Departamento de Ciencias de la Computación

Doctoral Programme in

Communication, Information and Technology in the Web Society

DOCTORAL THESIS

Detecting Communities and Analysing Interactions with Learning Objects in Online Learning Repositories

Author:

Sahar Yassine

Supervisors:

Prof. Miguel Ángel Sicilia

Prof. Seifedine Kadry

Alcalá de Henares, Spain

Abstract

The widespread use of online learning object repositories has raised the need of studies that assess the quality of their contents, and their user's performance and engagement. The present research addresses two fundamental problems that are central to that need: the need to explore user interaction with these repositories and the detection of emergent communities of users.

The current dissertation approaches those directions through investigating and mining the Khan Academy repository as a free, open access, popular online learning repository addressing a wide content scope. It includes large numbers of different learning objects such as instructional videos, articles, and exercises. In addition to a large number of users.

Data was collected using the repository's public application programming interfaces combined with Web scraping techniques to gather data and user interactions. Different research activities were carried out to generate useful insights out of the gathered data. We conducted descriptive analysis to investigate the learning repository and its core features such as growth rate, popularity, and geographical distribution. A number of statistical and quantitative analysis were applied to examine the relation between the users' interactions and different metrics related to the use of learning objects in a step to assess the users' behaviour. We also used different Social Network Analysis (SNA) techniques on a network graph built from a large number of user interactions. The resulting network consisted of more than 3 million interactions distributed across more than 300,000 users. The type of those interactions is questions and answers posted on Khan Academy's instructional videos (more than 10,000 video). In order to analyse this graph and explore the social network structure, we studied two different community detection algorithms to identify the learning interactions communities emerged in Khan Academy then we compared between their effectiveness. After that, we applied different SNA measures including modularity, density, clustering coefficients and different centrality measures in order to assess the users' behaviour patterns and their presence.

Using descriptive analysis, we discovered many characteristics and features of the repository. We found that the number of learning objects in Khan Academy's repository grows linearly over time, more than 50% of the users do not complete the

watched videos, and we found that the average duration for video lessons 5 to 10 minutes which aligns with the recommended duration in literature. By applying community detection techniques and social network analysis, we managed to identify learning communities in Khan Academy's network. The size distribution of those communities found to follow the power-law distribution which is the case of many real-world networks. Those learning communities are related to more than one domain which means the users are active and interacting across domains. Different centrality measures we applied to focus on the most influential players in those communities.

Despite the popularity of online learning repositories and their wide use, the structure of the emerged learning communities and their social networks remain largely unexplored. Our findings could be considered initial insights that may help researchers and educators in better understanding online learning repositories, the learning process inside those repositories, and learner behaviour.

Dedication

To the memory of my father, Hussein Yassine

-to the person who always believed in me

You are gone but your belief in me

has made this journey possible

Acknowledgments

First and foremost, I would like to express my sincere appreciation to Professor Miguel Ángel Sicilia for his wise guidance, for the valuable research ideas and for all your constructive suggestions. I want to express my gratitude to you for giving me the chance to be part of the University of Alcalá, for your patience, understanding and your ongoing support. Truly, nothing could be done without your support and assistance. Thank you.

This whole work would never be existed without the help and assistance of Professor Seifedine Kadry. I am deeply grateful for having you as a great mentor who enlightened my path with his intelligent ideas. There were many times I reached the crossroads and each time you were there to direct me to the right path. Thank you for your enthusiastic encouragement during all the ups and downs in this journey. Thank you for the useful critiques, the advice and for keeping me on schedule. I am lucky to have you as my co-supervisor, my sincere colleague, and my true friend.

Finally, I would like to thank my family to whom I owe a great gratitude. To my mother Ghada for her love, encouragement, and support. My special thanks go to my husband Ghayath for his patience, encouragement and understanding. Also, I want to express my thanks to my daughter Rawan and my son Rayan. You will always be my great inspiration and motivation.

I am sure this thesis would never have been accomplished without all of you.

Thank you.

Summary

List of Figures

List of Tables

1. Introduction

In 2017, the US department of Education encouraged all the educators, the administrators, and the professional development programs to include the online learning and the digital tools and resources in their practices (U.S. Department of Education, 2017). It stressed on the collaboration with the families and researchers to try to reduce the gap, reach outside the walls of traditional classrooms, and build strong partnerships to support the learning everywhere and at any time.

Nowadays, as we are facing unprecedented times and almost 300 million students[1] suffer from educational disruption, online learning becomes a genuine necessity. Many research efforts attempted to define online learning and focus on its main features. A wide broad definition described the online learning as a learning delivery method built on exchanging the resources over a communication network which improves access to educational opportunities (Moore, Dickson-Deane, and Galyen, 2011). The main features of online learning include accessibility, flexibility, communication, and the ability to promote varied interactions.

Online learning repositories with all of their various types are very popular, and they are widely used by different types of users including different levels of learners, educators, and parents (De Medio et al., 2019). Those stakeholders interact with each other and with the learning resources published on those repositories in a daily basis in order to learn, practice, share experiences and information, and connect with others. All those interactions develop the learners' engagement and their presence in the learning process which is the fundamental of the dynamic learning environment (Tablatin, Patacsil, and Cenas, 2016). This vast number of interactions can be converted to valuable information if assessed and analysed thoroughly which opens the door to study the movements of the emerging online communities, their social

[1] According to data published in UNESCO's website on March 2020:

https://en.unesco.org/news/290-million-students-out-school-due-covid-19-unesco-releases-first-global-numbers-and-mobilizes

networks, and their interactions to evaluate their emergent patterns (Lockyer, Heathcote, and Dawson, 2013).

Data science methods and specifically social network analysis (SNA) techniques can play a powerful role in understanding the structure and the dynamics of online learning communities (Cela, Sicilia, and Sánchez, 2015). In doing so, there is a need to extract the networks (Huda et al., 2018) generated by the interactions in online learning settings, and there is also a need to assess those learning communities emerged from interacting with the learning objects to investigate learners' engagement and presence. All of which promote the collaboration between the educators who are concerned about learners' engagement, presence, and performance and the data scientists who are concerned about assessing the big data emerged (Macià and García, 2016).

The sustainability of open learning repositories is a fundamental key to ensure their long-term viability. Limited research attempts have tried to understand this sustainability through analysing the engaged learning communities, their created content, interactions, characteristics, and preferences (De Medio et al., 2019). The present work investigates online learning repositories to explore their structure and analyse the main features. Concretely, it examines users' interactions inside repositories, detects their emergent online learning communities and assesses their properties. Our analysis were done on data collected from Khan Academy's repository which is a free, open source and wide-reaching learning repository (Kelly and Rutherford, 2017). It includes different types of learning objects but the most frequent one is the instructional video lessons which are widely spread and attract significant amounts of interactions (Rao, Hilton, and Harper, 2017). The overall motivation of our work is to form useful and clear visions of the mechanism of online learning repositories and their users' interactions, which may lead researchers to a better comprehension and can use them in their practical applications and future research directions.

The rest of this document is organized as follows. In the next chapter, we will present the research objectives. Then, in chapter 3, the literature review is presented. In chapter 4, the research methodology is described, followed by the data acquisition process (chapter 5). The results of our research work are presented across the next

three chapters (6, 7, and 8). The work presented in those chapters is derived entirely or partially from the materials that already published and mentioned in the publication sections. Conclusions are provided in chapter 9 and potential future directions are outlined in chapter 10.

2. Research Objectives

The main objective of our research revolves around exploring online learning repositories and analysing their characteristics. It also aimed to assess and study users' interactions with those repositories and examine their patterns in order to detect emergent online communities and understand their properties. We intended to investigate the repositories, examine their online communities, and analyse their patterns to generate useful insights that may lead the researchers to a better comprehension and may be used in their practical applications and future research directions.

In order to achieve to our objectives, we used descriptive analysis, quantitative analysis, community detection algorithms and social network analysis (SNA) techniques. In alignment with our main objective, we addressed the following sub-objectives:

1. To review the previous work on the contents, structure, and evolution of online learning repositories.
2. To use different descriptive statistical analysis in order to investigate a representative example of online learning repositories and to assess the relation between learning materials and users' interaction. This helps in understanding its characteristics, the potential impact on learning process and gives some useful insights to help in assessing the instructional resources and evaluating their quality.
3. To apply different community detection techniques to identify the learning communities that are generated from users' interactions inside learning repositories. To study the structure of those communities and to assess their movements and engagements in order to discover their presence and develop a better understanding to the online learning setting.
4. To examine users' interactions with the learning repository by applying SNA techniques and measures to produce valuable information for educators and researchers that may increase their comprehension to online learning through interactions.

The objectives just describe require the empirical analysis of relevant data. In that direction, the methodology section details the selection and criteria used for choosing a target and the procedures to obtain the data.

3. Literature Review

In this chapter, we discuss the literature in order to provide theoretical background for studies on online learning repositories and on different ways to assess them, their quality, and their users' interactions. The review provides a background on online learning repositories and the different ways it can be assessed. This chapter is divided into 6 sub-sections to focus on the following: (1) a general historical background on online learning repositories and their most trending research, (2) studies that assess the online learning repositories using descriptive analysis, (3) studies that assess the repositories using statistical and quantitative methods, (4) research studies that apply social network analysis (SNA) techniques to assess and analyse learning repositories, (5) classifications of different community detection techniques as a main process in SNA, and (6) research efforts that tried to detect and identify communities in online learning repositories. This approach to the revision of the literature is guided by the objectives set in the doctoral research.

3.1 Historical Background on Online Learning Repositories

In (Yassine, Kadry, and Sicilia, 2016a), we reported a thorough review on the literature of learning objects repositories (LORs). We analysed the research related to learning objects and their repositories with the aim to understand their evolution, characteristics, analysed the applied quality approaches, and identify future directions. In the last three decades, a vastly growing amount of learning objects repositories has become available on the internet for the use of learners and educators. In 1998, Wiley was the first researcher to announce the open content license which adopted the idea of offering the educational content free, sharable and available for reuse (Caswell et al., 2008). This was the main concept of online learning objects repositories which were identified as multi-functional digital databases designed to enhance the access to different formats of reusable objects (Downes, 2001). Some operative functionalities (Higgs, Meredith, and Hand, 2003) must be presented in those repositories to provide secure access to the learning objects, such as: search, request, submit, store and publish. However, the massive number of learning objects collections and their heterogeneity created some limitations in their use and utilization (Tsakonas et al., 2013). This has encouraged many research attempts to classify the

learning objects repositories according to different principals. Here we will demonstrate some of those classifications:

- (Krämer, 2010) classified the learning repositories into four types based on their infrastructure: (1) centralized LOs with centralized metadata where all the learning objects and their metadata are located on a central server. (2) Centralized Learning Objects (LOs) with metadata distributed on several servers: in this type, LOs are located on a central server while their metadata are distributed away to minimize the processing cost. (3) Distributed LOs with centralized metadata: where the metadata and indexing are centralized and they provide links to the learning objects which are stored on distributed servers. (4) Distributed LOs with distributed metadata where all the architecture is distributed on different servers across the network.
- Another classification was proposed by (Clements, Pawlowski, and Manouselis, 2014) which was determined by the type of content and by their providers: (1) National Repositories: owned by ministries of education and main users are schools' teachers and students. (2) Thematic Repositories: where the provided learning objects focus on a certain topic such as science, math or music. (3) Federated International Repositories: This type of repositories is built on harvesting the metadata of other repositories and collecting critical masses of learning objects available.
- Despite ambiguities in distinguishing between the types of learning object repositories, Ochoa in his dissertation (Ochoa, 2011) identified four different types of learning objects repositories and compared between them. Learning object repository or "referatory" (LORP or LORF) which stores the metadata while the objects are stored somewhere else on different servers or on the web. The second type was open courseware initiative (OCW) which provides open and free digital versions of high quality educational materials arranged as courses that can be reused and shared in other learning settings. The learning management system (LMS) was the third type identified by Ochoa which stores a huge amount of learning materials and share them among a small group consists of the course users such as its students and teachers (Yassine, Kadry, and Sicilia, 2016a). The last identified type was institutional repository

(IR) which contains a set of digital learning materials developed and offered by institution to serve the members of its community.

- Some other studies tried to differentiate between the massive open online courses (MOOCs) and the open courseware initiatives (OCW). Some authors (Bonk et al., 2015), considered the MOOC as the recent release in the continuous evolution of open educational resources (OERs) while the open courseware initiative (OCW) was considered as the earlier version. In another study, (Martinez, 2014) the researchers identified OCW as open educational resources that can be always accessed everywhere and can be used freely without any conditions. However, they identified the MOOC as a new online educational tool which is more interactive, dynamic and social than OCW. The authors differentiated between OCW and MOOC from different perspectives. OCW was considered as a static resource which is a product of individual work, it is always free accessible, it does not require assessment nor accreditation, and it never threats the universities. On the other hand the MOOC was described as a dynamic resource that is produced by collaborative work, accessible while the course is open, it requires some kind of assessment and accreditation, and it is considered a competitor to the universities.

Other studies were engaged in assessing the quality of learning objects inside those repositories. In 2014, (Ochoa, Carrillo, and Cechinel, 2014) demonstrated an overview of quality assessments inside different types of learning repositories. They exhibited the different adopted evaluation strategies and they concluded that they are inadequate to cover the massive growing number of open educational resources. An assessment model was proposed by (Kay and Knaack, 2008). It was called the Learning object evaluation model (LOEM). It uses an assessing rubric to review of the details of the instructional design based on five main criteria: design, content, engagement, usability, and interactivity. Another quality evaluation model (MECOA) was developed by (Eguigure et al., 2011) which assess learning objects from a pedagogical perspective. This model works by evaluating a group of six indicators: content, competence, representation, creativity, signification, and self-management. An additional successful attempt to create a quality assurance framework for learning repositories (LORQAF) was done by (Clements, Pawlowski, and Manouselis, 2015) which works as a full approach to comprehend the full picture of learning repositories

quality approaches. In our study (Yassine, Kadry, and Sicilia, 2016b), we exhibited the importance of measuring the learning outcomes for online courses as a mean in identifying its quality and we highlighted the role of learning analytics in ensuring continuous quality improvements in learning environments. As an another step in the research (Yassine, Kadry, and Sicilia, 2016c), we proposed a framework for integrated learning analytics tool that assesses the learning outcomes of a learning management system (LMS) activities and relates them to the learning object's design and its quality. Lately, a recent research (Marín, Orellana, and Peré, 2019) tried to evaluate different educational resources for research training purpose using criteria extracted from the well-known Learning Object Review Instrument (LORI). The authors recommended to develop some other dimensions about the collaborative evaluation and the use of the instrument such as supporting the choice of the relevant criteria, providing more guides to educators, and sharing the results of evaluations would add more value.

3.2 Online Learning Repositories and Descriptive Analysis

An important step at the beginning of each study is to apply some descriptive analysis to describe the data of the population, simplify it, organize it, and present it in a meaningful manner (Loeb et al., 2017). This is to enforce the understanding of the nature, properties, and features of the data. A thorough descriptive analysis of the usage of ICT in higher education was presented in (Iniesta-Bonillo, Sánchez-Fernández, and Schlesinger, 2013) research. They described the influencing characteristics and focused on the influence of gender-related features on the usage of ICT in education. They demonstrated different measurements such as the use of university's website for search, the use of virtual classrooms, and the use of different electronic resources (databases and journals).

Some other studies provided descriptive analysis to different multiple learning objects repositories in order to compare between them. A study that was done by (Santos-Hermosa, Ferran-Ferrer, and Abadal, 2017) analysed a group of content's indicators related to learning objects that were extracted from 110 different learning repositories. Those learning repositories shared some common criteria such as they serve the higher education level, updated from 2011 and still operating, and contain minimum 50 learning objects. They analysed some main features such as the

disciplines of the learning objects, storage place, educational level, geographical origin, metadata standards, and reuse. This research determined that a lot of the repositories used in the higher education are institutional repositories (IRs) intended for educational purposes (Yassine, Kadry, and Sicilia, 2020a). Those are using open licenses and social networks to guarantee reusing and sharing their learning objects. Another study (Tzikopoulos, Manouselis, and Vuorikari, 2007) assessed the most important common characteristics of well-known learning objects repositories such as MERLOT, ARIADNE, and CAREO. The analysis described three types of characteristics: (1) General and content such as geographical coverage, interface language, and discipline subjects. (2) Technical characteristics such as the offered technical services and metadata specifications. (3) Quality characteristics such as quality control policy, resource rating policy, and copyright policy.

Some studies provided descriptive analysis to learning objects repositories or referatories such as MERLOT which is a learning object refaratory (LORF) (Sicilia et al., 2013) that contains meta-data of free open educational resources (OERs). It was designed primarily to serve faculty and students in higher education (Shmueli, 2017). In 2010, (Cechinel et al., 2010) gathered data from more than 20,000 learning objects in MERLOT and analysed them using some descriptive analysis. The analysis assessed different material types from different perspectives such as growth over time, different ratings given, categories of discipline and personal collections. Ochoa in his study (Ochoa, 2010) provided as well some descriptive analysis of Connexions which is a free and open learning objects repository launched at Rice University. The analysis assessed the contents' growth over time, objects popularity over time and objects reuse distribution.

Other studies delivered descriptive analysis to one or more learning management systems (LMS). A study done by (Song and McNary, 2011) with the objective of assessing online interaction patterns in online courses posted on Blackboard LMS. The authors applied some descriptive data analysis to assess the different types of posts, the changes in posts over time, and the students' differences in amount, type, and pattern of

Posts. In another study, the authors (Costa, Alvelos, and Teixeira, 2012) described the functionalities and tools of Moodle (LMS) that applied and used in a Portuguese

university. They reported some descriptive analysis about the usage of Moodle activities to assess the characteristics of users, the patterns of use and the importance level of each Moodle tool.

Different studies covered descriptive analysis for some Massive open online courses (MOOCs). In one of those studies (Tsai et al., 2018), the researchers investigated the answers of 126 students enrolled in a MOOC for learning Chinese as a second foreign language. They conducted descriptive analysis to assess the relation between the metacognition of the students and their learning interests (enjoyment, engagement, and liking). Those analysis assessed the demographic distribution of the students as well. The study concluded that raising the students' metacognition can help in increasing their learning interests hence it increases their passion to keep learning with MOOCs. In another study (Mclaren and Donaldson, 2018), the authors aimed to evaluate the MOOC as a pedagogical approach in teaching the essentials of healthcare-related subjects. The researchers collected quantitative data for 957 participants through the MOOCs demographic database. They assessed their demographics, learning activities and outcomes. The applied analysis reported many findings such as: 46% of the sample took the course to help their career while 32% took it to help with their academic studies, most of the learners were females with 88% of the sample, around 60% of the participants took the course from their home environment, the majority of them liked the overall learning activities.

Some open courseware initiatives (OCWs) also were assessed using different descriptive analysis. A study was done by (Sheu and Shih, 2017) to evaluate an open courseware initiative implemented in the National Taiwan University (NTU) and to examine its usage. In order to do so the researchers performed some descriptive analysis to assess the courses' characteristics such as number of added courses over time and growing disciplines and to assess the users' characteristics such as their age, gender, type, and number of accessed sessions over time. Another study was done in 2018 by (Balbay, 2018) to assess the behaviour of 50 students in an open courseware (OCW) designed and launched specifically to English speaking skills course in a Turkish University. The author reported some descriptive analysis to assess the distribution of the students among the course units, the types of the course content, the clicks activity, the view duration, and likes and dislikes.

3.3 Online Learning Repositories and Statistical Inferential Analysis

While statistics is the science of learning from data, the statistical inferential analysis focuses on providing an approximation for an unknown that is difficult to be measured through comparisons, tests and data prediction (Ali and Bhaskar, 2016). It aims to draw conclusions about the population, and it tries to give meaning to the meaningless numbers. In 2009, Ochoa and Duval (Ochoa and Duval, 2009) reported the first study that performed quantitative analysis of several types of learning repositories. They analysed multiple learning repositories with different types: (LORP) which is learning object repository, (LORF) which is learning object referatory, (OCW) which is open courseware initiative, and (LMS) which is learning management system. Their extensive analysis covered the following: (1) Size analysis where they applied different statistical and distribution tests such as Lotka, Exponential, and Log-Normal to find the typical size of the repository. This was identified by the number of learning objects published inside the repository and specifically inside each course and the average number of learning objects imbedded in a course then they compared between the repositories. After this step they concluded that the learning objects distribution is very unequal and their concentration is a sequence of the power law distribution while lower average size values were observed to be in LORPs and LORFs. (2) The second type of analysis was the growth analysis. They considered two different ways to measure the growth rate of learning repositories: measuring the content growth by measuring the average growth rate for the added learning objects per day and measuring the contributor base growth. The findings demonstrated that all types of repositories grow linearly with an initial low growing phase. (3) The third type was measuring the contribution distribution which assessed the typical number of contributors that the repository has. They found the size of contributor base is not necessarily related to the repository's size and most of the contributors published only one object. In another study (Costley and Lange, 2016), the researchers aimed to investigate the variables that influence learners' satisfaction, their engagement, and their relationships among three different online learning environments in an online graduate program. They collected data from 216 graduates enrolled in this program using a scaled questionnaire and examined them using some statistical analysis to explore instructor presence, students' relationships and satisfaction. They applied ANOVA test to compare between the mean, variance

of satisfaction and learning scores. The study concluded that student interactions are not significantly impacted by their satisfaction while the rest of the variables were positively correlated with each other. In another study (Lin, Zhang, and Zheng, 2017), the researchers examined students' motivation and learning strategies of group of online language courses. They collected data through a survey and applied exploratory factor analysis to determine the most influence ones. Then they assessed those factors using some quantitative and statistical tests such as confirmatory factor analysis (CFA) which is used to assess the fitness of the factors in generating students' motivation and Chi-square to examine the goodness of the fit. This study found that online learning outcomes cannot be predicted by motivation, but it can be predicted significantly by the implemented self-regulated learning strategies.

Cechinel focused on one type of repositories which is LORFs. He tried in different studies to assess MERLOT repository thoroughly using multiple statistical methods and quantitative measures. In one of his studies (Cechinel, Sánchez-Alonso, and García-Barriocanal, 2011), he obtained information of 35 metrics were extracted from 6,470 learning objects in MERLOT. Those metrics were categorized to measure different classes such as link measures, text measures, and multimedia measures. Multiple analysis were conducted to contrast metrics against the learning objects in the repository including linear discriminant analysis to predict objects' quality classifications. The study managed to classify learning objects to three different statistical profiles which are good, average and poor. They concluded that the type of profile should be determined by the type of rating either peer-review or users' ratings. In a further experiment (Cechinel et al., 2014), he used the same metrics to conduct two experiments with the purpose of developing automated model that assess the quality of learning objects inside repositories according to their intrinsic properties and the available metadata. This method can be used to automatically deliver quality information internally for any new learning resource.

Some research efforts applied different inferential analysis to assess the usage of the learning management system (LMS). Ghilay (Ghilay and Ph, 2019) examined the attitude of lecturers who have different levels of activity towards the main characteristics of Moodle (LMS). The characteristics were examined using Cronbach's alpha to measure the reliability and one-way ANOVA to identify significant differences between the assessments of lecturers. The study found a

significant difference between two groups of lecturers with medium-level and high-level activities. Another study (Mahali, Changilwa, and Anyona, 2019) investigated the impact of the users training level on the LMS utilization in some public universities in Tanzania. Some inferential statistics were applied using SPSS including the correlation matrix, the model fitness, ANOVA, and regression coefficient for the level of training on the utilization of LMS. The study found a significant, positive relationship between level of training and LMS utilization. It found that most of the students in public universities have certificate in LMS training which helped them to utilize LMS in learning without seeking for assistance.

Some researchers focused on analysing other type of repositories which is MOOCs. In a study (Ren, Rangwala, and Johri, 2016), the researchers performed sets of experiments on students' data obtained from three EdX MOOCs in order to develop new approach in predicting assessment scores using personalized linear multiple regression (PLMR) model. This model was built on some properties extracted from the logs of MOOCs server and grouped into: session features such as login and logout, quiz-related features such as number of taken quizzes, video-related features including number of video sessions and number of pause actions per video, homework-related features, and time-related features such as time of playing video. In another study in 2018 (Lee, 2018), the researchers assessed the relation between the uninterrupted time-on-task and the student's academic success. This was done by analysing statistically some variables that represent the un-interruption. For instance, the number of uninterrupted learning activities and their duration which were collected from an EdX MOOC log files. The study analysed activities for more than 4,000 students. This study used the variables as predictors for different nine models of logistic regression. Those models were implemented to compare the accuracy and predictive power metrics (AUC and AUPRC) in order to estimate the likelihood of students to get certificate by the end of the course. The conclusion was that probability of gaining a course certificate is getting higher when the same number of learning activities occurred in fewer learning periods (uninterrupted periods). Recently, some researchers (Oh, Chang, and Park, 2019) reviewed the design of 40 MOOCs specialized in computer science. They were selected from Coursera and EdX in order to assess their pedagogical design using evidence-based e-learning principles. They used a developed instrument with 50 items of e-learning principles to evaluate the

quality of the design for each MOOC. Then they analysed the results using sets of ANOVA tests. The study concluded that the application of the principles is relatively low and most of them tend to lack meaningful interaction and feedback.

Some open courseware initiatives (OCWs) were assessed using inferential statistics. In one of the studies (Wang and Chen, 2012), Wang and his group proposed a theoretical framework called the Theory of User Acceptance of OCW and investigated the factors that affect the user intention of using OCW. They performed correlation and multiple regression analysis to test the significance of the correlation between the perceived behavioural control of using OCW and some external factors such as knowledge and experience, community influence, and channels to raise computer literacy. The study concluded that those factors influence the behavioural attitude and the user intention of using OCW. Another research effort (Yang and Sun, 2013) was done to investigate the influential factors affecting vocabulary acquisition through different OCW lectures. They examined three lectures from MIT and Yale open courses. They analysed the duration and the text of the lectures and they examined the difficulty of the three lectures by performing ANOVA with repeated measures. Results of the analysis highlighted the following: the learners were able to gain the knowledge of vocabulary by watching the lecture once, the level of vocabulary, which includes academic, technical, and low-frequency vocabulary, is the most influential factor that affects the vocabulary acquisition, and the frequency of occurrence had small positive effects on learners' vocabulary gain.

3.4 Online Learning Repositories and Social Network Analysis (SNA)

Network science domain has been developed significantly through its research evolution. It is highly used to discover and assess the features of large scale networks (Newman and Girvan, 2004). Social network analysis (SNA) is considered the application of network science in social networks (Brandes, 2015) which can be defined as groups of members (nodes) tied by one or more types of relations (Lazega, Wasserman, and Faust, 1995). By applying social network analysis, the network structure can be comprehended as the patterned organization of users and their relationships that will lead to explain the impact of such patterns on behaviours and attitudes (Wellman and Gulia, 2018). Applying social network analysis in online

learning environment can enrich the understanding of learners' behaviours, the nature and type of their interactions and help in optimizing the instructional design (Cela, Sicilia, and Sánchez, 2015).

A methodological SNA framework was developed (Corallo et al., 2010) to monitor the change over time in a virtual learning community in a master's online course. It was built on two dimensions: (1) individual growth-level which was assessed through the following SNA metrics: betweenness centrality, degree centrality and contribution index. Group growth-level which was assessed through the metrics: Group's betweenness centrality, degree centrality, density, core/periphery structure. The purpose of this framework was to perform as a warning system to leaders to improve mentors' availability to support learners in their learning path. While in another recent research attempt (Christoforos et al., 2019), the authors introduced a social network analysis (SNA) toolkit that supports understanding the dynamics of the classroom social network. This toolkit performs by finding different social network measures then developing network maps for each classroom. The toolkit was tested and validated on a data collected from grade-8 classroom social network. More applications of social network analysis were performed to discover types and patterns of interactions and to understand the social dimension of the learning in online learning contexts. SNA techniques were applied on an undergraduate Biology online classroom (Grunspan, Wiggins, and Goodreau, 2014) to discover students' network and assess the relation between their position in the network and their success on exams. The study applied several SNA measurements on two exam study networks for the same group then compared between them using correlation analysis, degree and betweenness centralities. The study found a social influence on the exam's performance when the actors revised their network positions by changing the studying patterns between the two exams.

Some research efforts attempted to analyse and explore LORs using social network analysis techniques. Sicilia and his research group (Sicilia et al., 2009) used social network analysis tools to analyse the whole community structure of a dataset obtained from MERLOT repository. They assessed the structure of the network using some basic measurements such as number of ties, diameter, and density and they examined the central actors in the network using centrality measures that include in-degree centrality, out-degree centrality, betweenness and closeness. Another research group

(Zervas, Alifragkis, and Sampson, 2016) attempted to study the co-tagging networks in a LOR named OpenScienceResources (OSR) by applying social network analysis (SNA) measures. They used different centrality metrics including degree centrality, closeness, betweenness, and eigenvector. The study managed to identify the taggers with high centralities who can strongly impact the tagging contribution and create crowded clusters of taggers.

Other efforts focused on analysing interactions and data collected from learning management systems (LMS). In 2010, (Dawson, 2010) one of the pioneer researchers reported and visualized the peer to peer students' interactions in a BlackBoard (LMS). He used SNA techniques to identify patterns of network behaviour that affect the learning. SNA analysis investigated the patterns on interactions, the degree of connectedness and the ego-networks of actors who participated in a Blackboard online course in which their data was extracted from the communication logs. The study highlighted the clear appearance of the teaching presence in the network. Results and procedures of this study helped in developing the social networks adapting pedagogical practice (SNAPP) which is SNA tool developed for online learning environments (Bakharia and Dawson, 2011). Another theoretical model was developed by (Paredes and Chung, 2012) to understand the engagement of the knowledge of learners in influencing the learning and performance. The proposed model investigated interactions collected from WebCT (LMS). It was built on social learning and social network theories and it was driven by content-based measures and social network analysis (SNA). The experiment analysed the egocentric network properties such as structure, position and tie for a network of industry professionals' students enrolled in an online course using different SNA measures such as density, contribution index, external-internal index, content richness score and average tie strength. This study concluded that both the performance and the social learning are highly impacted by the learners' network of contacts. A research group of another study (Saqr, Fors, and Nouri, 2018) assessed the online interaction data collected from Moodle (LMS) database of four online courses using different mining and SNA techniques in an attempt to investigate the online problem-based learning and to predict the learners performance. The findings demonstrated a strong association between the students' performance and their centrality measures. By using SNA

indicators, the authors were able to categorise students based on their achievements with high accuracy.

Some application for SNA techniques was implemented to analyse MOOCs networks. In a research effort was done by Norman and his research group (Norman et al., 2018), they developed a blended instructional design that combines xMOOCs, cMOOCs and social network analysis. The xMOOC was the course material and its learning objects, and the cMOOC was social media platform used as a discussion space. Different SNA tools were imbedded as the analytics that analyses and assesses the students in each phase of this course. In another recent research (Lu, Liu, and Zhang, 2020), the authors used SNA techniques to discover the structure, information diffusion potential, and the vulnerability in the network of four MOOC courses. They compared between the networks in terms of the number of nodes and edges, and their average degree. They used text classifiers to analyse the users post in those courses to categorize the posts based on their relation to the course content. The results indicated that analysing learners' social behaviour is crucial, and it help in establishing different guiding mechanisms. Also, they concluded that the network structure is determined by the course features, the guidance of the teacher, and the learner behaviour and background.

Other applications of SNA techniques were assessing open courseware (OCW) initiatives and their learners' interactions. A research study (Tovar et al., 2013) used SNA techniques and measures to examine the impact of learning through OCW initiatives in Spain and Latin America. The authors assessed the networks structure using different metrics including density, diameter, and average path length and they measured degree and betweenness centralities to different tags and knowledge areas to detect the most influential labels in Spain and Latin America. The results concluded that SNA analysis techniques can be used to analyse the current OCW's state and the potential one that it may have through analysing implicit and explicit associations. Another research group (Piedra et al., 2015) proposed a fundamental method to describe, analyse, and visualize knowledge sharing on OCW initiatives. They used semantic technologies and linked data guidelines. They applied different SNA measures to explain the social interactions between individuals. Those measures include closeness, betweenness, and eigenvector centralities.

3.5 Community Detection Algorithms and Techniques

Community detection process is an effective processes of social network analysis (SNA) which performs by detecting clusters and communities in a social network. Those networks are built up of connected nodes which can be classes, users, learning objects, or any other entity. Those nodes are linked in a graph presentation where they define the vertices of this graph and their connections define the edges. Community detection (Kelley et al., 2012) is the process of identifying those clusters of interacting nodes according to their structural features. Until now, the ultimate challenge in this field is that there is no universal definition for the community structure (Fortunato and Hric, 2016) and that is why community detection techniques methods and algorithms are still attracting the research efforts in order to explore them and investigate their use in many fields such as healthcare, computer science, social sciences, and education.

Community detection techniques gained a lot of research interests. Most of them have been derived from various research efforts applied in several study fields. Many research efforts tried to group the techniques identified in the literature based on different perspectives. We will demonstrate briefly one of the widest grouping attempt to categorize community detection techniques based on their mechanism and work dimensions (Javed et al., 2018):

- Traditional techniques: Those are clustering, and partitioning based methods used to explore the disjoint communities. The time complexity is the main pitfall of those techniques especially if performed on large complex networks. Traditional techniques include graph partitioning, partitional clustering such as k-mean, k-sum, and k-median, hierarchal clustering such as agglomerative and divisive algorithms and spectral clustering such as Laplacian spectral partitioning.
- Modularity-based techniques: those methods operate by optimizing the quality function for communities' approximation. This function is called modularity. Optimizing the modularity means the higher modularity value the better partitioning is. Modularity techniques include: Greedy Optimization method (GN) proposed by (Newman, 2004) This was the first algorithm applied to optimize modularity. Another method proposed by (Blondel et al., 2008) and it was called

Louvain method. This is a heuristic method to detect uncovering communities in complex weighted graphs (Khan and Niazi, 2017).

- Overlapping community detection techniques: most real networks contain nodes belonging to more than one community at a time, which is why they are called overlapped communities. The popular and effective overlapping community detection techniques include: Clique percolation method which performs by forming subgraphs of cliques that are connected through internal edges (Derényi, Palla, and Vicsek, 2005) and Label propagation algorithm (LPA) (Raghavan, Albert, and Kumara, 2007) which is effective because it performs through associating the vertices with same labels to form community.
- Dynamic community detection techniques: Those techniques revise the evolving and changing of the nodes in the network over time (Shang et al., 2016). Random walk techniques such as PageRank, WalkTrap, and Infomap algorithms are famous examples for dynamic community detection techniques.

3.6 Online Learning Repositories and Community Detection Methods

In a process to complete the whole picture of our research, we explored the literature in a systematic review which is the final revision phase in order to identify the volume of research in detecting communities in online learning environments. In this study we searched different well-known databases for research efforts in detecting communities in online learning repositories during the period from 2011 until June 2020. We identified 65 studies that met our criteria. We investigated their application to community detecting methods and defined their main research goals. We highlighted the important role that identifying learning communities plays in several educational research topics. Our findings revealed a great potential in using community detection techniques to identify learning communities which help in developing and improving the educational research and in utilising the massive amount of data generated from interacting with online learning environments. During our assessments, we found that the most investigated types of learning environments were Learning Management Systems (LMS), MOOCs, and various social learning environments (SLE). While the top widely used techniques to identify communities

and clusters were K-means Clustering, Clique Analysis, and Louvain Method. We will demonstrate here some of the identified research efforts.

In 2018, a study was implemented (Adraoui et al., 2018a) to predict at-risk learners through evaluating the community of learners in some Facebook learning groups (SLE) based on their relations and interactions. They simulated a database of different interactions for 2,000 learners. Those interactions included 10,675 comments' interactions and 40,632 reactions' interactions (likes). They applied Louvain Blondel method to detect communities then they visualized them using Gephi software[2], Force Atlas2 layout. They identified 31 communities via comments interactions and 9 communities via likes and analysed them through applying some centrality measures (degree, betweenness, and closeness). The results of this research were assessing the learners' communities and identifying the safe and at-risk ones. Another attempt was done by the same group of researchers (Adraoui et al., 2018b) to evaluate students in a Moodle (LMS) online discussion forum. They extended the same previous techniques to include one more layout algorithm (Fruchterman-Reingold algorithm) to detect communities and cluster a network that include 117,988 interactions for 4,000 learners. The study identified the status of each community (safe or at-risk). Recently, this research group proposed a new algorithm (Adraoui et al., 2019) used to detect communities and assess their features (EDCA). This algorithm performs in two phases: (1) detecting safe learners in the network who can easily interact and exchange information with others. This can be done using a new defined centrality measure also proposed by them and called "safely centrality". (2) Identifying the communities by detecting the neighbours of those safe nodes through many iterations based on their modularity. The published study (Adraoui et al., 2020) represents an application of the new proposed algorithm to evaluate and detect learning communities in two different datasets of learning networks: the first one includes learners' interactions from a German school and the other one includes learners' interactions from a computer engineering online course. First, they detected the safe learners then they identified their communities with their status to detect the at-risk groups. The experiment compared between this algorithm and other three known community

[2] Gephi Software: https://gephi.org

detection techniques (leading eigenvector algorithm, Infomap, and Fast greedy). Comparison criteria included modularity and other performance metrics.

In another research effort (Jimoyiannis and Angelaina, 2012), the authors developed a framework for assessing students' interactions for 21 grade-9 students engaged in blog-based learning activities. They applied some content analysis on the students' publications to determine their categories of presence (social, learning, and cognitive). Then they applied cohesion analysis using Cyram NetMiner software[3] to explore the network's structure and detect the students' cliques. In 2013, the same main author (Jimoyiannis, Tsiotakis, and Roussinos, 2013) extended the previous research to include a larger network of the blog's users. He applied the same previous cohesion analysis on them and expanded the analysis to include degree centrality as a power analysis to present the power distribution among the communities of users. In 2017, (Tsiotakis and Jimoyiannis, 2017) investigated the teachers' communities or cliques in a computer science online course for master's degree using the same framework and techniques applied previously with betweenness centrality as a power analysis technique. The study detected 58 teachers' sub-groups. Most of their users had a significant contribution in those cliques.

More applications of community detection and clustering techniques were done to help in identifying the learners' behaviour. The research effort presented by (Wang, 2018), proposed Eigenvector Label Propagation Algorithm (ELPA) which is an upgraded version of the famous overlapping communities detection algorithm Label Propagation Algorithm (LPA). The proposed algorithm operates using the eigenvector feature as the node label which represents a combination of 8 features. It is designed to help in discovering interactive learning students' communities based on their social networks in m-learning environments. The study demonstrated an experimental comparison between applying the proposed algorithm with two other algorithms (GN and BMLPA) on a social learning network. The results revealed that ELPA performed better in terms of time and space complexity and its detected students' groups have higher fit degree. Another effort (Kovanović et al., 2019) applied clustering techniques to assess the learning strategies and the students' learning experiences in

[3] Cyram NetMiner software: http://www.netminer.com/product/overview.do

a computing EdX MOOC. They extracted course engagement measurements and other cumulative ones from the trace data, and they examined them with the students' final grades. The study applied the agglomerative hierarchical clustering algorithm to detect communities in the network. Then, some statistical analysis such as ANOVA and MANOVA were implemented to assess the differences between the clusters in responding to some identified measures. This study found significant differences in the students' commitment to learning and in the perceived level of cognitive presence. While it failed to find major differences in the social presence and the learning presence. K-means algorithm was used in this study (Wang and Zhang, 2019) in order to cluster the online learning behaviour of students enrolled in 'Principles of Database' blended course offered by the school of information in Beijing. This course was built on online teaching platform including instructional videos, online exercises, tests and discussions. A correlation analysis was performed to detect the relation between the students' results and six learning characteristics indicators extracted from the dataset. Then, k-means clustering algorithm was carried out to detect the differences in behaviour in different groups and the influence of those differences on the learning setting. In a recent study (Wang and Wang, 2019), the researchers proposed a binary graph community detection algorithm (BGCD) that uses the bigram conditional probability to analyse the relationships and interactions between the learners to predict whether they will drop out the course or not. To validate the performance of this new algorithm, an experiment that covered four MOOCs was carried out and its effectiveness was compared to other algorithms in terms of accuracy, recall, precision and F1 values. This study showed that the prediction accuracy of the new algorithm is considerable, and it can be reliable so the teachers can depend on it in guiding the learners during their learning path.

4. Methodology

This research was prepared and developed according to Creswell's framework (Creswell, 2003) for research design which outlines characteristics of quantitative method research approach. This approach includes problem statement, hypothesis formation, literature review, and quantitative data analysis. According to Creswell, the quantitative research is a process of collecting, analysing, interpreting, and writing the results of a study. This type of research utilizes different investigative strategies and collects data on pre-defined instruments that produce statistical data. Whereas, the findings can be predictive, explanatory, and confirming. We followed his framework in identifying the research problem, reviewing the literature, stating the research objectives, collecting data, analysing the obtained data, reporting the findings, and evaluating the research.

In 2016, we started to review the literature extensively to understand the scope of the research in assessing and investigating online learning repositories and their user interactions. During this phase, we reviewed the history of the learning repositories, their types and the several ways and techniques used to assess their quality and user interactions. Then, we selected to study a popular learning repository which is Khan Academy to be our experimental repository where we apply all our analysis, investigations, and examinations. Khan Academy is a free and open initiative that was introduced in 2008 by Salman Khan (Thompson, 2011). The main purpose of it was to construct a set of online tools to help in educating students by addressing a wide content scope. At the beginning, this initiative started by publishing short video lessons on YouTube. After that, Khan Academy evolved its own dedicated platform in 2009 which provides full different educational lessons. The decision of selecting Khan Academy was built due to the following reasons:

- Khan Academy's repository is a free, open source, wide-reaching and recognized repository. It serves a varied range of users' segmentations starting from KG user-level to college user-level in addition to the parents, teachers, and other stakeholders.
- This repository consists of large number of learning objects in different fields. Those learning objects can be reused and shared in different sites to support learning (Hodgins and Wiley, 2002). It includes different types of learning

objects such as instructional videos lessons, articles, and exercises. Those can be used for presentation and practice purposes (Churchill, 2007).

- Khan Academy's repository is a concrete learning initiative (Sicilia et al., 2013) that has some common features with the different learning repository types such as a referatory, LMS, MOOC, or OCW but still it cannot be considered as any of them. This initiative can be used as a complete separate online learning environment as well as it can be used as an interactive tool in a learning blended setting or in a flipped classroom.
- In the literature, there is a lack of detailed and in-depth studies and analysis for Khan Academy's instructional videos repository, their data structure and all the interactions related to them.
- This repository has a large and diverse users' base. All their interactions with the instructional videos are public, accessible and can be gathered using scraping techniques as well as it has a public API which was used in building the structure of our dataset.

To be able to assess and investigate Khan Academy's repository, we started the data acquisition phase after planning, designing, and creating the tools needed for that. More details regarding the data acquisition phase are demonstrated in chapter 5.

Based on the gathered dataset, a combination of descriptive analysis was implemented in order to describe the data, simplify it, organize it, and present it in a meaningful manner (Loeb et al., 2017). The descriptive analysis presented in chapter 6, tries to enforce the comprehension of the nature of the data, its characteristics, and intrinsic features. It tries to identify how the repository, its learning objects and its users-base evolved over the years. Also, we performed inferential analysis techniques including Pearson test, Phi & Cramer's V test, and ANOVA regression to investigate the relation of learning materials and users' interactions. We tried to assess the behaviour of users' interactions toward some video-related and interactions-related metrics and to examine the association between them. In order to be able to do so, we classified the instructional videos (learning objects) into different profiles according to their user's interaction levels then we examined the significance of association against those profiles.

A deeper investigation through the users' interactions was presented in chapter 7 by drawing their network graph and applying two different community detection techniques which are Parallel Louvain method (PLM) and Parallel Label Propagation (PLP) Algorithm in order to identify the emerged online communities, study their characteristics and compare between them. The decision of choosing these two techniques came after investigating the literature for the most applied effective techniques. Parallel Louvain method (PLM) is the parallel implementation of the well-known Louvain method which was found to be one of the highly used techniques to detect disjoint communities in online learning settings. It demonstrated effectiveness specially with the large populations because of its concept of optimizing the modularity (Ghosh et al., 2019). Parallel Label Propagation is also the parallel implementation of Label Propagation Algorithm which is widely applied to detect the overlapping learning communities in online learning networks. This algorithm proved its efficiency especially in large-scale complex social networks (Garza and Schaeffer, 2019). In chapter 8, we examined the network structure and the characteristics of the emerged learning communities using different SNA measures including density, modularity, network diameter, and average path length. We assessed the users learning behaviour and identified the influential actors in those communities by performing some centrality measures such as degree and eigenvector centralities.

The literature review is presented in Chapter 3. It covers the previous research work on the contents, structure, and evolution of online learning repositories, their quality, and their types. Also, it includes different studies attempted to assess the different types of repositories (referatories, LMS, MOOCs, and OCWs) using descriptive analysis, inferential analysis, community detection methods, and different SNA measures. The work demonstrated in Chapter 9 presents conclusions according to our research objectives. Chapter 10 points out some possible future insights. Based on our work, we can conclude that our research demonstrates one of the first studies that sheds light on investigating and examining the characteristics and processes inside open learning repositories and detect emerging online communities through users' interactions.

Our analysis aimed to help us and other researchers to comprehend the features of the learning process inside learning repositories, the evolution of online communities of interactions and comprehend their behaviour. The three chapters from 6 to 8, will summarize our findings throughout all the stages of our research. Those results were disseminated in several published works. Our results meant to explore the objectives mentioned above. Thus, those chapters will cover three sub-objectives as explained in the **Table 1** below:

Table 1. Research Results vs. Research Objectives

Results' Chapter	Research Objective
R1. Assessing Online Learning Repository with Descriptive Statistical Analysis	O2. To use different descriptive statistical analysis in order to investigate a representative example of online learning repositories and to assess the relation between learning materials and users' interaction. This helps in understanding its characteristics, the potential impact on learning process and gives some useful insights to help in assessing the instructional resources and evaluating their quality.
R2. Detecting Communities in Online Learning Repository	O3. To apply different community detection techniques to identify the learning communities that are generated from users' interactions inside learning repositories. To study the structure of those communities and to assess their movements and engagements in order to discover their presence and develop a better understanding to the online learning setting.
R3. SNA Measures and Users' Interactions	O4. To examine users' interactions with the learning repository by applying SNA techniques and measures to produce valuable information for educators and researchers that may increase their comprehension to online learning through interactions.

5. Data acquisition

Part of the material of this chapter was published in (Yassine, Kadry, and Sicilia, 2020a) and (Yassine, Kadry, and Sicilia, 2020b). In May 2016, we started to design a PHP script to scrap Khan Academy's (KA) repository. It took around three months to finalize a clean and a functional version of our PHP scraping tool. This tool managed to traverse Khan Academy's structure deeply from the top level of it which is called domain until it reached their instructional videos and gathered all the non-authenticated data related to them such as the skills, video duration, date added, users' interactions which are questions and answers posted on the videos, and some details about the users' who post them. The tool is available at: https://github.com/SaharYassine/KA_Scraper. We created MySQL database to start scraping the repository and gathering all the data in it. The structure of our database was built based on Khan Academy's public API. **Figure 1** shows ER diagram form Khan Academy's database structure.

Figure 1. KA Database ER Diagram

The database was updated every six months to gather the new additions. By August 2019, we managed to extract 3,284,510 user interactions which are questions and answers posted on instructional videos. Those were posted by 359,163 users on Khan Academy's instructional videos.

5.1 Dataset

The scraping mechanism which also represents the database structure begins with Khan Academy's eight domains which are related to various study fields and different learning purposes: math, science, computing, humanities and arts, economics, partner content, Test- Prep and college, careers, and more. Each of them covers different subjects while each subject includes several topics. This repository contains 985 topics related to 127 different subjects in the eight domains. Each topic consists of various related instructional videos, exercises, and articles which are called skills. All the gathered data are public and can be reached by any user without any kind of authentication. An important part of the collected data is users' posts and interactions with instructional videos which are their questions and answers posted on the video lessons. We focused on this type of users' interactions and we analysed it disregarding the content factor and considering that all posts are relevant to the video's content. We managed to collect more than 3M users' interactions in the form of questions and answers. Those were posted by more than 300,000 users and we collected the non-authenticated details for 22,000 users out of them. The gathered interactions were posted on several Khan Academy's videos which are related to 985 different topics in Khan Academy. **Table 2** shows the details scrapped from Khan Academy's website while **Table 3** shows some examples of the scraped subjects, topics and skills. The gathered data were posted on the website in the period between February 2011 and August 2019.

Table 2. Number of the subjects, topics, videos, and scraped interactions

Domain Name	Subjects	Number of Topics	Number of Videos	Total Collected Posts	Number of Questions	Number of Answers
Math	57	446	11,364	2,242,408	1,175,703	1,066,705
Science	11	146	3,068	546,462	262,776	283,686
Test Prep	7	44	2,037	178,063	91,102	86,961
Arts and humanities	12	91	1,755	140,895	61,299	79,596
Economics and finance	5	40	750	87,390	43,037	44,353
Partner content	28	169	1,353	42,530	20,955	21,575
Computing	3	17	74	30,677	16,102	14,575
College, careers, and more	4	32	498	16,085	8,414	7,671
Total	**127**	**985**	**20,899**	**3,284,510**	**1,679,388**	**1,605,122**

Table 3. Examples of Subjects, Topics, Sub-topics, and Skills in KA-domains

Domain	Subjects (Sample)	Topics (Sample)	Sub-Topics (Sample)	Skills
Math	• Early math • 1st grade • 2nd grade… • Basic geometry • Pre-algebra • Algebra basics	In Early math Subject: • Counting • Addition & subtraction • In Pre-Algebra Subject: • Arithmetic properties • Fractions & Decimals	In Counting Topic: • Counting • Numbers 0 to 120 • Counting objects • Comparing small numbers	In Counting Sub-Topic: 3-Video Skills 2-Exercise Skills
Science	• Physics • Chemistry • Biology • Health & medicine	In Physics Subject: • Fluids • Magnetic forces • In Chemistry Subject: • Atoms & compounds	In Fluids topic: • Density and Pressure • Buoyant Force and Archimedes Principle • Fluid Dynamics	In Newton's law of gravitation Sub-Topic: 6-Video Skills 3-Exercise Skills
Computing	• Computer programming • Computer science • Hour of Code	In Hour of Code Subject: • Drawing with code • Creating webpages • Creating SQL databases	In Algorithms Topic: • Intro to algorithms • Binary search • Asymptotic notation • Selection sort	In Make your webpages interactive Sub-topic: 1-Video Skills 2-Exercise Skills
Arts and humanities	• US history • World history • Art history • Music	In Art History Subject: • Prehistoric art • Medieval Europe & Byzantine	In Prehistoric art Topic: • Paleolithic art • Neolithic art • Quiz: prehistoric art	In Early colonization projects Sub-topic: 5-Video Skills 2-Exercise Skills
Economics and finance	• Macroeconomics • Microeconomics • Finance and capital markets	In Microeconomics: • Supply, demand, & market equilibrium • Consumer and producer surplus	In Basic economics concepts Topic: • Scarcity • Opportunity cost • Demand & Supply	In Scarcity Sub-topic: 5-Video Skills 1-Exercise Skills
Partner content	• The Museum of Modern Art • Asian Art Museum • Stanford School of Medicine	In Asian Art Museum: • South Asia, Southeast Asia • The Himalayas & Tibetan Buddhist • Hinduism	In American Museum of Natural History Topic: • Dinosaurs • The Universe • Human Evolution	In the Sun and solar storms Sub-topic: 10-Video Skills 8-Exercise Skills
Test Prep	• SAT • MCAT • NCLEX-RN • GMAT • CAHSEE	In SAT Subject: • Full-length SAT • SAT Math practice • SAT Reading & Writing practice	In SAT Reading and Writing practice Topic: • Reading • Writing: Passages • Writing: Grammar	In Principles of bioenergetics Sub-topic: 6-Video Skills 2-Exercise Skills
College, careers, and more	• College admissions • Careers • Personal finance • Entrepreneurship	• In College Admissions: • Getting started • Exploring college options • Applying to college	In Getting started Topic: • Introduction: College admissions • Importance of college	In Importance of college Sub-topic: 5-Video Skills 0-Exercise Skills

5.2 Data Preparation

We needed to clean the data and prepare it for our analysis. For that reason, we created a Python data preparation script to execute several pre-processing steps to enhance the data and represent it in a more convenient format. Some Python's libraries were used to clean the dataset such as Pandas and NumPy. We cleaned the data from NULL, duplicates, and misleading values.

We used the extracted 3,284,510 users' interactions to create our data frame and we encoded the variables for consistent processing. Each user who posted a question or an answer represents a node. Total number of nodes (users) involved in this dataset is 359,163 users while the total number of edges between them is 621,226 with different weights. The weight of the edge represents the total number of posts (interactions) took place between two nodes (users).

NetworkX library (*NetworkX*, 2008), which is a well-known Python library used for studying real-world networks and graphs, was used to create the network graph. We defined our graph as $G = (V, E, W_E)$, where V is the set of nodes (users) that were extracted from the dataset, E is the group of edges that represents the interaction or the relationship between the users (nodes), and W_E is the weight of those edges which reflects number of interactions (questions and answers) between each two nodes.

After creating the graph, we tried to keep working with NetworkX to perform the community detection techniques and social network analysis, but it was highly consuming for both time and resources due to the big number of the processed data. This is due to the fact the NetworkX has pure Python implementations which makes it not suitable for such large network (Staudt, Sazonovs, and Meyerhenke, 2016).

Therefore, we searched for a more capable tool which has the ability to work with large-scale complex networks more effectively. NetworKit (*NetworKit*, 2013) is a growing open source toolkit for largescale network analysis. We found it to be capable to process large networks in parallel computation for multiple of analytics kernels very fast and with a lower memory footprint than NetworkX (Staudt, Sazonovs, and Meyerhenke, 2014). NetworKit was designed and implemented as a two- layer hybrid of performance- aware code written in C++ with additional functionality and interface written in Python. It is distributed as a Python package and can be used from a Python shell. This hybrid architecture allows NetworKit to process quickly in

parallel threads and to integrate easily with other data analysis Python's libraries such as Pandas and NumPy.

In order to present, visualize and analyse our data we used several software packages and applications. **Table 4** summarize the used tools.

Table 4. Summary of the used tools in the research

Tool	**Used for**
Tableau[4]	Visualize Descriptive Analysis
R-Studio[5]	Visualize Descriptive Analysis
MATLAB[6]	Apply Statistical Tests
NetworkX [7]+ Libraries	Data preparation, creating a graph and applying SNA measures
NetworKit [8]+ Libraries	Data preparation, creating a graph and applying SNA measures
Gephi[9]	Drawing the graph, the detected communities and SNA measures

[4] Tableau: https://www.tableau.com

[5] R-Studio: https://rstudio.com/products/rstudio

[6] MATLAB: https://www.mathworks.com/products/matlab.html

[7] NetworkX: https://networkx.org

[8] NetworKit: https://networkit.github.io

[9] Gephi: https://gephi.org

6. Assessing Online Learning Repository with Descriptive Statistical Analysis

This chapter is based entirely or partially on the material that has been already published in (Yassine, Kadry, and Sicilia, 2020a). We conducted descriptive analysis to describe Khan Academy's repository, organize the collected data, and present it in a meaningful manner that gives indication and useful information regarding its growth, its features, and the characteristics of its users. Our descriptive statistical analysis was allocated in three categories: General Descriptive analysis which is related to the performance and evolution of KA's repository, interactions-related analysis which investigates the users' interactions with Khan Academy's videos, and inferential analysis which assesses the relation between learning objects and users' interactions.

6.1 General Descriptive Analysis

6.1.1 Growth of KA repository

Khan Academy started in 2008 through YouTube and Yahoo Doodle Images but its videos' repository became well known in 2011. Most of the field-related domains such as 'math', 'science', 'economics & finance', and 'arts & humanities' started in 2011 while other domains that are designed for specific purpose were added in 2014. Those 3-domains are 'Test Prep', 'partner content', and 'college, careers, more'. Despite that Khan Academy's team started building those domains and adding videos to them before 2014 but the official announcement for those domains was in 2014.

Following to Ochoa's work in applying growth analysis to different types of learning repositories that was reported earlier in the literature review. We tried to measure the growth of Khan Academy using two different dimensions: measuring the content growth and measuring the user-base growth.

Content Growth over time:

To measure the growth of number of learning objects, all the videos (skills) that were gathered in our dataset were included and their publication dates were tested using four fitted models linear($at + b$), logarithmic($b + \ln(at)$), polynomial($at^d + bt + c$) and exponential($b * e^{at}$). We found that the growth of the learning objects in

Khan Academy follows a linear trend over time (**Figure 2**) which is similar, according to Ochoa's study, to the growth of other types of repositories. But interestingly, the publishing trend in Khan Academy started with the largest amount of the added learning objects in 2011 which is unlike what was found by Ochoa as most the learning repositories tend to start with initial low growing.

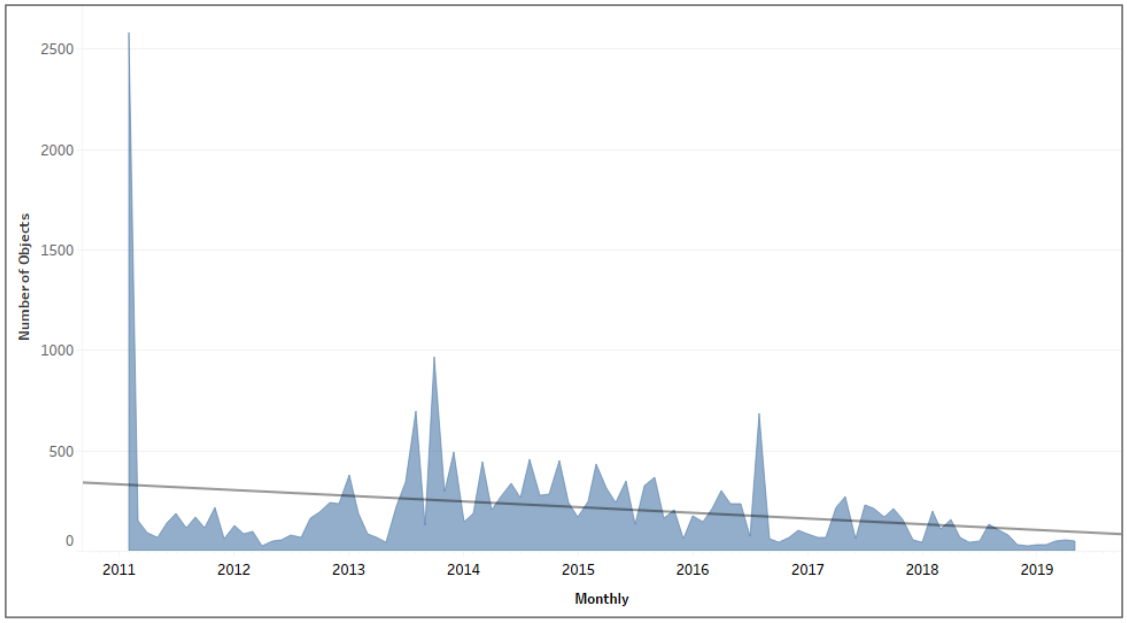


Figure 2: Khan Academy's Content Growth over Time

Growth rate of this repository can be figured out from the number of added videos over years (**Table 5**). This rate differs from one domain to another. Below we will have a brief explanation to the trend of each domain:

Table 5. Khan Academy's Repository Growth over time (by Domains)

	Year Added								
Domain Name	2011	2012	2013	2014	2015	2016	2017	2018	2019
Arts and humanities	54	98	283	202	108	286	349	366	44
College- careers- and more	31		26	146	5	17	193	80	
Computing		28	10	21	13	1		1	
Economics and finance	373	218	33	9		9	6	102	10
Math	2,667	699	2,690	1,084	1,420	1,417	1,096	291	119
Partner content	5	1	71	596	365	173	43	99	
Science	636	209	488	729	526	403	32	45	44
Test prep	125	151	292	772	557	21	80	39	

- Math is the oldest and largest domain; the first big addition of videos was in 2011 when Khan Academy's team added 2,667 videos to the repository. They added 2,690 math videos in 2013. From 2014 to 2017, KA team maintained the same rate of addition yearly which is around 1,000 to 1,400 videos. In 2018, they added only 291 videos while only 119 videos were added during 2019. This growth pattern gives an indication that the math domain reached to its maturity level and there is only a minimal room for any extra addition.
- Science domain started in 2011 by posting 636 videos related to different science subjects. Another big addition was implemented in 2014 when Khan Academy's published more 729 videos. This was the peak in the publishing rates. After that, the addition started to decline until 2019 when they added only 44 videos.
- Arts and humanities domain can be considered as one of the fast growing domains. It started by posting 54 videos in 2011. Khan Academy's team continued adding videos to this domain in an increasing rate over the years until 2018 when they added 366 videos. Then it declined in 2019 when they added only 44 videos.
- Economics and finance domain's growth can be considered as a declining rate. The domain started by publishing 373 videos in 2011 which was a good start. Then the addition started to decline until the publishing stopped in 2015. Very humble additions were made through 2016 and 2017. In 2018, Khan Academy's team tried to boost up this domain again by adding 102 videos, while they added only 10 videos in 2019.
- Computing domain is not growing at all. In 2012, Khan Academy's team started to add very small number of videos in this domain. They stopped adding videos in 2017 then they added only one video in 2018. This slow addition is not only in the videos learning objects but it covers also the exercises which means that the whole domain is not promising in Khan Academy's learning environment and it cannot be considered a competitor online learning provider in the computing field.
- Partner content domain was officially announced by NASA in May 2014. This domain demonstrated a new cooperation between NASA and Khan Academy to deliver STEM opportunities to online learners through adding dynamic educational materials to Khan Academy's repository (Loff, 2014). Although we can find some addition before 2014 and that was due to preparation and reusing reasons, but we can notice that the growth in publishing videos started from 2014.

- In 2015, Khan Academy and College Board Organization announced their cooperation in creating free SAT study tools to be published as a new domain called 'Test Prep' on Khan Academy's repository (*Free SAT Practice from Khan Academy*, 2018). Many 'Test Prep' videos were added earlier than 2015. This was to serve learners in different study fields as well.
- 'College, careers, more' domain was announced in 2014. This was created in order to serve the high school students and college counsellors and to provide them with college admissions resources. The purpose of this domain is to guide the learners through their university application processes and to assist them in navigating different academic options.

User-Base Growth over time

Another dimension to consider the repository growth is to measure the user-base growth over the time. The user-base can be described using two different ways: describing the number of students joined the repository over time and describing their activities and interactions over time. **Figure 3**, demonstrates the joining year for more than 38,000 users who joined Khan Academy from 2011 until 2019. The result of the fitting indicates that the distribution was best explained by the 2-degree polynomial distribution model.

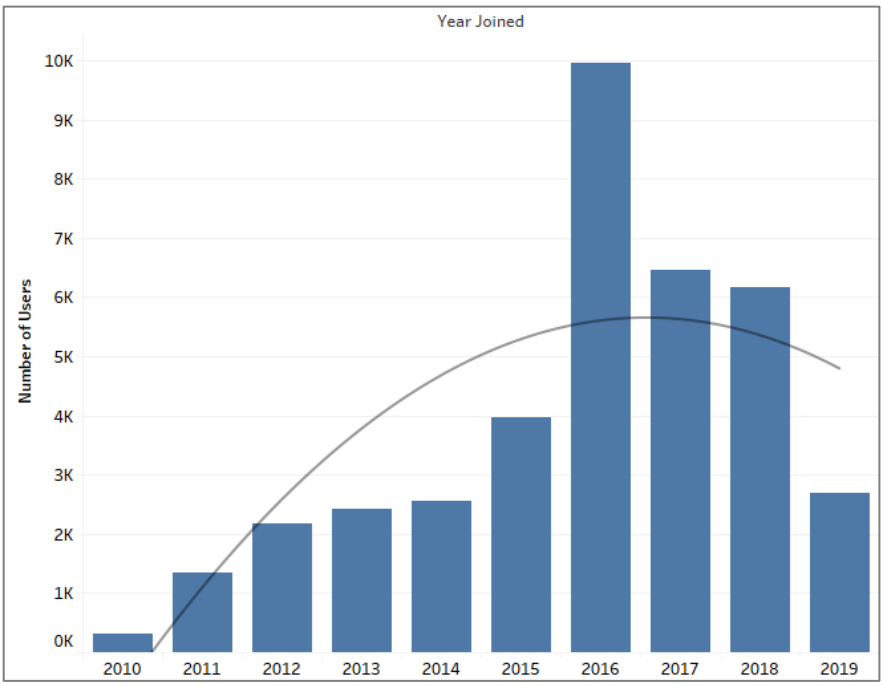


Figure 3. Describing the user-base growth by the number of joined users in KA over time

In our dataset, we found 319 users who joined the repository in 2010 who were most probably testing cases. In 2011, 1340 users joined the repository and still active. In 2012, the number of joined users started to increase gradually until 2015. The exponential increase happened in 2016 when the number of joiners jumped from 3,971 users in 2015 to 9,949 users in 2016. This was the peak, after that the number of the joined users started to decrease gradually until it reached 2,698 users in 2019. This drop in the number of joiners may have many reasons such as the availability of wide variety of alternative open repositories which increase the competition and other possible reason is raising the interest of having institutional online learning initiatives in a big range of academic institutions which requires involvement from their users.

The second way that can be useful in measuring the use-base growth is describing their activities and interactions over time. As we mentioned previously, we gathered more than 3.2M users' interactions grouped in questions and answers

Figure 4. After applying several fitting procedures, the best fitting was found to be the normal distribution. Users started interacting with Khan Academy's videos since 2011 as we collected 82,833 posts which were added in that year. The observed increase in interactions started after that gradually until they reached their maximum in 2016 (647,129 posts). In 2017 interactions level started to decline until 2019 when it reached 140,824 posts.

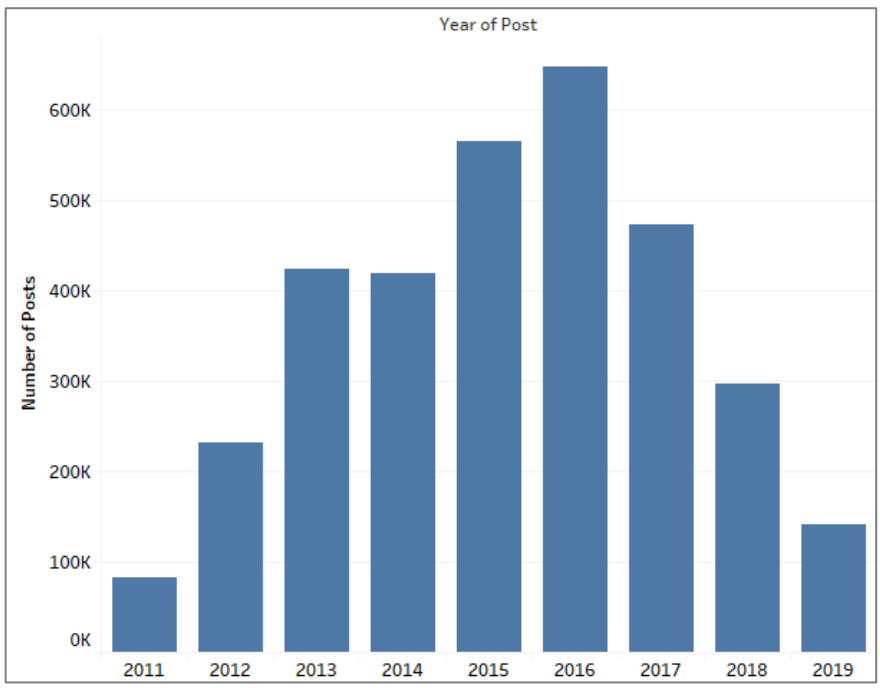


Figure 4. Describing the user-base growth by the number of users' interaction over time

Both the above two ways used to describe the user-base growth gave the same indications that the growth reached the peak point of it in 2016 and we can consider that Khan Academy's gained its highest level of popularity and recognition at that time. After that there was a drop which can be related to many reasons. One of those reasons could be the increase in competition and having different alternatives.

6.1.2 Geographical Distribution of KA Users:

It is important to know the geographical spots for the interacting users. This helps in identifying their needs and how to satisfy them. In our dataset, we detected 11,614 users who stated their location in their profiles publicly (**Figure 5**). The top country where Khan Academy's users are in is the United States. 40.8% of those users (4,744 users) are in US. This shows that Khan Academy is widely used in the U.S. A lot of those learners use math domain as most of them were detected from interacting with math videos. The second country where the learners use Khan Academy's videos widely is Canada which represents 4.9% of the detected users (579 users). India is ranked third with 3.8% of the users (443 users). Khan Academy's videos can also be considered as popular in UK with 2.9% and in Australia with 2.66%. New Zeeland also represents a good potential region for the use of Khan Academy's videos.

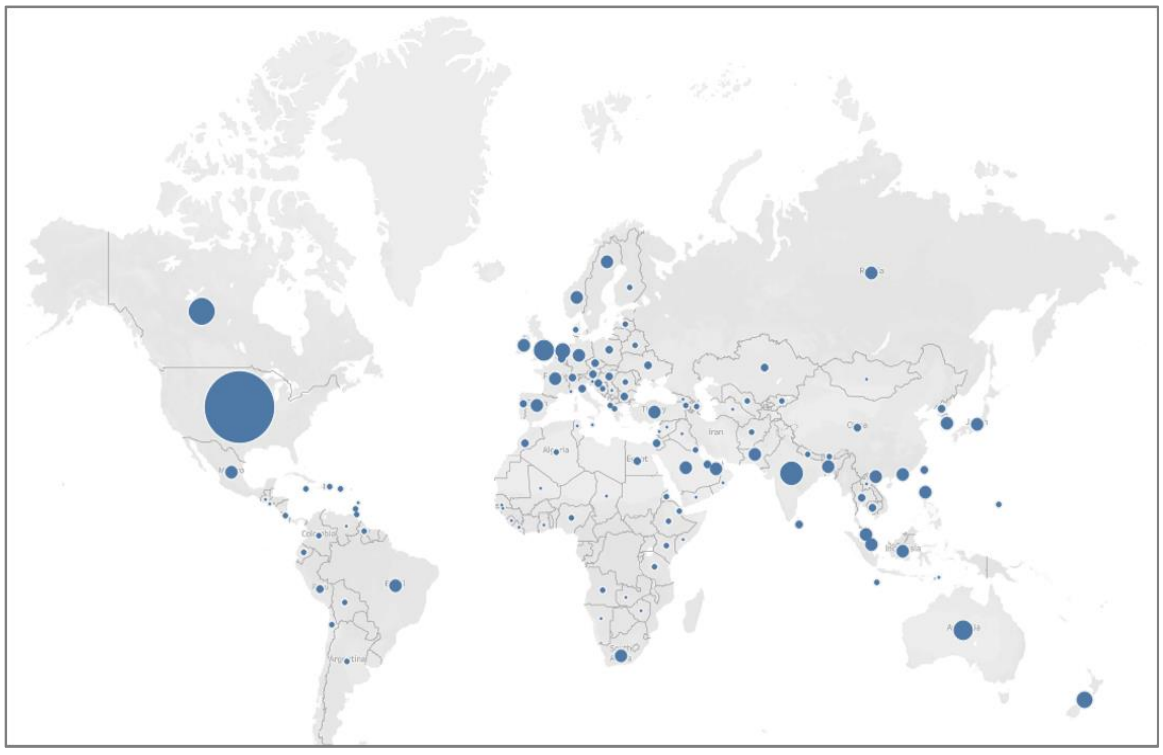


Figure 5. Geographical Distribution of KA Users

6.1.3 Average Duration of KA Videos:

Video's duration is an important feature that may reflect on users' interactions and behaviours. In Khan Academy's repository, the average video length fluctuates from domain to another. **Figure 6** shows the videos' duration (in seconds) related to each domain. We can notice some outliers in Arts and Humanities, College and Career, Partner content and 'Test Prep' domains.

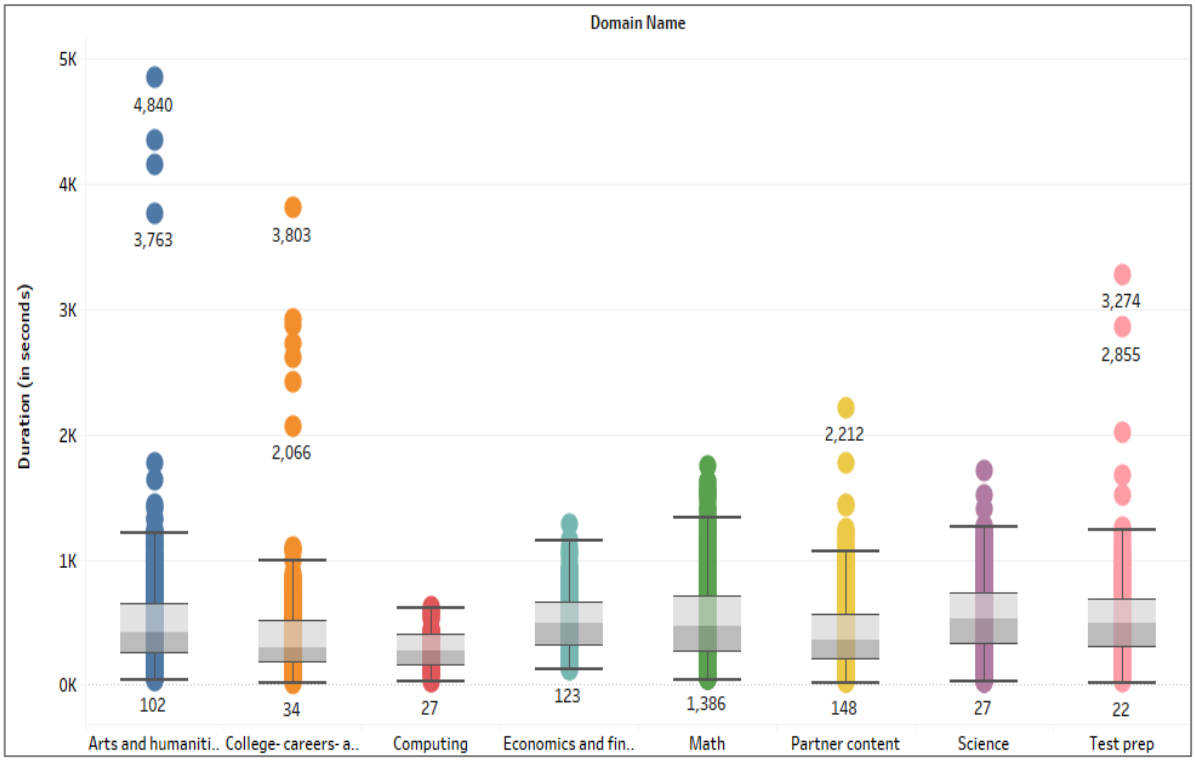


Figure 6. Videos' Duration in Each Domain

In **Table 6,** we measured the average video duration (in seconds), the maximum and minimum durations for the videos related to each domain separately. The longest video found in the repository was (Art Making Programs for Individuals with Dementia) which belongs to Arts and Humanities domain. This video lasts for 1 hour, 20 minutes, 40 seconds (4,840 seconds). On the other hand, the shortest video (Student story: Work study as a study hall) was found in the "College, Careers, and more" domain. The duration of this video is 17 seconds only. The typical average video length in the repository is around 412 seconds (7 minutes).

Table 6. Average Video Duration in Each Domain			
Domain Name	**Avg. Duration**	**Max. Duration**	**Min. Duration**
Arts and humanities	452	4,840	36
College- careers- and more	361	3,803	17
Computing	290	616	27
Economics and finance	484	1,287	123
Math	334	1,739	44
Partner content	322	2,212	19
Science	542	1,706	27
Test Prep	485	3,274	22

6.1.4 Number of Videos Completed by Users:

The number of videos watched completely by each user is also an important indicator for the user behaviour patterns. In **Figure 7** we represented the total number of videos watched completely by users. In our dataset, we found that around 45% of the users who never completed any video. While we found around 35% of them who completed number of videos from 1 to 50. The distribution of the number of completed videos seems to follow Power Law and an inverse relation between the number of users and their number of their completed videos can be detected. This shows that users often feel bored immediately or have no desire to watch the whole video. Many of them focus only on the part they are interested in.

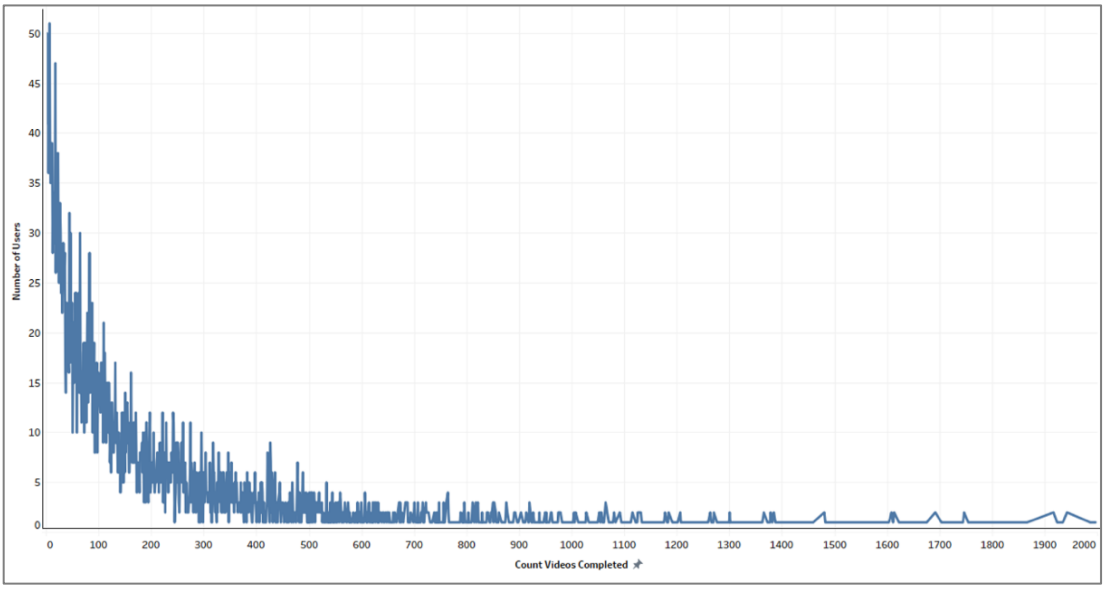


Figure 7. Number of Videos Completed by Users

6.2 Interactions-related Analysis

6.2.1 Evolution and distribution of interactions around KA contents:

More than 2.2M of the gathered users' interactions were posted on math domain videos which is expected especially because Khan Academy's team gave more attention and posted more videos in this domain. The rest 1,042,102 users' interactions were gathered from the rest of the domains. This is an indicator to the high users' engagement with the math domain which reflects the importance of the video contents and shows how much they are attractive.

Figure 8 shows a comparison between Khan Academy's domains in terms of users' interactions posted on each domain's videos. While the largest amount of interactions were gathered from the math domain (68.27%), the science domain became the second highest one in attracting users' engagement. Users' interactions collected from this domain represents 16.63% of the total gathered ones. Surprisingly, the third domain is 'Test Prep' where its interactions represent 5.42% of the total gathered. This domain is one of the newest domains which was added in 2014 and it serves a limited segment of users who are preparing themselves to the college phase. Arts & humanities ranked the fourth with 4.29% of the shares. Finally, the rest of the domains share around 5% of the scraped users' interactions.

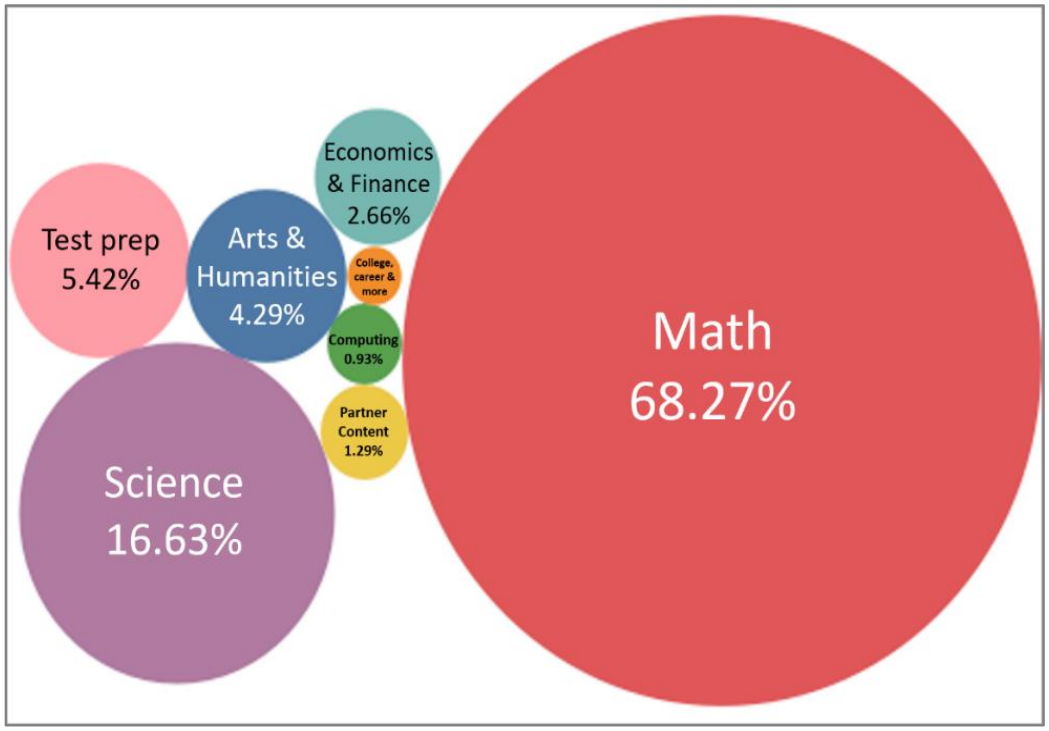


Figure 8. Number of Users Interactions per Domain

If we take a deeper look to check what are the subjects that attracted the most of user's engagement (**Figure 9),** we can find that Pre-Algebra is the most attractive subject while preparing for the medical test MCAT came in the second place. This can be a good indicator that the users are looking for real assistance to serve them during the pre-college level.in order to get college acceptance and determine their choices. This indicator shows the potential in serving such segment and focusing on their needs. If we searched for the science subjects, we can find that the most attractive science subject is Biology which is ranked as the 7th subject in terms of gaining users' interactions with 104,251 interactions.

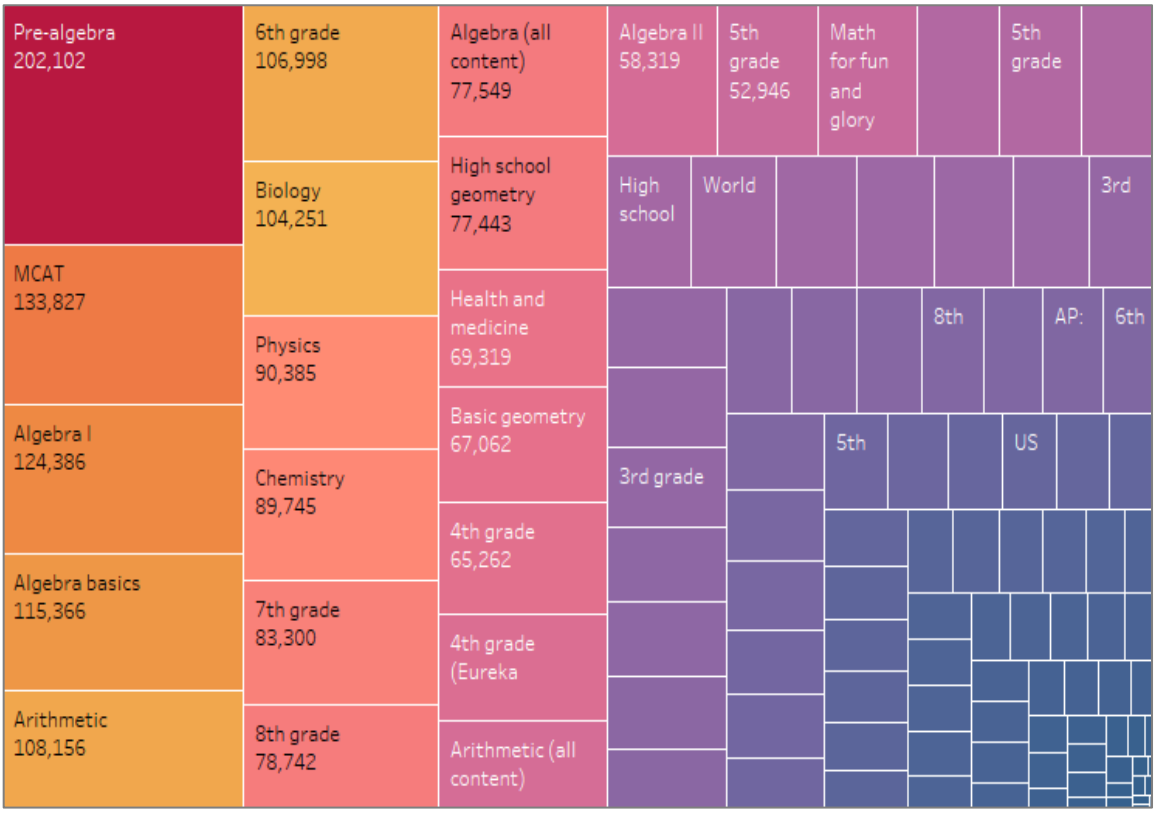


Figure 9. Number of Users Interactions per Subject

In **Figure 10** we demonstrated another distribution for the interactions which shows a fluctuation in the number of user interactions over the years in each domain. This also ensures that the math domain and over the years remains the main domain that attracts the most users.

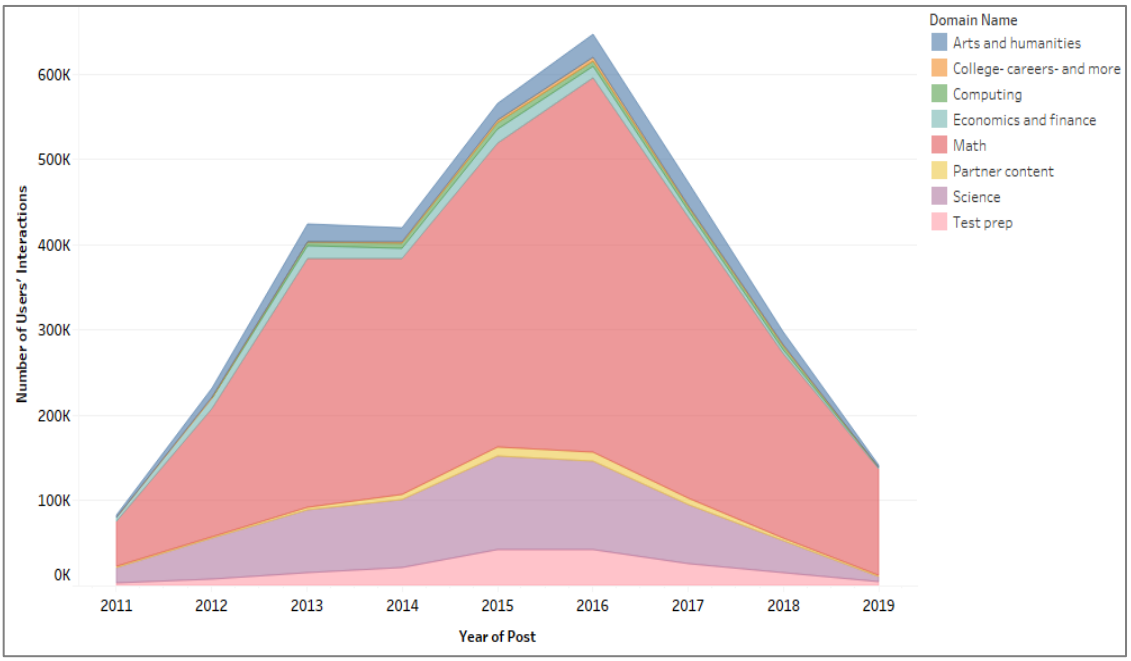


Figure 10. Number of Users' Interactions per Domain over Years

6.2.2 Number of Users' Interactions per Video

The users' engagement with the specific videos can also give indication on and describe the videos' attractiveness. The most interesting video that gained the highest user's attention is called (Radius, diameter, circumference, Pi) which belongs to the math domain. This video has received more than 20K posts. While the first non-math video that gets high user engagement is called (Elements and Atoms) from the science domain and it collected more than 16K posts. This video is ranked the 4th video according to the number of interactions. This analysis can help in identifying the importance of the video and to what extent it is interesting and gaining user's attention (**Figure 11**).

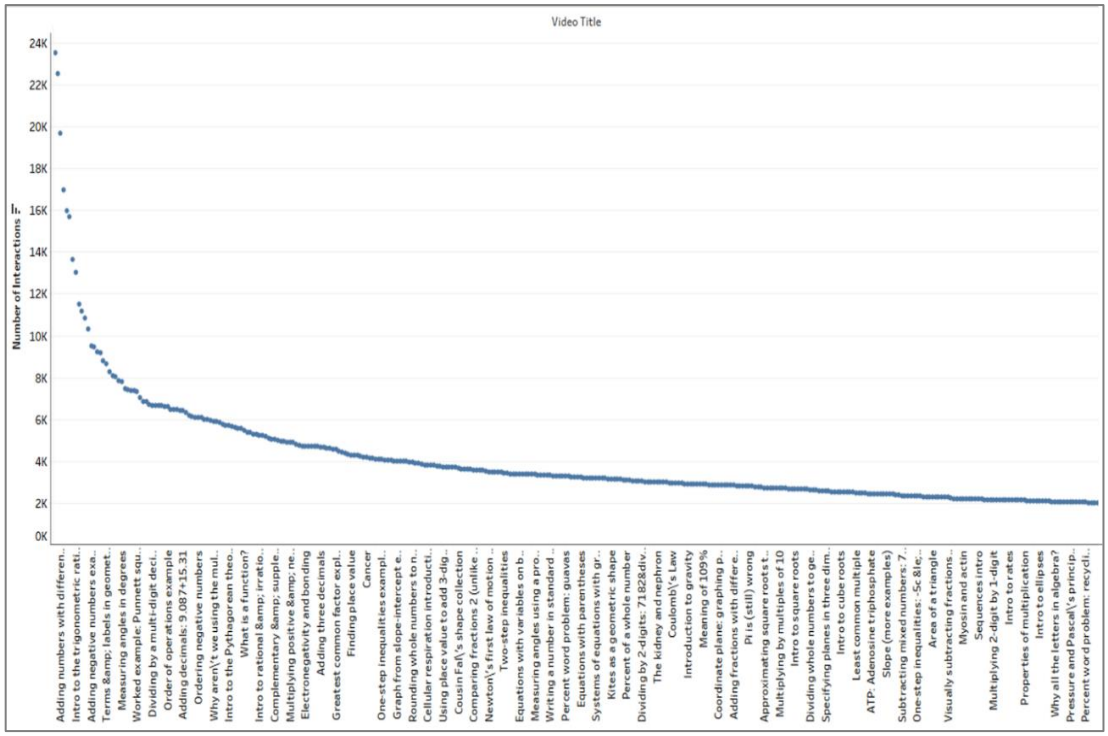


Figure 11. Number of Users Interactions per Video

6.3 Assessing the Relation between Learning Objects and Users' Interactions Using Inferential Analysis

We conducted some analysis to investigate the relation between the user interaction and the learning objects which are the instructional videos in Khan Academy. In this step, we were trying to provide some approximations to identify and characterize the association. To do so, we collected different metrics from the repository that can describe the learning objects. Then we classified the instructional videos according to the level of their users' interactions. After that, we tested the significance of the association between the metrics and the interactions profiles.

6.3.1 Identifying the Collected Metrics

In our dataset, we identified different 20 metrics which can help in constructing some quality indicators to the learning materials. Those metrics were classified to three different classes of measure based on the described object. Those classes are: Video-content-related measures, interaction-related measure and user-related measures **Table 7.**

Table 7. Metrics Collected for the study

Class of Measure	Metrics
Video-content-related measures	Publishing year, video duration, download size, number of reusing the video in different subjects, number of questions posted per video, number of answers posted per video, topic's related, subject's related, and domain's related.
Interaction-related measures	Type of interaction (question or answer), date of interaction, number of votes per interaction, number of comments per interaction, content of the interaction
User-related measures	Joined date, streak length, energy points, city, country, number of completed videos

In this section, our analysis will focus on the video-content-related measures in order to find the relation between the learning materials and the users' interactions. Some of these metrics are numerical data such as video duration, download size, number of reuses, number of keywords, and numbers of questions and answers related. Other video-content metrics are categorical ones such as publishing year, domain type, subject type, and topic type. In our analysis, we tried to assess both types of metrics. We chose video duration and number of reuses as numerical metrics and we chose publishing year and domain type as categorical ones. We did not consider the rest of the video-content metrics for several reasons:

- We used the number of interactions (questions and answers) to classify the learning materials in the next step.
- In Khan Academy's repository, the user does not need to download the video to be able to watch it. So that, the download size cannot be considered in terms of the relation with users' interactions.
- The relation with the domain type can give an indication to the relation with the subject and topics types.

6.3.2 Classifying Learning Material into Interaction Profiles

We used Tableau software to visualize the large number of user interactions associated with 10,565 videos related to different domains. **Table 8** shows the number of scraped videos with their interactions in each domain. Those 3.2M interactions have been posted over the years. In the Box–Whiskers plots **Figure 12** we applied different plots to demonstrate the distribution patterns of interactions in each domain separately over years.

Table 8. Number of scraped videos with their interactions in each domain

Domain Name	Total number of published videos	Number of scraped videos	Number of users' interactions
Arts and humanities	1,755	1,131	140,895
College- careers- and more	498	253	16,085
Computing	74	74	30,677
Economics and finance	750	477	87,390
Math	11,364	3,414	2,242,408
Partner content	1,353	1,187	42,530
Science	3,068	2,369	546,462
Test Prep	2,037	1,660	178,063
Total	**20,899**	**10,565**	**3,284,510**

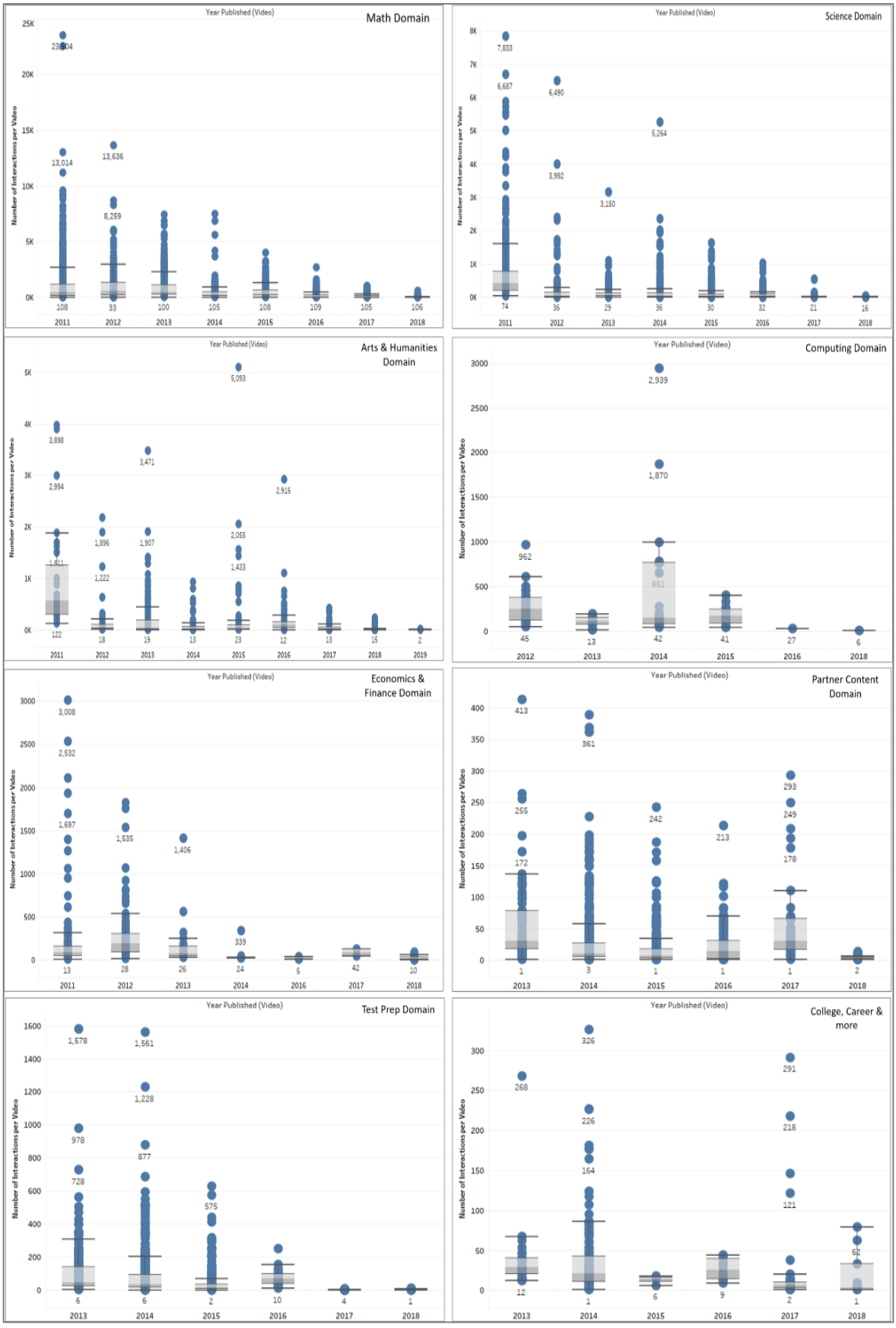


Figure 12. Box Whiskers Plots showing Distributional Patterns of Users Interactions in each Domain

The plots that are related to math and science domains show that the distribution of interactions per video over years follows the power law with some outliers in 2011. The plot of Arts & humanities follows the polynomial distribution with some fluctuations over time. The median of the year 2016 has been increased to almost the triple of the median of 2015. Then it declined again in 2017 to reach 22 interactions per video. The user interactions associated with all the published Computing videos are demonstrated in the Computing plot. This plot shows that the materials published in 2014 were the highly interactive ones with a median of 163 interactions. It displays also that only one video was published in 2017 and 2018 which gives a sign that computing domain is losing its popularity. "Economics and Finance" plot follows the Poisson distributional pattern. The interactions with Economics videos reached the peak point with videos that were published in 2012 and has a median of 189 interactions. Partner Content plot follows a multimodal distributional pattern by reaching to the top interaction level twice in 2013 and in 2017 where the median in both is 30 interactions. The level of interactions with 'Test Prep' domain reached the top level with videos published in 2016 when their median reached to 68 interactions. Finally, 'College, Careers and more' domain also has a multimodal distribution where their videos in 2014 and 2018 are gaining the most of attention.

Also, we plotted Box Whiskers plot to illustrate the overall distributional pattern of the users' interactions associated each domain **(Figure 13).** According to that, the overall median for users' interactions across all domains equals to 60 interactions. The upper-quartile is 242 interactions and the lower-quartile is 16. After analysing those findings, we classified Khan Academy's videos into three different profiles according to the number of users' interactions. They were built up based on the identified quartiles. Those three profiles are: high interaction profile, medium interaction profile and low interaction profile **Table 9**.

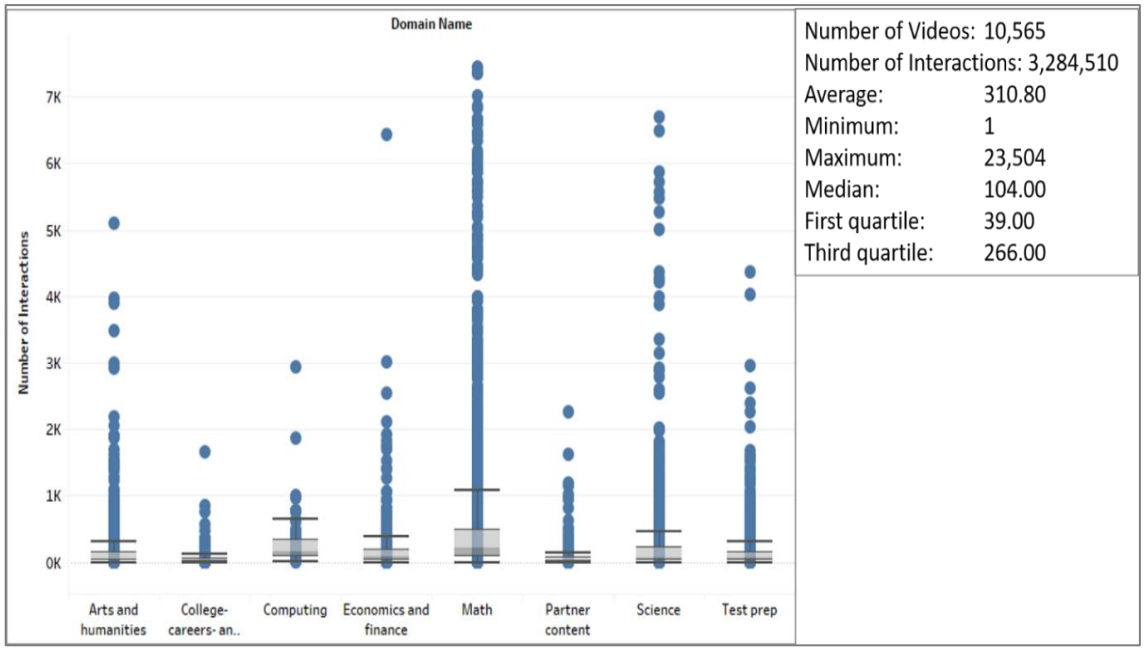


Figure 13. Box Whiskers Plot showing total Users Interactions

Table 9. Number of KA Learning Objects per Domain per Statistical Profile

Domain Title	Low-interaction Profile	Medium-interaction Profile	High-interaction Profile	Total
Math	543	1,226	1,645	3,414
Science	952	971	446	2,369
Arts and humanities	620	406	105	1,131
Economics and finance	120	287	70	477
Computing	5	45	24	74
Partner content	975	193	19	1,187
Test Prep	936	574	150	1,660
College, careers, and more	174	65	14	253
Total	**4,325**	**3,767**	**2,473**	**10,565**

To consider a video in the Low interaction profile, the number of users' interactions associated with it must fall below the interquartile range (less than 39 interactions). This profile was found to include most of learning objects in our dataset (4,325 videos). All the videos that belong to the medium interaction profile must be associated the interquartile number of interactions (included 39 and 266). This group contains 3,767 videos out of the scraped ones. The high interaction profile includes all the videos that have interactions more than the interquartile range (above 266). We found only 2,473 videos that belong to this profile.

6.3.3 Relationship between Domain Type and Interaction Profiles

After categorising the learning materials in to three interactions profiles above, we will find the correlation between those profiles and the metrics determined previously. Pearson Chi Square test and Phi & Cramer's V test (**Table 10)** was performed to examine the relationship between the videos in the three profiles and the domain that they belong to. Pearson Chi Square is a very useful statistics tool used for testing hypotheses when the variables are categorical (McHugh, 2012). While Cramer's test is a very common strength test applied to test the data when Chi-square result is significant. These are very popular and proved their efficiency. We performed these tests to examine the strength of relationship between the domain of the instructional videos and the users' interactions associated with those videos. By applying Chi Square test on the low-interaction profile videos, a significant relationship between the number of users' interactions and the domain type of those videos was demonstrated. The Phi value of (0.417, 0.000) indicates that they have a strong positive relationship.

Table 10. Testing Domain Type vs. Interaction Profiles

Chi-Square Tests

	Low Interaction Profiles			Medium Interaction Profiles			High Interaction Profiles		
	Value	df	Asymptotic Significance (2-sided)	Value	df	Asymptotic Significance (2-sided)	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	741.382	259	.000	1564.395	1589	.665	4203.365	4333	.919
Likelihood Ratio	733.839	259	.000	1481.854	1589	.973	1583.111	4333	1.000
N of Valid Cases	4325			3767			2473		

Symmetric Measures

		Low Interaction Profiles		Medium Interaction Profiles		High Interaction Profiles	
		Value	Approximate Significance	Value	Approximate Significance	Value	Approximate Significance
Nominal by Nominal	Phi	.417	.000	.657	.665	1.884	.919
	Cramer's V	.158	.000	.248	.665	.712	.919
N of Valid Cases		4325		3767		2473	

On the other hand, when we applied Chi Square test on the medium-interaction and high-interaction profiles, it exhibited that there is no statistically significant

association between the domain's type and number of users' interactions. The results of this investigation indicate that in general, the domain type does not affect the attractiveness or popularity of the learning material.

6.3.4 Relationship between the publishing Year of learning material and the interaction Profiles

Another Pearson Chi Square test and Phi & Cramer's V test has been applied in **Table 11** to examine the association between the video's publishing year which is a categorical variable as well, and the videos related to the different interaction profiles.

Table 11. Testing Publishing Year vs. interaction Profiles

Chi-Square Tests

	Low Interaction Profiles			Medium Interaction Profiles			High Interaction Profiles		
	Value	df	Asymptotic Significance (2-sided)	Value	df	Asymptotic Significance (2-sided)	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1197.363	259	.000	1689.763	1589	.039	3129.799	3714	1.000
Likelihood Ratio	1130.017	259	.000	1753.494	1589	.002	2038.576	3714	1.000
Linear-by-Linear Association	541.100	1	.000	164.162	1	.000	27.984	1	.000
N of Valid Cases	4325			3767			2473		

Symmetric Measures

		Low Interaction Profiles		Medium Interaction Profiles		High Interaction Profiles	
		Value	Approximate Significance	Value	Approximate Significance	Value	Approximate Significance
Nominal by Nominal	Phi	.530	.000	.683	.039	1.626	1.000
	Cramer's V	.200	.000	.258	.039	.664	1.000
N of Valid Cases		4325		3767		2473	

As we are testing the relation between the year and the users' interactions, we can assume and expect that the older the video the more interactions it have. But after performing the above test, we found that the low-interaction profile videos, showed a strong, positive, and significant relationship between the two variables (publishing year and their users' interactions. While the videos that are related to the medium-interaction profile showed a less significant correlation than the one in the low-interaction profile.

On the other hand, Chi square results for the videos that are related to the high-interaction profile and who have more users' interactions showed that there is no significant relationship between video's publishing year and the number of users'

interactions associated with it. This is not surprising, as it indicates that the popularity and attractiveness of the learning material does not have to be related to its age of availability to the users.

6.3.5 Relationship between interaction profiles, video length and reuse rate

As we are dealing with the remaining two variables which are numerical variables (video length and reuse rate), the most suitable ways to examine the association is to use the analysis of variance. Analysis of variance (ANOVA) is one of the most frequently used statistical tests in social sciences (Silk, 1981). In **Table 13** and **Table 14**, we conducted the analysis to test the relationship between the number of users' interactions as a dependent variable and two different variables which are video length and reuse rate acting as predictors. We applied ANOVA test which demonstrated a significant association for all the interaction profiles. Our regression model that resulted for each profile have the following appearance **Table 12:**

Table 12. Results of regression models for each profile	
Low Interaction Profile:	$\hat{\gamma} = 11.928 + 0.003x_0 + 0.684x_1$
Medium Interaction Profile:	$\hat{\gamma} = 82.159 + 0.014x_0 + 7.601x_1$
High Interaction Profile:	$\hat{\gamma} = 406.498 + 0.315x_0 + 37.198x_1$

Where $\hat{\gamma}$ represents the estimated value of users' interactions associated with the video from the specific profile, x_0 represents the video duration and x_1 represents the number of video reuses in different subjects. The different coefficients of x_0 in the different equations demonstrated a significant but overall weak correlation with the video length. On the other hand, the different coefficients of x_1 exhibited a significant positive relationship between the video's reuse rate and the number of users' interactions in all types of profiles. This relationship is stronger in the videos of highly-interaction profiles than in the ones related to the low-interaction profiles. If we take a look to the F-value in the table below, we can find that the high-interaction profile videos demonstrate the best ones to fit our regression model (Cuevas, Febrero, and Fraiman, 2004).

Table 13. Testing Video Duration and Number of Reuses vs. Number of interactions ANOVA Test

ANOVA[a]						
Low Interaction Profiles						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3237.718	2	1618.859	4.765	.000[b]
	Residual	467410.742	4323	109.644		
	Total	470648.460	4325			
Medium Interaction Profiles						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	383789.462	2	191894.731	2.815	.000[b]
	Residual	12236919.450	3765	3377.565		
	Total	12620708.910	3767			
High Interaction Profiles						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10514728.470	2	5257364.237	1.991	.000[b]
	Residual	517789342.100	2471	438432.974		
	Total	528304070.600	2473			

a. Dependent Variable: Total interactions
b. Predictors: (Constant), Reuses (for the video), Duration

Table 14. Testing Video Duration and Number of Reuses vs. Total Number of interactions Coefficients

Coefficients						
Low Interaction Profiles						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11.928	.528		22.579	.000
	Duration	.003	.001	.067	4.368	.000
	Reuses (for the video)	.684	.194	.054	3.528	.000
Medium Interaction Profiles						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	82.159	2.939		27.953	.000
	Duration	.014	.004	.062	3.678	.000
	Reuses (for the video)	7.601	.716	.179	10.611	.000
High Interaction Profiles						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	406.498	60.560		6.712	.000
	Duration	.315	.074	.130	4.237	.000
	Reuses (for the video)	37.198	9.853	.116	3.775	.000

7. Detecting Communities in Online Learning Repository

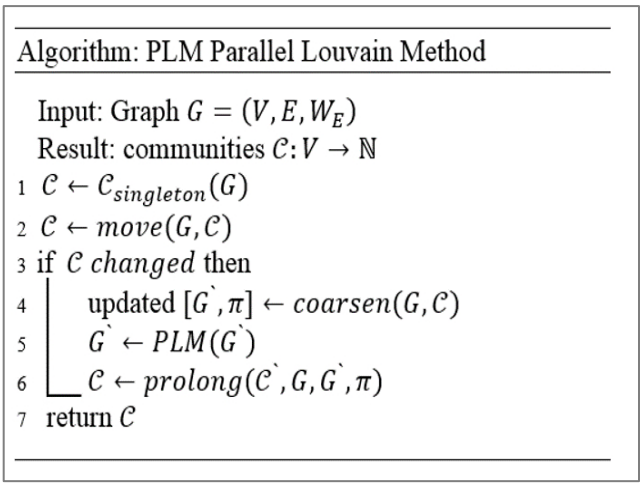
This chapter is derived from the material published in (Yassine, Kadry, and Sicilia, 2020b). We performed two different community detection algorithms to identify the cohesive clusters in our created graph. To find the implemented codes and algorithms, please follow this link (https://github.com/SaharYassine/KA-Analysis-Files).

As we mentioned previously in the data acquisition section, we used NetworkX and NetworKit libraries to prepare the data and create our graph $G = (V, E, W_E)$, where V is the set of users (nodes) that were extracted from the dataset, E is the group of edges that represents the relationship between the users which are the posts (questions and answers) between them, and W_E is the weight of those edges which reflects the number or the volume of interactions (questions and answers) between each two nodes. So that the relationship represented in our graph is a user to user relationship.

During this phase, we investigated the literature for the convenient methods used to identify the emerged communities in learning repositories. Based on our systematic review (Yassine, Kadry, and Sicilia, systematic review, in process, 2021), we revealed a great potential in implementing community detection techniques to identify learning communities which help in developing and improving the educational research and in utilising the massive amount of data generated from the learning interactions. We found that the most applied technique to detect communities in learning settings is K-means clustering which effective with small datasets. While the most applied technique with large learning datasets was found to be Louvain method which is used to detect the disjoint communities. We were looking for community detection algorithms that are applied conveniently and effectively in studying real-world networks. However, the choice of the best algorithm for this task cannot be straightforward (Linhares et al., 2020). Also, we investigated extensively many studies that compared between applying algorithms with different principles in different large-scale, complex, and real-world networks. We decided to perform different detection methods according to the following two different factors: the nature and the size of the collected dataset, and the functionality provided by the NetworKit library. We compared between the performance, goodness, and

effectiveness of the selected algorithms as each method is based on a different principle and applies different technique in identifying learning communities:

- Parallel Louvain Method (PLM) which is the parallelized version of the Louvain method. The Louvain method which is presented by Blondel (Blondel et al., 2008) is a multi-phase, iterative, and a modularity-based algorithm. It is a heuristic method for uncovering communities in complex weighted graphs (Khan and Niazi, 2017). Louvain method works by optimizing the modularity which is the quality function of communities' approximation and it proved more time efficiency and higher modularity than Newman's method (Newman, 2004). The parallel Louvain method (PLM) uses a shared-memory parallelization where the nodes' movements are assessed and completed in parallel. It operates by performing two major steps (**Figure 14**) (Staudt, 2016). The first step is executing phases one at a time with running multiple iterations in each phase to perform a parallel evaluation of the nodes using the community information gained from the previous iteration. Those phases will keep going until the modularity gain between iterations is higher than the threshold. The second step is graph rebuilding which takes place after each successive phase by using the new identified community assignment to construct a new graph as an input for the next phase. This is done by adding the new communities to the new graph as nodes and the edges are built based on their connection to other communities (Forster, 2018).

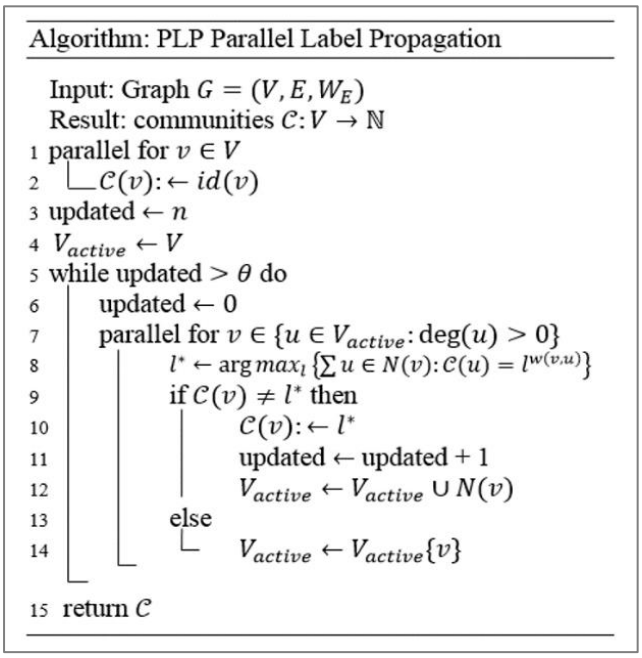

Algorithm: PLM Parallel Louvain Method

Input: Graph $G = (V, E, W_E)$
Result: communities $\mathcal{C}: V \to \mathbb{N}$
1 $\mathcal{C} \leftarrow \mathcal{C}_{singleton}(G)$
2 $\mathcal{C} \leftarrow move(G, \mathcal{C})$
3 if $\mathcal{C}$ *changed* then
4 updated $[G`, \pi] \leftarrow coarsen(G, \mathcal{C})$
5 $G` \leftarrow PLM(G`)$
6 $\mathcal{C} \leftarrow prolong(\mathcal{C}`, G, G`, \pi)$
7 return $\mathcal{C}$

Figure 14. PLM Algorithm

- Parallel Label Propagation Algorithm (PLP) which is based on the parallelization of label propagation algorithm. Label propagation algorithm proposed by Raghavan (Raghavan, Albert, and Kumara, 2007) is considered one of the fastest graph clustering techniques which proved to be efficient because of its linear time-complexity. It identifies communities by labelling each node in the set of nodes with a unique label indicating the community it belongs to. In every label

propagation iteration, the node adopts the most common label in its neighbourhood. Number of performed iterations depend on the graph structure and not on its size. Finally, every connected group of nodes with a common label forms a community. The parallel label propagation (PLP) starts first by initializing the labels, then it proceeds to the main loop where the major computations are performed to count the occurrences of the adjacent labels **(Figure 15**) (Staudt, 2016). The loop of counting the adjacent labels consists of four steps: gathering the labels, sorting the adjacent labels by nodes, extracting boundaries of the adjacent labels, and computing their frequencies. PLP uses multiple processors to build a common label array of nodes and it applies efficient parallelization to accumulate the labels of the adjacent nodes by exploiting multiple data parallel primitives. This variant of LPA is an effective method used to detect those overlapping communities in dynamic real networks which contain nodes that belong to more than one community at a time. (Kozawa, Amagasa, and Kitagawa, 2017).


Algorithm: PLP Parallel Label Propagation

Input: Graph $G = (V, E, W_E)$
Result: communities $\mathcal{C}: V \to \mathbb{N}$
parallel for $v \in V$
  $\mathcal{C}(v) \leftarrow id(v)$
updated $\leftarrow n$
$V_{active} \leftarrow V$
while updated $> \theta$ do
  updated $\leftarrow 0$
  parallel for $v \in \{u \in V_{active} : \deg(u) > 0\}$
    $l^* \leftarrow \arg max_l \{\sum u \in N(v) : \mathcal{C}(u) = l^{w(v,u)}\}$
    if $\mathcal{C}(v) \neq l^*$ then
      $\mathcal{C}(v) \leftarrow l^*$
      updated $\leftarrow$ updated + 1
      $V_{active} \leftarrow V_{active} \cup N(v)$
    else
      $V_{active} \leftarrow V_{active}\{v\}$
return $\mathcal{C}$

Figure 15. PLP Algorithm

- Using NetworkX libraries, we ran the original variants of Louvain method and label propagation algorithm and we included them in our comparison.

Regarding the volume of our dataset, it was crucial and more efficient to use parallelism to scale our data. Options and alternatives to do so are still limited and need more research and investigation efforts (Staudt, Sazonovs, and Meyerhenke, 2014). NetworKit as a social analysis tool found to be the most suitable one to perform the social analysis with our dataset. It provides scalable solution techniques such as parallelization, heuristics for computationally expensive problems, and efficient data structures (Staudt, Sazonovs, and Meyerhenke, 2016). NetworKit is still under development and it doesn't have a diverse library such as NetworkX, but it proved its

effectiveness with large scale networks and it provides the parallel heuristics of Louvain method and label propagation algorithm in addition to different centrality measures such as degree, eigenvector, and closeness.

7.1 Applying Parallel Label Propagation Algorithm (PLP)

PLP algorithm was applied to our graph (G). It was able to identify 34,032 communities in our graph (G) (**Figure 16)** within 1.786 seconds. Those communities emerged from users interacting with each other by asking questions regarding the instructional videos and answering them. By looking to PLP communities, we found that 38% of those communities consist of two users (nodes) only. The big and coloured nodes represent the largest communities that were detected. The big grey node represents the largest PLP community which includes 31,397 users. The average size of the community according to PLP-solution is 10.5537 users. 55.44% of PLP communities are strongly connected to each other while the rest can be considered disconnected or isolated ones. PLP network's diameter which calculates the shortest path between the two most distant communities is 9 while the average path length between all the pairs is 2.953.

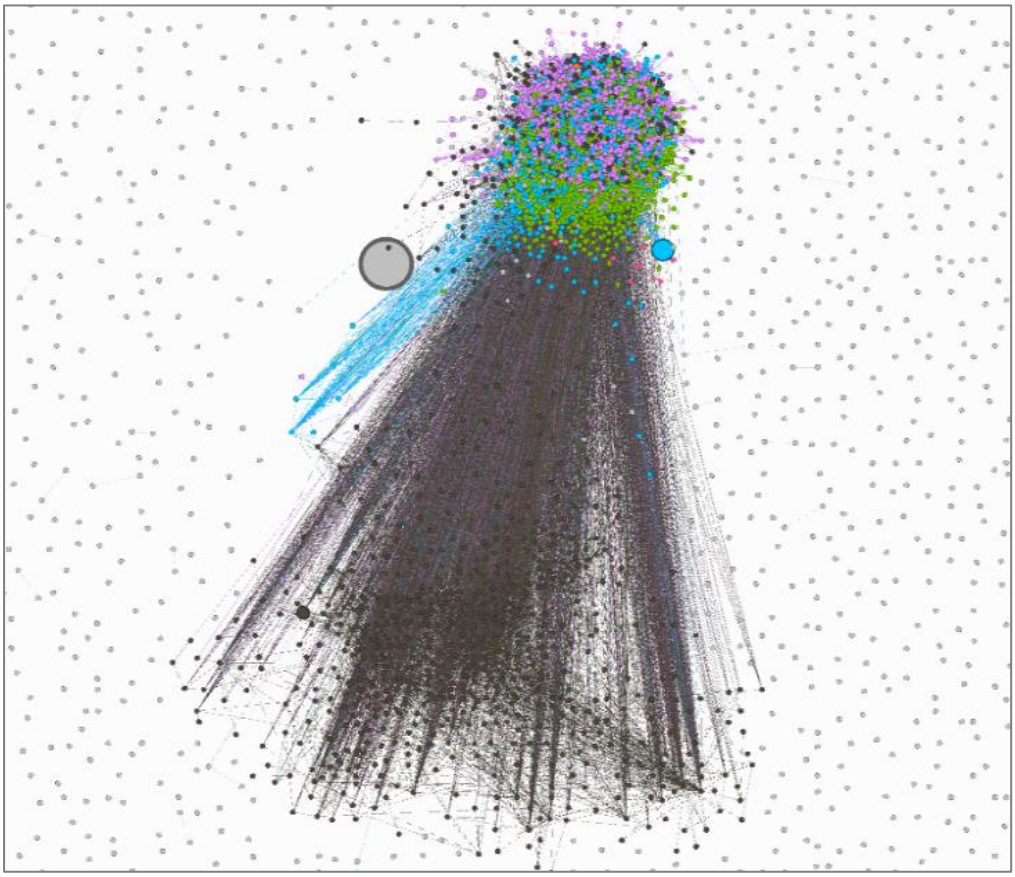


Figure 16. PLP Detected Communities

7.2 Applying Parallel Louvain Algorithm (PLM)

We applied PLM Algorithm to our graph (G) to identify the communities (**Figure 17)**. It was able to detect 11,315 communities which emerged from interacting the users around the instructional videos in Khan Academy's dataset within 8.436 seconds. We found that 18% of the identified communities consist of 2 users only. Whereas the average size of the community according to PLM-method is 31.7422 users which indicates that the communities emerged from PLM method are bigger in terms of size than the ones detected by PLP. The big purple node represents the largest PLM community detected and it includes 40,074 users. We also found that the detected communities are very well-connected to each other as the number of the strongly-connected communities is 11,107 which means 98% of the identified ones. This means that PLM was able to detect more well-connected communities than PLP and that is obvious by comparing the communities detected in **Figure 16** and **Figure 17.** The shortest path between the two most faraway communities is 3 whereas the average path length between all the pairs is 2.324.

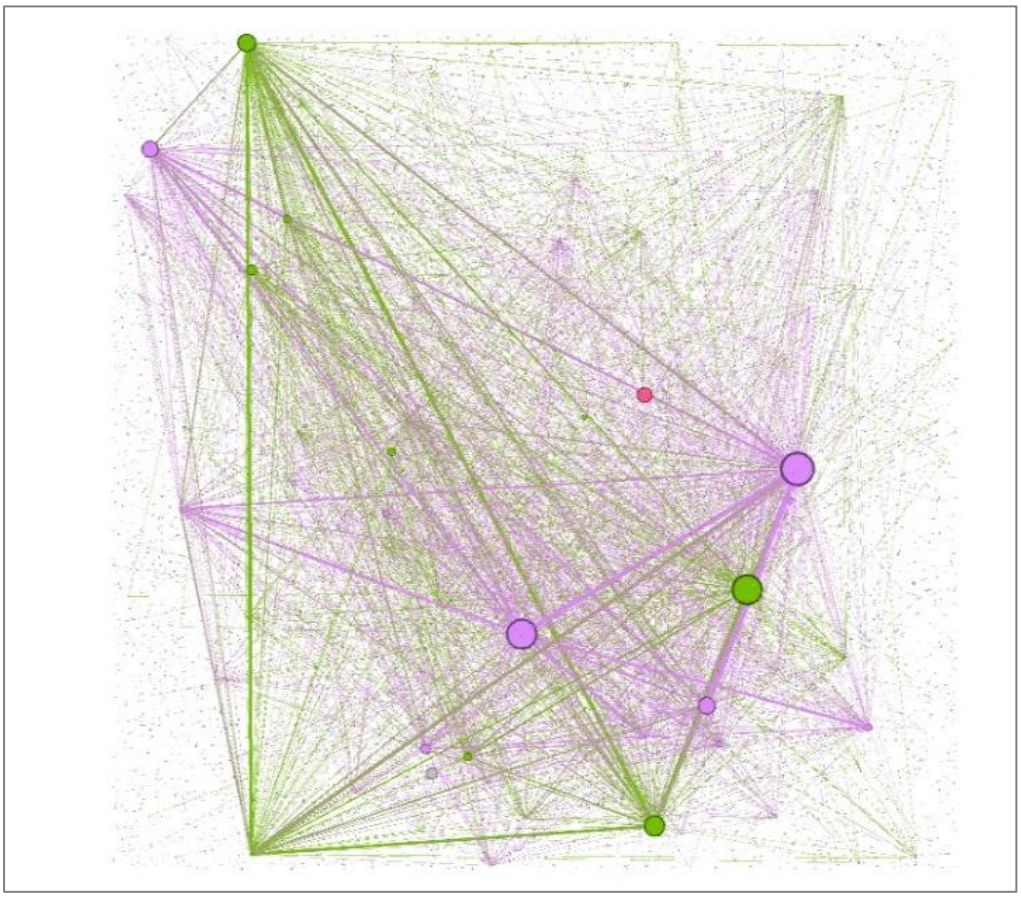


Figure 17. PLM Detected Communities

7.3 Applying Louvain Method and Label Propagation Algorithm (LPA) using NetworkX

We used NetworkX libraries to apply both Louvain method and label propagation algorithm on our graph (G) to identify the communities. As we mentioned earlier that the application was performed but with higher consuming for time and technical resources and that is why at later stages we will continue with PLP and PLM only. Louvain method took around one hour to detect 11,107 communities with average size of 32.21 users. While the label propagation algorithm consumed more than two hours to detect 76,151 communities with average size of 4.7 users. **Table 15** shows a comparison between the network properties for each method:

Table 15. Comparison between the properties of the applied community detection methods:

Properties	PLP	PLM	Louvain	LPA
Number of communities	34,032	11,315	11,107	76,151
Smallest community size	2	2	2	2
Largest community size	31,397	40,074	11,148	11,534
Average community size	10.5537	31.7422	32.214	4.716
Modularity	0.577	0.687	0.674	0.449
Network Diameter	9	3	3	7
Average Path Length	2.953	2.324	2.15	2.732
Well-Connected Components	18,868	11,107	10,945	14,810
Time score (seconds)	1.786	8.436	3848	8171

In a way to evaluate the implemented community detection techniques, we demonstrated the results of two main efficiency factors in a two-dimensional Pareto Front diagram (Hosseinpoor et al., 2020) (**Figure 18)**. Those two main factors are the running time and the quality of the applied technique. In the diagram, the time score represents the running time of the algorithm that was needed to detect the communities, while the modularity score represents the quality score as it measures the strength of division of a network into modules or communities and it indicates to the good partition. According to the diagram, it is obvious that PLP is the supreme in terms of running time and it shows a very good response in terms of modularity while PLM shows the best response in terms of modularity score and the second best in terms of time score.

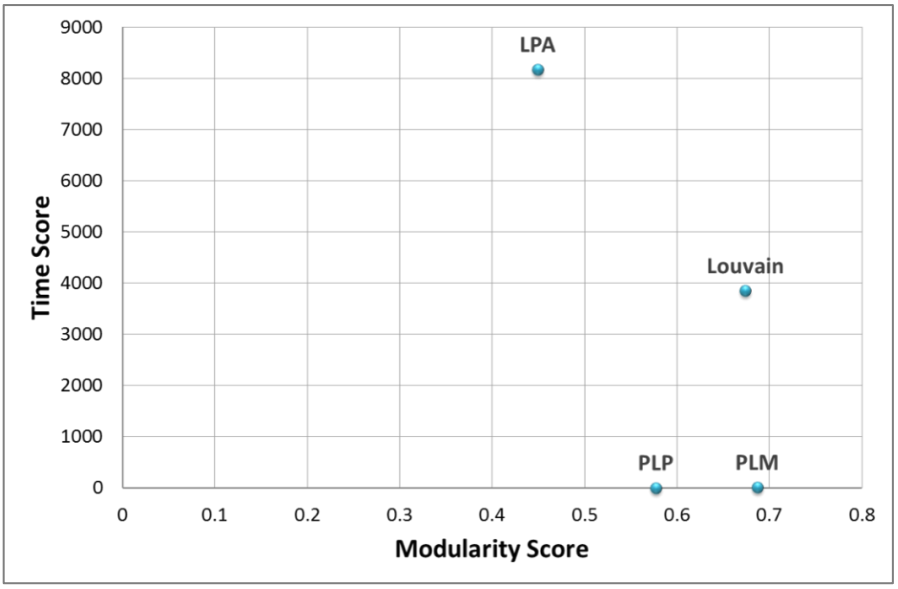


Figure 18: Pareto Front Diagram for applied community detection algorithms

We examined the relationship between the detected communities and the domain type according to the associated users' interactions. As we know, that PLP tends to identify the overlapped communities where the users can belong to more than one community. Some of the detected communities found to be related to more than one domain which means that their users are active and interacting across different domains. In **Table 16,** we demonstrated the number of PLP and PLM detected communities related to each domain based on their highest number of interactions. Most of the detected PLP-communities are mainly related to Science domain while 'Test Prep' is ranked the second domain. While, this is not the case in the PLM detected communities where the majority are related to Math domain then come the Science one. Interestingly, PLP detected communities are not aligned with number of interactions gathered from each domain (**Figure 8).** In fact those communities did not show correlation with the domain type.

Table 16. Number of communities related to each domain

Domain Title	PLP-Communities	PLM-Communities
Science	**13,229**	2,263
Test Prep	8,180	1,411
Math	7,155	**6,530**
Arts and humanities	5,946	505
Partner content	4,450	437
Economics and finance	3,617	647
Computing	3,320	386
College- careers- and more	2,035	300

8. SNA Measures and Users' Interactions

In order to comprehend the detected network structure, the pattern of users' organization, their relationships, and their behaviours we assessed the detected communities using different social network analysis techniques and measures. Applying SNA measures and techniques in online learning environments is important to evaluate the structural significance and learners' behaviour and engagement (Cela, Sicilia, and Sánchez, 2015). We examined the effectiveness of the emerged online communities and we assessed the users' interactions with the online learning objects. **Table 17** demonstrates summary statistics of social network analysis measures for the communities detected by PLM and PLP methods. The calculated measures display the mean across the identified communities. Data presented in this chapter is based on the material published in (Yassine, Kadry, and Sicilia, 2020b).

Table 17. Summary statistics of network properties according to PLP and PLM methods

	Khan Academy's Communities Detected by	
	PLP Method	PLM Method
Number of Detected Communities	34,032	11,315
Modularity	0.5765	0.6873
Density	0.6449	0.9168
Clustering coefficient	0.284	0.928
Degree Centrality	1.8953	2.4802
Eigenvector Centrality	0.1785	0.0483
Closeness Centrality	0.2363	0.1369
Betweenness Centrality	2.668	8.031

8.1 Community size distribution

The distribution of communities' sizes is a key indicator to the effectiveness of the applied algorithm. We performed Kolmogorov Smirnoff (KS) test to examine the distribution of the size of the communities produced by PLP and PLM methods. For both groups of communities, the size distribution found to follow the power-law distribution. This is the case in many of the real-world networks (Labatut and Balasque, 2013). **Table 18** exhibits the number of PLP and PLM communities with sizes from 2 to 10 members. According to our results, PLP Algorithm tends to generate a regular power-law distribution, while PLM Algorithm generated a long tailed one. The community size distribution generated by each algorithm is demonstrated in **Figure 19** (the log-size is used).

Table 18. Comparison between PLP & PLM communities in terms of size

Community size	Number of Communities	
	PLP Method	PLM Method
2	13,237	9,152
3	6,821	1,415
4	3,455	351
5	2,058	132
6	1,396	68
7	1,032	32
8	752	21
9	592	13
10	506	10

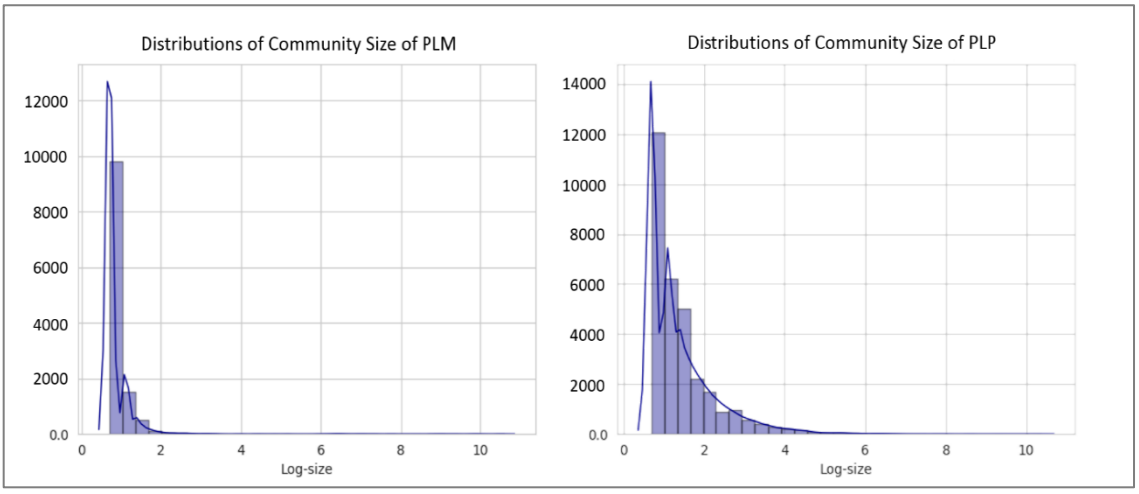


Figure 19. Distributions of Community Size for PLM & PLP Detected Communities

8.2 The Largest Community

In this section we will investigate the largest detected community according to both methods PLP and PLM.

As mentioned earlier, PLP-algorithm managed to detect 34,032 online communities which are built up from users' interactions. The average size of those communities was 10 users per community while the largest community in terms of size was 31,397 users. In the largest PLP-community, the dominant domain type is Science which is associated with more than 149K users' interactions which were posted on 130 different science topics. On the other hand, the "Physical Processes" topic which belongs to Test-Prep domain was the most attractive topic by engaging more than 7,000 users in posting on it more than 24K of questions and answers. In

Figure 20, we exhibited the largest community that was detected by applying PLP algorithm and **Table 19** demonstrated the top five topics in attracting users to engage with according to PLP communities' detection.

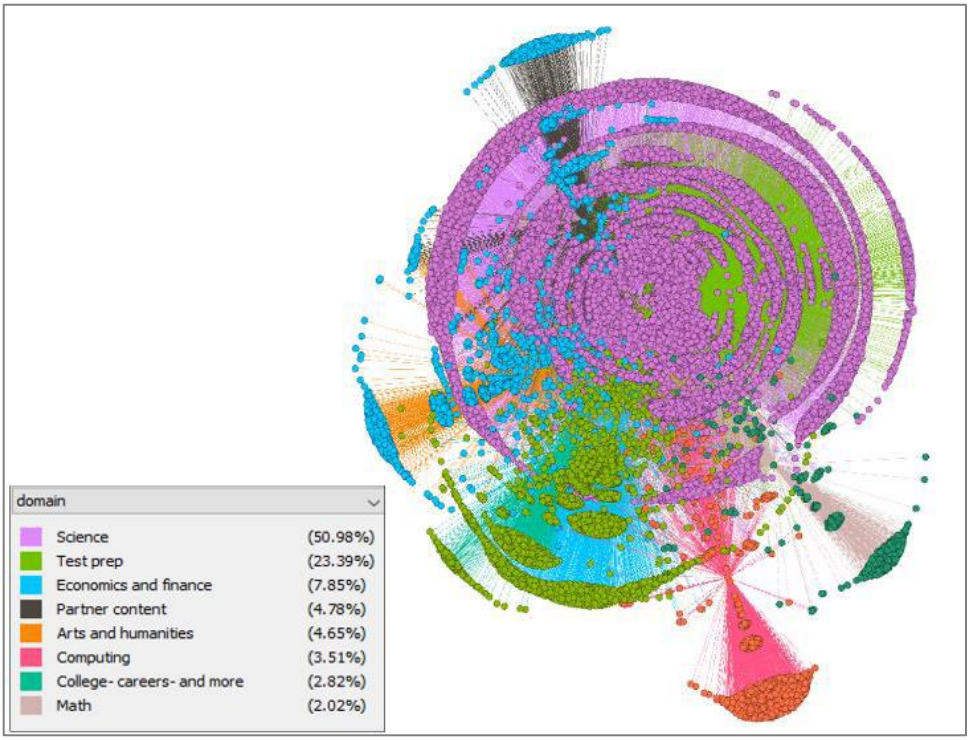


Figure 20. Largest PLP Community

Table 19. TOP 5 Attractive Topics - According to PLP Algorithm

Topic Title	Domain	Number of Posts
Physical processes	Test Prep	24,078
One-dimensional motion	Science	9,510
Forces & Newton's laws of motion	Science	8,833
Scale of the universe	Science	7,533
Chemistry of life	Science	6,693

PLM-algorithm managed to detect 11,315 communities and this number is almost one third of the number that has been detected by applying PLP-algorithm. The average community size was 31 users while PLM largest community in terms of size consisted of 40,074 users. In this largest community, the Science domain is the dominant one which is associated with more than 218K users' interactions posted on 128 different topics, but this community includes other different domains as well. Also, "Physical Processes" topic which belongs to Test-prep domain is the most attractive topic which gained more than 34.9K questions and answers posted by

around 11K users. In **Figure 21** we illustrated the largest community that was detected using PLM algorithm and **Table 20** demonstrated the top five topics that attract users' engagement according to PLM detecting communities Algorithm.

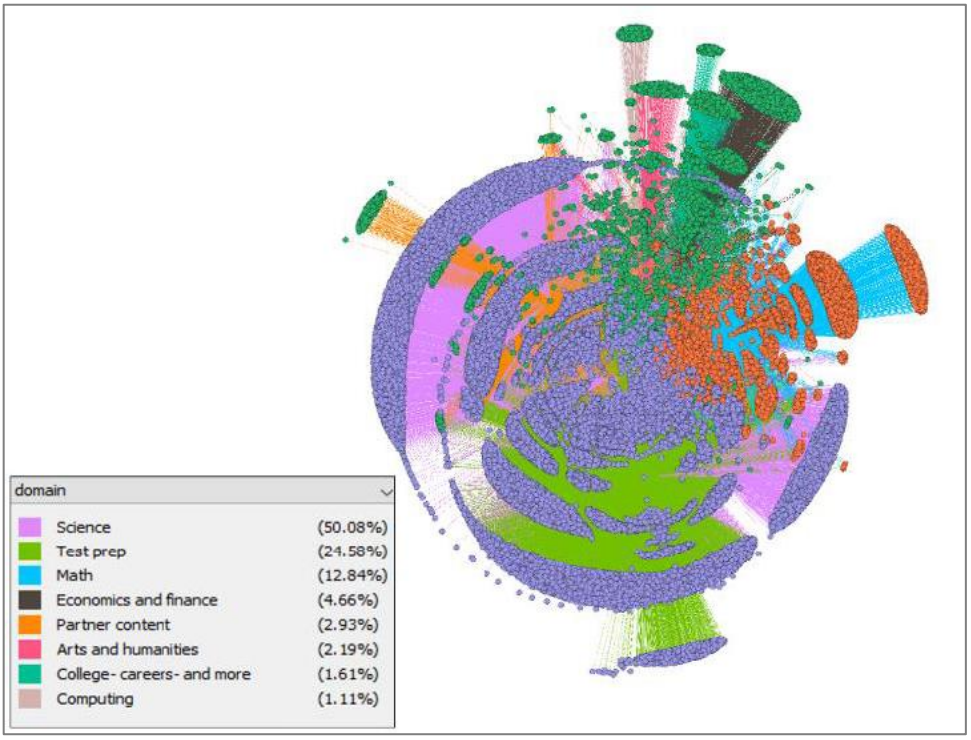


Figure 21. Largest PLM Community

Table 20. TOP 5 Attractive Topics - According to PLM Algorithm

Topic Title	Domain	Number of Posts
Physical processes	Test Prep	34,991
One-dimensional motion	Science	12,134
Chemistry of life	Science	11,523
Forces & Newton's laws of motion	Science	10,967
Chemical processes	Test Prep	10,523

8.3 Modularity

Modularity measure is applied to assess the effectiveness of a network cluster. Newman & Girvan (Newman and Girvan, 2004) figured out that modularity measures the quality of a particular segment of a network which means that it is used to measure the strength of partitioning a network into clusters or communities. This metric recognizes the good community as the one which is strongly connected internally and isolated from the rest of the network (Zhao et al., 2018) which makes it a good indicator to the intensity of users' interactions as well. The equation (1) below was used to define the modularity:

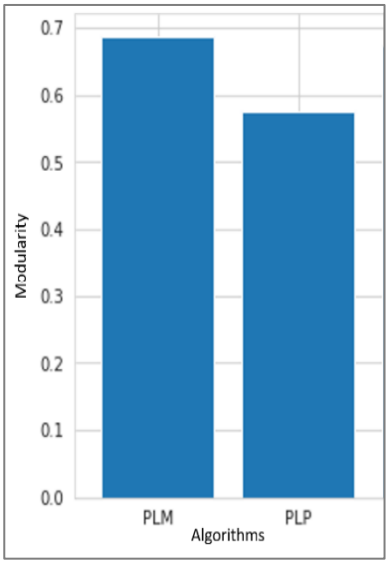


Figure 22. Modularity According to PLM & PLP Algorithms

$$Q = \frac{1}{2m}\sum_{vw}(A_{vw} - \frac{k_v k_w}{2m}) \qquad (1)$$

Where: (m) is the number of edges in the graph, (vw) are nodes (users) within the network, $(A_{vw)}$ is the adjacency matrix and $(k_v k_{w)}$ is the probability that a random edge would go between (v) and (w).

Figure 22 exhibits a comparison between modularity values for online communities that have been detected by applying both PLP and PLM algorithms on our dataset**.** The results showed that PLM algorithm is able to identify more strongly connected communities even though they are disjointed ones.

8.4 Density

Density measure is the proportion of direct ties or edges to the number of total possible ties between nodes (Harenberg et al., 2014). It is used to evaluate the architecture of the network and the quality of users' interrelations within communities and that is why it acts as a good indicator to assess the goodness of the applied algorithms. To calculate the density, we identified the maximum number of the possible edges as $\frac{1}{2}n(n-1)$ where (n) is the number of nodes (Metcalf and Casey, 2016). The following equation (2) defines the density of the graph$D(G)$:

$$D(G) = \frac{2|E|}{|V|(|V|-1)} \quad (2)$$

Where: $|E|$ is the number of edges and $|V|$ is the number of vertices. According to the results demonstrated in **Table 17** above, applying PLM algorithm helped in identifying high strongly connected communities with density 0.9168. Whereas applying PLP algorithm detected a relatively lower dense communities (0.6449) but they are still considered strong connected ones. The higher density reveals that users have deeper interrelations between them. Density is a good indicator that those users usually interact heavily by asking and answering to many questions. Those users have a strong engagement with their communities which exposes their cohesiveness. In other words, they demonstrate a high social presence in the learning process as they can interact online to develop the feeling of trust and engagement (Peacock and Cowan, 2019).

8.5 Clustering Coefficient

Clustering coefficient is another measure to assess the effectiveness of the detected communities. It measures the ability of the nodes (users) to group together. It works by identifying the triangles of the node which are the number of closed triplets in the node's neighbourhood (Davis et al., 2014) then by assessing the density of those triangles in the network. Equation (3) was used to calculate clustering coefficient for each node then formula (4) was applied to find the average clustering coefficient of our undirected graph for both PLP and PLM algorithms:

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i,\ e_{jk} \in E\}|}{k_i(k_i-1)} \quad (3)$$

$$C(G) = \frac{1}{|V|}\sum_{v \in V} c(v) \quad (4)$$

Where: (C_i) is the coefficient of the node (v_j), (N_i) is the neighbourhood of the same node, (e_{jk}) is the edge that connects (v_j) and (v_k), (k_i) represents the number of neighbours of a node while $C(G)$ is the clustering coefficient of the graph G which is the average over the clustering coefficients of its nodes and $|V|$ is the number of vertices.

Clustering coefficient is considered as the measure of neighbourhood connectivity and that is why the higher clustering coefficient indicates to the higher probability of

the inter-connections inside the detected communities which increase their dense and their goodness as well (Harenberg et al., 2014). After applying the above equations, we figured that the clustering coefficients of PLM detected communities is higher (Avg. 0.928) than the one of PLP detected communities (Avg. 0.284). This indicates the PLM communities are connected better and more strongly than PLP ones and the structure their learners tend to be fast in building up communities (Pham, Derntl, and Klamma, 2012).

8.6 Centrality measures

To examine the central roles played in the network and within communities, we applied the most commonly used centrality measures that capture various aspects of network structural properties (Lazega, Wasserman, and Faust, 1995). Those centrality measures were applied to detect the most significant roles and hubs in the network. **Table 21** shows descriptive statistics of the network centrality measures for the high performing communities detected by both PLP and PLM methods. In the following sub-sections we will demonstrate the performed centrality measures:

Table 21. Summary statistics of the network centrality measures for the high achieving PLP & PLP communities

Community	Degree Centrality	Eigenvector Centrality	Closeness Centrality	Betweenness Centrality
PLM.15	3,250	0.858218	0.646853	33,873.39
PLM.9	2,390	0.856615	0.592101	21,978.29
PLM.1	2,122	0.859679	0.482301	14,848.22
PLM.0	1,924	0.807328	0.402101	4,990.88
PLM.10	1,567	0.780971	0.201012	5,059.88
PLP.21	1,424	0.909184	0.719844	362.75
PLP.134	1,352	0.901959	0.520782	292.37
PLP.216	1,220	0.888815	0.507008	176.72
PLP.315	989	0.678914	0.484697	159.10
PLP.145	950	0.546276	0.483994	145.77

8.6.1 Degree centrality

One of the main power analysis measures is the degree centrality which is the simple count of neighbours. It works by simply counting the number of edges of the edge. Relatively speaking, the more edges within the graph, the more important the node. Those nodes with higher edges (in this case the communities) usually looks like

a “hub” of activity (McKnight, 2014). We calculated the degree centrality according to Freeman’s general equation (5):

$$C_D(G) = \sum_{v \in G} \frac{|\deg(v_*) - \deg(v)|}{(|V| - 1)(|V| - 2)} \quad (5)$$

Where: (v_*) is the node with the highest degree and $|V|$ is the number of nodes. When we applied PLM algorithm, the detected communities reflected slightly higher collaboration between them (2.4802) than the ones detected by PLP algorithm (1.8953). Although the top 5 PLP communities (in terms of degree centrality) are higher than the top PLM ones but the mean degree centrality demonstrates that PLM communities have higher communication activities between them.

8.6.2 Eigenvector Centrality (EVC)

Eigenvector centrality is a powerful tool to identify the central hubs because it measures the influence of the node in the network. It evaluates the node’s importance while giving consideration to the importance of its neighbours (Golbeck, 2013). This measurement is defined as a recursive function of the strength, centrality and number of neighbours’ connections (Ilyas and Radha, 2011). We applied the measure efficiently using NetworKit because it is based on power iteration (Staudt, Sazonovs, and Meyerhenke, 2014). Eigenvector centrality score (x_v) for the node v was found using the following formula (6):

$$x_v = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t \quad (6)$$

Where: (t) is a neighbour node to (v), $(a_{v,t})$ is the adjancey matrix which equals 1 if the two nodes are linked together otherwise it is equal to 0 and (x_t) is the eigenvector score of the node (t).

Table 17 and Table 21 above, exhibited that PLP communities have higher Eigenvector centrality than PLM communities which indicates that they demonstrated more importance, and influence in their network.

8.6.3 Closeness Centrality

Closeness centrality is a measure that estimates how fast the flow of information would be through a given node to other nodes (Ghalmane et al., 2019). It is calculated

as the average of the shortest distance from the node to every other node in the network. Closeness centrality score for the node v can be found using the following formula (7):

$$C(v) = \sum_{w \in G} \frac{N}{d(v,w)} \quad (7)$$

Where: $d(v,w)$ is the distance between nodes (v) and (w) and N is the number of nodes in the graph.

Our analysis showed that closeness centrality of PLP communities are higher than the PLM ones which indicates that PLP communities have closer relationships between them and that specifies their ability to reach any other node within a few hops even if it is very distant in the graph (Perez and Germon, 2016).

8.6.4 Betweenness Centrality

Betweenness centrality shows how much the node is between others. It measures the number of shortest paths between any couple of nodes that passes through the target node (Perez and Germon, 2016). It reflects the expected benefits for the node that bridge two or more distinct parts of the network (Lazega, Wasserman, and Faust, 1995). We calculated the betweenness centrality score using the following formula (7):

$$C(u) = \sum_{u \neq v \neq w} \frac{\sigma_{v,w}(u)}{\sigma_{v,w}} \quad (8)$$

Where: $\sigma_{v,w}$ is the total number of shortest paths between nodes (v) and (w), and $\sigma_{v,w}(u)$ is the number of those paths that pass through the node (u). According to our results, PLM communities play a bridging role in the network more than PLP ones. They depend on each other to be connected and they have higher amount of influence over the flow of information in the graph.

8.7 Centrality measures within communities

Using NetworKit, we tried to perform centrality measures to all the users (nodes) in our graph according to their locations in the detected communities using both methods (PLP and PLM). Those measures can be used to evaluate the users' activities inside the communities to understand their interaction patterns and the structure of their relationships. Due to the computational resource limitations, we managed to run degree and eigenvector centralities only. Closeness and betweenness required more processors and kernels to run in parallel.

In **Figure 23** we demonstrated the degree centrality for each user according to communities' distributions identified by both PLM and PLP algorithms. This measure identifies the highly connected users in their communities (Metcalf and Casey, 2016). The highly connected users are the highly interacted ones which means asking more questions and having the most answers. This identifies those users as the main hubs in their communities who participate effectively in their communities to extend understandings through constructive communications and that defines their cognitive presence (Garrison, 2016). The user with the highest degree centrality is a member in the both largest PLM and PLP communities. This member has connections with more than 11K users. Most of the associated interactions are answers to other users' questions. Most of those answers are related to Science and test-prep domains. We can assume that this user plays a facilitator role in different topics and domains and acts as a proactive participant especially in science.

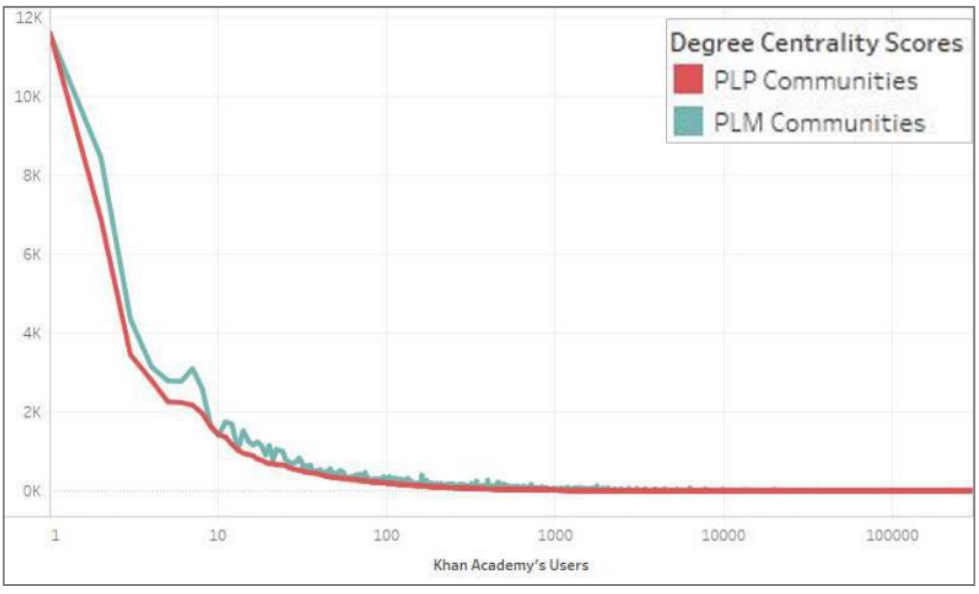


Figure 23. Degree Centrality for Users According PLP & PLM Algorithms

Also, we calculated the eigenvector centrality scores for all users in our dataset twice and according to PLM and PLP detected communities. In **Table 22** we identified that the top five central users or the main hubs in PLM communities are totally different than the ones in PLP communities. The top central user according to PLM communities had joined Khan Academy's repository in 2013. This user posted 209 answers and 30 questions on different topics mostly related to Math domain and some of them were posted on Science and Test-Prep domains. This user is connected to more than 60 different users in the same community which indicates that this user is an active learner which gives him/her a hub role. In contrary, the top central user according to PLP communities had joined Khan Academy in 2009. He posted more than 20,000 answers on different topics mainly related to Computing domain with some other answers related to Math as well. The most influencer PLP user interacted with more than 1,000 different users in the same community which gives him/her a facilitator role and a main hub as well.

Table 22. TOP 5 Central Users in PLM Communities & in PLP Communities

	PLM Detected Communities			PLP Detected Communities		
User	Community ID	EVC Score	Degree Score	Community ID	EVC Score	Degree Score
1st-PLM-EVC	32	0.95942	80	73769	0.18575	68
2nd-PLM-EVC	1923	0.95753	12	100518	0.95397	11
3rd-PLM-EVC	87	0.95549	72	146916	0.97845	55
4th-PLM-EVC	790	0.92900	11	169783	0.94281	8
5th-PLM-EVC	3726	0.92783	13	129052	0.85615	11
1st-PLP-EVC	46	0.25608	367	119907	0.99632	218
2nd-PLP-EVC	10	0.03940	333	116782	0.99538	214
3rd-PLP-EVC	9	0.00248	295	152813	0.99531	171
4th-PLP-EVC	24	0.02208	285	102567	0.99485	229
5th-PLP-EVC	39	0.01837	363	22360	0.99442	343

Those users are highly connected to other users in their communities which makes them high influencers as well. Moreover, they play a hub role which indicates that they have the power to act as opinion leaders (Yuqing Ren, Kraut, and Kiesler, 2007). One information worth to be notified here is neither the 1st-PLM-EVC user nor the

1st-PLP-EVC user has the top degree score and that is aligned with (Bonacich, 2007). In **Figure 24** we demonstrated the communities of the top five central users according to each algorithm.

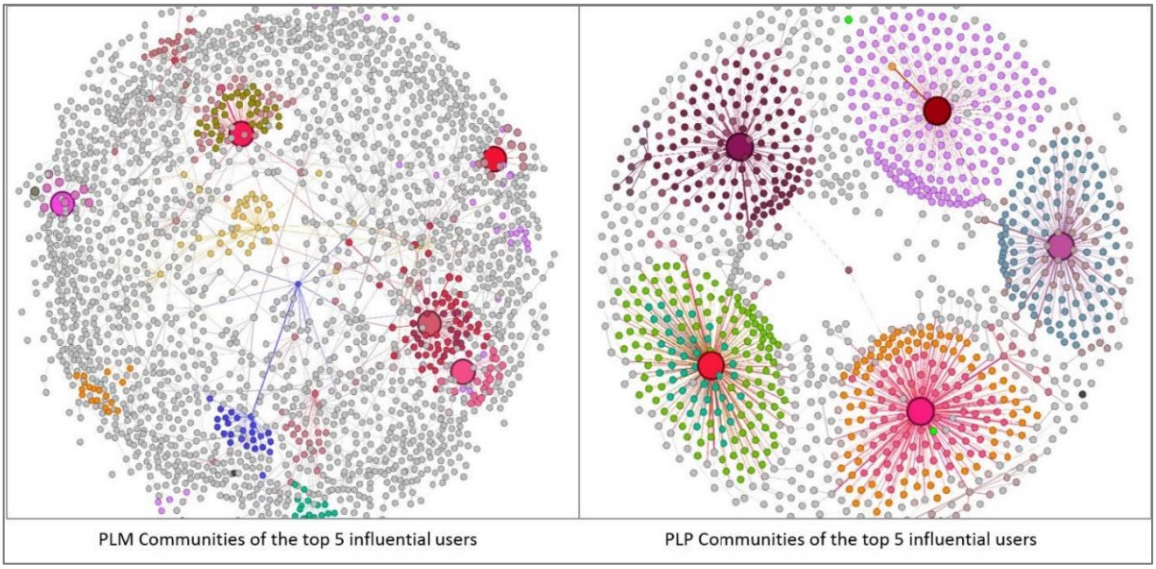


Figure 24. PLM & PLP Communities of the Top 5 Influential Users

9. Conclusions

On this dissertation we explored a representative case of online learning repositories, analysed its core features, and examined the patterns of their users' interactions in order to detect online communities and understand their properties using a large scale of data-driven analysis. Our range of analysis included descriptive analysis, inferential statistical analysis, community detection algorithms and social network analysis (SNA) techniques. We selected Khan Academy's repository to be our case of study because it has different structure and properties than the ones found in literature and there is a lack in investigating such concrete initiatives. We investigated this repository and collected a large dataset of their instructional video lessons, their characteristics, and their users' interactions using scraping techniques. Our study covered only one type of the provided learning objects which is the instructional video and we did not cover other types in the repository such as exercises and articles. This was because we aimed to focus on the most common and popular learning object in the repository to give deeper insights that may provide more focused and useful insights

The whole project presented in this dissertation provides valuable methodological support for investigating learning repositories and their users' interactions using community detection and SNA techniques to explore useful insights in the social structure and the users' learning behavioural patterns. Our findings can serve as a guide for researchers, educators and developers in assessing their online learning resources (Dawson, 2010) and in exploring emerging user communities to detect their movements, engagement and learning behaviour.

As a first step in our investigation, we performed descriptive analysis in order to discover the data of the population, organize it, and explore its core features. Then we performed some inferential analysis to investigate the associations and relationships between the learning objects and the users' interactions with them. We developed multi-level interaction profiles to classify the instructional videos according their interaction level and we examined their interactions behaviour to discover the main features that associated with them. In a later stage, we applied different community detection methods to identify the emerged online interactions communities. Finally, we used SNA techniques to assess those interactions and communities to generate

useful understandings. During our research we were able to achieve our objectives presented in chapter 2.

R1: Assessing Online Learning Repository with Descriptive Statistical Analysis

Chapter 6 presents descriptive statistical analysis to explore Khan Academy's dataset and investigate its features, properties, and the association between its learning objects and the users' interactions. We performed general descriptive analysis that is related to the performance, evolution, and characteristics of the repository itself and we implemented some analysis to investigate and describe the users' interactions with the video lessons inside the repository. The main findings in this chapter are:

The growth of Khan Academy's repository was measured using two dimensions: content growth over time and user-base growth over time. We found that the number of learning objects grows linearly but with an initial fast start-up phase by publishing the largest amount of the materials in the first year. This finding differs from what was found by Ochoa in his study which concluded that the number of learning objects in repositories grows linearly with a slow growth initial phase which may take from 1 to 3 years.

The growth of the user-base found to follow 2-degree polynomial distribution. This makes sense as the number of users joining the repository increased at the beginning as a result of increasing the popularity of the repository, the peak was in 2016 with the highest number of users joining the repository. Interestingly, the maturity level didn't last long, the number of joined users dropped in 2017. This is maybe related to the increasing competition between online learning repositories and emerging new substitutes and alternatives. Another possible reason is raising the interest of having institutional online learning initiatives inside the academic institutions which requires involvement of their users.

After analysing the geographical distribution of the users of the repository, we realized that more than 45% of the users are located in North America (United States and Canada). Other regions such as India, UK, and Australia represent a good potential markets for Khan Academy to grow and to focus on their needs and demands. The dataset was collected from the main Khan Academy's platform which

is the English version and that explains why most of the users are from English-speaking countries. Khan Academy is working on developing localized platforms with content translated to different language. Those can be potential to further investigations in the future.

We found that the average video lessons duration in most of the domains is between 300 and 500 seconds which means from 5 to 10 mins. This is aligned with the recommendation in the literature to segment the educational videos into short chunks, with length that should adhere to the 10 minute attention span (Lagerstrom, Johanes, and Ponsukcharoen, 2015, Chauvet et al., 2020). Also, we discovered that more than 50% of the gathered users did not complete any of the watched videos. This behaviour can be a response to many obstructions that may divert the users' attention and increase the distraction. This can be eliminated and minimized by focusing on three elements during the design phase: using strategies to increase the cognitive load such as signalling and weeding information, using methods to increase the student engagement such as speaking with enthusiasm, and using strategies that concentrate on the active learning (Brame, 2016),.

The largest domain and the most growing one is found to be the Math domain which makes sense because it attracts the largest and most diversified segment of the users which includes users with different age and educational levels starting from KG level to college level. On the other hand, Test Prep domain found to be an attracting one with a good potential to grow more and focus on. This could be figured out by tracking the user's engagement and interactions with the videos related to this domain. Knowing that more than 30% of the students entering the higher education in U.S. are not ready and found to be under-prepared for the college-level work, this field can act as a competitive advantage, as it is a new interesting one and highly demanded (Bailey et al., 2016, Perin and Holschuh, 2019)

We also presented a quantitative study to investigate the relation between the users' interactions and the learning objects which are the instructional videos in the repository through performing correlation and regression analysis. We proposed a group of metrics related to the content of the instructional videos and their usage and we examined the behaviour of the user's interactions toward those metrics. We proposed a multi-level profile classification for the instructional videos according to

their user's interactions level (low, medium, and high) in a step to group the level of interactions.

We found that one of the proposed metrics which is the domain type does not associate significantly with the number of user interactions especially with the ones posted on the medium and high profiles videos. This was interesting, because it was expected that the users interact more with the most growing and popular domains such as the math domain but in fact there is no correlation.

Another examined metric is the publishing year, which represents the age of the learning material. We found that the number of user interactions associated with videos following the low profile group are strongly and positively correlated with the publishing year of those videos. Interestingly, most of those videos are related to the recently added domains which are relatively new ones. On the other hand, there is no significant correlation found between the user interactions in the medium and high profile videos with their publishing year although most of them were published long time ago. This indicates that the attractiveness and popularity of the instructional video is not related to its age.

Two of the proposed metrics (video length and reuse rate) correlate well with the user interactions and the popularity of the video. Video length found to have a significant weak inverse relationship with the number of users interactions in videos related to all profiles. That means the shorter videos are more popular and attractive to user interactions than the longer ones although users will watch longer videos if it is justified (Alpert and Hodkinson, 2019). This finding is also supported in the literature several times (Yu et al., 2006, Chen et al., 2017). The other substantial metric that is associated strongly and positively with user interactions in all profiles is the reuse rate. Reusing the video in different subjects and sites raises its viewership thereby the user interactions will increase. The same result was supported widely in the literature (Frank and Suzuka, 2016, Frango Silveira, 2016, Pérez-Sanagustín et al., 2017, Kim and Suzuka, 2020).

R2: Detecting Communities in Online Learning Repository

Another experimental study in our research was applied to detect and investigate communities in online learning repository is presented in chapter 7. We created a graph from the users' interactions collected in Khan Academy's dataset. Our

undirected graph represented the relationship between users who are interacting around Khan Academy's contents, specifically videos. We examined the graph using different clustering techniques and tools. We performed four different algorithms using NetworKit and NetowrkX libraries and compared between them in terms of efficiency. The main contributions of this chapter are:

NetworKit libraries found to be more effective and efficient in utilising time and computational resources than NetworkX libraries especially with large-scale datasets. That's why we continued working with it during the next phase.

Parallel label propagation algorithm (PLP) found to be the fastest algorithm in detecting communities with 1.78 seconds running time. It decomposed the graph into more than 34K overlapped communities which is the largest number of communities detected. Around 55% of those identified communities are well-connected to each other.

Parallel Louvain method (PLM) found to be the best method in terms of modularity. It detected around 11K communities with a modularity of 0.687 which shows the strength of partitioning the network. The network diameter of PLM communities is the lowest one which indicates that all the communities are in proximity and the graph is more compact (Scardoni and Lau, 2012). 98% of the detected communities by PLM are strongly connected to each other which rank it the highest one.

PLP and PLM demonstrated the most efficient community detection methods of the applied ones in terms running time and quality scores. The detected communities are analysed in more details to find more insights inside the network structure of the learning repository.

The detected communities found to be related to more than one domain. Most of the PLP detected communities are evolving around Science and Test prep videos while PLM communities are mainly evolving around math and science. It means that their users are active and interacting across different domains. This is aligned with our findings in the previous chapter that the users interact with each other in the repository without any correlation to the domain type.

R3: SNA Measures and Users' Interactions

In chapter 8 of this research we used different social network analysis (SNA) measures to assess the emerged online learning communities and analyse the user engagement, behaviour, and movement to detect their presence and roles and to enhance the understanding of learning through interactions. Our analysis included different measures used to examine the intensity, goodness, and effectiveness (Wagenseller, Wang, and Wu, 2018) such as community size distribution, density, modularity, and, clustering coefficients. We also performed different centrality measures across communities and inside communities to identify the central, important, and powerful roles in the network who are responsible for the communities' cohesion. The main findings of this chapter are:

The PLP community size distribution found to follow a regular power-law distribution and this is the case in most of the real world networks (Labatut and Balasque, 2013) while the PLM community sizes generated a long tailed one.

Modularity scores for both PLP and PLM communities indicated that although PLM method produces disjoint communities, but they perform stronger connections than PLP ones.

Density scores evaluated the network's architecture and the quality of user interrelation in the communities. Our results demonstrated that PLM method managed to identify stronger connected communities with higher interrelations inside those communities and that can act as an indicator to the high social presence presented through the strong engagement in asking and answering within communities.

Also PLM method proved more goodness through having higher clustering coefficient which shows the higher probability to connect the neighbours inside communities. The structure of the learners in PLM communities tend to be fast in building communities and creating cohesiveness groups.

Centrality measures applied to the detected communities proved that PLM method identified communities with higher degree centralities which demonstrates higher communication activities across those groups. Also, they have a higher Betweenness scores which reflects that they have more groups acting as information bridges across the network and they have higher influence over the flow of information. On the other

hand, PLP communities have more groups acting as central influencing actors with higher eigenvector scores. While the higher closeness scores demonstrated that those PLP groups provide a faster flow of the information between them.

The performed centrality measures on the users within communities helped in identifying the highly connected users and the most influential users in those groups. The most connected user and the central to other nodes is the same user according to both methods. This user had connections with more than 11K other users which gives an indication that he has a facilitator role in multiple topics and domains. This user can be considered as a proactive participant especially in science.

According to the eigenvector centrality scores, the most influential user in PLM communities is different than the one identified in PLP communities and they are playing different influencing roles. PLP highest influential user interacted with more than 1000 different users in different domains which gives the impression that this user plays a facilitator role. PLM highest influential user had connections with around 60 users mainly in math domain asking and answering questions. This gives the impression that the user is an active learner and plays the role of the hub of information. Such central users are the ones who are responsible for holding up their communities.

10. Future Work

As the field of online learning unfolds, ample research opportunities are provided to understand emerging learning networks, the learning environment, and learners' behaviour. Here are some interesting research directions:

While the results demonstrate the potential of retrieving user data from learning repositories to understand their emerging networks and communities and to systematically define learning characteristics and social behaviour in any online learning environment, additional work would be needed to assess whether the conclusions of this study could be extrapolated to other online learning repository types.

The massive number of interactions produces from interacting with online learning initiatives can be converted in to a valuable information to educators, instructional designers, decision makers, and learners. This is an opportunity to implement dynamic analysis process to study and analyse such interactions' datasets in regular basis (Giannakos, Chorianopoulos, and Chrisochoides, 2015) to enhance the whole learning process and specifically to enhance the quality of learning objects.

There is a research opportunity in studying other forms of interactions that can be found and gathered from interacting with learning initiatives in different ways including reusing, sharing, posting interactions, voting, clicking, hitting interactive buttons and practising exercises and simulations. In our research the posts type (questions and answers) was the key identifier of the relationship between users. Future work can analyse different types of interactions and from other perspectives.

Another interesting future direction will be applying sentiment analysis to the contents of the user interactions which helps in detecting their emotional tone, their opinions behind the use of the learning objects and any hidden opportunity to improve and develop.

A potential research idea can attempt investigating the integrated interactive environment provided by the online learning repository such as the whole virtual classroom which includes many types of learning objects and materials such as instructional videos, interactive buttons, lecture notes, slides, and exercises. Joining the parts to assess the big picture and its impact on education will contribute in

building new innovative educational strategies and in enhancing the learning performance, efficiency and the whole experience (Yip et al., 2019).

Further research collaboration between data scientists and instructional designers is needed to study the users interactions and comprehend their learning behaviours and reactions in order to enhance designing educational content and learning technologies with higher quality (Saurabh and Gautam, 2019).

Another collaboration between data scientists and educators is needed to examine the user interactions with online learning repositories by applying various community detection methods especially for dynamic networks and SNA techniques to identify the emerging learning communities, analyse their movements, and assess the learners' behaviours in order to help in developing better educational strategies that aim to increase the learners' engagement and elevate their performance in online learning processes.

Online learning repositories need to prove their effectiveness and the educators need to monitor and focus more on the quality inside those learning environments. Nowadays, researchers and educators are thinking thoroughly, planning and implementing new educational strategies that entrench online learning in the core of the educational process. Decision makers need to understand that an optimal educational strategy that enables the flexibility to maintain teaching and learning anywhere, anytime, for anyone and under all circumstances must embed more well-planned online learning experiences. This is very critical and it points to the same goal of our research which is the possibility of creating a framework for studying online learning repositories, monitoring their quality, analysing their learning communities, and evaluating user interactions with them using different analysis techniques to enhance learning experiences and maximize learning outcomes.

12. References

Adraoui, M., Retbi, A., Idrissi, M.K., and Bennani, S. (2020). A New Approach to Detect At-Risk Learning Communities in Social Networks. *EMENA-ISTL 2019: Innovation in Information Systems and Technologies to Support Learning Research.*, *7*(1), 75–84. **DOI:10.1007/978-3-030-36778-7_9**

Adraoui, M., Retbi, A., Idrissi, M.K., and Bennani, S. (2019). A New Algorithm to Detect and Evaluate Learning Communities in Social Networks: Facebook Groups. *International Journal of Emerging Technologies in Learning*, *14*(23), 165–179. **DOI:10.3991/ijet.v14i23.10889**

Adraoui, M., Retbi, A., Idrissi, M.K., and Bennani, S. (2018a). Evaluate Learning Communities in the Online Social Media. *Proceedings of the 12th International Conference on Intelligent Systems: Theories and Applications* ACM, 1–6. **DOI:10.1145/3289402.3289505**

Adraoui, M., Retbi, A., Idrissi, M.K., and Bennani, S. (2018b). Network Visualization Algorithms to Evaluate Students in Online Discussion Forums: A Simulation Study. *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)* IEEE, 1–6. **DOI:10.1109/ISACV.2018.8354020**

Ali, Z. and Bhaskar, S.B. (2016). Basic Statistical Tools in Research and Data Analysis. *Indian Journal of Anaesthesia*, *60*(9), 662–669. **DOI:10.4103/0019-5049.190623**

Alpert, F. and Hodkinson, C.S. (2019). Video Use in Lecture Classes: Current Practices, Student Perceptions and Preferences. *Education and Training*, *61*(1), 31–45. **DOI:10.1108/ET-12-2017-0185**

Bailey, T., Bashford, J., Boatman, A., Squires, J., Weiss, M., Doyle, W., Valentine, J.C., LaSota, R., Polanin, J.R., Spinney, E., Wilson, W., Yeide, M., and Young, S.H. (2016). Strategies for Postsecondary Students in Developmental Education–A Practice Guide

for College and University Administrators, Advisors, and Faculty. *What Works Clearinghouse*,

Bakharia, A. and Dawson, S. (2011). SNAPP: A Bird's-Eye View of Temporal Participant Interaction. *ACM International Conference Proceeding Series* Association for Computing Machinery, 168–173. **DOI:10.1145/2090116.2090144**

Balbay, S. (2018). Educational Analytics on an Opencourseware. *International Online Journal of Education and Teaching*, *5*(May), 673–685

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10). **DOI:10.1088/1742-5468/2008/10/P10008**

Bonacich, P. (2007). Some Unique Properties of Eigenvector Centrality. *Social Networks*, *29*(4), 555–564. **DOI:10.1016/J.SOCNET.2007.04.002**

Bonk, C.J., Lee, M.M., Reeves, T.C., and Reynolds, T.H. (2015). MOOCs and Open Education around the World. *MOOCs and Open Education Around the World*, *33*, 1–358. **DOI:10.4324/9781315751108**

Brame, C.J. (2016). Effective Educational Videos: Principles and Guidelines for Maximizing Student Learning from Video Content. *CBE Life Sciences Education*, *15*(4), 6.1-6.6. **DOI:10.1187/cbe.16-03-0125**

Brandes, U. (2015). Social Network Algorithms and Software. in *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier, 454–460. **DOI:10.1016/B978-0-08-097086-8.43121-1**

Caswell, T., Henson, S., Jensen, M., and Wiley, D. (2008). Open Educational Resources: Enabling Universal Education. *International Review of Research in Open and Distance Learning*, *9*(1). **DOI:10.19173/irrodl.v9i1.469**

Cechinel, C., Camargo, S.D.S., Sànchez-Alonso, S., and Sicilia, M.Á. (2014). Towards

Automated Evaluation of Learning Resources inside Repositories. in *Recommender Systems for Technology Enhanced Learning: Research Trends and Applications*. Springer, 25–46. **DOI:10.1007/978-1-4939-0530-0_2**

Cechinel, C., Sánchez-Alonso, S., and García-Barriocanal, E. (2011). Statistical Profiles of Highly-Rated Learning Objects. *Computers and Education*, *57*(1), 1255–1269. **DOI:10.1016/j.compedu.2011.01.012**

Cechinel, C., Sánchez-Alonso, S., Sicilia, M.-Á., and de Mattos, M.C. (2010). Descriptive Analysis of Learning Object Material Types in MERLOT. in *Communications in Computer and Information Science*, Springer Berlin Heidelberg, 331–341. **DOI:10.1007/978-3-642-16552-8_30**

Cela, K.L., Sicilia, M.Á., and Sánchez, S. (2015). Social Network Analysis in E-Learning Environments: A Preliminary Systematic Review. *Educational Psychology Review*, *27*(1), 219–246. **DOI:10.1007/s10648-014-9276-0**

Chauvet, P., Botchorishvili, R., Curinier, S., Gremeau, A.-S., Campagne-Loiseau, S., Houlle, C., Canis, M., Rabischong, B., and Bourdel, N. (2020). What Is a Good Teaching Video? Results of an Online International Survey. *Journal of Minimally Invasive Gynecology*, *27*(3), 738–747. **DOI:10.1016/j.jmig.2019.05.023**

Chen, Z., Cui, L., Jiang, Y., and Wang, Z. (2017). Understanding Viewing Engagement and Video Quality in a Large-Scale Mobile Video System. *Proceedings - IEEE Symposium on Computers and Communications*. 1271–1277. **DOI:10.1109/ISCC.2017.8024699**

Christoforos, M., J., D.A., Charlotte, S., Irene, K., and George, M. (2019). Learning, Friendship and Social Contexts: Introducing a Social Network Analysis Toolkit for Socially Responsive Classrooms. *International Journal of Educational Management*, *33*(6), 1255–1270. **DOI:10.1108/IJEM-03-2018-0103**

Churchill, D. (2007). Towards a Useful Classification of Learning Objects. *Educational Technology Research and Development*, *55*(5), 479–497. **DOI:10.1007/s11423-006-**

9000-y

Clements, K., Pawlowski, J., and Manouselis, N. (2015). Open Educational Resources Repositories Literature Review – Towards a Comprehensive Quality Approaches Framework. *Computers in Human Behavior*, *51*, 1098–1106. **DOI:10.1016/j.chb.2015.03.026**

Clements, K., Pawlowski, J., and Manouselis, N. (2014). Why Open Educational Resources Repositories Fail - Review of Quality Assurance Approaches. *EDULEARN14 Proceedings. 6th International Conference on Education and New Learning Technologies Barcelona, Spain* 929–939. **DOI:10.1093/shm/hkr167**

Corallo, A., Maggio, M. De, Grippa, F., and Passiante, G. (2010). A Methodological Framework to Monitor the Performance of Virtual Learning Communities. *Human Factors and Ergonomics In Manufacturing*, *20*(2), 135–148. **DOI:10.1002/hfm.20205**

Costa, C., Alvelos, H., and Teixeira, L. (2012). The Use of Moodle E-Learning Platform: A Study in a Portuguese University. *Procedia Technology*, *5*, 334–343. **DOI:10.1016/j.protcy.2012.09.037**

Costley, J. and Lange, C. (2016). The Effects of Instructor Control of Online Learning Environments on Satisfaction and Perceived Learning. *Electronic Journal of E-Learning*, *14*(3), 169–180

Creswell, J.W. (2003). A Framework for Design. in *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Third Edit. Sage, 9–11

Cuevas, A., Febrero, M., and Fraiman, R. (2004). An Anova Test for Functional Data. *Computational Statistics and Data Analysis*, *47*(1), 111–122. **DOI:10.1016/j.csda.2003.10.021**

Davis, J., Ojeda, T., Murphy, S.P., Bengfort, B., and Dasgupta, A. (2014). *Practical Data Science Cookbook*,. Packt Publishing Ltd

Dawson, S. (2010). '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. **DOI:10.1111/j.1467-8535.2009.00970.x**

Derényi, I., Palla, G., and Vicsek, T. (2005). Clique Percolation in Random Networks. *Phys. Rev. Lett.*, *94*(16), 160202. **DOI:10.1103/PhysRevLett.94.160202**

Downes, S. (2001). Learning Objects: Resources for Distance Education Worldwide. *International Review of Research in Open and Distance Learning*, *2*(1), 66–93. **DOI:10.19173/irrodl.v2i1.32**

Eguigure, Y.A., Zapata, A., Menendez, V., and Prieto, M. (2011). Quality Evaluation Model for Learning Objects From Pedagogical Perspective. a Case of Study. *Iberoamerican Journal of Applied Computing*, *1*(2)

Forster, R. (2018). Parallel Louvain Community Detection Optimized for GPUs. *ArXiv*,

Fortunato, S. and Hric, D. (2016). Community Detection in Networks: A User Guide. *Physics Reports*, *659*, 1–44. **DOI:10.1016/j.physrep.2016.09.002**

Frango Silveira, I. (2016). OER and MOOC: The Need for Openness. *Proceedings of the 2016 InSITE Conference*, *13*, 924. **DOI:10.28945/3485**

Frank, R.D. and Suzuka, K. (2016). Examining the Reuse of Qualitative Research Data: Digital Video in Education. *Archiving Conference*. 146–151

Free SAT Practice from Khan Academy, (2018). **https://collegereadiness.collegeboard.org/about/benefits/khan-academy-practice** [Retrieved 8 February 2020]

Garrison, D.R. (2016). *E-Learning in the 21st Century: A Community of Inquiry Framework for Research and Practice, Third Edition*,. Routledge. **DOI:10.4324/9781315667263**

Garza, S.E. and Schaeffer, S.E. (2019). Community Detection with the Label Propagation Algorithm: A Survey. *Physica A: Statistical Mechanics and Its Applications*, *534*,

122058. **DOI:10.1016/j.physa.2019.122058**

Ghalmane, Z., Cherifi, C., Cherifi, H., and Hassouni, M. El (2019). Centrality in Complex Networks with Overlapping Community Structure. *Scientific Reports*, *9*(1), 10133. **DOI:10.1038/s41598-019-46507-y**

Ghilay, Y. and Ph, D. (2019). Effectiveness of Learning Management Systems in Higher Education : Views of Lecturers with Different Levels of Activity in LMSs. *Journal of Online Higher Education*, *3*(2), 29–50

Ghosh, S., Halappanavar, M., Tumeo, A., and Kalyanarainan, A. (2019). Scaling and Quality of Modularity Optimization Methods for Graph Clustering. *2019 IEEE High Performance Extreme Computing Conference, HPEC 2019*. 1–6. **DOI:10.1109/HPEC.2019.8916299**

Giannakos, M.N., Chorianopoulos, K., and Chrisochoides, N. (2015). Making Sense of Video Analytics: Lessons Learned from Clickstream Interactions, Attitudes, and Learning Outcome in a Video-Assisted Course. *International Review of Research in Open and Distance Learning*, *16*(1), 260–283. **DOI:10.19173/irrodl.v16i1.1976**

Golbeck, J. (2013). Network Structure and Measures. in *Analyzing the Social Web*, Boston: Morgan Kaufmann, 25–44. **DOI:10.1016/b978-0-12-405531-5.00003-1**

Grunspan, D.Z., Wiggins, B.L., and Goodreau, S.M. (2014). Understanding Classrooms through Social Network Analysis: A Primer for Social Network Analysis in Education Research. *CBE Life Sciences Education*, *13*(2), 167–178. **DOI:10.1187/cbe.13-08-0162**

Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., Padmanabhan, K., and Samatova, N. (2014). Community Detection in Large-Scale Networks: A Survey and Empirical Evaluation. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(6), 426–439. **DOI:10.1002/wics.1319**

Higgs, P.E., Meredith, S., and Hand, T. (2003). Technology for Sharing: Researching Learning Objects and Digital Rights Management. *Flexible Learning Leader Report*

2002,

Hodgins, W. and Wiley, D.A. (2002). *The Future of Learning Objects; The Instructional Use of Learning Objects: Online Version*, Agency for Instructional Technology

Hosseinpoor, M., Parvin, H., Nejatian, S., Rezaie, V., Bagherifard, K., Dehzangi, A., Beheshti, A., and Alinejad-Rokny, H. (2020). Proposing a Novel Community Detection Approach to Identify Co-Interacting Genomic Regions. *Mathematical Biosciences and Engineering*, *17*(3), 2193–2217. **DOI:10.3934/mbe.2020117**

Huda, M., Maseleno, A., Atmotiyoso, P., Siregar, M., Ahmad, R., Jasmi, K.A., Muhamad, N.H.N., Mustari, M.I., and Basiron, B. (2018). Big Data Emerging Technology: Insights into Innovative Environment for Online Learning Resources. *International Journal of Emerging Technologies in Learning*, *13*(1), 23–36. **DOI:10.3991/ijet.v13i01.6990**

Ilyas, M.U. and Radha, H. (2011). Identifying Influential Nodes in Online Social Networks Using Principal Component Centrality. *IEEE International Conference on Communications*. 1–5. **DOI:10.1109/icc.2011.5963147**

Iniesta-Bonillo, M.A., Sánchez-Fernández, R., and Schlesinger, W. (2013). Investigating Factors That Influence on ICT Usage in Higher Education: A Descriptive Analysis. *International Review on Public and Nonprofit Marketing*, *10*(2), 163–174. **DOI:10.1007/s12208-013-0095-7**

Javed, M.A., Younis, M.S., Latif, S., Qadir, J., and Baig, A. (2018). Community Detection in Networks: A Multidisciplinary Review. *Journal of Network and Computer Applications*, *108*(C), 87–111. **DOI:10.1016/j.jnca.2018.02.011**

Jimoyiannis, A. and Angelaina, S. (2012). Towards an Analysis Framework for Investigating Students' Engagement and Learning in Educational Blogs. *Journal of Computer Assisted Learning*, *28*(3), 222–234. **DOI:10.1111/j.1365-2729.2011.00467.x**

Jimoyiannis, A., Tsiotakis, P., and Roussinos, D. (2013). Social Network Analysis of Students' Participation and Presence in a Community of Educational Blogging.

Interactive Technology and Smart Education, *10*(1), 15–30. **DOI:10.1108/17415651311326428**

Kay, R.H. and Knaack, L. (2008). A Multi-Component Model for Assessing Learning Objects: The Learning Object Evaluation Metric (LOEM). *Australasian Journal of Educational Technology*, *24*(5), 574–591. **DOI:10.14742/ajet.1192**

Kelley, S., Goldberg, M., Magdon-Ismail, M., Mertsalov, K., and Wallace, A. (2012). Defining and Discovering Communities in Social Networks. in *Springer Optimization and Its Applications*, vol. 57. Boston, MA: Springer US, 139–168. **DOI:10.1007/978-1-4614-0754-6_6**

Kelly, D.P. and Rutherford, T. (2017). Khan Academy as Supplemental Instruction: A Controlled Study of a Computer-Based Mathematics Intervention. *International Review of Research in Open and Distance Learning*, *18*(4), 70–77. **DOI:10.19173/irrodl.v18i4.2984**

Khan, B.S. and Niazi, M.A. (2017). Network Community Detection: A Review and Visual Survey. *ArXiv*, *abs/1708.0*

Kim, J. and Suzuka, K. (2020). Reusing Qualitative Video Data: Matching Reuse Goals and Criteria for Selection. *Aslib Journal of Information Management*, *72*(3), 395–419. **DOI:10.1108/AJIM-08-2019-0215**

Kovanović, V., Joksimović, S., Poquet, O., Hennis, T., de Vries, P., Hatala, M., Dawson, S., Siemens, G., and Gašević, D. (2019). Examining Communities of Inquiry in Massive Open Online Courses: The Role of Study Strategies. *Internet and Higher Education*, *40*, 20–43. **DOI:10.1016/j.iheduc.2018.09.001**

Kozawa, Y., Amagasa, T., and Kitagawa, H. (2017). GPU-Accelerated Graph Clustering via Parallel Label Propagation. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* ACM, 567–576. **DOI:10.1145/3132847.3132960**

Krämer, B.J. (2010). Learning Objects: Standards, Metadata, Repositories, and LCMS - Edited by Keith Harman & Alex Koohang. *British Journal of Educational Technology*, *41*(6), 973–973. **DOI:10.1111/j.1467-8535.2010.01135_1_4.x**

Labatut, V. and Balasque, J.M. (2013). Detection and Interpretation of Communities in Complex Networks: Practical Methods and Application. in *Computational Social Networks: Tools, Perspectives and Applications*. vol. 9781447140. Springer, 81–113. **DOI:10.1007/978-1-4471-4048-1_4**

Lagerstrom, L., Johanes, P., and Ponsukcharoen, U. (2015). The Myth of the Six Minute Rule: Student Engagement with Online Videos. *ASEE Annual Conference and Exposition, Conference Proceedings*. 14–17

Lazega, E., Wasserman, S., and Faust, K. (1995). *Social Network Analysis: Methods and Applications*,. vol. 36. Cambridge university press. **DOI:10.2307/3322457**

Lee, Y. (2018). Effect of Uninterrupted Time-on-Task on Students' Success in Massive Open Online Courses (MOOCs). *Computers in Human Behavior*, *86*, 174–180. **DOI:10.1016/j.chb.2018.04.043**

Lin, C.H., Zhang, Y., and Zheng, B. (2017). The Roles of Learning Strategies and Motivation in Online Language Learning: A Structural Equation Modeling Analysis. *Computers and Education*, *113*, 75–85. **DOI:10.1016/j.compedu.2017.05.014**

Linhares, C.D.G., Ponciano, J.R., Pereira, F.S.F., Rocha, L.E.C., Paiva, J.G.S., and Travençolo, B.A.N. (2020). Visual Analysis for Evaluation of Community Detection Algorithms. *Multimedia Tools and Applications*, **DOI:10.1007/s11042-020-08700-4**

Lockyer, L., Heathcote, E., and Dawson, S. (2013). Informing Pedagogical Action: Aligning Learning Analytics With Learning Design. *American Behavioral Scientist*, *57*(10), 1439–1459. **DOI:10.1177/0002764213479367**

Loeb, S., Dynarski, S., McFarland, D., Morris, P., Reardon, S., and Reber, S. (2017). Descriptive Analysis in Education: A Guide for Researchers. *U.S. Department of*

Education, Institute of Education Sciences. National Center for Education Evaluation and Regional Assistance, (March), 1–40. **DOI:10.1094/PDIS.2003.87.5.550**

Loff, S. (2014). *NASA, Khan Academy Collaborate to Bring STEM Opportunities to Online Learners*, **https://www.nasa.gov/content/nasa-khan-academy-collaborate-to-bring-stem-opportunities-to-online-learners/** [Retrieved 8 February 2020]

Lu, X., Liu, X.W., and Zhang, W. (2020). Diversities of Learners' Interactions in Different MOOC Courses: How These Diversities Affects Communication in Learning. *Computers & Education*, *151*, 103873. **DOI:10.1016/j.compedu.2020.103873**

Macià, M. and García, I. (2016). Informal Online Communities and Networks as a Source of Teacher Professional Development: A Review. *Teaching and Teacher Education*, *55*, 291–307. **DOI:10.1016/j.tate.2016.01.021**

Mahali, D.B., Changilwa, P., and Anyona, J. (2019). The Influence of Level of Training in LMS and Student Utilization of LMS in Public Universities in Tanzania. *Journal of Education*, *2*(4), 19–46

Marín, V.I., Orellana, M.L., and Peré, N. (2019). Open Educational Resources for Research Training: Quality Assurance through a Collaborative Evaluation. *Research in Learning Technology*, *27*(0 SE-Original Research Articles). **DOI:10.25304/rlt.v27.2271**

Martinez, S. (2014). OCW (OpenCourseWare) and MOOC (Open Course Where?). *Proceedings of OpenCourseWare Consortium Global*,

McHugh, M.L. (2012). The Chi-Square Test of Independence. *Biochemia Medica*, *23*(2), 143–149. **DOI:10.11613/BM.2013.018**

McKnight, W. (2014). Graph Databases. in *Information Management*, Boston: Morgan Kaufmann, 120–131. **DOI:10.1016/b978-0-12-408056-0.00012-6**

Mclaren, J. and Donaldson, J. (2018). Learning Analytics Suggest a Positive Experience. *17th European Conference on E-Learning* 670–678

De Medio, C., Limongelli, C., Marani, A., and Taibi, D. (2019). Retrieval of Educational Resources from the Web: A Comparison Between Google and Online Educational Repositories. in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11841 LNCS. Cham: Springer International Publishing, 28–38. **DOI:10.1007/978-3-030-35758-0_3**

Metcalf, L. and Casey, W. (2016). *Cybersecurity and Applied Mathematics*, Elsevier. **DOI:10.1016/C2015-0-01807-X**

Moore, J.L., Dickson-Deane, C., and Galyen, K. (2011). E-Learning, Online Learning, and Distance Learning Environments: Are They the Same?. *The Internet and Higher Education*, *14*(2), 129–135. **DOI:10.1016/j.iheduc.2010.10.001**

NetworKit, (2013). **https://networkit.github.io/** [Retrieved 27 November 2019]

NetworkX, (2008). **https://networkx.github.io/** [Retrieved 27 November 2019]

Newman, M.E.J. (2004). Fast Algorithm for Detecting Community Structure in Networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *69*(6), 5. **DOI:10.1103/PhysRevE.69.066133**

Newman, M.E.J. and Girvan, M. (2004). Finding and Evaluating Community Structure in Networks. *Phys. Rev. E*, *69*(2), 26113. **DOI:10.1103/PhysRevE.69.026113**

Norman, H., Nordin, N., Yunus, M.M., and Ally, M. (2018). Instructional Design of Blended Learning with MOOCs and Social Network Analysis. *Advanced Science Letters*, *24*(11), 7952–7955. **DOI:10.1166/asl.2018.12464**

Ochoa, X. (2011). Learnometrics: Metrics for Learning Objects. *ACM International Conference Proceeding Series*, 1–8. **DOI:10.1145/2090116.2090117**

Ochoa, X. (2010). Connexions: A Social and Successful Anomaly among Learning Object Repositories. *Journal of Emerging Technologies in Web Intelligence*, *2*(1), 11–22. **DOI:10.4304/jetwi.2.1.11-22**

Ochoa, X., Carrillo, G., and Cechinel, C. (2014). Use of a Semantic Learning Repository to Facilitate the Creation of Modern E-Learning Systems. *Proceedings of the XV International Conference on Human Computer Interaction*. 92

Ochoa, X. and Duval, E. (2009). Quantitative Analysis of Learning Object Repositories. *IEEE Transactions on Learning Technologies*, *2*(3), 226–238. **DOI:10.1109/TLT.2009.28**

Oh, E.G., Chang, Y., and Park, S.W. (2019). Design Review of MOOCs: Application of e-Learning Design Principles. *Journal of Computing in Higher Education*, **DOI:10.1007/s12528-019-09243-w**

Paredes, W.C. and Chung, K.S.K. (2012). Modelling Learning & Performance: A Social Networks Perspective. *ACM International Conference Proceeding Series* Association for Computing Machinery, 34–42. **DOI:10.1145/2330601.2330617**

Peacock, S. and Cowan, J. (2019). Promoting Sense of Belonging in Online Learning Communities of Inquiry in Accredited Courses. *Online Learning Journal*, *23*(2), 67–81. **DOI:10.24059/olj.v23i2.1488**

Pérez-Sanagustín, M., Hilliger, I., Alario-Hoyos, C., Kloos, C.D., and Rayyan, S. (2017). H-MOOC Framework: Reusing MOOCs for Hybrid Education. *Journal of Computing in Higher Education*, *29*(1), 47–64. **DOI:10.1007/s12528-017-9133-5**

Perez, C. and Germon, R. (2016). Graph Creation and Analysis for Linking Actors: Application to Social Data. in *Automating Open Source Intelligence: Algorithms for OSINT*, Boston: Syngress, 103–129. **DOI:10.1016/B978-0-12-802916-9.00007-5**

Perin, D. and Holschuh, J.P. (2019). Teaching Academically Underprepared Postsecondary Students. *Review of Research in Education*, *43*(1), 363–393. **DOI:10.3102/0091732X18821114**

Pham, M.C., Derntl, M., and Klamma, R. (2012). Development Patterns of Scientific Communities in Technology Enhanced Learning. *Educational Technology and Society*, *15*(3), 323–335

Piedra, N., Chicaiza, J., López, J., and Tovar Caro, E. (2015). Towards a Learning Analytics Approach for Supporting Discovery and Reuse of OER an Approach Based on Social Networks Analysis and Linked Open Data. *IEEE Global Engineering Education Conference, EDUCON*. 978–988. **DOI:10.1109/EDUCON.2015.7096092**

Raghavan, U.N., Albert, R., and Kumara, S. (2007). Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks. *Physical Review E, 76*(3), 036106. **DOI:10.1103/PhysRevE.76.036106**

Rao, A., Hilton, J., and Harper, S. (2017). Khan Academy Videos in Chinese: A Case Study in OER Revision. *International Review of Research in Open and Distance Learning*, *18*(5), 305–315. **DOI:10.19173/irrodl.v18i5.3086**

Ren, Z., Rangwala, H., and Johri, A. (2016). Predicting Performance on MOOC Assessments Using Multi-Regression Models. *Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016*, 484–489

Santos-Hermosa, G., Ferran-Ferrer, N., and Abadal, E. (2017). Repositories of Open Educational Resources: An Assessment of Reuse and Educational Aspects. *International Review of Research in Open and Distance Learning*, *18*(5), 84–120. **DOI:10.19173/irrodl.v18i5.3063**

Saqr, M., Fors, U., and Nouri, J. (2018). Using Social Network Analysis to Understand Online Problem-Based Learning and Predict Performance. *PLoS ONE*, *13*(9). **DOI:10.1371/journal.pone.0203590**

Saurabh, S. and Gautam, S. (2019). Modelling and Statistical Analysis of YouTube's Educational Videos: A Channel Owner's Perspective. *Computers & Education*, *128*(1), 145–158. **DOI:10.1016/j.compedu.2018.09.003**

Scardoni, G. and Lau, C. (2012). Centralities Based Analysis of Complex Networks. in *New Frontiers in Graph Theory*, InTech, 323–348. **DOI:10.5772/35846**

Shang, J., Liu, L., Li, X., Xie, F., and Wu, C. (2016). Targeted Revision: A Learning-Based

Approach for Incremental Community Detection in Dynamic Networks. *Physica A: Statistical Mechanics and Its Applications*, *443*, 70–85.

DOI:10.1016/j.physa.2015.09.072

Sheu, F.R. and Shih, M. (2017). Evaluating NTU's OpenCourseWare Project with Google Analytics: User Characteristics, Course Preferences, and Usage Patterns. *International Review of Research in Open and Distance Learning*, *18*(4), 100–122. **DOI:10.19173/irrodl.v18i4.3025**

Shmueli, E. (2017). MERLOT - A Reliable Framework for OER. *Proceedings - International Computer Software and Applications Conference*. 697–699. **DOI:10.1109/COMPSAC.2017.280**

Sicilia, M.A., Ochoa, X., Stoitsis, G., and Klerkx, J. (2013). Learning Object Analytics for Collections, Repositories & Federations. *ACM International Conference Proceeding Series* Association for Computing Machinery, 285–286.

DOI:10.1145/2460296.2460359

Sicilia, M.A., Sánchez-Alonso, S., García-Barriocanal, E., and Rodríguez-García, D. (2009). Exploring Structural Prestige in Learning Object Repositories: Some Insights from Examining References in MERLOT. *International Conference on Intelligent Networking and Collaborative Systems, INCoS 2009*. 212–218. **DOI:10.1109/INCOS.2009.12**

Silk, J. (1981). *The Analysis of Variance.*,. vol. 72. John Wiley & Sons.

DOI:10.2307/2986901

Song, L. and McNary, S.W. (2011). Understanding Students' Online Interaction: Analysis of Discussion Board Postings. Gibson, D. and Dodge, B. (eds.) *Journal of Interactive Online Learning* Association for the Advancement of Computing in Education (AACE), 1–14

Staudt, C.L. (2016). *Algorithms and Software for the Analysis of Large Complex Networks* ,

Karlsruher Institut für Technologie (KIT). **DOI:10.5445/IR/1000056470**

Staudt, C.L., Sazonovs, A., and Meyerhenke, H. (2016). NetworKit: A Tool Suite for Large-Scale Complex Network Analysis. *Network Science*, *4*(4), 508–530. **DOI:10.1017/nws.2016.20**

Staudt, C.L., Sazonovs, A., and Meyerhenke, H. (2014). NetworKit: A Tool Suite for Large-Scale Complex Network Analysis. *Arxiv*, *1403.3005*, 1–25

Tablatin, C.L.S., Patacsil, F.F., and Cenas, P. V (2016). Design and Development of an Information Technology Fundamentals Multimedia Courseware for Dynamic Learning Environment. *Journal of Advances in Technology and Engineering Research*, *2*(6), 202–210. **DOI:10.20474/jater-2.6.5**

Thompson, C. (2011). How Khan Academy Is Changing the Rules of Education. *Wired Magazine*, July, *126*, 1–5

Tovar, E., Lopez-Vargas, J.A., Piedra, N.O., and Chicaiza, J.A. (2013). Impact of Open Educational Resources in Higher Education Institutions in Spain and Latin Americas through Social Network Analysis. *ASEE Annual Conference and Exposition, Conference Proceedings*, *23*, 1. **DOI:10.18260/1-2--19700**

Tsai, Y., Lin, C., Hong, J., and Tai, K. (2018). The Effects of Metacognition on Online Learning Interest and Continuance to Learn with MOOCs. *Computers & Education*, *121*, 18–29. **DOI:10.1016/j.compedu.2018.02.011**

Tsakonas, G., Mitrelis, A., Papachristopoulos, L., and Papatheodorou, C. (2013). An Exploration of the Digital Library Evaluation Literature Based on an Ontological Representation. *Journal of the American Society for Information Science and Technology*, *64*(9), 1914–1926. **DOI:10.1002/asi.22900**

Tsiotakis, P. and Jimoyiannis, A. (2017). Investigating the Role of Structure in Online Teachers' Communities of Learning. in *Research on E-Learning and ICT in Education*, Cham: Springer International Publishing, 161–174. **DOI:10.1007/978-3-319-34127-**

9_12

Tzikopoulos, A., Manouselis, N., and Vuorikari, R. (2007). An Overview of Learning Object Repositories. in *Learning Objects for Instruction*, IGI Global, 29–55. **DOI:10.4018/978-1-59904-334-0.ch003**

U.S. Department of Education (2017). Reimagining the Role of Technology in Education: 2017 National Education Technology Plan Update. *Office of Educational Technology, U.S. Department of Education*

Wagenseller, P., Wang, F., and Wu, W. (2018). Size Matters: A Comparative Analysis of Community Detection Algorithms. *IEEE Transactions on Computational Social Systems*, *5*(4), 951–960. **DOI:10.1109/TCSS.2018.2875626**

Wang, C.H. and Chen, C.P. (2012). An Analysis of Factors Influencing the User Acceptance of OpenCourseWare. in *Communications in Computer and Information Science*. vol. 352 CCIS. Springer, 15–22. **DOI:10.1007/978-3-642-35603-2_3**

Wang, J. and Zhang, Y. (2019). Clustering Study of Student Groups Based on Analysis of Online Learning Behavior. *ACM International Conference Proceeding Series* Association for Computing Machinery, 115–119. **DOI:10.1145/3341042.3341065**

Wang, Q. and Wang, H. (2019). Study on MOOC Withdrawal Rate Based on Graph Community Detection Model. *2019 10th International Conference on Information Technology in Medicine and Education (ITME)* IEEE, 526–529. **DOI:10.1109/ITME.2019.00124**

Wang, Z. (2018). Eigenvector Label Propagation Algorithm for Interactive Learning in Student Groups Based on Student Social Network. *Proceedings of 2017 6th International Conference on Computer Science and Network Technology, ICCSNT 2017*. 247–250. **DOI:10.1109/ICCSNT.2017.8343696**

Wellman, B. and Gulia, M. (2018). Net-Surfers Don’t Ride Alone: Virtual Communities as Communities. *Networks in the Global Village: Life in Contemporary Communities*,

10(3), 331–366. **DOI:10.4324/9780429498718**

Yang, H.C. and Sun, Y.C. (2013). It Is More than Knowledge Seeking: Examining the Effects of OpenCourseWare Lectures on Vocabulary Acquisition in English as a Foreign Language (EFL) Context. *Computer Assisted Language Learning*, *26*(1), 1–20. **DOI:10.1080/09588221.2011.624523**

Yassine, S., Kadry, S., and Sicilia, M.A. (2020a). Statistical Profiles of Users' Interactions with Videos in Large Repositories: Mining of Khan Academy Repository. *KSII Transactions on Internet and Information Systems*, *14*(5), 2101–2121. **DOI:10.3837/tiis.2020.05.013**

Yassine, S., Kadry, S., and Sicilia, M.A. (2020b). Application of Community Detection Algorithms on Learning Networks. The Case of Khan Academy Repository. *Computer Applications in Engineering Education*, 1– 14. **DOI:10.1002/cae.22212**

Yassine, S., Kadry, S., and Sicilia, M.A. (2016a). Learning Analytics and Learning Objects Repositories: Overview and Future Directions. in *Learning, Design, and Technology*, Cham: Springer International Publishing, 1–29. **DOI:10.1007/978-3-319-17727-4_13-1**

Yassine, S., Kadry, S., and Sicilia, M.A. (2016b). Measuring Learning Outcomes Effectively in Smart Learning Environments. *2016 Smart Solutions for Future Cities* IEEE, 1–5. **DOI:10.1109/SSFC.2016.7447877**

Yassine, S., Kadry, S., and Sicilia, M.A. (2016c). A Framework for Learning Analytics in Moodle for Assessing Course Outcomes. *2016 IEEE Global Engineering Education Conference (EDUCON)* IEEE, 261–266. **DOI:10.1109/EDUCON.2016.7474563**

Yip, J., Wong, S.-H., Yick, K.-L., Chan, K., and Wong, K.-H. (2019). Improving Quality of Teaching and Learning in Classes by Using Augmented Reality Video. *Computers & Education*, *128*, 88–101. **DOI:10.1016/j.compedu.2018.09.014**

Yu, H., Zheng, D., Zhao, B.Y., and Zheng, W. (2006). Understanding User Behavior in Large-

Scale Video-on-Demand Systems. *ACM SIGOPS Operating Systems Review*, *40*(4), 333–344. **DOI:10.1145/1218063.1217968**

Yuqing Ren, Kraut, R., and Kiesler, S. (2007). Applying Common Identity and Bond Theory to Design of Online Communities. *Organization Studies*, *28*(3), 377–408. **DOI:10.1177/0170840607076007**

Zervas, P., Alifragkis, C., and Sampson, D.G. (2016). Studying Co-Tagging Networks in Learning Object Repositories. *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)* IEEE, 8–12. **DOI:10.1109/ICALT.2016.117**

Zhao, Z., Zheng, S., Li, C., Sun, J., Chang, L., and Chiclana, F. (2018). A Comparative Study on Community Detection Methods in Complex Networks. *Journal of Intelligent & Fuzzy Systems*, *35*(1), 1077–1086. **DOI:10.3233/JIFS-17682**

Publications

Yassine, S., Kadry, S., and Sicilia, M.A. (2021). Detecting Communities Using Social Network Analysis in E-Learning Environments: Systematic Literature review. (In process)

Yassine, S., Kadry, S., and Sicilia, M.A. (2020a). Application of Community Detection Algorithms on Learning Networks. The Case of Khan Academy Repository. *Computer Applications in Engineering Education*, 1– 14.

DOI:10.1002/cae.22212

Yassine, S., Kadry, S., and Sicilia, M.A. (2020b). Statistical Profiles of Users' Interactions with Videos in Large Repositories: Mining of Khan Academy Repository. *KSII Transactions on Internet and Information Systems*, *14*(5), 2101–2121.

DOI:10.3837/tiis.2020.05.013

Yassine, S., Kadry, S., and Sicilia, M.A. (2016a). Learning Analytics and Learning Objects Repositories: Overview and Future Directions. *Learning, Design, and Technology*, Cham: Springer International Publishing, 1–29.

DOI:10.1007/978-3-319-17727-4_13-1

Yassine, S., Kadry, S., and Sicilia, M.A. (2016b). Measuring Learning Outcomes Effectively in Smart Learning Environments. *2016 Smart Solutions for Future Cities* IEEE, 1–5.

DOI:10.1109/SSFC.2016.7447877

Yassine, S., Kadry, S., and Sicilia, M.A. (2016c). A Framework for Learning Analytics in Moodle for Assessing Course Outcomes. *2016 IEEE Global Engineering Education Conference (EDUCON)* IEEE, 261–266.

DOI:10.1109/EDUCON.2016.7474563