JOURNAL OF LEARNING ANALYTICS

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

Role Modelling in MOOC Discussion Forums

Tobias Hecking
University of Duisburg-Essen, Germany

Irene-Angelica Chounta
Human–Computer Interaction Institute, Carnegie Mellon University, USA

H. Ulrich Hoppe
University of Duisburg-Essen, Germany

hecking@collide.info

ABSTRACT: To further develop rich and expressive ways of modelling roles of contributors in discussion forums of online courses, particularly in MOOCs, networks of forum users are analyzed based on the relations of information-giving and information-seeking. Specific connection patterns that appear in the information exchange networks of forum users are used to characterize user roles. Additionally, semantic roles are derived by identifying thematic areas in which an actor (learner) looks for information (problem areas) and the areas of interest in which an actor provides information to others (areas of expertise). The interplay of social and semantic roles is analyzed using a socio-semantic blockmodelling approach. The results indicate that social and semantic roles are not strongly interdependent. The methodological contribution is in combining traditional blockmodelling with semantic information to characterize participant roles. Furthermore, we use sequential pattern analysis techniques to analyze the posting activity of users over time in terms of categories of cognitive engagement. The combination of the different approaches reveals that user roles derived from the analysis of engagement patterns are strongly related to socio-semantic user roles.

Keywords: MOOCs, discussion forums, social network analysis, temporal data

1 INTRODUCTION

In online learning courses, there are often limited possibilities for immediate and synchronous interaction between learners and tutors. This is especially the case with massive open online courses (MOOCs) where, in the absence of individual support by a tutor, discussion forums are often the only provided channel for information exchange and peer-to-peer-support. On the one hand, only a small fraction of all MOOC participants use the forum to communicate (Abnar, Takaffoli, Rabbany, & Zaïane, 2015). However, forum activity often goes along with higher engagement in the course and higher completion rates (Anderson, Huttenlocher, Kleinberg, & Leskovec, 2014; Engle, Mankoff, & Carbrey, 2015). Discussion activities between learners in MOOCs offer the potential of involving more learners in sustainable collaborative knowledge building in a social context when adequately supported (c.f. Ferschke, Howley, Tomar, Yang, & Rosé, 2015; Rosé, Goldman, Zoltners, & Resnick, 2015). Further

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

insights into the information exchange in discussion forums utilizing analytical methods can contribute to improving the design and application of such forums and to the development of new types of communication channels for online courses.

This paper is primarily focused on analysis methods combining network and content analytics approaches. The mixed approach is applied to the characterization of learners' in MOOC discussion forums in terms of role patterns in different dimensions. Here, the notion of "role" refers to groupings of users with similar characteristics coupled to inherent expectations regarding their activity and influence in the forum. Such roles can be modelled in at least three different aspects.

1. **Structure**: In information exchange networks as they emerge from forum discussions, positional analysis subsumes a set of social network analysis methods — such as blockmodelling (Doreian, Batagelj, Ferligoj, & Granovetter, 2004) — that models user roles based on their connection patterns i.e., network position.
2. **Content**: By investigating the contributions of users and discussions they are participating in, one can derive interest profiles of users, especially regarding topics of expertise and problem areas.
3. **Activity**: Apart from network positions and content-related characteristics, learners can be classified according to their level of engagement in forum activities.

Most of the existing research on discussion forums in a learning context focuses on one of these dimensions. However, a combined analysis of all three dimensions is necessary to get a complete picture of the course community that uses the discussion forums. One goal of this work is to combine the aforementioned aspects and to discover interrelationships between different role models.

As a prerequisite, the raw forum data — i.e., listings of discussion threads — have to be preprocessed in order to identify information-seeking and corresponding information-giving posts. Then, the coded posts are used to model a directed network of forum users where an edge represents a "provides information" relation between two users. This detailed procedure is described in Section 3.

Section 4 describes our socio-semantic blockmodelling approach, which combines network and content analysis. The results of coding the posts into information-giving and information-seeking, in conjunction with the discovery of thematic areas of discussion threads, are used to represent users by the thematic areas of interest in which they seek for information (problem areas) and the areas of interest in which they provide information to others (expertise). While a social network of forum users can essentially be characterized through person-to-person relations, similar semantic interests of two users do not necessarily signify a communication between the users. Blockmodelling (cf. Doreian et al., 2004) is a network analysis technique that allows for grouping actors in a network in terms of similar connection patterns to other groups of users. Here, the groups stand for "social positions" or roles and are not necessarily strongly connected internally. Socio-semantic blockmodelling extends this approach by incorporating semantic models of users based on their information-seeking and information-giving

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

interests. On this basis, a role can be interpreted as a group of participants with similar connection patterns in a thematic context (in the sense of communities with a congruence of interests).

With respect to activity and engagement, Section 5 outlines our approach to classifying forum users into different categories of engagement in forum discussions during the progress of the course. The categories used in this analysis are inspired by the ICAP taxonomy (interactive, constructive, active, passive) of cognitive engagement proposed by Chi and Wylie (2014). The engagement over time constitutes another type of role characterization of forum participation, apart from network and content based approaches.

In Section 6, we analyze and compare the results of applying the different analysis approaches to forum data of two MOOCs. The first course, Introduction to Corporate Finance, took place over a six-week period from November 2013 to December 2013. The second course, Global Warming: The Science of Climate Change, was offered over eight weeks from October 2013 to December 2013. Both courses were offered on the Coursera[1] platform. For the sake of simplicity, these courses will be referred to as "Corporate Finance" and "Global Warming," respectively.

The triangulation of different methods leads to insights into different roles taken by users in MOOC discussion forums from different point of view. Thus, the results can contribute valuable information for the successful development or redesign of collaboration support in large online courses to better fit the needs of different types of users.

2 BACKGROUND AND RELATED WORK

The collection of posts in MOOC discussion forums accounts for the information flow between learners from various knowledge backgrounds and is an indicator of collaborative knowledge building between learners with diverse knowledge backgrounds (Sharif & Magrill, 2015). Related research has shown that user engagement in MOOC discussion forums tends to differ. While many users are not active at all or use the forum only for purpose-specific activity (i.e., asking for assignment solutions or rapid and trustworthy responses to specific questions; Onah, Sinclair, Boyatt, & Foss, 2014), MOOC forums are usually dominated by few, highly active users who can influence other participants and stimulate and sustain the discussions (Huang, Dasgupta, Ghosh, Manning, & Sanders, 2014; Wong, Pursel, Divinsky, & Jansen, 2015). This diverse behaviour results in different user roles that can be described in various aspects using various analysis techniques.

Techniques used for the analysis of MOOC discussion forums can be characterized as content-related or communication-related. Content analysis aims to uncover the nature of forum contributions from the post content (Rossi & Gnawali, 2014). The nature of forum discussions can be very diverse and the discussions are not necessarily related to the actual course content, even in online courses structured by an outline or syllabus. A crucial step in the analysis of forum discussions in online courses is the

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

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

identification of the discussion threads relevant for course-related information exchange between participants. In a recent study (Wise, Cui, & Vytasek, 2016), a combination of content analysis and machine learning was used to distinguish forum threads in which participants discuss the course content from those merely socializing or discussing organizational matters. Content analysis is also used to characterize forum users based on the types of contributions they make (Arguello & Shaffer, 2015; Liu, Kidzinski, & Dillenbourg, 2015). Social network analysis is commonly applied for communication-related analytic approaches. Social networks of users based on common discussion threads can serve to investigate the coherence of the underlying social network (Gillani, Yasseri, Eynon, & Hjorth, 2014), detection of communication patterns (Gillani & Eynon, 2014) and community support (Malzahn, Harrer, & Zeini, 2007). However fine grained network modelling is required to adequately reflect and represent the concrete post/reply communication between participants. In discussion forums with nested threads, these relations can be observed directly from the thread structure (Rabbany, Takaffoli, & Zaïane, 2011). However, in forums with a more linear thread structure, such as the Coursera forums investigated in this paper, the identification of direct communication between users requires content-analytic approaches such as discussion act tagging (Arguello & Shaffer, 2015; Liu et al., 2015). In this case, user roles can be inferred based on the communication behaviour of an actor, i.e., the position of the actor in the social network. Abnar et al. (2015) use centrality measures in subcommunities to identify roles, such as leaders and mediators, in a forum communication network. Apart from that, role models in communication networks can also be based on discussion content (McCallum, Wang, & Corrada-Emmanuel, 2007) or connection patterns (Rossi & Ahmed, 2015).[2]

Our work combines techniques of network and content analysis to characterize roles of users by blending the position in the information exchange network with semantic similarity based on content analysis of the threads they were active in (see Section 4). From a network-analytics perspective, the notion of roles (or "positions") denotes certain relational patterns between classes of nodes in the network. This will be further explained below under the notion of blockmodelling. This notion of role may appear to be weak in comparison to the notion of "role" in pedagogical theories and models. However, based on the help-seeking/help-giving distinction, or on the characterization of types of contributions, it may lead to characterizations that are clearly relevant from a pedagogical perspective.

Regarding the combination of structural (network-based) measures with content analysis, our approach is similar to the work of Yang, Wen, Kumar, Xing, and Rosé (2014) who combine network data with post content in a single model to identify subcommunities of learners based on discussion topics and reply relations in the forum. However, Yang et al. (2014) assume that there is interplay between users' interests and social relations that is inherently encoded in the model. In our work, we investigate the possible interdependence between social relations and semantic similarity more closely with respect to user roles in a network that represents course-related information exchange more explicitly. In addition, using the blockmodelling approach, we do not assume that users with the same role have to form a cohesive subcommunity.

[2] For a detailed survey on role modelling in social networks, refer to Forestier, Stavrianou, Velcin, and Zighed (2012) and Rossi and Ahmed (2015).

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

3 NETWORK EXTRACTION FROM FORUM DATA

The first dataset contains forum posts from the Corporate Finance MOOC. Overall, there were 8336 posts by 1540 different users in 870 different threads. Anonymous users made 1436 posts. Many of the discussion threads were used by the course participants to introduce themselves, to seek for learning groups with peers of the same mother tongue, etc. We explicitly restricted the analysis to discussion threads dedicated specifically to issues regarding lectures, exercises, and quizzes since we are only interested in tracking information-giving and information-seeking related to the course content. This resulted in a dataset of 540 threads with 5533 posts from 945 different users. It is important to note that all anonymous posts were counted as posts of a single, artificial, "anonymous" user.

The second dataset collected from the Global Warming MOOC is smaller in terms of number of users (1007) but with a higher number of discussion threads (1020). By removing non-course-related threads, we ended up with a dataset of 4216 posts by 434 users in 546 threads.

Table 1: Example of a Discussion Thread with Three Users

Post	User	Content	Post type
P1	User A	I have a problem with …	Information-seeking
P2	User B	Have you tried the following?	Information-giving
P3	User A	That helps. Thank you.	Other
P4	User C	An alternative solution …	Information-giving

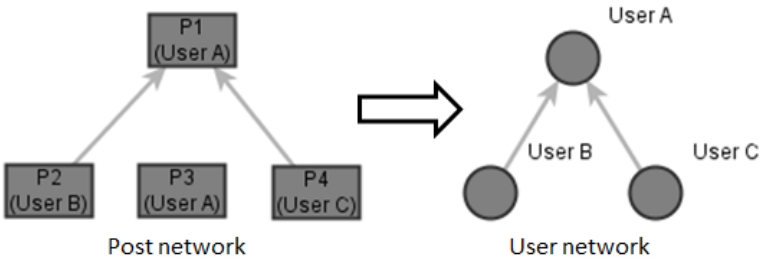

Figure 1: Basic scheme for the network extraction from forum posts.

The starting point of the analysis is the set of threads in the discussion forum. These threads contain a sequence of posts where the unique identifier, post content, and author's identity are available for each post. The analysis relies on the social network of users who participated in content-related, knowledge exchange in the discussion forum. In contrast to most of the existing studies (see Section 2), the network should reflect the directed relations between users who ask for information and users who reply to these specific information requests. The initial task is to extract this network from the raw forum thread

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

data. This task can be structured in three successive steps: 1) post classification, 2) post linking, 3) transformation to a social knowledge exchange network. An example of the procedure for the example discussion thread presented in Table 1 is shown in Figure 1.

3.1 Post Classification

In order to identify the information-giving and information-seeking relations between the users, the first step is to identify the posts as information-giving or information-seeking. Previous studies have identified different types of posts in MOOC discussion forums (Arguello & Shaffer, 2015; Kim, Wang, & Baldwin, 2010; Liu et al., 2015). In this study we generalize the classification schemes for MOOC discussion forums described by Arguello and Shaffer (2015) and the classification schema of Liu et al. (2015) to three classes of posts: information-seeking (all types of questions, clarification requests, report of an issue), information-giving (answers, issue resolutions, hints and recommendations), and other posts. For classification purposes, we trained an automated classification model using 500 posts that were hand-classified by three experts. The high interrater agreement among all three experts according to Fleiss-Kappa ($\kappa = .78, p < .005$) ensured the validity of the classification.

The organization of the course forum into sub-forums is used to filter the dataset prior to the automatic post classification. In previous work (Hecking, Hoppe, & Harrer, 2015), we proposed forum post classification on the entire dataset incorporating threads that, likely, do not contain content-related discussions. Social posts, like self-introduction or requests for study groups, were also classified with considerable accuracy since the sub-forum is a good predictor for those posts. In this study, we used information on the sub-forum in which a discussion thread occurs to restrict the analysis to those explicitly dedicated to content-related issues, such as assignments and lectures. Posts were encoded by structural features (position in the thread, number of votes) and content-related features (text length, occurrences of questions words, question/exclamation marks, and specific phrases such as “need help” or “helps you”). The best results, based on 10-fold cross validation, were obtained by a random forest classifier (Breiman, 2001) when 10-iteration bagging was applied. Information-seeking posts can be classified with high F1-scores (F1-score = 0.77). For information-giving posts, the F1-score is also moderately high (F1-score = 0.66). However, “other” posts often led to misclassifications, as the confusion matrix in Table 2 shows.

Table 2: Confusion Matrix for Post Classification

	True info. seeking	True info. giving	True other	Class precision
Predicted info. seeking	75	16	2	0.81
Predicted info. giving	27	88	30	0.61
Predicted other	0	17	36	0.68
Class recall	0.74	0.73	0.53	

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

In order to reduce the effect of misclassified other posts, the final classification was improved using the iterative classification algorithm described by Ó Duinn and Bridge (2014). This algorithm uses the results from the classifier described above to compute the number of preceding posts of each class for each post. Then, an additional classifier is trained to incorporate this information update into the initial classification. This leads to improved results since misclassifications, such as the classification of information-giving posts without a preceding information-seeking post, can be avoided. This addition increases the F1-score for information-seeking posts to 0.79 (from 0.77) and for information-giving posts to 0.71 (from 0.66) based on the evaluation of another 200 hand-classified posts.

3.2 Network Extraction

Based on the classified posts, we initialized the network of information-seeking and related information-giving posts. As a first step, we removed the anonymous user and isolated users (users who did not receive a reply to their posts). This resulted in a network of 647 of the original 1540 users for the Corporate Finance MOOC and 343 of the original 1007 users for the Global Warming MOOC, showing that even though many users contribute to forum posts, they are not involved in information exchange.

Next, all posts classified as "other" were filtered out from each thread such that only the "information-seeking" and "information-giving" posts remained. Additionally, we had to build a network of posts (see Figure 1) as an intermediate step before creating the social network between users. For that, we took into account that the users in Coursera discussion forums usually maintain the structure of a thread themselves, such that the relations between posts are recognizable. Most content-related threads start with a request for information. This initial request is either directly answered by another user or more questions follow until an information-giving post occurs in the sequence. After a sequence of information-giving posts, sometimes further questions are posted. Comments are attached to a single post. This helps to relate posts to previous posts even if the discussion has proceeded and other unrelated posts occurred in between. Sequences of comments attached to a parent post can be seen as sub-threads that may contain both types of posts (information-seeking and information-giving) with the parent post as the initial post. Consequently, a forum thread and the corresponding sub-threads can be decomposed into alternating sequences of information-seeking and information-giving posts. This structure enables the linking of information-giving to previous information-seeking posts by linking the posts of each information-giving sequence to the posts of the most recent sequence of information-seeking posts in a thread.

In the resulting forum post network, each post node is annotated with the author of the post and a timestamp. In the final step, each post node labelled with the same author is collapsed into a single node representing the user, resulting in the final knowledge exchange network between forum users (see Figure 1).

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

4 METHODS 1: SOCIO-SEMANTIC BLOCKMODELLING

Blockmodelling (Doreian et al., 2004) is used to reduce a network to a macro structure by grouping actors groups based on their connection patterns and modelling relations. Those groups are commonly interpreted as roles or positions since it is assumed that similar connection patterns indicate similar functions. Figure 2 gives an example of a blockmodel with three roles and relations between them that reflect the hierarchical structure of the network. In this section, we describe the existing techniques for blockmodelling based on similarities of connection patterns of users. The extensions we made incorporate the semantic similarity of users based on their interest in thematic areas. This new approach is described in subsections 4.3 and 4.4.

4.1 Blockmodelling Foundations

In general, blockmodelling groups actors based on a certain notion of similarity. These groups reflect the roles of the actors but are not necessarily cohesive in the sense that actors of the same role are necessarily densely interconnected among themselves. A blockmodel fitted to the network structure can be used to infer relations between those groups of actors. In a generalized blockmodelling approach (Doreian et al., 2014) one distinguishes between various types of relations that can exist between two groups/roles, indicating different types of connection patterns between the actors of the roles. A complete directed relation between two groups, A and B, is given if all actors in A have an outgoing relation to all actors in B. This indicates the strongest possible relationship between two groups. Regular relations can be seen as a relaxation of a complete relation. If a regular relation from group A to group B exists, all actors in A point to at least one actor in B and all actors in B have at least one ingoing relation to actors in A. Regular relations are very important for this work since they reflect information flow. For information-giving relations between actors, regular relations between groups can be interpreted as existing information flows from group A to group B. Note that complete relations are a special case of regular relations. If no relations between actors in group A and group B are present, the relationship between the groups is considered null.

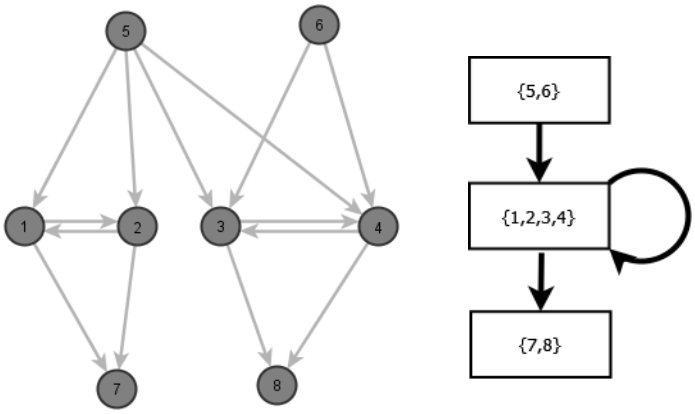

Figure 2: Example network with regular and structural equivalences.

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

In forum networks, there is often no perfect fit in the relations between groups of users with the mentioned relation types. For example, if groups A and B both contain more than one member and there is only one relation from an actor in A to an actor in B, the group relation is far from being regular or complete, but it can also not be considered as a null-relation (no ties between the groups). In cases where none of the described relations are applicable, we chose the relation that can be applied with minimal modifications to the links between the actors in A and B. The total number of such modifications is referred as the blockmodel error.

An important fact, often ignored, is that blockmodelling can clearly be distinguished from the more common subcommunity detection (Fortunato, 2010) in social network analysis. Even though both blockmodelling and subcommunity detection group users into clusters, the objectives of these methods are quite different. Community detection methods aim to find densely connected substructures in the network by clustering such that the number of connections within the cluster, as much as possible, exceeds the number of connections between actors of different clusters. Blockmodelling does not require any connections between actors of the same cluster at all, although they are not forbidden (see group {1,2,3,4} in Figure 2). Moreover, users in a blockmodel belong to the same group since they have similar connection patterns with users in other groups. Thus, a cluster can be interpreted as users with a similar position or role in the network. In order to highlight this difference compared to sub-communities based on dense intra-cluster relations, in the following the groups found by user similarity are referred to as roles.

4.2 Graph-based Actor Similarity

Graph-based similarity derives actor similarity directly from the graph structure. This is the traditional approach for blockmodelling. The benefit of this approach is that actors are grouped into roles/positions such that the previously described relations between groups of actors are inherently induced by the groupings. Graph-based similarity measures commonly applied to blockmodelling are structural and regular similarity.

4.2.1 Structural similarity

Structural similarity (Lorrain & White, 1971) relates to the position of the actors within the network. Structural similarity can be assessed by the correlations between the connections of each pair of actors. If two actors are structurally equivalent (maximum structural similarity), they have ingoing relations from the same set of actors and outgoing relations to the same set of actors. For example, actors 3 and 4 in Figure 2 are structurally equivalent. This means they have the same position and can be replaced by a single node without information loss. A perfect assignment based on structural similarity — i.e., all actors in one role are structurally equivalent — leads to a perfectly fitting blockmodel with only complete and null blocks. However, finding such a model in forum networks is quite unlikely. Thus, this type of similarity is not used in the blockmodels described later in favour of regular similarity described next.

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

4.2.2 Regular similarity

Regular similarity (White & Reitz, 1983) between two actors, in contrast to structural similarity, does not explicitly take into account mutual connections to concrete instances of actors in the network. Moreover, the regular similarity between two actors measures to what extent these two have the same connections to classes of actors. Actors with a high regular similarity are considered to have the same role in the network. The problem then is to assign appropriate roles to actors such that same-role actors are also similar with respect to the roles of the actors to whom they are connected. If there is an assignment of actors to roles such that actors within a role are regularly equivalent (maximum regular similarity), the fitted blockmodel has only regular and null blocks without any errors. For example, in Figure 2 a perfectly fitting blockmodel would result from the regular equivalence classes {{1,2,3,4},{5,6},{7,8}}. In order to compute regular similarity in this work, the REGE algorithm (Borgatti & Everett, 1993) is applied.

4.3 Semantic Similarity

In contrast to graph-based similarity, semantic similarity is not computed from the connection patterns in the social network. Users can have certain properties like interests, age, gender, etc. The similarity of two users is calculated based on the distance of the users' property set or vector in a certain feature space. Thus, blockmodels based on this type of similarity can be considered as feature based (Rossi & Ahmed, 2015). In those blockmodels, roles are induced to the social network from external observations instead of direct inference from the network structure.

In our approach, the semantic similarity of users is calculated from the thematic areas in which they provide information and the thematic areas in which they seek for information (Figure 3). More formally, the notion of semantic similarity in MOOC discussion forums can be described as follows:

Given two users u_x and u_y. Each user provides (P) information in subsets of all forum threads $T_x^P, T_y^P \subseteq T$ and seeks (S) for information in $T_x^S, T_y^S \subseteq T$. The similarity regarding the information providing interests or expertise can then be calculated as in equation 1.

$$sim_{sem}^{P}(u_x, u_y) = \frac{\sum_{t_{x,i} \in T_x^P} \max\left(sim(t_{x,i}, T_y^P)\right)}{\max\left(|T_x^P|, |T_y^P|\right)} \quad (1)$$

The term $sim(t_{x,i}, T_y^P)$ corresponds to the similarities between the *i*th thread in which user u_x provides information and the set of threads in which user u_y provides information. The calculation for the similarity of their information-seeking interest $sim_{sem}^{S}(u_x, u_y)$ of two users can be calculated by their sets of threads in which they ask for information accordingly.

The final semantic similarity of users u_x and u_y will be defined as the average of their expertise similarity and the similarity of their information-seeking interests, as given in equation 2.

$$sim_{sem}(u_x, u_y) = \frac{sim_{sem}^{P}(u_x, u_y) + sim_{sem}^{S}(u_x, u_y)}{2} \quad (2)$$

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

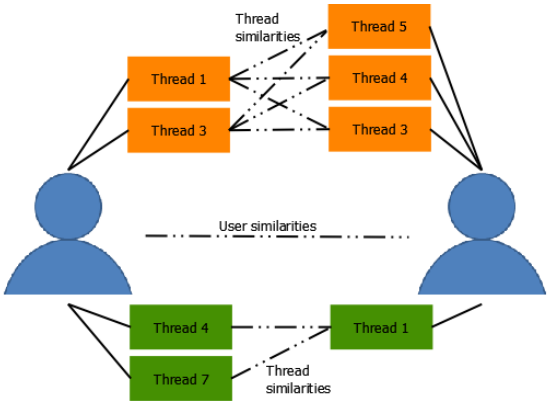

Figure 3: Semantic similarity of two users based on the similarity of threads in which they provide information (orange) and seek for information (green).

The distinction between information-giving and information-seeking interests is crucial in determining role in semantic modelling. A role, in terms of thematic interests, can be interpreted as users who are information providers for the themes X and pull information from themes Y. Furthermore, if the distinction between information-giving and information-seeking had not be made, the resulting blockmodel is likely to contain relations from a certain role to the role itself, hardly allowing for a distinction between social and semantic roles. Communication in one thematic area implies corresponding connections in the information exchange network.

For the calculation of the similarity between threads, which is a prerequisite for the calculation of the semantic similarity between users, one has several options. Forum threads can be considered as documents. One possibility then would be to calculate their semantic similarities based on latent semantic indexing (LSI; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), a well-known technique from information retrieval. LSI, in general, derives the similarities between threads based on a principal component analysis of the columns of a term-document matrix. An alternative approach, used in this work, is to extract meaningful concepts from the forum threads first and then calculate the similarity of threads from the average semantic similarity of the assigned concepts. Concept similarity is calculated by the UMBC semantic similarity service (Han, Kashyap, Finin, Mayfield, & Weese, 2013), which combines latent semantic analysis on large corpora with word net similarity of the assigned concepts. The concept extraction is done by the Social Tagging Engine provided by Thompson Reuthers' Open Calais.[3] It extracts concepts from textual documents by comparing the documents to Wikipedia pages. This has several benefits compared to other approaches for keyword extraction. First, the concepts do not have to be mentioned exactly in the thread posts. The assigned concepts generalize the keywords to higher-order concepts using Wikipedia page titles as a controlled vocabulary, which can be seen as an inherent resolution of synonyms, polysemy, and disambiguation. This also solves the problem of short text and inexact language, which is common in discussion forums. Additionally, this approach

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

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

has the advantage of simultaneously assigning meaningful concepts to the threads, which is very helpful for the interpretability of the semantic clusters found in later steps.

4.4 Socio-semantic Blockmodelling

Next, we show how regular similarity (social role modelling) and semantic similarity (semantic role modelling) can be combined into a hybrid approach that we call socio-semantic blockmodelling. The goal, given an allocation of users to roles, is to identify regular relations between semantically coherent (but not necessarily socially coherent) roles in the knowledge exchange network extracted from the forum data. A directed regular relation from a role A to a role B in a regular similarity blockmodel indicates information flow from role A to role B since all users in A give information to at least one user in B and all users in B receive information from at least one user in A (c.f. Section 4.2.2).

Semantic similarity, as described in Section 4.3, identifies semantically coherent roles but with possibly heterogeneous communication patterns. For example, a graph-based role summarizes people who have many outgoing connections (information providers) to people with many ingoing connections (information consumers). A semantic role can characterize users who have problems with topic X or who have expertise on topic Y. The combination of both can then be seen as a social role in the semantic context.

On the one hand, if the semantic structure of the community is not strongly interleaved with the structure of information exchange, it might be very hard to find regular relations between roles and the resulting blockmodel is very inaccurate. On the other hand, if the blockmodel is solely created from role assignments based on regular similarity, the resulting blockmodel is likely to be more accurate than a blockmodel derived from semantic similarity since the roles are discovered using the same criterion used to identify role relations. However, regular similarity identifies role relations based on communication patterns while ignoring the interests and semantic coherence of users within a role. The problem is to find a good assignment of users to roles such that the resulting blockmodel is as accurate as possible in terms of regular role relations (information flow) and high semantic coherence within a role. To achieve this, our socio-semantic approach to blockmodelling combines regular and semantic similarity in the assignment of users to roles. The roles in this context can be interpreted differently. For example, information providers for topic X discovered by the semantic approach could be subdivided into different types based on their connection patterns in the network discovered based on regular similarity.

Combining user features with network structure (Rossi & Ahmed, 2015), and identifying the optimal blockmodel with respect to multiple objectives by optimizing role allocations is a hard problem (Brusco, Doreian, Steinley, & Satornino, 2013; Harrer & Schmidt, 2012). An indirect approach where regular and semantic similarities can be “mixed” into a joint similarity by weighted average (equation 3) gives good results and is feasible for big datasets. Further, varying the values for the weighting factors allows for investigating the interdependency between both semantic and social (regular) similarity, which will be reported in Section 6.

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

$$sim_{socsem}(x,y) = \frac{\sigma_{reg} * sim_{reg}(x,y) + \sigma_{sem} * sim_{sem}(x,y)}{(\sigma_{reg} + \sigma_{sem})} \quad (3)$$

Based on this formulation of similarity, a blockmodel is derived as follows:

1. Build a hierarchical clustering based on $sim_{socsem}(x,y)$ for each pair of users.
2. Determine the number of roles by cluster bootstrapping (Fang & Wang, 2012), a method that estimates the optimal number of clusters given distances/similarities of objects and a clustering function by minimizing cluster instability.
3. Assign the role relations such that the blockmodel error is minimal as described in Section 4.1.

The sparsity of the network is a problem since it biases the inference towards null relations (see Section 4.1). If the density of a network is too small, assigning null relations always gives a small blockmodel error. For this reason, the acceptable error for introducing a regular relation between two roles is enhanced in relation to the network density as suggested in Ziberna (2013).

5 METHODS 2: CHARACTERISTICS OF ACTIVITY AND ENGAGEMENT OVER TIME

5.1 Classification of Learners by Activity/Engagement

The previously introduced method (socio-semantic blockmodelling) is of an "interactional" or "relational" nature in that it derives network structures from interactions (especially knowledge exchange) between actors. Alternative characterizations can be based on certain activities or contribution types attributed to individual learners. According to the ICAP framework (Chi & Wylie, 2014), learners can be progressively classified into four modes (or states) of cognitive engagement: passive < active < constructive < interactive. The core claim of ICAP is that the quality of learning increases in this order. In more detail, the ICAP categories can be characterized as follows:

- **The Passive Type**: learners who receive information but do not engage in any other observable activity that could be related to learning
- **The Active Type**: learners who receive information and additionally perform or engage in observable motor or mental activities that relate to, support, or scaffold the learning process
- **The Constructive Type**: learners who create content or build knowledge on top of the information they have received, either during self-reflection or while engaging in some constructive activity
- **The Interactive Type**: learners who demonstrate a constructive behaviour in combination with interpersonal activities such as constructive dialogues over a period that signifies meaningful engagement beyond mere information sharing

The ICAP framework has been used to characterize learner behaviour in MOOC discussions (Wang, Yang, Wen, Koedinger, Rosé, 2015). In this approach, the classification was based on human judgement. We

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

have tried to operationalize these characterizations in order to associate them with learner data available in the MOOC discussion forums, resulting in the following classifications:

- **Passive:** Users who do not contribute to discussions.
- **Active**: Users who have at least one content-related contribution based on the forum post classification described in Section 3.1, i.e., information-giving or information-seeking posts.
- **Constructive**: Constructive learners contribute on top of course materials. Typical activities are reflection, posting questions, and forming hypotheses. Active users who additionally start discussion threads can be considered as constructive since they take the initiative to further work on the course content together with others. For the majority of cases, thread starters post information-seeking posts such as questions on the course material or challenge concepts of the course topic by requesting others opinions. Furthermore, users with at least five posts in one week can also be considered constructive. While the number of posts in one week per user is typically below two, we argue that someone who has five or more content-related contributions in one week should be someone who at least reflects on the course material, posts several questions and answers, and thus, contributes to the knowledge construction in the forum community.
- **Interactive**: Constructive users do not necessarily engage in interactive knowledge construction. For example, a user can be constructive by posting single help-giving posts to different threads without engaging in actual conversations. Consequently, users are classified as interactive if they are constructive and additionally engage in content-related dialogue with others. We define that a user is in an interactive engagement state if they participate in a forum thread on a certain topic and have at least three content-related contributions interleaved with contributions of others in this discussion. The threshold of three can be justified considering possible forum activity that leads to this classification. For example, a user can trigger interactive knowledge construction by posting a question, receiving an answer, posting one of the follow-up questions, receiving another answer, and then either providing information to other information seekers in the same thread, or keep challenging others with questions. Another example would be someone whose role is more of an information provider (expert) on the discussion topic interacting with several information seekers. Accordingly, interactive discourse can be characterized by sequential patterns of participation. Requiring a minimal threshold of three posts per single user with intermediate posts of others implies that the discussion threads have to be sufficiently long, which is another indication of interactive discussions.

5.2 Temporal Patterns of Engagement and Activity

In contrast to the analysis of structural and content aspects, as in described in Section 4, activity based characterization of forum users can also be based on sequential patterns of the aforementioned states of engagement. Forum users with similar patterns can be considered as users who adopt similar roles in this aspect.

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

To assess the similarity of these state sequences, we use a sequence alignment approach originating in bioinformatics, for example, to discover similar gene sequences (Mount, 2004). More generally, the similarity of any pair of state sequences can be measured by their optimal matching distance (Abbott & Tsay, 2000). The optimal matching distance between two sequences results from the number of substitution and insertion operations necessary to unify the two sequences. The insertion of states or gaps is only necessary for sequences of unequal length. In our case, the length of all sequences is equal to the number of course weeks, therefore only substitution costs matter. Furthermore, different substitutions can have different costs. This again is inspired by biology-related research where the substitution of one gene by another gene with a similar function is not as expensive as the substitution of genes that are very different. Here, the substitution cost can be directly inferred from the ICAP framework, since it already introduced an ordering of engagement states (I > C > A > P). This results in the substitution cost matrix given in Table 3.

Table 3: Substitution Cost Matrix of ICAP States

	Interactive	**Constructive**	**Active**	**Passive**
Interactive	0	1	2	3
Constructive	1	0	1	2
Active	3	1	0	1
Passive	3	2	1	0

For example, the two-state sequences given below can be unified at a cost of three by substituting the second state at a cost of one and the forth state at a cost of two.

Sequence 1	passive	Passive	active	constructive
Sequence 2	passive	Constructive	active	interactive

For the distances, respectively, the similarities of the activity sequences can be used by a clustering algorithm to group users based on their forum activity over time.

6 RESULTS

In the next subsections, the approaches for the characterization of user roles in MOOC discussion forums described in Sections 4 and 5 are reported for both courses separately. First, we aim to answer to what extent the social and semantic structure of the forum community is interleaved. More concretely, how well does role assignment based on semantic similarity induce a blockmodel that has a small error and are such roles also semantically coherent in the sense of discussion topics? The results are further contrasted with typical engagement patterns based on the ICAP sequences introduced in Section 5.

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

6.1 Results I: Corporate Finance MOOC

6.1.1 Semantic vs. social structuring

In the following, we investigate the relation between the social structure of the social information exchange network and the semantic structure based on the similarity of interests/expertise of the users in thematic areas in the discussion forum.

Table 4: Correlations between Different Types of Similarities (Corporate Finance MOOC)

	Structural	Regular	Semantic
Structural	1*	–0.19*	–0.16*
Regular	–0.19*	1*	0.36*
Semantic	–0.16*	0.36*	1*

*Statistical significance level: p<<0.05

First, we conducted a correlation analysis between the graph-based (social) similarities described in Section 4.2 and the semantic similarity of users (Section 4.3). If social and semantic structure were highly correlated, role assignment based on graph-based and semantic similarity would result in very similar blockmodels. Thus, the parameter settings in equation 3 would have no strong effect on the result. The Spearman rank correlations between the different types of user similarities are reported in Table 4 for the Corporate Finance MOOC. All correlations are statistically significant ($p << .05$). There is a moderate positive correlation between regular and semantic similarity. This means that there is no strong interdependence between the semantic structure based on the information-giving and information-seeking interests (semantically induced roles) and the information flow between roles based on connection patterns (regular similarity induced roles) in the discussion forum of the Corporate Finance MOOC. This indicates that direct communication between users does not influence their interests significantly and, vice versa, interests do not affect the social structure of the community. Structural equivalence correlates on a very low level negatively with the other similarity measures. Thus, concrete connections between users can be considered as independent from the regular role structures and users' interests.

In order to further investigate the relations between social and semantic role structures, we generated blockmodels with different emphasis of regular (social) and semantic similarity by varying the parameters σ_{reg} and σ_{sem} (equation 3). For each blockmodel, the normalized blockmodel error (*bm_err* — i.e., the deviation from an optimal model, see Section 4.1) is provided. The semantic dissimilarity of a role is evaluated by the ratio of the average semantic distance of users within the same role and the average distance of users of different roles (*wb_ratio*). Consequently, a "*good*" blockmodel should have a low values for *bm_err* and *wb_ratio*.

The results are presented in Figure 4. For both, the *bw_ratio* and the *bm_err*, a state transition between role assignments emphasizes social similarity and role assignment more than the semantic similarity of users. The results are compared to the average *wb_ratio* and *bm_err* of 50 blockmodels based on a random assignment of users to roles. Even if the social and semantic structure of the community is not strongly related, there is at least some influence such that, even for the extreme cases, pure semantic

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

and pure social blockmodels are still better than random role assignment. These findings support the assumption that socio-semantic coevolution takes place in the discussion forum to some extent. Furthermore, this shows that the community bears a structure in both the social dimension and the semantic dimension.

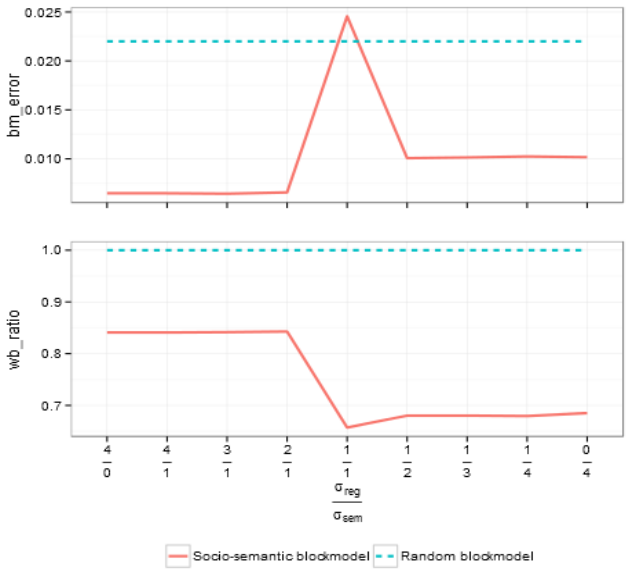

Figure 4: Blockmodel error (top) and ratio of average semantic distance within roles and between roles (bottom) for different ratios of σ_{reg} and σ_{sem}.

6.1.2 Socio-semantic blockmodelling

In the following, the socio-semantic structure of the forum communication is analyzed based on a hybrid blockmodel. For our analysis, we took into account the semantic coherence of roles as well as the blockmodel error in terms of regular relations. In order to do this, we had to establish a good trade-off between semantic and regular similarity according to equation 3 such that the semantic similarity of users within the same role is high and the blockmodel error is small. Figure 5 depicts the ratio between the blockmodel error *bm_error* and the coherence of the roles (1 — *wb_ratio)* for different values for σ_{reg} and σ_{sem}. As (1 — *wb_ratio)* has to be as large as possible and *bm_error* as small as possible, a good "mixture" is given for σ_{reg}*=1* and σ_{sem}*=2.*

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

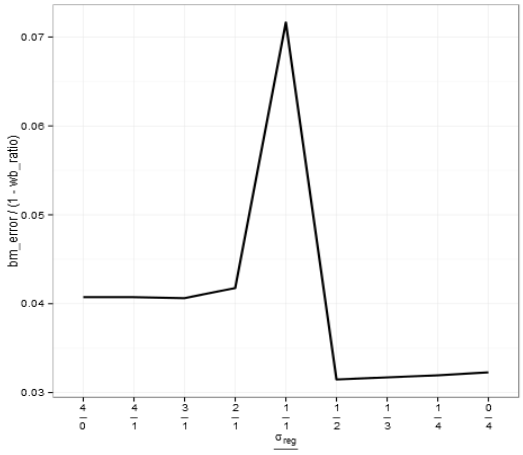

Figure 5: Ratio between blockmodel error and semantic coherence of the roles in the Corporate Finance MOOC.

The resulting blockmodel is depicted in Figure 6. The nodes represent the three discovered roles and the edges represent regular relations between them. The node size corresponds to the number of users assigned to the role and the edge width to the number of links present between the roles.

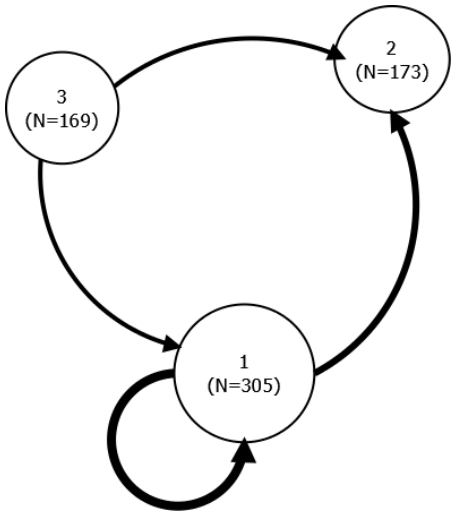

Figure 6: Blockmodel for the forum discussion in the Corporate Finance MOOC.

For the Corporate Finance MOOC, there is one dominant role (role 1) consisting of 305 users. It has regular relations not only with the other roles but also with itself. This means that there is information flow from role 1 to role 2 and also information flow within the role indicated by the self-loop. The two smaller roles 2 and 3 have different connection patterns. Role 2 has only ingoing regular relations to the other roles and role 3 has only outgoing relations. This indicates a smaller set of users who can be characterized as information seekers (role 2) and others as information-providers (role 3). This is further validated by the mean inreach and outreach of the users (columns 3 and 4 of Table 5). Inreach and outreach are measures that characterize forum users according to information-seeking and information-

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

giving behaviour (Hecking et al., 2015). Users with high inreach receive many answers from different users and users with high outreach have many information-giving posts having a high diversity of targeted users. The value for the mean outreach is very small for role 2 and the value for inreach is small for role 3. However, the largest values for both measures can be found for role 1. Role 1 can be seen as the core community comprising information providers and information seekers as well as users who are both. In this sense, roles 2 and 3 can then be seen as peripheral users who are more specialized in their communication behaviour. On the semantic level, the roles can be differentiated with respect to the thematic areas in which they provide information (expertise) and areas in which they seek for information (first two columns of Table 5). For role 1, there is no clear semantic distinction between information-giving and information-seeking interests, which is also reflected by the self-loop in the blockmodel (Figure 6). The concept "Mathematical finance" is associated with every role since it is an important general concept assigned to many threads by the concept extraction described in Section 4.3. This is reasonable since many of the assignments in the course deal with calculations of various values related to corporate finance. Consequently, this concept cannot be used to characterize the particular roles.

Table 5: Properties of the Discovered Roles for the Corporate Finance MOOC

Role	Top info. giving	Top info. seeking	Mean inreach	Mean outreach
1	1. Mathematical finance 2. Investment 3. Depreciation 4. Taxation	1. Mathematical finance 2. Investment 3. Depreciation 4. Taxation	8.38	8.45
2	None	1. Mathematical finance 2. Investment 3. Depreciation 4. Rate of return *5. Question*	3.58	0.43
3	1. Mathematical finance 2. Investment 3. Rate of return 4. Net present value	*1. Ambiguity* 2. *Decision theory*	0.28	3.08

Other most frequent concepts extracted from forum threads in which users of role 1 appear as information seekers and givers are "Investment," "Depreciation," and "Taxation" — some of the main concepts covered during the course. Participants had to calculate depreciation and investment rates as part of their assignments. Issues regarding the calculation itself and formal requirements (such as the rounding of real numbers) were discussed among the participants. In particular, the correct formulas were heavily discussed, such that users of role 1 appear as both information-givers and information-seekers.

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

Information-seekers, role 2, have no key concepts assigned to their information-giving interests. These users seek information, especially in areas related to investments. They receive help from users in roles 1 and 3 on this topic. The nature of role 2 is further underlined by the fact that many threads in which they are active are additionally annotated with the keyword "question." Role 3 users can be interpreted as experts for the topics related to investment appraisal. Although this is a relatively small role in terms of number of users, the mean outreach is moderately high. Possible reasons for this could be that these users either 1) provide information on a course topic in their field of expertise and then stop participating in the forum, or 2) show a kind of "elder statesman" behaviour by occasionally contributing to the information exchange in the forum as experts on topics of wide interest to the whole community.

In summary, the three discovered roles can be interpreted as role 1 = "core users," role 2 = "peripheral information seekers," and role 3 = "peripheral information givers."

6.1.3 Activity characteristics of users

User roles based on ICAP state sequences. As described above, the users of the discussion forums can also be characterized by their activity over time. Therefore, the contributors were first classified into the ICAP states interactive, constructive, active, and passive for each week of the course, as described in Section 5.1. The state distribution of the actors for the Corporate Finance discussion forum is depicted in Figure 7.

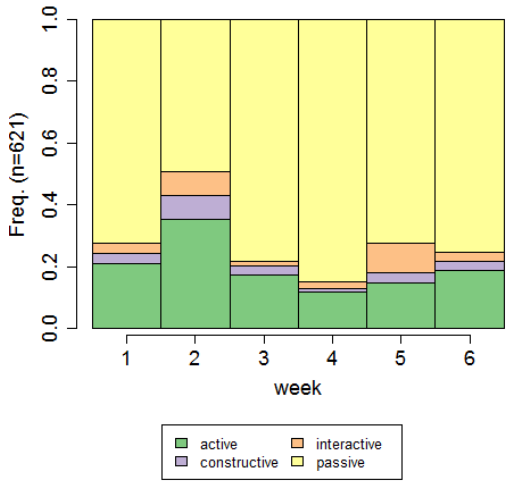

Figure 7: Distribution of ICAP classes over course weeks in the Corporate Finance MOOC.

As expected, the majority of users are in the passive state and, possibly due to course dropouts, the fraction of passive users is higher after the first two course weeks. Only a very small subset of forum users show constructive behaviour and an even smaller fraction are engaged in collaborative knowledge construction in the form of dialogues on the course content (interactive state).

For a more in-depth analysis, and in order to model roles of users from the perspective of engagement, users were clustered based on the similarity of their ICAP state sequences over course weeks according to the optimal matching distance between each two sequences (Section 5.2). A proper number of

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

clusters was determined by cluster bootstrapping (Fang & Wang, 2012), the same as for the socio-semantic blockmodels, which results in a partitioning of the ICAP sequences into 8 clusters.

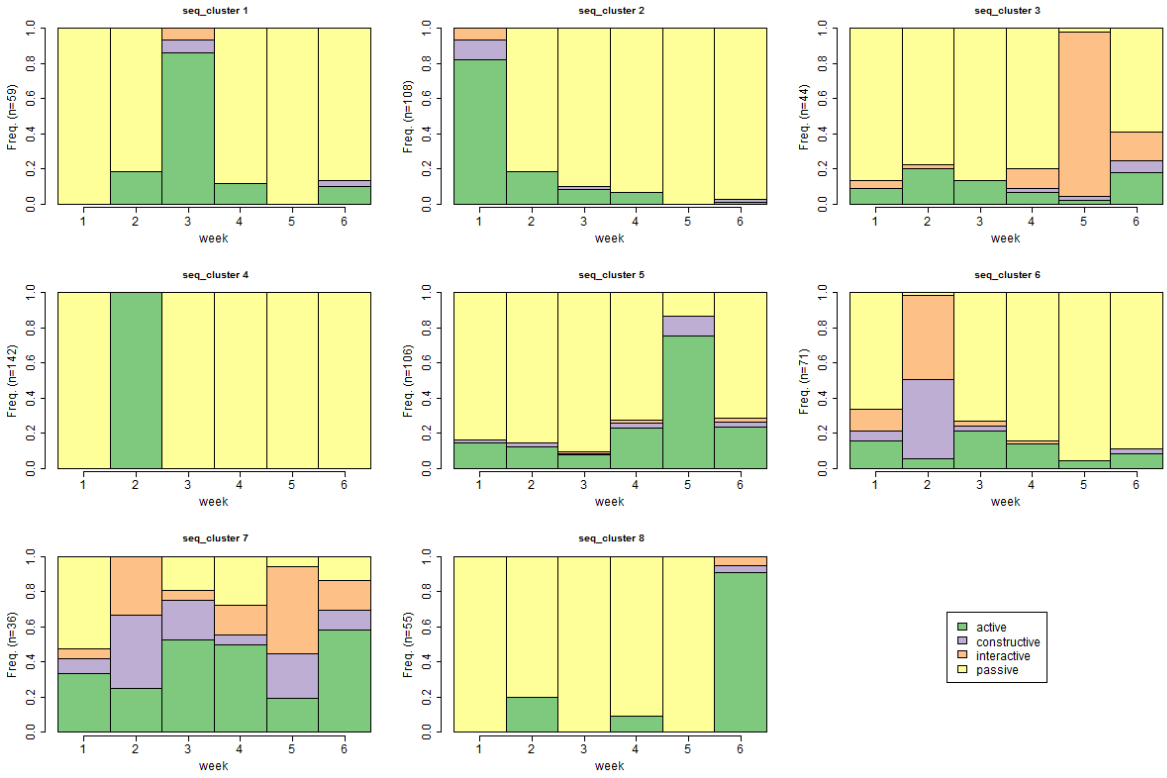

Figure 8: ICAP sequence clusters for the Corporate Finance MOOC.

Figure 8 depicts the state distributions over the course weeks of the eight ICAP sequence clusters determined for the Corporate Finance MOOC. The vertical axes of the diagrams denote the number of users/sequences in each cluster. These results show that the activity of many users is restricted to a limited period, i.e., one or two weeks. In this sense the temporal activity based roles of users can be distinguished between “early dropouts” i.e., users who are more active in the beginning of the course (sequences clusters 1, 2, and 4) and “late starters” (sequence clusters 3, 5, and 8). Moreover, it can also be seen that sequence cluster 7 especially comprises users with more evenly distributed forum contributions over the course weeks. This type of role can be named “all time actives.” Consequently, one can state that collaborative knowledge exchange in a constructive or even interactive manner takes place among the small number of users who use the forum frequently during the course. Sequence clusters 3 and 6 denote cases in which semi-active users only engage in interactive dialogue or constructive contributing in particular weeks. While cluster 6 comprises users who start as passive or active and then become constructive and interactive towards the end of the course, cluster 3 shows the opposite case: users who are interactive or constructive at the beginning of the course and become increasingly passive in subsequent weeks.

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

Relating temporal activity and socio-semantic roles. The next evaluation step is to relate the user characteristics based on engagement over time and the different groups of users (or user roles) discovered by the blockmodelling approach.

A quantitative comparison of the blockmodel roles and the groups induced by the ICAP sequence clustering (seq_cluster_x) described in the previous section is reported in Table 6. The cells of the table show the probability of picking a user with a certain blockmodel role given an ICAP sequence cluster *P(bm_role|seq_cluster)*. The ICAP sequence clusters 3, 6, and 7 that comprise the sequences that stretch the posting activities over several weeks are also those that include the most sequences with constructive and interactive states. These clusters are dominated by core users according to the socio-semantic blockmodels. This is not surprising given the results above. The peripheral users, according to socio-semantic connection patterns, are users who are active at the beginning of the course (early dropouts) or to the end of the course (late starters). This suggests that the periphery of the information exchange network consists mainly of users who either use the discussion forum opportunistically — as a means to get an understanding of specific course topics at the beginning of the course — or for support for finishing the course. This interpretation is also supported by their more specific discussion topics in contrast to those of core users (c.f. Table 5).

Table 6: Probabilities of Particular Blockmodel Roles (Rows) Given Sequence Clusters (Columns) *P(bm_role|seq_cluster)* for the Corporate Finance MOOC

	Core Users	Periph. Info. Seeker	Periph. Info. Giver
seq_cluster_1	***0.39***	0.32	0.29
seq_cluster_2	***0.35***	0.3	***0.35***
seq_cluster_3 (late starters)	**0.88**	0.07	0.05
seq_cluster_4 (early dropouts)	0.15	0.41	***0.44***
seq_cluster_5	**0.49**	0.32	0.19
seq_cluster_6 (early interactives/ constructives)	**0.81**	0.07	0.12
seq_cluster_7 (all time actives)	**1**	0	0
seq_cluster_8 (late actives)	***0.37***	0.28	0.35

6.2 Results II: Global Warming MOOC

This section reports the results of the analysis steps for the Global Warming MOOC and relates it to the results reported above.

6.2.1 Semantic vs. social structuring

As for the Corporate Finance MOOC, in the Global Warming MOOC there is also a moderate Spearman

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

correlation ($p << 0.05$) between regular (relation based) similarity and semantic similarity (Table 7). However, there is also a correlation between the structural (common neighbour based) similarity and the semantic similarity. This indicates that in the Global Warming MOOC the discussion topics may be established more by direct communication than externally induced by the course outline.

Table 7: Correlations between Different Types of Similarities (Global Warming MOOC)

	Structural	Regular	Semantic
Structural	1	0.12	0.3
Regular	0.2	1	0.32
Semantic	0.3	0.32	1

The relation between social and semantic role structures in the Global Warming MOOC is depicted in Figure 9. Similar to the results for the Corporate Finance MOOC reported in Section 6.1.1, we generated blockmodels with a different emphasis for regular (social) and semantic similarity by varying the parameters σ_{reg} and σ_{sem} (equation 3).

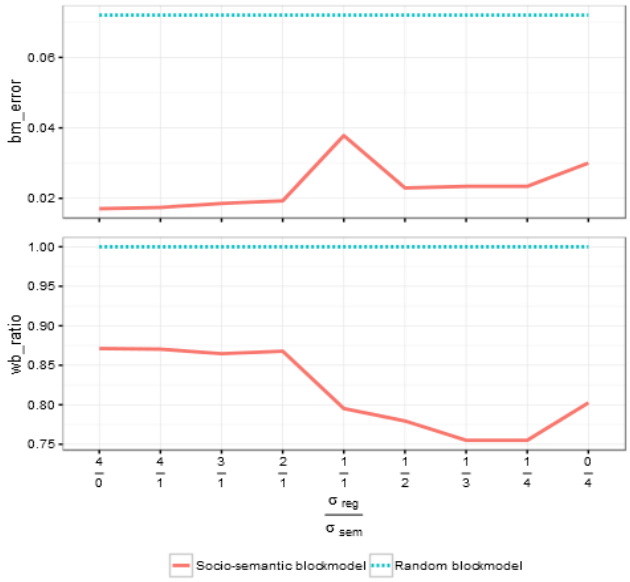

Figure 9: Blockmodel error (top) and ratio of average semantic distance within roles and between roles (bottom) for different ratios of σ_{reg} and σ_{sem}.

Again, a state transition between role assignments emphasizes more social similarity and role assignments than semantic similarity of users. Compared to the average *wb_ratio* (semantic within cluster distance/semantic between cluster distance) and the blockmodel error (*bm_err*) of 50 random blockmodels, the socio-semantic role assignments are much better. Thus, socio-semantic co-evolution can also be assumed for the Global Warming MOOC, which is a prerequisite for the following analyses.

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

6.2.2 Socio-semantic blockmodelling

The weighting parameters $\sigma_{reg}=1$ and $\sigma_{sem}=3$ for the combination of social (regular) similarity and semantic similarity of users (equation 3) to derive the roles of the socio-semantic blockmodel depicted in Figure 10 were determined in the same way as in Section 6.1.2 by determining the best combination according to low blockmodel error and high semantic cohesion of the roles.

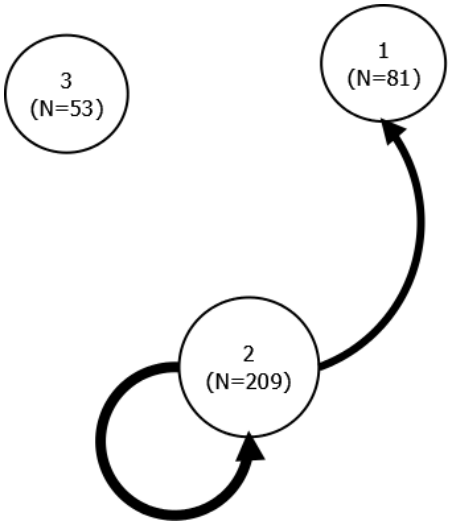

Figure 10: Blockmodel for the forum discussion in the Global Warming MOOC.

The discovered socio-semantic role pattern for the Global Warming MOOC (Figure 10) is similar to that of the Corporate Finance MOOC (Section 6.1.2). However, the fraction of users who can be considered as the core community (role 2) is much larger compared to the other course. This may be a result of the smaller number of users in the Global Warming forum, resulting in a more densely connected information exchange network. Similar to role 3 in the Corporate Finance MOOC, role 3 in the Global Warming MOOC consists mainly of peripheral information givers, as the outreach values reported in Table 8 show. However, this group is not connected to other roles by a regular relation. According to the definition of regular relations (see Section 4.2.2), this does not mean that these users do not have any relations with other users. However, there are no regularities in their relations with other roles since the connection patterns and discussion interests within this role are more diverse than in the other roles. The second peripheral role (role 1) in the Global Warming MOOC comprises mainly information seekers, as can be seen in Table 8. In contrast to the corresponding role in the Corporate Finance MOOC, peripheral information seekers in the Global Warming forum seem not to be specialized in their thematic interests since general terms like “atmosphere” and “climate forcing” are among the most frequent information-seeking topics.

In general, the discovered socio-semantic role structure in both courses reflects the general assumptions about MOOC discussion forums very well. A “core” community (role 1 in Corporate Finance and role 2 in Global Warming) is more engaged in the main discussion topics than other roles, which can be seen by the higher values for in- and outreach. There is also communication within this role. The other roles (roles 2 and 3 in Corporate Finance and roles 1 and 3 in Global Warming) correspond to the users who

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

participate in forum communication occasionally and can thus be considered either "peripheral information-givers" or "peripheral information-seekers" on certain topics.

Table 8: Properties of the Discovered Roles for the Global Warming MOOC

Role	Top Info. Giving	Top Info. Seeking	Mean Inreach	Mean Outreach
1	None	1. Atmosphere 2. Climate change 3. Climate forcing	4.27	0.9
2	1. Climate change 2. Global warming 3. Climate history 4. Electromagnetic radiation	1. Climate change 2. Global warming 3. Climate history 4. Electromagnetic radiation	19.81	21.22
3	1. Global warming 2. Climate change policy 3. Energy development 4. Renewable energy	None	0.12	2.74

6.2.3 Activity characteristics of users

User roles based on ICAP states. The distribution of ICAP states for the Global Warming MOOC can be seen in Figure 11. Similar to the Corporate Finance MOOC, most users are inactive in particular course weeks. However, the fraction of users who are in the constructive or interactive state is slightly higher, suggesting that in this smaller course the discussion forum is used more for collaborative knowledge construction. The course is two weeks longer than the Corporate Finance MOOC, but in the last weeks (7 and 8) almost 20% of the forum users are still in a non-passive state.

The ICAP sequence clusters for the Global Warming MOOC can be seen in Figure 12. Cluster bootstrapping suggests eight clusters, the same as for the Corporate Finance MOOC. Here, similar patterns of users who are active only for few course weeks can be observed. However, this pattern is not as salient as in the Corporate Finance MOOC. Late starting is not as frequent and a large number of users show "early dropout behaviour" with decreasing or no contributions after the first weeks of the course (especially sequence clusters 4 and 5). Furthermore, in this smaller course some small sequence clusters have a very high presence of constructive and interactive states. This may also result from the overall higher relative forum activity, which suggests higher engagement in interactive information exchange (c.f. Figure 11). Especially the small sequence clusters 7 and 8 comprise a majority of constructive and interactive users who can be considered as "all time actives" contributing regularly during the course. This result shows that a small discussion forum such as this may even be a potential opportunity for community building among the most active forum users.

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

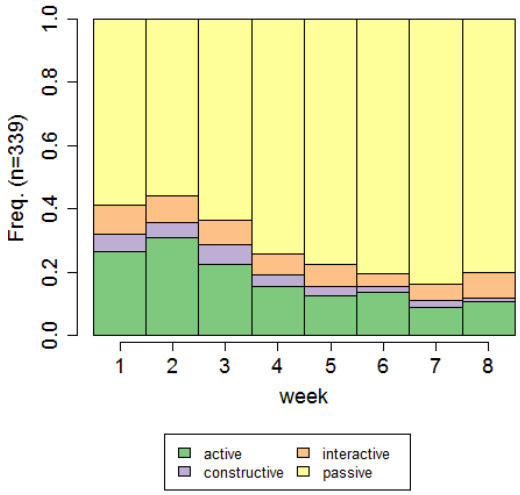

Figure 11: Distribution of ICAP classes over course weeks in the Global Warming MOOC.

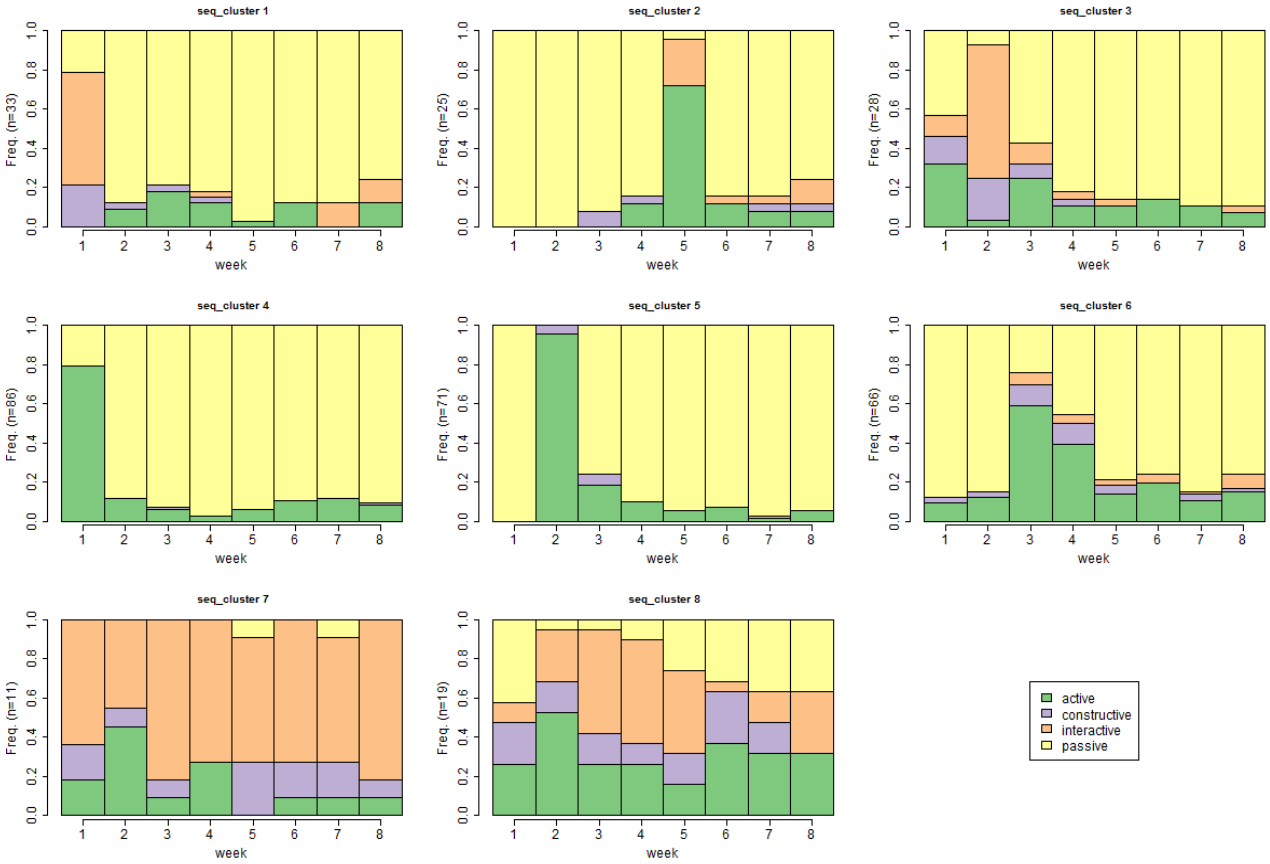

Figure 12: ICAP sequence clusters for the Global Warming MOOC.

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

Relating temporal activity and socio-semantic roles. The quantity of different blockmodel roles in the ICAP sequence clusters for the Global Warming MOOC are described in Table 9, expressed as the probability of picking a certain blockmodel role given a sequence cluster *P(bm_role|seq_cluster)*. The fraction of core users is high in all clusters, which results from the overall higher number of users who belong to the core. However, in the sequence clusters that comprise the most non-passive states (3, 6, 7, and 8) this overrepresentation is extreme. Peripheral information seekers are well represented in sequence clusters 2 and 4. While cluster 4 shows the typical early dropout pattern, cluster 2 has a high level of engagement in week 4 of the course. This suggests that peripheral help seekers can be divided into those who seek help early in the course and then stop contributing to the forum and those who appear as information seekers in one or two particular weeks, leading to the assumption that in mid-course, participants needed to gather information from the forum to progress. The role of peripheral information givers, which is not well represented in the Global Warming MOOC, also comprise users who drop out early from the discussion or are active in week 5 of the course.

Table 9: Probabilities of Particular Blockmodel Groups (Rows) Given Sequence Clusters (Columns) *P(bm_role|seq_cluster)* for the Global Warming MOOC

	Core Users	Periph. Info. Seeker	Periph. Info. Giver
seq_cluster_1 (early dropouts)	**0.43**	0.28	0.29
seq_cluster_2	***0.38***	***0.38***	0.24
seq_cluster_3	**0.89**	0.11	0
seq_cluster_4 (early dropouts)	***0.49***	0.36	0.14
seq_cluster_5 (early dropouts)	**0.67**	0.22	0.11
seq_cluster_6	**1**	0	0
seq_cluster_7 (all time actives)	**1**	0	0
seq_cluster_8 (all time actives)	**0.81**	0.08	0.11

7 CONCLUSION

To discover implicit user roles in MOOC discussion forums, we introduced and applied a mix of methods. First, we analyzed the social and semantic structure of a community of learners participating in a discussion forum. In the social dimension, users were assigned to roles based on the similarity of their position in the information exchange network ("regular similarity"). In the semantic dimension, roles were modelled based on the thematic areas in which users were actively providing or seeking information. Those semantic roles can also be interpreted as expertise and information-seeking for specific themes respectively. The third modelling approach concerns the activity and engagement of forum users. As a basis, we needed an adequate taxonomy for classifying states of user/learner engagement based on the forum data. In this respect, we operationalized the ICAP framework to

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

characterize forum users related to their activity over time. We believe that such a combination of methods is necessary to highlight different interdependent aspects and to further increase the understanding of the user/learner roles in MOOC discussion forums. This, in turn, is a prerequisite for the development of advanced concepts for community support.

The approaches have been demonstrated on data from forums that accompanied two Coursera MOOCs: Introduction to Corporate Finance and Global Warming. Our first objective was to investigate the interdependencies between the positions of users in the information exchange network and their topics of expertise and problem areas in the semantic dimension.

Our results clearly support the presence of social and semantic role structures in the investigated discussion forums (Sections 6.1.1 and 6.2.1). The semantic coherence of user roles with respect to the semantic similarity of users scores far better than a random assignment of users to semantic roles. The same can be stated about the error of a blockmodel based on regular similarity of users in terms their connection patterns in the information exchange network. Consequently, the community in the discussion forum did not evolve completely at random as might be suggested by the known differences of behaviour and engagement of participants in MOOC discussion forums.

Moreover, we have shown that the social and semantic roles of the user are not completely congruent. We discovered a moderate correlation between the regular similarity of users in the network and their semantic similarity. Our resulting hybrid blockmodels that combine both types of similarity for role assignment had a better fit with respect to semantic coherence of roles. The same holds for the blockmodel error with respect to regular role relations compared to random models, even in extreme cases (only regular similarity or only semantic similarity). However, semantic roles and social roles are also not interchangeable, which means that forum communication has only limited influence on the interests of users and vice versa. External factors such as individual experience and personal communication preferences might also impact the evolution of forum communication.

For our dataset, hybrid social-semantic blockmodelling revealed three different roles. The most dominating role characterized users who discussed the main course content and additionally communicated heavily with other users of the same role. Apart from that, we also identified two smaller roles that could be considered to contribute less to forum communication: 1) providing information on specific course topics and 2) seeking out information on very concrete issues. Occasional information exchange took place among users who belonged to these role groups.

Furthermore, we explored the relation between user roles identified through blockmodelling and states of engagement over time (Section 6.1.2 for Corporate Finance and Section 6.2.2 for Global Warming). Users who belong to the core (according to the blockmodels) are also the ones who showed a higher level of activity and engagement in interaction with others. This pinpoints that core users are more likely to form social communities and bonds with other users and to exert higher-level cognitive engagement and deeper learning as compared to peripheral users who circumstantially seek specific information. Further evidence shows that core users are also more likely to develop a deeper understanding of the

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

course matter than peripheral users. According to the ICAP framework, they engage in different cognitive, knowledge-related processes (related to constructive and interactive behaviours instead of only passive or active engagement) resulting in increased learning. This finding is not surprising; however, it is critical because it provides insight regarding how the posting activity of MOOC users and their online behaviour on discussion boards relates to learning gains and better learning. Thus, we can gain valuable information to improve the design and orchestration of online courses and to provide effective and meaningful feedback to users.

All of these findings suggest a need for better support of information exchange between peers in MOOCs. Advances in the design of asynchronous communication in online courses should consider better adaptivity to different needs of different user roles. In order to increase the engagement in collaborative activities, it would be desirable to help participants move from the periphery of the information exchange network to the core, i.e., to help them become more communicative and thus achieve a higher level of collaboration. As shown, expertise and information needs in thematic areas are not well reflected in the social communication structure of the discussion forum. Results from socio-semantic role modelling can be used to provide social support; for example, recommendations that help students find proper communication partners for certain thematic areas. This together with explicit stimulation using conversational agents, as described by Ferschke et al. (2015) may enhance the engagement of learners in sustainable knowledge building dialogues and information exchange in the discussion forum. Furthermore, the low level of engagement of the majority of users could be counteracted by incorporating incentives into the discussion forums to motivate learners to participate in discussions in a constructive and interactive manner. Initial steps in this direction have already been taken (Anderson et al., 2014).

Our study has certain inherent limitations in that it was focused on the characterization of users with respect to their online behaviour, such as their posting habits and engagement patterns. However, we have not explored how possible external factors drive the evolution of the community and how these factors might affect the emergence of different user roles. For example, it would be interesting to explore how tutorial intervention (i.e., posts or messages from MOOC instructors) can affect the behaviour of MOOC learners with respect to the different user roles that we identified; for example, whether users of particular roles are likely to change or adopt a different role after receiving feedback from the instructor. Additionally, it would be interesting to explore the extent to which cultural characteristics affect the emergence and formation of user roles.

In our future work, we aim to investigate more MOOC discussion forums in order to find out whether the structures we have found for the course described in this paper can be considered as general patterns of forum communication in such online courses. In this paper, we focused on user activity in discussion forums. However, we plan to extend this research and study how users of different roles are engaged in other activities of the online course. On the methodological level, it would also be interesting to incorporate resource usage patterns into the role modelling.

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

REFERENCES

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology review and prospect. *Sociological Methods & Research, 29*(1), 3–33. https://dx.doi.org/10.1177/0049124100029001001

Abnar, A., Takaffoli, M., Rabbany, R., & Zaïane, O. (2015). SSRM: Structural social role mining for dynamic social networks. *Social Network Analysis and Mining, 5*(1). http://dx.doi.org/10.1007/s13278-015-0292-y

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). Engaging with massive online courses. *Proceedings of the 23rd International Conference on World Wide Web* (WWW '14), 7–11 April 2014, Seoul, Republic of Korea (pp. 687–698). New York: ACM. http://dx.doi.org/10.1145/2566486.2568042

Arguello, J., & Shaffer, K. (2015). Predicting speech acts in MOOC forum posts. *Proceedings of the 9th International AAAI Conference on Web and Social Media* (ICWSM '15), 26–29 May 2015, Oxford, UK (pp. 2–11). Palo Alto, CA: AAAI Press.

Borgatti, S. P., & Everett, M. G. (1993). Two algorithms for computing regular equivalence. *Social Networks, 15*(4), 361–376. http://dx.doi.org/10.1016/0378-8733(93)90012-A

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324

Brusco, M., Doreian, P., Steinley, D., & Satornino, C. (2013). Multiobjective blockmodeling for social network analysis. *Psychometrika, 78*(3), 498–525. http://dx.doi.org/10.1007/s11336-012-9313-1

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. http://dx.doi.org/10.1080/00461520.2014.965823

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391–407.

Doreian, P., Batagelj, V., Ferligoj, A., & Granovetter, M. (2004). *Generalized blockmodeling* (Structural analysis in the Social Sciences). New York: Cambridge University Press.

Engle, D., Mankoff, C., & Carbrey, J. (2015). Coursera's introductory human physiology course: Factors that characterise successful completion of a MOOC. *The International Review of Research in Open and Distributed Learning, 16*(2). http://dx.doi.org/10.19173/irrodl.v16i2.2010

Fang, Y., & Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis, 56*(3), 468–477. http://dx.doi.org/10.1016/j.csda.2011.09.003

Ferschke, O., Howley, I., Tomar, G., Yang, D., & Rosé, C. P. (2015). Fostering discussion across communication media in massive open online courses. *Proceedings of the 11th International Conference on Computer Supported Collaborative Learning* (CSCL 2015), 7–11 June 2015, Gothenburg, Sweden (pp. 459–466). International Society of the Learning Sciences.

Forestier, M., Stavrianou, A., Velcin, J., & Zighed, D. A. (2012). Roles in social networks: Methodologies and research issues. *Web Intelligence and Agent Systems, 10*(1), 117–133. http://dx.doi.org/10.3233/WIA-2012-0236

(2017). Role modelling in MOOC discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116. http://dx.doi.org/10.18608/jla.2017.41.6

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*(3), 75–174. http://dx.doi.org/10.1016/j.physrep.2009.11.002

Gillani, N., & Eynon, R. (2014). Communication patterns in massively open online courses. *The Internet and Higher Education, 23*, 18–26. http://dx.doi.org/10.1016/j.iheduc.2014.05.004

Gillani, N., Yasseri, T., Eynon, R., & Hjorth, I. (2014). Structural limitations of learning in a crowd: Communication vulnerability and information diffusion in MOOCs. *Scientific Reports, 4*, 6447. http://dx.doi.org/10.1038/srep06447

Han, L., Kashyap, A. L., Finin, T., Mayfield, J., & Weese, J. (2013). UMBC_EBIQUITY-CORE: Semantic textual similarity systems. *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics* (SEM 2013), 13–14 June 2013, Atlanta, GA, USA (pp. 44–52). Association for Computational Linguistics.

Harrer, A., & Schmidt, A. (2012). An approach for the blockmodeling in multi-relational networks. *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM 2012), 26–29 August 2010, Istanbul, Turkey (pp. 591–598). New York: ACM. http://dx.doi.org/10.1109/ASONAM.2012.100

Hecking, T., Hoppe, H. U., & Harrer, A. (2015). Uncovering the structure of knowledge exchange in a MOOC discussion forum. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM 2015), 25–28 August 2015, Paris, France (pp. 1614–1615). New York: ACM. http://dx.doi.org/10.1145/2808797.2809359

Huang, J., Dasgupta, A., Ghosh, A., Manning, J., & Sanders, M. (2014). Superposter behavior in MOOC forums. *Proceedings of the 1st ACM Conference on Learning @ Scale* (L@S 2014), 4–5 March 2014, Atlanta, Georgia, USA (pp. 117–126). New York: ACM. http://dx.doi.org/10.1145/2556325.2566249

Kim, S. N., Wang, L., & Baldwin, T. (2010). Tagging and linking web forum posts. *Proceedings of the 14th Conference on Computational Natural Language Learning* (CoNLL '10), 15–16 July 2010, Uppsala, Sweden (pp. 192–202). Stroudsburg, PA: Association for Computational Linguistics.

Liu, W., Kidzinski, L., & Dillenbourg, P. (2015). Semi-automatic annotation of MOOC forum posts. *Proceedings of the 2nd International Conference on Smart Learning Environments* (ICSLE 2015), 23–25 September 2015, Sinaia, Romania (pp. 339–408). Springer. http://dx.doi.org/10.1007/978-981-287-868-7_48

Lorrain, F., & White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology, 1*(1), 49–80. http://dx.doi.org/10.1080/0022250X.1971.9989788

Malzahn, N., Harrer, A., & Zeini, S. (2007). The fourth man: Supporting self-organizing group formation in learning communities. In C. Chinn, G. Erkens, & S. Puntambekar (Eds.), *Proceedings of the 7th International Conference on Computer-Supported Collaborative Learning* (CSCL 2007), 16–21 July 2007, New Brunswick, NJ, USA (pp. 547–550). International Society of the Learning Sciences.

McCallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research, 30*(1), 249–272. http://dx.doi.org/10.1613/jair.2229

Mount, D. W. (2004). *Bioinformatics: Sequence and genome analysis*, 2nd ed. Cold Spring Harbour, NY: Cold Spring Harbour Laboratory Press. http://dx.doi.org/10.1373/clinchem.2005.053850

Ó Duinn, P., & Bridge, D. (2014). Collective classification of posts to internet forums. *Case-Based Reasoning Research and Development, 8765*, 330–344. http://dx.doi.org/10.1007/978-3-319-11209-1_24

Onah, D. F. O., Sinclair, J., Boyatt, R., & Foss, J. G. K. (2014). Massive open online courses: Learner participation. *Proceedings of the 7th International Conference of Education, Research and Innovation* (iCERi2014), 17–19 November 2014, Seville, Spain (pp. 2348–2356). IATED Academy.

Rabbany, R., Takaffoli, M., & Zaïane, O. R. (2011). Analysing participation of students in online courses using social network analysis techniques. In M. Pechenizkiy et al. (Eds.), *Proceedings of the 4th Annual Conference on Educational Data Mining* (EDM2011), 6–8 July 2011, Eindhoven, Netherlands. International Educational Data Mining Society. http://educationaldatamining.org/EDM2011/wp-content/uploads/proc/edm2011_paper20_full_Rabbany.pdf

Rosé C., Goldman, P., Zoltners S. J., & Resnick, L. (2015). Supportive technologies for group discussion in MOOCs. *Current Issues in Emerging eLearning, 2*(1), 5.

Rossi, R. A., & Ahmed, N. K. (2015). Role discovery in networks. *IEEE Transactions on Knowledge and Data Engineering, 27*(4), 1112–1131. https://dx,doi.org/10.1109/TKDE.2014.2349913

Rossi, L. A., & Gnawali, O. (2014). Language independent analysis and classification of discussion threads in Coursera MOOC forums. *Proceedings of the 15th IEEE International Conference on Information Reuse and Integration* (IRI 2014), 13–15 August 2014, Redwood City, CA, USA (pp. 654–661). IEEE. http://dx.doi.org/10.1109/IRI.2014.7051952

Sharif, A., & Magrill, B. (2015). Discussion forums in MOOCs. *International Journal of Learning, Teaching and Educational Research, 12*(1).Retrieved from https://www.ijlter.org/index.php/ijlter/article/view/368

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating how students' cognitive behavior in MOOC discussion forums affect learning gains. In O. C. Santos et al. (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 226–233). International Educational Data Mining Society.

White, D. R., & Reitz, K. P. (1983). Graph and semigroup homomorphisms on networks of relations. *Social Networks, 5*(2), 193–234. http://dx.doi.org/10.1016/0378-8733(83)90025-4

Wise, A. F., Cui, Y., & Vytasek, J. (2016). Bringing order to chaos in MOOC discussion forums with content-related thread identification. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 188–197). New York: ACM. http://dx.doi.org/10.1145/2883851.2883916

Wong, J.-S., Pursel, B., Divinsky, A., & Jansen, B. J. (2015). An analysis of MOOC discussion forum interactions from the most active users. *Social Computing, Behavioral-Cultural Modeling, and Prediction, 9021*, 452–457. http://dx.doi.org/10.1007/978-3-319-16268-3_58

Yang, D., Wen, M., Kumar, A., Xing, E. P., & Rosé, C. (2014). Towards an integration of text and graph clustering methods as a lens for studying social interaction in MOOCs. *The International Review of Research in Open and Distributed Learning, 15*(5). http://dx.doi.org/10.19173/irrodl.v15i5.1853

Ziberna, A. (2013). Generalized blockmodeling of sparse networks. *Metodološki zvezki, 10*, 99–119.