Dynamics of MOOC Discussion Forums

Mina Shirvani Boroujeni
CHILI Group, EPFL, RLCD, Station 20
1015, Lausanne, Switzerland
mina.shirvaniboroujeni@epfl.ch

Tobias Hecking
COLLIDE group,University of Duisburg-Essen
47048 Duisburg, Germany
hecking@collide.info

H. Ulrich Hoppe
COLLIDE group,University of Duisburg-Essen
47048 Duisburg, Germany
hoppe@collide.info

Pierre Dillenbourg
CHILI Group, EPFL, RLCD, Station 20
1015, Lausanne, Switzerland
pierre.dillenbourg@epfl.ch

ABSTRACT

In this integrated study of dynamics in MOOCs discussion forums, we analyze the interplay of temporal patterns, discussion content, and the social structure emerging from the communication using mixed methods. A special focus is on the yet under-explored aspect of time dynamics and influence of the course structure on forum participation. Our analyses show dependencies between the course structure (video opening time and assignment deadlines) and the overall forum activity whereas such a clear link could only be partially observed considering the discussion content. For analyzing the social dimension we apply role modeling techniques from social network analysis. While the types of user roles based on connection patterns are relatively stable over time, the high fluctuation of active contributors lead to frequent changes from active to passive roles during the course. However, while most users do not create many social connections they can play an important role in the content dimension triggering discussions on the course subject. Finally, we show that forum activity level can be predicted one week in advance based on the course structure, forum activity history and attributes of the communication network which enables identification of periods when increased tutor supports in the forum is necessary.

Categories and Subject Descriptors

K.3.1 [**Computer Uses in Education**]: collaborative learning; H.1.2 [**User/Machine Systems**]: Human Factors

Keywords

Massive open online courses, MOOCs, Discussion forum, Social network, Temporal analysis, Content analysis

1. INTRODUCTION

Massive Open Online Courses (MOOCs) offer high quality education provided by domain experts of various subjects to a massive number of participants with very different backgrounds and constraints. Thereby, flexibility in planning and organizing learning activities is often considered as one of the main benefits offered by MOOCs [17]. Recently, importance of time factor in MOOCs analysis has been highlighted in several studies [7, 4, 3]. The timing of learning activities is a key issue in any educational situation, and it is even more critical in MOOCs, where different time structures have affordances and constraints [7]. In [3] timing of learners activities over different weekdays and different times of the day and its relation to external factors such as employment status has been explored. Moreover, in a recent work shirvani et al. [4] proposed quantitative methods for analyzing the timing patterns of learners' activities and measuring the regularity level of students in terms of following a particular weekly study plan.

Furthermore, social learning is considered an important element of scalable education in MOOCs [5] and the potential to establish collaboration on massive scale has been argued. However, in most MOOCs, with their lack of individual support for students by tutors, discussion forums are the only channel for support and for information exchange between peers. Many studies point out limitations of discussion forums such as low overall participation [15, 19], and sometimes a lack of responsiveness [30]. Consequently, there is a discrepancy between the goal of establishing a learning community and the actual implementation of collaboration mechanisms. Given this, it has been argued that collaboration in MOOCs under consideration of the limitations of asynchronous communication and heterogeneous population of participants has to be much better supported [22, 33]. This includes personalization, support in finding peers for information exchange, and formation of learning groups [22].

In order to foster the development of sophisticated support mechanisms for peer exchange, a deep understanding of the current situation and learners interactions within the discussion forums is inevitable. For this reason, analyses of MOOCs discussion forums have received much attention in recent years. Previous studies in the literature have investigated discussion forums from different perspectives such as learners engagement and activities [15, 19, 1], discussion themes and topics or linguistic properties of written messages [28, 23, 16], structure of the communication network,

LAK '17, March 13-17, 2017, Vancouver, BC, Canada
© 2017 ACM. ISBN 978-1-4503-4870-6/17/03…$15.00
DOI: http://dx.doi.org/10.1145/3027385.3027391

group formation and social interactions among forum participants [9, 10, 18]. A detailed overview of previous works on forum analysis is presented in section 2.

The goal of this work is to extend this body of research by providing an integrated study on all the mentioned aspects combining different analysis methods. Moreover, considering the importance of time in online discussions, a particular focus in this paper is on dynamics and temporal patterns. Time dynamics are often neglected by existing works which consider aggregated variables over time to describe forum communications. In particular this work covers three main dimensions of discussion forums: Time, content and social. In **time dimension** we consider daily timeline of course duration with the main course related events, namely video openning time and assignment deadlines, and track the evolution of forum activity with respect to the course timeline. In **content dimension**, we investigate the topics of forum discussions and in **social dimension**, we study the underlying social structure of discussion forums (global level) and learners' roles in the communication network (individual level). By contrasting the results of these analyses insights on the interrelation between forum activity, discussion content, and social communication structure can be shown. Concerning the aforementioned aspects we aim to answer the following research questions:

- **RQ1.** How does the overall activity in discussion forums evolve over time and is it influenced by course structure? [*time dimension*]
- **RQ2.** How do the discussion topics evolve over time and is it related to the course structure? [*content + time dimension*]
- **RQ3.** Does the course structure influence the structure of information exchange network? [*social + time dimension*]
- **RQ4.** How do the students structural roles in discussion forum evolve over time? [*social + time dimension*]
- **RQ5.** How are the students structural roles in the communication network related to discussion content? [*content + social dimension*]
- **RQ6.** Is the overall forum activity predictable?

The rest of the paper is organized as follows. Section 2 reviews related work and section 3 present the dataset. In Section 4 overall forum activity across course timeline is investigated (*RQ1*) Section 5 explores the dynamics of discussion topics (*RQ2*). In section 6 dynamics of social communication structure at global (*RQ3*) and individual level (*RQ4*) is being explored. Section 7 investigates the relation between social and content aspects (*RQ5*). In section 8, extracted features from several of previous sections are integrated into a machine learning model to predict the forum activity level (*RQ6*). Section 9 provides a comprehensive discussion of results and concludes the paper.

2. RELATED WORK

Existing studies on discussion forum analysis focus on different aspects such as users' activity, produced content, and social structure. The question how engaged different MOOC users are in discussion forums was addressed in various studies. The fact that the discussion forums are commonly used only by a small fraction of the course participants [15] is meanwhile commonly known. Furthermore, the fraction of users who use the forums intensively is even smaller [19] and discussion volume often represents a continuous declines over the duration of the course [5]. These findings contradict the intention of discussion forums to foster collaborative knowledge building between course participants from various knowledge backgrounds [24]. On the other hand there is evidence for a relation between engagement in discussion forums and different styles of engagement with respect to other course activities [1] and that forum activity goes along with completion rates [1, 9]. Forum participation features have also been employed to predict performance and engagement in the course [20, 31].

Apart from questions about the activity of course participants in the discussion forum the actual content of the discussions is of interest as well. This typically requires natural language processing to analyse the textual contributions of forum users. The types and themes of discussion forums can be very diverse and are not necessarily related to the actual course subject [19], for example, non-course subject related discussions like search for learning groups or personal introductions, technical and organisational support. Since collaborative knowledge building and information exchange is of great interest Wise and Cui [28] proposed content-based indicators for subject related discussions. Similarly, Rossi et al. [23] build supervised classifiers to predict the type of discussion of forum threads. Another strand in content-based analysis is concerned with the nature of forum posts. Classification of speech-acts in MOOC discussion forums [2, 16], such as questions, answers or issue resolution, provides insights into the composition of discussion forum from the perspective of contribution types. Apart from speech acts, contributions can also be classified according to constructs of conceptual and operational learning levels according to the Anderson and Krathwohl's taxonomy [29].

In the aspect of the social and communication structures emerging from interaction in course forums, social network analysis is applied to networks of interconnected forum users to investigate structural patterns and the underlying relational organisation of a course community. Gillani et al. [9, 10] analyzed networks of forum users connected by co-contribution to the same discussion threads. They argue that the coherence of the social structure mainly depends on a small set of central users and the forum users can be rather be considered as a loosely connected crowd rather than a strongly connected learning community. These difference between regular forum users and occasional posters was explicitly taken into account by Poquet and Dawson [18] showing that regular users shape a denser and more centralised communication network since they have more opportunities to establish connections. In the context of structural analysis of forum communication networks, different studies used exponential random graph models (ERGMs) [21] or related statistical network analysis models to identify factors that influence the emergence of the observed network characteristics [18, 13, 32, 14]. In general these results reveal an effect of reciprocated ties and a lack of centralization of the networks to few influential users. On the level of individuals, social network analysis is further applied to identify different roles of users based on their social connections and thematic affiliations [12, 11]. This will be explicitly taken up in this work later on in Section 6.

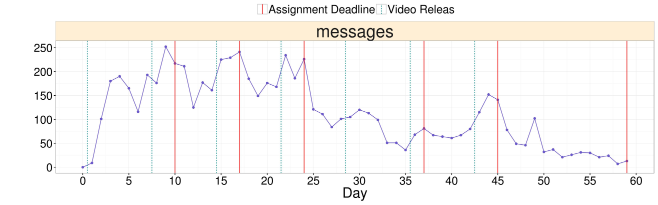


Figure 1: Number of message per day in *Scala*.

3. DATASET

The dataset used in this study comprises of two engineering MOOCs offered by Coursera entitled: "*Functional Programming Principles in Scala*" and "*Principles of Reactive Programming*". Hereafter referred to as *Scala* and *Reactive*. Both courses were eight weeks long and videos were released in a weekly basis. The final grade was based on a weighted average of six assignments corresponding to different weeks and passing grade was 60 out of 100. Discussion forums in both courses were structured several into sub-forum such as general discussions, search for learning group, questions and clarifications about course lectures and assignments. We restricted our analysis to lectures and assignments sub-forums as our focus is in tracking the evolution of discussions related to the course content. This resulted in 7,699 messages (posts or comments) by 1,175 different participants in 939 threads for *Scala* course and 12,283 messages by 1,902 participants in 1,702 threads in *Reactive* course.

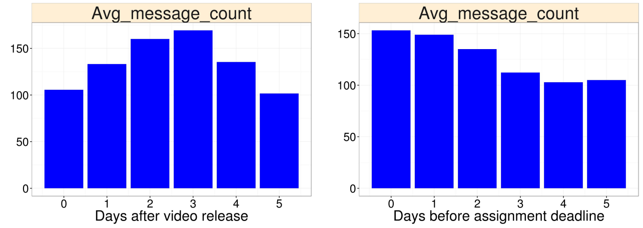
4. FORUM ACTIVITY OVER TIME

To investigate the time dynamics of forum activity and its relation with the course structure (**RQ1**), we extracted number of messages (posts or comments), number of forum contributors (participants who wrote a message) and number of threads added to the content forums on each day. As an example, Figure 1 represents daily count of messages in the discussion forum in *Scala* MOOCs with respect to main course events: video release (dashed blue lines) and assignment deadlines (solid red lines). Contributors and threads charts follow a similar trend. As it can be perceived from the charts, despite the decline of forum activity over time, at several points close to the video release or assignment deadlines there is an increment of messages in the discussion forum. This is better perceived from Figure 2 reflecting that the highest level of forum activity is associated with two to three days after the video release and activity level declines afterwards. Considering the proximity assignment deadlines, as represented in Figure 2b, the forum activity level increases as the deadline approaches. These observations further confirm the dependency between course structure and forum activity level suggesting that forum activity is a function of the weekly course structure.

5. DISCUSSION CONTENT OVER TIME

With respect to our research question on how the content of discussion forum evolves during the outline of the course (**RQ2**), in the following we present an in-depth analysis considering the posts' content over time.

5.1 Method



(a) (b)

Figure 2: Average number of new messages depending on the proximity to video release (a) and assignment deadline (b) in *Scala*.

To investigate the discussion content over time [RQ2], we compute the distributions of certain indicator phrases over time to investigate potential relations with the course events. Table 1 gives an overview of the used indicator phrases. The set of **subject related keywords** was created specifically for each course. First, the most frequent concepts in the discussion threads were determined using the Open Calais API[1]. This initial set of keywords was then manually refined based on the course outlines and detailed knowledge about the course topics (e.g. common concepts and tools in functional and reactive programming). Different spellings and synonyms were explicitly taken into account, for example, "lambda function" and "anonymous function" were mapped to the same concept. This resulted in a set of 25 subject related keywords for *Scala* and 19 for *Reactive*. In addition to subject related keywords that can directly be mapped to specific course topics, we can also identify terms that are of general nature but indicate content related discourse. These **general content related keywords** can appear without mentioning specific domain concepts (e.g. "I have no idea how to approach this problem. Can someone clarify?"), or they can be used in combination with domain terms (e.g. "Is there a *difference between* a lambda and an anonymous function?"). Such keywords as "difference_between" have been characterised by Daems et al. [6] as "signal concepts". Our overall distinction between types of indicators for certain categories of contributions has also been inspired by Wise and Cui's findings on the identification of forum threads related to the course content using indicator phrases [28]. Additionally, we also identify **resource related keywords** mentioning course material, i.e. videos and assignments as well as posts containing hyperlinks to other resources. Finally, since both of the courses analysed in this study are concerned with programming, **technical posts** which contain source code or error messages are identified based on "<code>" or "<error>" tags in the post markup.

5.2 Results

5.2.1 Course related posts over time

The ratio of posts containing general content related indicator phrases and mentions of course resources for each day in *Scala* can be seen in Figure 3. For the *Reactive* course there is a similar pattern so we do not report these diagrams for space reasons. In general, one can see that content re-

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

Table 1: Examples of indicator phrases used to track discussion topics over time

Subject related keywords
Keywords related to the main course topics.
General content related keywords
dont_understand, difference, no_idea, solution, feedback, clarify, grade, question, answer, example
Resource related keywords
video/lecture, assignment/quiz, submission, post contains a hyperlink
Technical content
Post contains source code, Post contains an error message

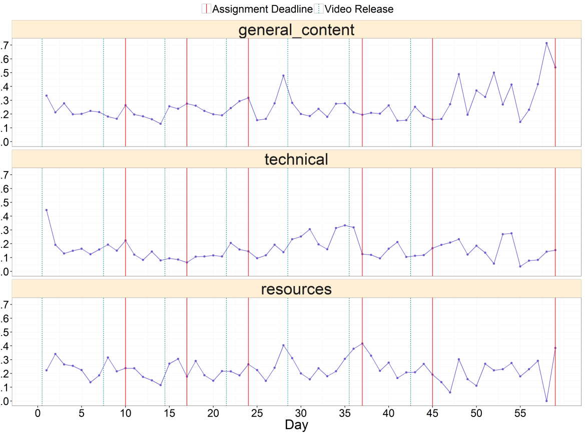


Figure 3: Genereal content, technical, and resource related keywords over time (days) in *Scala*

lated discussions are not strongly influenced by the course structure. Moreover, content is discussed throughout the entire period of the course. However, regarding the mentions of lecture videos and assignments (resource related keywords) there is often an increase short after video releases indicating that a new video is directly discussed after release.

5.2.2 *Subject related discussions over time*

As reported above general course content related keyphrases do not show a strong pattern that can be related to the course structure. Apart from the the occurrence of general content related keywords (c.f. Table 1), Figure 4 gives concrete examples for the occurrence terms in forum posts that specifically relate to the course subject of *Reactive* course (Same patterns could be identified for *Scala*). Certain terms cannot be definitely related to course events. This suggest that some participants of the MOOC have a certain background knowledge when they join the course, and thus, are able to discuss important concepts independently from the conveyed knowledge in the lecture videos. On the contrary there are terms that are discussed much more extensively after a video release. In given example of the *Reactive* course, terms like "promise" or mentions of the "akka" reactive programming framework are clearly introduced by the corresponding lecture videos. Interestingly the lecture introduced concepts remain in the discussion until the end of the course indicating that the discussion forums are to some extent useful for further discussion of lecture introduced knowledge that can be connected with the following course sections.

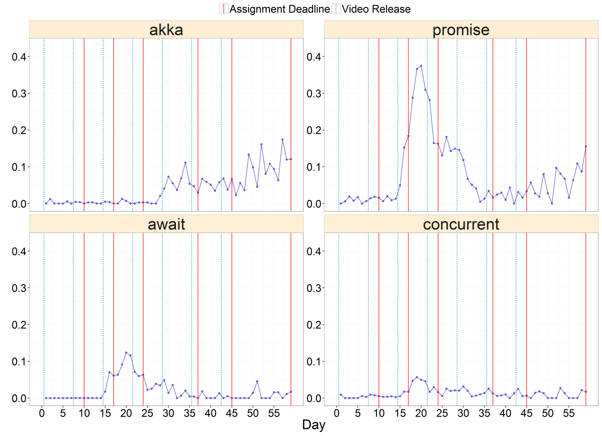


Figure 4: Examples for subject related keyphrases in *Reactive*

6. SOCIAL COMMUNICATION STRUCTURE

In this section we explore the social aspect of discussion forum and investigate the network of information exchange among forum contributors. In particular we study the information exchange network at two levels: global and individual. At the global level we explore the evolution of network over time (section 6.2.1), whereas at the individual level we focus on students' roles in the network (section 6.2.2).

6.1 Methods

6.1.1 *Network extraction*

In related work there are several approaches to model social networks from forum communication. The most simple approach is to build an undirected network by linking all forum users who contribute to the same discussion threads as in [10]. Another approach is to build reply networks where users who write in a discussion thread are linked to the thread initiator by directed outgoing relations [14]. However, since we are interested in the concrete information exchange relations between course participants a more complex network extraction method introduced in [12] was applied. This method incorporates three steps: **(1)** Classification of forum posts into three classes "information giving", "information seeking", and "others" using supervised classification models (bagged random forests). The models were trained on a set of 200 hand classified posts, where all posts that request information, for example, concrete questions on course topics or asking for advice are coded as "information seeking". Posts that provide any kind of information to information seekers are subsumed as "information giving". Posts classified as "others" cannot be associated to any of the other classes. As reported in [12] the accuracy of the classification model is considerable (F1=0.77). **(2)** Extraction of relations between posts. After deletion of all "other" posts, discussion thread (or a sub-thread comprising of comments to a parent post) can be decomposed into sequences of "information seeking" posts followed by sequences of "information giving" posts. In this sense a thread is considered as a sequence of alternating "information seeking" sequences and "information giving" sequences. In the most usual case, posts of "information giving" sequence refer to the most recent "information seeking" sequence. This allows to extract a network of posts. Since the time when forum posts was made is available, an edge between two post nodes carry a timestamp indicating when it was created. **(3)** The final information exchange network is derived by collapsing all nodes of the post network from step (2) with the same author into a single node. In this network there exist a directed edge between two nodes (representing forum users) if the first user provided some information to the second user. For more details about the network extraction process we refer to [12].

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

6.1.2 *Network structure over time*

In order to study the temporal dynamics of network, we consider network slices over one week periods using a sliding window approach. This results in one network slice for each day of the course ($d > 6$) corresponding to the forum activity during the past seven days ($[d - 6 : d]$). For each network slice a set of classic structural attributes were then computed, including number of nodes and edges, average node degree[2], network density[3], average path length[4], diameter[5] and global clustering coefficient[6].

6.1.3 *Role modeling*

Role modeling in social network analysis is often referred to as "positional analysis" [26], where the position of a node is determined by its connection patterns to other nodes. Blockmodeling [8] is a technique to decompose the set of nodes of a network into clusters of nodes with (almost) equivalent connection patterns. In network science those clusters are interpreted as users who have similar roles or positions in the community, hereby referred to as "structural roles". Various notions of equivalence can be used for clustering the nodes. The most common are structural equivalence and regular equivalence. While structural equivalence between two nodes requires that the nodes have exactly the same neighbours, regular equivalence relaxes this strict criterion such that equivalent nodes should have connections to nodes that are equivalent themselves. This recursive definition of regular equivalence can easily be understood as coloring the nodes of a network such that nodes that are equivalent have the same color, and nodes of the same color receive edges from the same set of colors and point to nodes of exactly the same set of colors. We refer the interested reader to [8] for mathematical details. Clustering of nodes according to structural and regular equivalences implies some interesting characteristics regarding the possible relations between the clusters. In a perfect structural equivalence clustering all relations between each pair of clusters $c1$ and $c2$ are either complete (all nodes in $c1$ point to all nodes in $c2$), or non-existent (there are no relations between nodes in $c1$ and $c2$). In the less strict notion of regular equivalence all nodes in $c1$ point to at least one node in $c2$ and all nodes in $c2$ receive arcs from at least on node in $c1$. In the case of communication networks such as the ones extracted from the discussion forums regular equivalence clustering are more reasonable because the requirement of structural equivalence is too strict for sparse networks and would result in many very small clusters. Furthermore, instead of perfect regular equivalence an approximate notion of regular similarity is used in conjunction with hierarchical clustering.

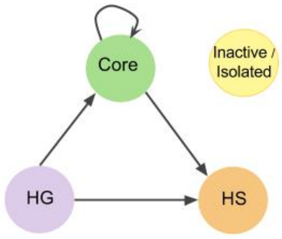


Figure 5: Core-periphery role structure

A blockmodel depicts clusters of nodes and the relational patterns between these clusters, and thus, reduces the possibly very complex social network into an interpretable macro structure. This allows for uncovering the inherent organisation of a network, such as hierarchical communication structures having nodes assigned to different levels, or core-periphery structures. In this study the core-periphery pattern based on regular similarity of nodes is of particular importance. While related works report that the overall forum communication network in MOOCs resemble a core-periphery structure [12, 14], we discovered that this is also the case for each time slice of the evolving forum communication network of the investigated courses.

Figure 5 depicts a typical core-periphery role structure that can be found in communication networks. There are four roles (clusters) represented as nodes and connections between them. The "*core*" users form a cohesive subgroup in the sense that they have many communication relations within their cluster. Furthermore, there are two peripheral roles that are not cohesive but are connected to other clusters. These two peripheral roles can be characterized as "*help givers (HG)*" or "*help seekers (HS)*" respectively since they have mainly outgoing or ingoing relations. There is also a fourth role "*inactive/isolated*", which comprises all users who do not have any connections to others in a particular time slice. This can have two reasons, either they were not active during the time span for which the model was created (inactive) or their posts could not be linked to other posts (isolated), for example, a help-seeking post without replies or posts not related to information exchange (Section 6.1.1).

Later on in Section 6.2.2 a core-periphery blockmodel is

[2] average number of connections that a node has to other nodes

[3] ratio of the number of edges and the number of possible edges

[4] average number of steps along the shortest paths for all possible pairs of connected nodes

[5] length of the longest shortest path (geodesic distance) between each two nodes

[6] Fraction of closed triangles (cliques of three) and possible triangles

derived from each time slice of the evolving information exchange network representing the role structure in different periods of the course and used to investigate role changes of course participants over time.

6.2 Results

6.2.1 Evolution of network structure over time

Considering our research question on the evolution of networks structure (**RQ3**), based on the trends observed in section 4, we hypothesize that course schedule influences discussion forum network structure. For instance with the increment of contributors and messages in the forum close to the video release or assignment deadlines, new nodes or edges could appear in the network influencing the network size, degree, density or other attributes.

Figure 6 for *Scala* is representative for the evolution of network attributes over weekly network slices (extracted using sliding window as described in section 6.1.2) for both courses. The overall decrease of forum activity towards the end of the course is also reflected by network size metrics (nodes and edges count). The networks are very sparse since the low average degree in relation to the network size results in a low density (< 0.02). Despite the comparatively larger network size in *Reactive*, in both courses average path length are relatively small (< 3 in *Scala* and < 6 in *Reactive*) throughout the course. Short path length though sparsity of the connections are a typical property of small-world networks [27] but in contrast to classical small-world network models the clustering coefficient is low. This indicates that the communication structure does not evolve into densely connected communities but rather into sparse parts interconnected via a few highly connected nodes.

On contrary to our hypothesis, course events do not show any direct influence on the network structure. One plausible explanation could be the structural limitations of the communication network such as the absence of persistent discussion groups throughout the course duration, which is also pointed out in previous studies such as [18] and [9]. Moreover, the increase of messages count in the discussion forum in a particular period of time, could be resulting from a sequence of information exchange messages between few students, which would not add new edges or nodes to the network. Since an edge in the networks aggregates possibly multiple communication events between a pair of users, such message sequences would not be reflected in the network structure.

6.2.2 Structural roles over time

To track the evolution of individuals' structural role in the discussion forum (**RQ4**), we consider successive bi-weekly slices of the network described in section 6.1.1 and extract the macro structure of each network slice as described in section 6.1.3. The resulting role models for all four time slices in both courses follow the same structure as in Figure 5. Based on the resulting role models, we then construct sequence of roles for each individual over the four time slices (phases). Additionally in each phase we differentiate latecomers or drop-outs from students who follow the course but have no contribution in the information exchange network. This results in a fifth role in sequences which we refer to as "*course inactive*". Figure 7 show role sequences of forum participants in both courses. In the sequence charts,

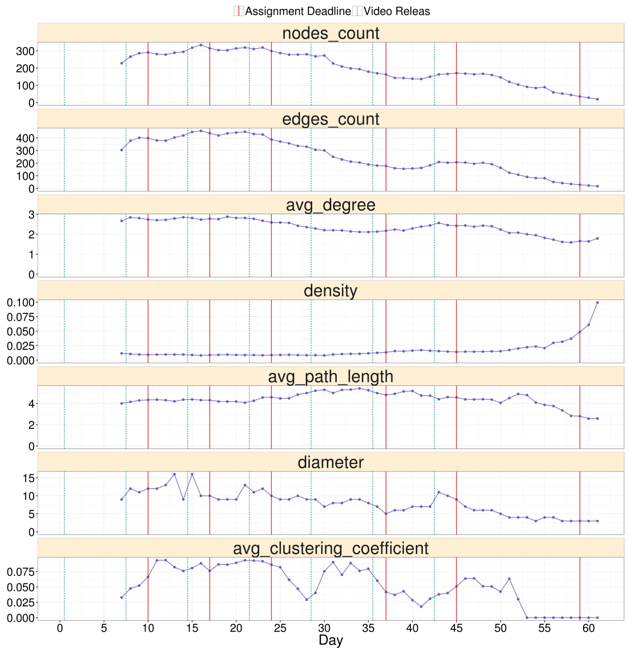
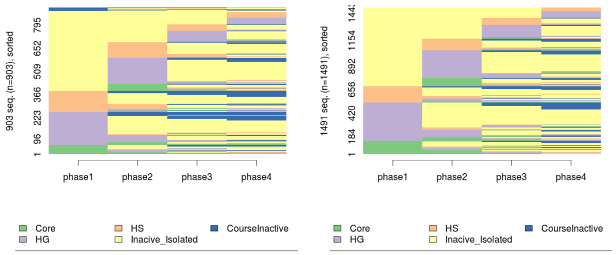


Figure 6: Network attributes over time (*Scala*)



(a) *Scala*

(b) *Reactive*

Figure 7: Role sequences of forum participants

each horizontal line corresponds to role sequence of a particular participant over the four biweekly phases. According to Figure 7, in both courses students are often active only in one or two phases and instances of active forum participation throughout the course are quite rare. Furthermore, a considerable portion of active students in each phase are new users (i.e. for the first time have a role different from *inactive/isolated*), which could also imply that persistent discussion groups in the forum are not very common.

Next, to identify common patterns in structural role sequences, we clustered role sequences using hierarchical clustering method with *optimal matching* as distance metric and substitution costs determined based on transition probabilities between states [25]. Number of clusters was determined based on the resulting dendograms.

Figure 8 and 9 represent the resulting clusters of role sequences for both courses. As it can be inferred from the figures, clusters in both courses can be characterized as *one time help seekers* (cluster 1, $N = 273$ in *Scala* and $N = 752$ in *Reactive*), *one time help givers* (cluster 2, $N = 383$ in *Scala* and $N = 276$ in *Reactive*), *active forum participants*

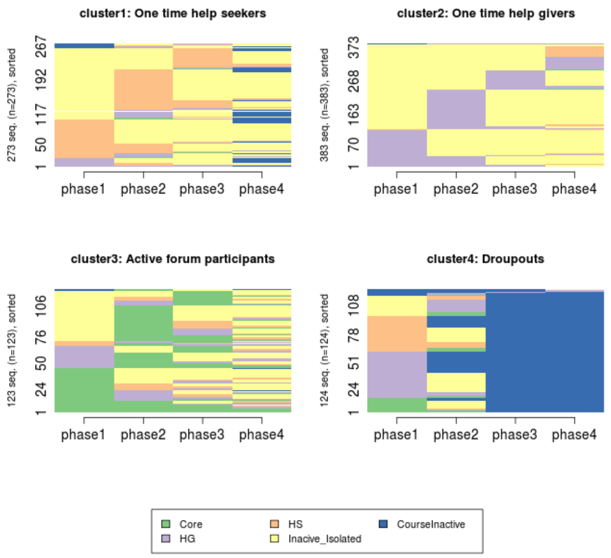


Figure 8: Clusters of role sequences (*Scala*).

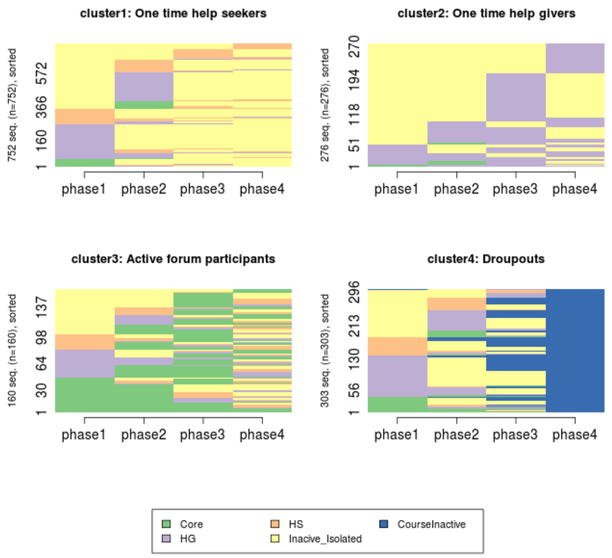


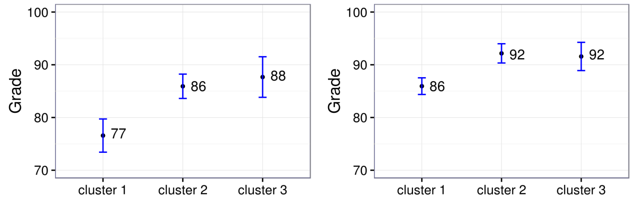
Figure 9: Clusters of role sequences (*Reactive*)

(cluster 3, $N = 123$ in *Scala* and $N = 160$ in *Reactive*) or *dropouts* (cluster 4, $N = 124$ in *Scala* and $N = 303$ in *Reactive*). The first cluster in *Reactive* course is slightly different from *Scala*, as it includes one time help seekers and also some help givers.

Comparison of average grade obtained by each cluster of participants (Figure 10) reveals that in in both courses, one time help seekers have significantly lower grades compared to one time help givers (77 vs 86, $F[1,1] = 23.29, p < 0.001$ in *Scala*, 86 vs. 92, $F[1,1] = 0.2, p = 0.002$ in *Reactive*). Additionally, despite the fact that one time help givers have lower forum participation compared to active participants, both groups achieve comparably high scores (86 and 88 in *Scala*, 92 in *Reactive*). One possible interpretation could be that active forum participants in this course, take advantage of discussion forum to advance their knowledge and resolve difficulties with respect to the course materials, whereas one time help givers are students with higher expertise level who only occasionally participate in the discussion forum, and when they do so, they provide answers to questions asked by other participants.

7. SOCIAL STRUCTURE AND DISCUSSION CONTENT

For the sake of investigating the relation between social structure and content (**RQ5**), in the following results of the previous section 6.2.2 are combined with the content analysis reported in Section 5. Tables 2 and 3 give statistics on the content of the posts made per user in the four different role sequence clusters (c.f. Section 6.2.2). Note that one post can contain keywords of different types. There is a significant relation between the different role sequence clusters and the distributions of different post types ($chi^2 = 19.29$, $p = 0.02$ in *Scala*, $chi^2 = 27.71$, $p = 0.001$ in *Reactive*). Especially for the *Scala* course, it is interesting to note that "one time help seekers", in comparison to the other groups, have relatively more posts mentioning course subject keywords or posts containing general content related phrases. Furthermore, in both courses these types of user have more technical posts and mention course resources more often. This finding is not obvious and gives interesting insights into the characteristics of forum users who are engaged in discussions in a limited time span. While related works suggest that the structural coherence of the forum communication mainly depends on the small set of very active users [10, 19], the content analysis of the posts shows that the other users can also have an important impact on the discourse by triggering focused discussions on specific subject areas and mentioning concrete problems.



(a) *Scala*

(b) *Reactive*

Figure 10: Average grades for clusters role sequences

8. PREDICTING FORUM ACTIVITY LEVEL

In order to predict the overall forum activity (**RQ6**) we trained three predictive models for number of new threads, messages and forum contributors on each day of the course.

8.1 Method

Table 4 provides an overview of the features considered in the predictive models. In particular three categories of features were extracted for each course day (d): *previous forum activity*, *structural features* and *network features*. **Previous**

Table 2: Number of keywords in posts per user by structural role clusters for *Scala*.

Cluster	Subject	General content	Technical	Resource	#Posts
One time help seekers	0.74	0.26	0.15	0.25	1148
One time help givers	0.59	0.23	0.12	0.22	1284
Active participants	0.65	0.21	0.16	0.23	2978
Dropouts	0.67	0.22	0.14	0.19	504

Table 3: Number of keywords in posts per user by structural role clusters for *Reactive*.

Cluster	Subject	General content	Technical	Resource	#Posts
One time help seekers	0.59	0.21	0.11	0.23	3405
One time help givers	0.59	0.2	0.08	0.22	1601
Active participants	0.56	0.21	0.08	0.22	4534
Dropouts	0.56	0.18	0.08	0.22	1947

forum activity features encode the volume and intensity of forum activity on each day and on the previous days. For instance T_0 represents count of threads created on the current day (d) and $T_{k>0}$ reflects the number of new threads on k days before. **Structural features** describe the properties of a day, related to the course structure such as time after video release or time before assignment deadlines. **Network features** describe the attributes of the network slice on k days before the current day (See section 6.1.2 for details on network partitioning and features description).

Additionally we consider a forth category of features describing the **initial forum activity** level, during the first week of the course. Such features could act as a normalization factor to compensate the difference in intrinsic popularity of discussion forums in different MOOCs.

Using the described features, we built three regression models for estimating number of forum threads (T_0), messages (M_0) and contributors (C_0) on a day. Several machine learning methods such as support vector regression models (SVR) with linear and RBF[7] kernels, random forests and neural networks were applied for training the models. Data from *Scala* and *Reactive* courses was randomly partitioned into train (70%) and test (30%) sets and 10-fold cross validation was used to tune the models' parameters. Highly correlated ($r > 0.7$) and linearly dependant features were removed from the features set prior to model training.

8.2 Results

In all cases SVR with linear kernels resulted in smallest prediction error, reported in Table 5. Predictive models capture 81% to 83% variance of the dependant variable and provide quite accurate predictions as reflected by low values of normalized root-mean-square errors (NRMSE) on the test data. Based on the analysis of the variable importance in the obtained models the first six features include count and average time between previous forum activity, ratio of the passed course length (Dr), number of days after latest video release (Dv), average network degree on previous days and number of open assignments (Na). Proposed models are capable of predicting the forum activity level one week in advance as forum activity history and network features included in the models correspond to seven days before the prediction day ($k = 7$). This information provided to the teaching team, could enable them to prepare the logistics to efficiently support students in discussion forums, mainly during the high activity periods.

Table 4: Description of features for each day

	Previous forum activity
T_k	number of new threads initiated on day $d-k$.
M_k	number of new messages created on day $d-k$.
C_k	number of forum contributors on day $d-k$.
TM_k	mean time between successive forum writing events
TT_k	mean time between successive thread initiating events
	Structural features
Dv	number of days after the latest video release.
Da	number of days after the latest assignment release.
Dd	number of days left to the next assignment deadline.
Dr	ratio of the current day (d) to the course length encoding what percentage of the course is passed.
Na	number of assignments open for submission.
	Network features
Net_k	network features on day $d-k$, including nodes and edges count, diameter, average degree, path length and clustering coefficient
	Initial forum activity (first week)
$W1T$	mean and standard deviation of threads count per day, mean time between new threads (3 features)
$W1M$	mean and standard deviation of messages count per day, mean time between messages (3 features)
$W1C$	mean and standard deviation of forum contributors per day (2 features)

Table 5: Overview of predictive models and results

DV	Features	R^2 (train)	NRMSE* (train)	NRMSE (test)
T_0	Prev, Struct, $W1T$, Net_7	0.81	0.40	0.28
M_0	Prev, Struct, $W1M$, Net_7	0.80	0.27	0.23
C_0	Prev, Struct, $W1C$, Net_7	0.83	0.27	0.22

*Normalized RMSE by mean of observed values

9. DISCUSSION AND CONCLUSION

In this work we applied mixed methods to investigate the forum communication in two MOOCs in the time, content, and social dimension leading to insights regarding the research questions formulated in the beginning (Section 1).

In the time dimension, in Section 4 we investigated how the outline of the courses (video release and assignment deadlines) influences the overall forum activity (**RQ1**). We could observe an increase of the number of posts before deadlines and after video release days, and thus conclude that course events have an impact on the forum communication.

[7] radial basis function

Surprisingly, based on the content analysis of the posts over time (**RQ2**), there was no clear coupling between the course structure and quantity of general content related, resource related, and technical posts as reported in section 5. However, mentions of some specific terms regarding the course subject tend to increase after video releases indicating that some discussion topics are introduced by the course while others are brought into discussion by the participants themselves.

The temporal dynamic of the social structure emerging from the forum communication with respect to course events (**RQ3**) was analyzed in Section 6.2.1. Here we could show that the global organisation (network characteristics) of the communication network is independent of the course structure. One reason can be the absence of a sustainable forum community and the high fluctuation of the active contributors. Consequently, there is no inherent self-organisation of the network, which would require coordination and maintenance of social relations. This further supports the claim of Gillani and Eynon [9] that MOOC forums resemble decentralized crowd behaviour rather than a social community.

A more user centric perspective on the integrated analysis of the time and the social dimension was taken in section 6.2.2 to answer research question on how roles of forum users evolve (**RQ4**). Interestingly, the role structure of the information exchange network comprising of a small cohesive core of active contributors, and peripheral help-giving or help-seeking users that has also been reported for static snapshots of the network in [12] and [14], persist over several time slices. While the overall structural organisation of the networks is stable it could be shown that the association of users to roles changes drastically over time. Only a small subset of the most active (core) users retain an active role over time, and majority of learners are active in only very few (mostly one) time slice. It could be seen that the fluctuation of active users is so pervasive such that in each time slice even the majority of users are "newcomers" in the sense that they form connections to other users for the first time. This can be considered as a major obstacle for the emergence of a sustainable community and further explains the irregularities in the overall network structure mentioned above.

Regarding research question on the relation between students roles in the communication network and discussion topics (**RQ5**), in section 7 the combination of structural role models and discussion content was investigated. Results showed that even if peripheral users (one time help-givers and seekers) are genrerally not as important for the structural cohesion of the communication network as the core users, they make fewer but important contributions indicated by their high rates of general content related and course subject related posts. Especially the participants in the group of occasional (or one-time) help-seekers post similar or even more information requests related to course content, technical issues including source code, or mentions of concrete course materials. Consequently those users can often be notable as initiators of discussions even if their activity is limited.

The question about the predictability of forum activity (**RQ6**) was answered in section 8. It could be shown that the forum activity level in terms of the number of discussion threads, messages and participants is predictable one week in advance given the course structure (video and assignment dates), history of forum activity and the described network features. Further, the predictive models can be considered as a building block for teaching support tools to forecast periods when increased tutor support for forum discussions is needed.

In summary, while the majority of research works focusing on single aspects of on MOOC discussion forums point to the conclusion that the current implementation of discussion forums are only used intensively by a small amount of course participants, and further, only a subset of the discussions are relevant for the information exchange, the outcome of our study suggests that discussion forums are much more complex. Activity, content, and structural related analyses highlight different aspects of forum communication, while there are several interdependencies between the progress of the course, the contributed content and structural roles of participants that have to be taken into account to get a clearer picture and to foster the development of future collaboration support. Recommendation mechanisms to find the right information and adequate discussion partners [22, 30] can be one initial step, but in order to transform the loosely connected "crowd" of forum users into a sustainable community in the sense of social learning requires also support for maintaining social contacts. Furthermore, combination of predictive models proposed in section 8 with content analysis of the forum contributions can potentially support instructors to turn their attention to upcoming important discussions and enable interventions and community management.

10. ACKNOWLEDGEMENTS

This study is part of the leading house DUAL-T research project funded by the Swiss State Secretariat for Education, Research and Innovation (SERI).

11. REFERENCES

[1] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Engaging with massive online courses. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW'14, pages 687–698. ACM, 2014.

[2] J. Arguello and K. Shaffer. Predicting speech acts in mooc forum posts. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, ICWSM'15, 2015.

[3] M. S. Boroujeni, L. Kidzinski, and P. Dillenbourg. How employment constrains participation in moocs. In *Proceedings of the 9th International Conference on Educational Data Mining*, pages 376–377, 2016.

[4] M. S. Boroujeni, K. Sharma, Ł. Kidziński, L. Lucignano, and P. Dillenbourg. How to quantify student's regularity? In *European Conference on Technology Enhanced Learning*, pages 277–291. Springer, 2016.

[5] C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F. M. F. Wong. Learning about social learning in moocs: From statistical analysis to generative model. *IEEE transactions on Learning Technologies*, 7(4):346–359, 2014.

[6] O. Daems, M. Erkens, N. Malzahn, and H. U. Hoppe. Using content analysis and domain ontologies to check

learners' understanding of science concepts. *Journal of Computers in Education*, 1(2):113–131, 2014.

[7] P. Dillenbourg, N. Li, and Ł. Kidziński. *From Books to MOOCs? Emerging Models of Learning and Teaching in Higher Education*, chapter The complications of the orchestration clock. Portland Press, 2016.

[8] P. Doreian, V. Batagelj, A. Ferligoj, and M. Granovetter. *Generalized Blockmodeling (Structural Analysis in the Social Sciences)*. Cambridge University Press, 2004.

[9] N. Gillani and R. Eynon. Communication patterns in massively open online courses. *The Internet and Higher Education*, 23:18 – 26, 2014.

[10] N. Gillani, T. Yasseri, R. Eynon, and I. Hjorth. Structural limitations of learning in a crowd: communication vulnerability and information diffusion in moocs. *Scientific Reports*, 4:6447, 2014.

[11] T. Hecking, I. A. Chounta, and H. U. Hoppe. Analysis of user roles and the emergence of themes in discussion forums. In *Proceedings of the 2nd European Network Intelligence Conference*, ENIC'15, pages 114–121. IEEE, 2015.

[12] T. Hecking, I.-A. Chounta, and H. U. Hoppe. Investigating social and semantic user roles in mooc discussion forums. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, LAK'16, pages 198–207. ACM, 2016.

[13] S. Joksimović, A. Manataki, D. Gašević, S. Dawson, V. Kovanović, and I. F. de Kereki. Translating network position into performance: Importance of centrality in different network configurations. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, LAK'16, pages 314–323. ACM, 2016.

[14] S. Kellogg, S. Booth, and K. Oliver. A social network perspective on peer supported learning in moocs for educators. *The International Review of Research in Open and Distributed Learning*, 15(5), 2014.

[15] R. F. Kizilcec, E. Schneider, G. L. Cohen, and D. A. McFarland. Encouraging forum participation in online courses with collectivist, individualist and neutral motivational framings. *Experiences and best practices in and around MOOCs*, pages 17–26, 2014.

[16] W. Liu, Ł. Kidziński, and P. Dillenbourg. Semiautomatic annotation of mooc forum posts. In *State-of-the-Art and Future Directions of Smart Learning*, pages 399–408. Springer, 2016.

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

[18] P. Oleksandra and D. Shane. Untangling mooc learner networks. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, LAK'16, pages 208–212. ACM, 2016.

[19] D. F. Onah, J. Sinclair, R. Boyatt, and J. Foss. Massive open online courses: learner participation. In *Proceeding of the 7th International Conference of Education, Research and Innovation*, pages 2348–2356, 2014.

[20] A. Ramesh, D. Goldwasser, B. Huang, H. Daumé III, and L. Getoor. Modeling learner engagement in moocs using probabilistic soft logic. In *NIPS Workshop on Data Driven Education*, volume 21, page 62, 2013.

[21] G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173 – 191, 2007.

[22] C. P. Rosé and O. Ferschke. Technology support for discussion based learning: From computer supported collaborative learning to the future of massive open online courses. *International Journal of Artificial Intelligence in Education*, 26(2):660–678, 2016.

[23] L. A. Rossi and O. Gnawali. Language independent analysis and classification of discussion threads in coursera mooc forums. In *Proceedings of the 15th International IEEE Conference on Information Reuse and Integration*, IRI'14, pages 654–661. IEEE, 2014.

[24] A. Sharif and B. Magrill. Discussion forums in moocs. *International Journal of Learning, Teaching and Educational Research*, 12(1), 2015.

[25] M. Studer and G. Ritschard. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):481–511, 2016.

[26] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*, volume 1 of *Structural analysis in the social sciences*. Cambridge University Press, 1994.

[27] D. J. Watts and S. H. Strogatz. Collective dynamics of śmall-worldńetworks. *Nature*, 393(6684):440–442, 1998.

[28] A. F. Wise, Y. Cui, and J. Vytasek. Bringing order to chaos in mooc discussion forums with content-related thread identification. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, pages 188–197. ACM, 2016.

[29] J.-S. Wong, B. Pursel, A. Divinsky, and B. J. Jansen. Analyzing mooc discussion forum messages to identify cognitive learning information exchanges. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIST'15, pages 23:1–23:10. American Society for Information Science, 2015.

[30] D. Yang, D. Adamson, and C. P. Rosé. Question recommendation with constraints for massive open online courses. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 49–56. ACM, 2014.

[31] D. Yang, T. Sinha, D. Adamson, and C. P. Rosé. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*, volume 11, page 14, 2013.

[32] J. Zhang, M. Skryabin, and X. Song. Understanding the dynamics of mooc discussion forums with simulation investigation for empirical network analysis (siena). *Distance Education*, 37(3):270–286, 2016.

[33] S. Zheng, M. B. Rosson, P. C. Shih, and J. M. Carroll. Designing moocs as interactive places for collaborative learning. In *Proceedings of the 2nd ACM Conference on Learning @ Scale*, L@S'15, pages 343–346. ACM, 2015.