediction in e-learning courses through the combination of machine learning techniques". *Computers & Education*, 53(3):950–965, 2009.

[145] Therese H Macan, Comila Shahani, Robert L Dipboye, and Amanda P Phillips. "College students' time management: Correlations with academic performance and stress". *Journal of educational psychology*, 82(4):760–768, 1990.

[146] C MacCann and RD Roberts. "Do time management, grit, and self-control relate to academic achievement independently of conscientiousness". *Personality and individual differences: Current directions*, pages 79–90, 2010.

[147] Carolyn MacCann, Gerard J Fogarty, and Richard D Roberts. "Strategies for success in education: Time management is more important for part-time than full-time community college students". *Learning and Individual Differences*, 22(5):618–623, 2012.

[148] Leah P Macfadyen and Shane Dawson. "Mining LMS data to develop an "early warning system" for educators: A proof of concept". *Computers & education*, 54(2):588–599, 2010.

[149] Roberto Martinez Maldonado, Kalina Yacef, Judy Kay, Ahmed Kharrufa, and Ammar Al-Qaraghuli. "Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop". In *Educational Data Mining 2011*, 2010.

[150] Carlos Márquez-Vera, Alberto Cano, Cristobal Romero, Amin Yousef Mohammad Noaman, Habib Mousa Fardoun, and Sebastian Ventura. "Early dropout prediction using data mining: a case study with high school students". *Expert Systems*, 33(1):107–124, 2016.

[151] A Martınez, Yannis Dimitriadis, Bartolomé Rubia, Eduardo Gómez, and Pablo De la Fuente. "Combining qualitative evaluation and social network analysis for the study of classroom social interactions". *Computers & Education*, 41(4):353–368, 2003.

[152] Roberto Martinez Maldonado, Judy Kay, Kalina Yacef, and Beat Schwendimann. "An interactive teacher's dashboard for monitoring groups in a multi-tabletop learning environment". In *Intelligent Tutoring Systems*, pages 482–492. Springer, 2012.

[153] Kirsten McKenzie and Kathryn Gow. "Exploring the first year academic achievement of school leavers and mature-age students through structural equation modelling". *Learning and Individual Differences*, 14(2):107–123, 2004.

[154] Enric Mor and Julià Minguillón. "E-learning personalization based on itineraries and long-term navigational behavior". In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 264–265. ACM, 2004.

[155] Alex Mottus, Sabine Graf, Nian-Shing Chen, et al. "Use of dashboards and visualization techniques to support teacher decision making". In *Ubiquitous Learning Environments and Technologies*, pages 181–199. Springer, 2015.

[156] Patrick Mukala, JCAM Buijs, and WMP Van Der Aalst. "Exploring students' learning behaviour in moocs using process mining techniques", 2015.

[157] Stanley A Mulaik. *Foundations of factor analysis*. CRC press, 2009.

[158] Fionn Murtagh and Pierre Legendre. "Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?". *Journal of Classification*, 31(3): 274–295, 2014.

[159] Som Naidu. "Enabling Time, Pace, and Place Independence". *Handbook of Research on Educational Communications and Technology*, pages 259–268, 2004.

[160] Ilona Nawrot and Antoine Doucet. "Building engagement for MOOC students: introducing support for time management on online learning platforms". In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 1077–1082, 2014.

[161] John C Nesbit, Mingming Zhou, Yabo Xu, and P Winne. "Advancing log analysis of student interactions with cognitive tools". In *12th biennial conference of the european association for research on learning and insruction (EARLI)*, 2007.

[162] Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy. "A comparative analysis of techniques for predicting academic performance". In *Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*, pages T2G7–12. IEEE, 2007.

[163] Poquet Oleksandra and Dawson Shane. "Untangling MOOC Learner Networks". In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, LAK'16, 2016.

[164] Daniel FO Onah, Jane Sinclair, Russell Boyatt, and JG Foss. "Massive open online courses: learner participation". In *Proceeding of the 7th International Conference of Education, Research and Innovation*, pages 2348–2356, 2014.

[165] Yang Ouyang and Miaoliang Zhu. "eLORM: learning object relationship mining-based repository". *Online Information Review*, 32(2):254–265, 2008.

[166] Melissa C O'Connor and Sampo V Paunonen. "Big Five personality predictors of post-secondary academic performance". *Personality and Individual differences*, 43(5):971–990, 2007.

[167] Umesh Kumar Pandey and Saurabh Pal. "Data Mining: A prediction of performer or underperformer using classification". *arXiv preprint arXiv:1104.4163*, 2011.

[168] Zacharoula Papamitsiou and Anastasios A Economides. "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence". *Journal of Educational Technology & Society*, 17(4):49, 2014.

[169] Zach A Pardos, Ryan SJD Baker, Maria San Pedro, Sujith M Gowda, and Supreeth M Gowda. "Affective states and state tests: investigating how affect and engagement during the school year predict end-of-year learning outcomes". *Journal of Learning Analytics*, 1 (1):107–128, 2014.

[170] Zachary Pardos and Neil Heffernan. "KT-IDEM: introducing item difficulty to the knowledge tracing model". *User Modeling, Adaption and Personalization*, pages 243–254, 2011.

[171] Zachary A Pardos and Neil T Heffernan. "Using HMMs and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset". *Journal of Machine Learning Research W & CP*, 2010.

[172] Zachary A Pardos, Ryan SJD Baker, Maria OCZ San Pedro, Sujith M Gowda, and Supreeth M Gowda. "Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes". In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 117–124. ACM, 2013.

[173] Walter Christian Paredes and Kon Shing Kenneth Chung. "Modelling learning & performance: a social networks perspective". In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 34–42. ACM, 2012.

[174] Nirmal Patel, Collin Sellman, and Derek Lomas. "Mining Frequent Learning Pathways from a Large Educational Dataset". *arXiv preprint arXiv:1705.11125*, 2017.

[175] Reinhard Pekrun, Thomas Goetz, Anne C Frenzel, Petra Barchfeld, and Raymond P Perry. "Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ)". *Contemporary educational psychology*, 36(1):36–48, 2011.

[176] Donald B Percival and Andrew T Walden. *Spectral analysis for physical applications*. Cambridge University Press, 1993.

[177] Dilhan Perera, Judy Kay, Irena Koprinska, Kalina Yacef, and Osmar R Zaïane. "Clustering and sequential pattern mining of online collaborative learning data". *IEEE Transactions on Knowledge and Data Engineering*, 21(6):759–772, 2009.

[178] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. "Deep knowledge tracing". In *Advances in Neural Information Processing Systems*, pages 505–513, 2015.

[179] Arthur E Poropat. "A meta-analysis of the five-factor model of personality and academic performance.". *Psychological bulletin*, 135(2):322, 2009.

[180] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daumé III, and Lise Getoor. "Modeling learner engagement in MOOCs using probabilistic soft logic". In *NIPS Workshop on Data Driven Education*, volume 21, page 62, 2013.

[181] Martina A Rau and Richard Scheines. "Searching for variables and models to investigate mediators of learning from multiple representations". In *In International Conference on Educational Data Mining*, pages 110–117. Citeseer, 2012.

[182] Zhiyun Ren, Huzefa Rangwala, and Aditya Johri. "Predicting performance on MOOC assessments using multi-regression models". *arXiv preprint arXiv:1605.02269*, 2016.

[183] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. "An introduction to exponential random graph (p) models for social networks". *Social Networks*, 29:173–191, 2007.

[184] María Jesús Rodríguez-Triana, Luis P Prieto, Andrii Vozniuk, Mina Shirvani Boroujeni, Beat A Schwendimann, Adrian Holzer, and Denis Gillet. "Monitoring, awareness and reflection in blended technology enhanced learning: a systematic review". *International Journal of Technology Enhanced Learning*, 9(2-3):126–150, 2017.

[185] Cristóbal Romero and Sebastián Ventura. "Educational data mining: a review of the state of the art". *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[186] Cristobal Romero and Sebastian Ventura. "Data mining in education". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.

[187] Cristóbal Romero, Sebastián Ventura, Amelia Zafra, and Paul De Bra. "Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems". *Computers & Education*, 53(3):828–840, 2009.

[188] Vicente-Arturo Romero-Zaldivar, Abelardo Pardo, Daniel Burgos, and Carlos Delgado Kloos. "Monitoring student progress using virtual appliances: A case study". *Computers & Education*, 58(4):1058–1067, 2012.

[189] Carolyn Penstein Rosé and Oliver Ferschke. "Technology Support for Discussion Based Learning: From Computer Supported Collaborative Learning to the Future of Massive Open Online Courses". *International Journal of Artificial Intelligence in Education*, 26 (2):660–678, 2016.

[190] Lorenzo A Rossi and Omprakash Gnawali. "Language independent analysis and classification of discussion threads in Coursera MOOC forums". In *Proceedings of the 15th International IEEE Conference on Information Reuse and Integration*, IRI'14, pages 654–661. IEEE, 2014.

[191] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of computational and applied mathematics*, 20:53–65, 1987.

[192] Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. "The TETRAD project: Constraint based aids to causal model specification". *Multivariate Behavioral Research*, 33(1):65–117, 1998.

[193] Beat A Schwendimann, Alberto AP Cattaneo, Jessica Dehler Zufferey, Jean-Luc Gurtner, Mireille Bétrancourt, and Pierre Dillenbourg. "The 'Erfahrraum': A pedagogical model for designing educational technologies in dual vocational systems". *Journal of Vocational Education & Training*, 67(3):367–396, 2015.

[194] Beat A Schwendimann, María Jesús Rodríguez-Triana, Andrii Vozniuk, Luis P Prieto, Mina Shirvani Boroujeni, Adrian Holzer, Denis Gillet, and Pierre Dillenbourg. "Understanding learning at a glance: An overview of learning dashboard studies". In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 532–533. ACM, 2016.

[195] Beat A Schwendimann, Maria Jesus Rodriguez-Triana, Andrii Vozniuk, Luis P Prieto, Mina Shirvani Boroujeni, Adrian Holzer, Denis Gillet, and Pierre Dillenbourg. "Perceiving learning at a glance: A systematic literature review of learning dashboard research". *IEEE Transactions on Learning Technologies*, 10(1):30–41, 2017.

[196] John Scott. *Social network analysis*. Sage, 2017.

[197] SERI. "Vocational and professional education and training in Switzerland - Facts and figures 2017". `https://www.sbfi.admin.ch/dam/sbfi/en/dokumente/2017/04/Fakten_Zahlen_BB2017.pdf.download.pdf/Fakten_Zahlen_BB2017_en.pdf`, 2017.

[198] Afsaneh Sharif and Barry Magrill. "Discussion forums in MOOCs". *International Journal of Learning, Teaching and Educational Research*, 12(1).

[199] Shitian Shen and Min Chi. "Clustering Student Sequential Trajectories Using Dynamic Time Warping". In *Proceedings of the 10th International Conference on Educational Data Mining*, pages 266–271, 2017.

[200] Benjamin Shih, Kenneth R Koedinger, and Richard Scheines. "Unsupervised discovery of student strategies". In *Educational Data Mining 2010*, pages 201–210, 2010.

[201] Mina Shirvani Boroujeni and Pierre Dillenbourg. "Discovery and Temporal Analysis of Latent Study Patterns from MOOC Interaction Sequences". In *8th International Learning Analytics and Knowledge Conference (LAK18)*, number EPFL-CONF-232491, 2018.

[202] Mina Shirvani Boroujeni, Łukasz Kidzinski, and Pierre Dillenbourg. "How employment constrains participation in MOOCS". In *Proceedings of the 9th International Conference on Educational Data Mining*, pages 376–377, 2016.

[203] Mina Shirvani Boroujeni, Kshitij Sharma, Łukasz Kidziński, Lorenzo Lucignano, and Pierre Dillenbourg. "How to quantify student's regularity?". In *European Conference on Technology Enhanced Learning*, pages 277–291. Springer, 2016.

[204] Mina Shirvani Boroujeni, Tobias Hecking, Heinz Ulrich Hoppe, and Pierre Dillenbourg. "Dynamics of MOOC discussion forums.". In *Proceedings of the 7th International Conference on Learning Analytics & Knowledge*, LAK'17, pages 128–137, 2017.

[205] George Siemens and Ryan SJ d Baker. "Learning analytics and educational data mining: towards communication and collaboration". In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254. ACM, 2012.

[206] Christopher A Sims. "Seasonality in regression". *Journal of the American Statistical Association*, 69(347):618–626, 1974.

[207] Katrina Sin and Loganathan Muthu. "Applications of big data in education data mining and learning analytics–a literature review". *ICTACT journal on soft computing*, 5(4), 2015.

[208] Tanmay Sinha. "Together we stand, Together we fall, Together we win: Dynamic team formation in massive open online courses". In *Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the*, pages 107–112. IEEE, 2014.

[209] Christoph Sonnenberg and Maria Bannert. "Discovering the effects of metacognitive prompts on the sequential structure of SRL-processes using process mining techniques". *Journal of Learning Analytics*, 2(1):72–100, 2015.

[210] Christina M Steiner, Michael D Kickmeier-Rust, and Dietrich Albert. "Learning analytics and educational data mining: An overview of recent techniques". *Learning Analytics for and in Serious Games*, 6:61–75, 2014.

[211] Matthias Studer and Gilbert Ritschard. "What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):481–511, 2016.

[212] Kikumi K Tatsuoka. "Rule space: An approach for dealing with misconceptions based on item response theory". *Journal of educational measurement*, 20(4):345–354, 1983.

[213] Dirk T Tempelaar, Alexandra Niculescu, Bart Rienties, Wim H Gijselaers, and Bas Giesbers. "How achievement emotions impact students' decisions for online learning, and what precedes those emotions". *The Internet and Higher Education*, 15(3):161–169, 2012.

[214] Melody M Terras and Judith Ramsay. "E-Learning, Mobility, and Time: A Psychological Framework". *Human-Computer Interaction: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*, pages 396–416, 2015.

[215] Nguyen Thai-Nghe, Tomas Horvath, and Lars Schmidt-Thieme. "Context-aware factorization for personalized student's task recommendation". In *Proceedings of the international workshop on personalization approaches in learning environments*, volume 732, pages 13–18, 2011.

[216] Bruce Thompson. *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association, 2004.

[217] Mike Tissenbaum, Camillia Matuk, Matthew Berland, Felipe Cocco, Marcia Linn, CREATE Nik Hajny, CREATE Al Olsen, Beat Schwendimann, Mina Shirvani Boroujeni, Jonathan Vitale, et al. "Real-time visualization of student activities to support classroom orchestration". *Transforming Learning, Empowering Learners*, 2016.

[218] Sabrina Trapmann, Benedikt Hell, Jan-Oliver W Hirn, and Heinz Schuler. "Meta-analysis of the relationship between the Big Five and academic success at university". *Zeitschrift für Psychologie/Journal of Psychology*, 215(2):132–151, 2007.

[219] Nikola Trcˇka, Mykola Pechenizkiy, and Wil van der Aalst. *Process mining from educational data*. Chapman & Hall/CRC, 2010.

[220] Shubhendu Trivedi, Zachary A Pardos, and Neil T Heffernan. "Clustering students to generate an ensemble to improve standard test score predictions". In *International Conference on Artificial Intelligence in Education*, pages 377–384. Springer, 2011.

[221] Wil Van der Aalst, Ton Weijters, and Laura Maruster. "Workflow mining: Discovering process models from event logs". *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1128–1142, 2004.

[222] Katrien Verbert, Sten Govaerts, Erik Duval, Jose Luis Santos, Frans Van Assche, Gonzalo Parra, and Joris Klerkx. "Learning dashboards: an overview and future research opportunities". *Personal and Ubiquitous Computing*, 18(6):1499–1514, 2014.

[223] JP Verma. "Application of Factor Analysis: To Study the Factor Structure Among Variables". In *Data Analysis in Management with SPSS Software*, pages 359–387. Springer, 2013.

[224] W Paul Vogt. *SAGE quantitative research methods*. Sage, 2011.

[225] Andrii Vozniuk. *Enhancing Social Media Platforms for Educational and Humanitarian Knowledge Sharing: Analytics, Privacy, Discovery, and Delivery Aspects*. PhD thesis, EPFL, 2017.

[226] Andrii Vozniuk, María Jesús Rodríguez-Triana, Adrian Holzer, Sten Govaerts, David Sandoz, and Denis Gillet. "Contextual learning analytics apps to create awareness in blended inquiry learning". In *Information Technology Based Higher Education and Training (ITHET), 2015 International Conference on*, pages 1–5. IEEE, 2015.

[227] Wei Wang, Jui-Feng Weng, Jun-Ming Su, and Shian-Shyong Tseng. "Learning portfolio analysis and mining in SCORM compliant environment". In *Frontiers in Education, 2004. FIE 2004. 34th Annual*, pages T2C–17. IEEE, 2004.

[228] Wei Wang, Han Yu, and Chunyan Miao. "Deep Model for Dropout Prediction in MOOCs". In *Proceedings of the 2nd International Conference on Crowd Science and Engineering*, pages 26–32. ACM, 2017.

[229] Yuan Wang and Ryan Baker. "Content or platform: Why do students complete MOOCs?". *Journal of Online Learning and Teaching*, 11(1):17, 2015.

[230] Duncan J. Watts and Steven H. Strogatz. "Collective dynamics of small-world networks". *Nature*, 393(6684):440–442.

[231] AJMM Weijters, Wil MP van Der Aalst, and AK Alves De Medeiros. "Process mining with the heuristics miner-algorithm". *Technische Universiteit Eindhoven, Tech. Rep. WP*, 166: 1–34, 2006.

[232] Miaomiao Wen. *Investigating Virtual Teams in Massive Open Online Courses: Deliberation-based Virtual Team Formation, Discussion Mining and Support*. PhD thesis, Carnegie Mellon University, 2016.

[233] Miaomiao Wen and Carolyn Penstein Rosé. "Identifying latent study habits by mining learner behavior patterns in massive open online courses". In *Proceedings of the 23rd*

ACM International Conference on Conference on Information and Knowledge Management, pages 1983–1986. ACM, 2014.

[234] Douglas R White and Karl P Reitz. "Graph and semigroup homomorphisms on networks of relations". *Social Networks*, 5(2):193–234, 1983.

[235] Alyssa Friend Wise and David Williamson Shaffer. "Why theory matters more than ever in the age of big data". *Journal of Learning Analytics*, 2(2):5–13, 2015.

[236] Alyssa Friend Wise, Yi Cui, and Jovita Vytasek. "Bringing order to chaos in MOOC discussion forums with content-related thread identification". In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, pages 188–197. ACM, 2016.

[237] Alyssa Friend Wise, Yi Cui, and Wan Qi Jin. "Honing in on social learning networks in MOOC forums: examining critical network definition decisions". In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 383–392. ACM, 2017.

[238] Alyssa Friend Wise, Yi Cui, Wanqi Jin, and Jovita Vytasek. "Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling". *The Internet and Higher Education*, 32:11–28, 2017.

[239] Annika Wolff, Zdenek Zdrahal, Andriy Nikolov, and Michal Pantucek. "Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment". In *Proceedings of the third international conference on learning analytics and knowledge*, pages 145–149. ACM, 2013.

[240] Jian-Syuan Wong, Bart Pursel, Anna Divinsky, and Bernard J Jansen. "Analyzing MOOC discussion forum messages to identify cognitive learning information exchanges". volume 52, pages 1–10. Wiley Online Library, 2015.

[241] Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. "Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses". In *Proceedings of the 2013 NIPS Data-driven education workshop*, volume 11, page 14, 2013.

[242] Diyi Yang, David Adamson, and Carolyn Penstein Rosé. "Question recommendation with constraints for massive open online courses". In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 49–56. ACM, 2014.

[243] Cheng Ye, John S Kinnebrew, Gautam Biswas, Brent J Evans, Douglas H Fisher, Gayathri Narasimham, and Katherine A Brady. "Behavior prediction in MOOCs using higher granularity temporal information". In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 335–338. ACM, 2015.

[244] Jeongah Yoon, Anselm Blumer, and Kyongbum Lee. "An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality". *Bioinformatics*, 22(24):3106–3108, 2006.

[245] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. "Individualized bayesian knowledge tracing models". In *International Conference on Artificial Intelligence in Education*, pages 171–180. Springer, 2013.

[246] Norazah Yusof, Azizah Abdul Rahman, et al. "Integrating content analysis and social network analysis for analyzing asynchronous discussion forum". In *Information Technology, 2008. ITSim 2008. International Symposium on*, volume 3, pages 1–8. IEEE, 2008.

[247] Sam Zeini, Tilman Göhnert, Tobias Hecking, Lothar Krempel, and H Ulrich Hoppe. "The impact of measurement time on subgroup detection in online communities". In *State of the art applications of social network analysis*, pages 249–268. Springer, 2014.

[248] Jiahua Zhang and Jianping Zhang. "A case study on web-based knowledge construction in Moodle platform". In *Computer Science and Education (ICCSE), 2010 5th International Conference on*, pages 1110–1114. IEEE, 2010.

[249] Jingjing Zhang, Maxim Skryabin, and Xiongwei Song. "Understanding the dynamics of MOOC discussion forums with simulation investigation for empirical network analysis (SIENA)". *Distance Education*, 37(3):270–286, 2016.

[250] Liang Zhang, Xiumin Liu, and Xiujuan Liu. "Personalized instructing recommendation system based on web mining". In *Young Computer Scientists, 2008. ICYCS 2008. The 9th International Conference for*, pages 2517–2521. IEEE, 2008.

[251] Saijing Zheng, Mary Beth Rosson, Patrick C. Shih, and John M. Carroll. "Designing MOOCs As Interactive Places for Collaborative Learning". In *Proceedings of the 2nd ACM Conference on Learning @ Scale*, L@S'15, 2015.

[252] Mengxiao Zhu, Yoav Bergner, Yan Zhang, Ryan Baker, Yuan Wang, and Luc Paquette. "Longitudinal engagement, performance, and social connectivity: a MOOC case study using exponential random graph models". In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 223–230. ACM, 2016.

[253] Barry J Zimmerman. "Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects". *American Educational Research Journal*, 45(1):166–183, 2008.

Mina Shirvani Boroujeni

Data scientist and Machine Learning Researcher

No.34, Avenue de florissant, 1020 Renens, Switzerland

(+41) 789601139 | shirvani.mina@gmail.com | mina-shirvani-boroujeni-b0574885

Education

Ecole polytechnique fédérale de Lausanne — *Lausanne, Switzerland*

PhD in Computer and Communication Sciences — *Sep. 2013 - PRESENT*

- Thesis title: Learning Analytics: discovering interaction patterns in online learning environments
  - Developed novel analytics methods to model and analyze learners' participation patterns in Massive Open Onlin Courses (MOOC).
  - Developed interactive dashboards to support awareness and enable monitoring of learners activity in online learning platforms.

Amirkabir University of Technology (AUT) — *Tehran, Iran*

MSc in Artificial Intelligence — *Sep. 2009 - Sep. 2012*

- Thesis title: Automatic event detection in video sequences for vision-based surveillance
- GPA: 18.34 out of 20

Amirkabir University of Technology — *Tehran, Iran*

BSc in Computer Science, Software engineering — *Sep. 2005 - Sep. 2009*

- Thesis title: Implementation of a steganography system for binary images
- GPA: 17.96 out of 20

Research Experience

Doctoral Researcher at Computer-Human Interaction in Learning and Instruction Laboratory (CHILI) — *EPFL, Lausanne*

Research on machine learning and visualization techniques for large scale educational data — *Sep. 2013 - PRESENT*

- Research:
  - **Big data mining**: Analysis of users' online interactions and behavioural patterns in Massive Open Online Cources (MOOCs)
  - **Machine learning**: Development of methods for modeling and predicting users' behaviour based on supervised and unsupervised approaches
  - **Data visualization**: Development of web-based contextual interactive dashboards
  - **Eye tracking**: Performing user study and analyzing gaze patterns in tangible user interfaces
- Projects:
  - Leading House DUAL-T, technologies in vocational training (link)

Internship at Learning Algorithm and System Laboratory (LASA) — *EPFL, Lausanne*

Research on video-based object tracking systems — *Feb. 2013 - Sep. 2013*

- Development of a 3D pose estimation and object tracking algorithm for robotic applications

Research assistant at Image Processing and Pattern Recognition Laboratory — *Tehran, Iran*

Research on computer vision and image processing — *Sep. 2012 - Aug. 2013*

- Development of image processing algorithms for object detection, classification, tracking, and event detection.

Teaching experience

2015-2017	**Introduction to visual computing**, Teaching assistant during three spring semesters	*EPFL, Lausanne*
2015-2017	**Digital education and learning analytics**, Teaching assistant during three fall semesters	*EPFL, Lausanne*
2014	**Introduction to Java programming**, Teaching assistant during fall semester	*EPFL, Lausanne*
2011	**Digital Image Processing**, Teaching assistant during spring semester	*Tehran, Iran*
2010	**Statistical Pattern Recognition**, Teaching assistant during fall semester	*Tehran, Iran*
2008	**Principles of Database Design**, Teaching assistant during spring semester	*Tehran, Iran*

Honors & Awards

Dec. 2018	**Best paper nominee**, International Learning Analytics and Knowledge Conference (LAK)	*EPFL, Lausanne*
Dec. 2015	**Best teaching assistant team**, Support the design and running of *Introduction to visual computing*	*EPFL, Lausanne*
Sep. 2012	**Ranked 1st**, The master program in artificial Intelligence	*Tehran, Iran*
Sep. 2009	**Honorary admission**, The master program in Amirkabir University of Technology	*Tehran, Iran*
Sep. 2009	**Ranked 3rd**, The bachelor program in the computer science department	*Tehran, Iran*
Sep. 2005	**Entrance elite**, The bachelor program, ranked top 0.1% in among 400,000 participants	*Tehran, Iran*

Publications

- **M. Shirvani Boroujeni**, P. Dillenbourg. *Discovery and Temporal Analysis of Latent Study Patterns in MOOC Interaction Sequences*. 8th International Learning Analytics and Knowledge Conference (LAK18), Sydney, Australia, 2018. **[Best paper nominee]**
- **M. Shirvani Boroujeni**, T. Hecking, H. U. Hoppe and P. Dillenbourg. *Dynamics of MOOC Discussion Forums*. 7th International Learning Analytics and Knowledge Conference (LAK17), Vancouver, British Colombia, Canada, March 13-17, 2017.
- B. A. Schwendimann, M. J. Rodriguez Triana, A. Vozniuk, L. P. Prieto and **M. Shirvani Boroujeni** et al. *Perceiving learning at a glance: A systematic literature review of learning dashboard research*, IEEE Transactions on Learning Technologies, 10(1):30–41, 2017.
- M. J. Rodriguez Triana, L. P. Prieto, A. Vozniuk, **M. Shirvani Boroujeni** and B. A. Schwendimann et al. *Monitoring, Awareness and Reflection in Blended Technology Enhanced Learning: a Systematic Review*, International Journal of Technology Enhanced Learning, 9(2-3):126–150, 2017.
- **M. Shirvani Boroujeni**, K. Sharma, L. Kidzinski, L. Lucignano and P. Dillenbourg. *How to quantify student's regularity?* 11th European Conference on Technology Enhanced Learning, Lyon, France, September 13-16, 2016.
- **M. Shirvani Boroujeni**, L. Kidzinski and P. Dillenbourg. *How employment constrains participation in MOOCs?* 9th International Conference on Educational Data Mining, Raleigh, USA, June 30 - July 2, 2016
- K. Łukasz, K. Sharma, **M. Shirvani Boroujeni** and P. Dillenbourg. *On generalizability of MOOC models*. 9th International Conference on Educational Data Mining, Raleigh, North Carolina, USA., June 29 - July 2, 2016.
- B. A. Schwendimann, M. J. Rodriguez Triana, A. Vozniuk, L. P. Prieto and **M. Shirvani Boroujeni** et al. *Understanding learning at a glance: An overview of learning dashboard studies*. 6th International Learning Analytics and Knowledge Conference (LAK16), Edinburgh, UK, April 25-29, 2016.
- **M. Shirvani Boroujeni**, S. Cuendet, L. Lucignano, B. A. Schwendimann and P. Dillenbourg. *Screen or Tabletop: An Eye-Tracking Study of the Effect of Representation Location in a Tangible User Interface System*. 10th European Conference on Technology Enhanced Learning, EC-TEL 2015, Toledo, Spain, September 15- 18, 2015.
- L. Lucignano, S. Cuendet, B. Schwendimann, **M. Shirvani Boroujeni** and P. Dillenbourg. *My hands or my mouse: Comparing a tangible and graphical user interface using eye-tracking data*. Fablearn conferenc, Stanford University, CA, USA, October 25-26, 2014.

Skills

Data Analysis and Machine Learning

- **Machine learning**: Expert knowledge and several years of experience with broad range of supervised and unsupervised machine learning methods including clustering and classification, predictive models, deep learning, dimensionality reduction algorithms
- **Statistical inference**: Skilled in statistical models for data exploration, correlation and regression analysis
- **Network analysis**: Competent in graph modeling, visualization and analysis, role modeling, community detection
- **Time series analysis**: Proficient in sequential pattern mining, periodic pattern identification, behavioural pattern extraction
- **Natural language processing**: Experienced in text mining, discussion topic modeling, sentiment analysis
- **Big data**: Experienced in managing and analyzing large-scale data sets

Technical Skills

- **Statistical analysis**: R, Python
- **Programming**: Java, C++, Matlab
- **Data visualization**: Rshiny, Tableau, D3
- **Database**: SQL, MongoDB
- **Web development**: Javascript, Meteor, HTML
- **Miscellaneous**: LaTeX, Linux, Windows, Adobe suite, Microsoft Office suite

Language

- **Farsi** Native proficiency
- **English** Full professional proficiency (C1)
- **French** Limited working proficiency (B1)

Personal Interests

- Biking, Hiking, Swimming, Cooking and Baking