Investigating Social and Semantic User Roles in MOOC Discussion Forums

Tobias Hecking
University of Duisburg-Essen
Lotharstraße 63/65
47048 Duisburg, Germany
hecking@collide.info

Irene-Angelica Chounta
University of Duisburg-Essen
Lotharstraße 63/65
47048 Duisburg, Germany
chounta@collide.info

H. Ulrich Hoppe
University of Duisburg-Essen
Lotharstraße 63/65
47048 Duisburg, Germany
hoppe@collide.info

ABSTRACT

This paper describes the analysis of the social and semantic structure of discussion forums in massive open online courses (MOOCs) in terms of information exchange and user roles. To that end, we analyse a network of forum users based on information-giving relations extracted from the forum data. Connection patterns that appear in the information exchange network of forum users are used to define specific user roles in a social context. Semantic roles are derived by identifying thematic areas in which an actor seeks for information (problem areas) and the areas of interest in which an actor provides information to others (expertise). The interplay of social and semantic roles is analysed using a socio-semantic blockmodelling approach. The results show that social and semantic roles are not strongly interdependent. This indicates that communication patterns and interests of users develop simultaneously only to a moderate extent. In addition to the case study, the methodological contribution is in combining traditional blockmodelling with semantic information to characterise participant roles.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human Factors; K.3.1 [**Computer Uses in Education**]: collaborative learning; H.4.2 [**Types of Systems**]: Decision support

General Terms

Algorithms, Measurement, Experimentation, Theory.

Keywords

Discussion Forums, MOOCs, Blockmodeling, Socio-semantic analysis.

1. INTRODUCTION

For online learning courses with no direct interaction between learners and tutors, discussion forums are commonly used as communication channels for information exchange between peers. This is especially the case with massive open online courses (MOOCs) where, in the absence of individual support by a tutor, threaded discussion forums are often the only means for information exchange and peer-to-peer-support provided by the MOOC platform.

The use of discussion forums in MOOCs has to be considered in a differentiated way. On one hand, only a small fraction of all participants in a MOOC use the forum to communicate [23] but forum activity often goes along with higher engagement in the course and completion rates [2, 10]. On the other hand, supported learner discussions in MOOCs have the potential of involving a large community in sustainable collaborative knowledge building in a social context (c.f. [12, 30]). In order to provide the necessary support, there is a strong need for a better understanding of the structure and function of the existing discussion forums. Further insights regarding information exchange in discussion forums in online courses can contribute to improvements with respect to the design and application of discussion forums or to the development of new types of communication channels for learners in online courses.

In this paper we aim to explore the characteristics of structured information exchange in a MOOC discussion forum. In particular, a forum community can be structured at least in two dimensions, namely, the social dimension and the semantic dimension. The social dimension is represented as a social network with relations between the users based on their communication – i.e. "who is talking to whom?". The semantic dimension reflects the semantic content the actors discuss in the discussion forum – i.e. "who is talking about what?".

Most of existing research on discussion forums in the learning context focuses on either the social or the semantic dimension. However, in order to get a more complete picture of the community based on discussion forums, a combined analysis of both dimensions is necessary. This requires mixed techniques from social network analysis and content analysis. To that end, we investigate the discussion forum of a MOOC with respect to the social and semantic structure of information exchange and user roles. The course was named "Introduction to Corporate Finance", it took place over a six-week period (11/2013 to 12/2013) and was offered at the Coursera[1] platform.

As a first step, information-seeking and corresponding information-giving posts are identified using automatic post classification. The classified posts are used to model a directed network of forum users and the information-giving relations between them. In addition, on the semantic level a user is represented by the thematic areas of interest in which the actor

LAK '16, April 25 - 29, 2016, Edinburgh, United Kingdom

ACM 978-1-4503-4190-5/16/04$15.00
DOI: http://dx.doi.org/10.1145/2883851.2883924

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

seeks for information (problem areas) and the areas of interest in which the actor provides information to others (expertise). While a social network of forum users explicitly models person-to-person relations, on the semantic level similar interests of two actors do not necessarily imply a social relation. Blockmodelling as an existing technique [9] for role modelling of users based on connection patterns to other users in a social network is extended by incorporating the semantic models of the users based on their information-seeking and information-giving interests. We call this approach socio-semantic blockmodelling. A role can be interpreted as a group of participants with similar connection patterns in a thematic context. Inferred relations between those roles give insights to the main structure of information exchange between users of different roles in the MOOC discussion forum.

In particular the proposed approach is used to answer the following research questions:

- To what extent does the community of actors in the discussion forum exhibits a social role structure discoverable by blockmodels? To what extent is the community structured in semantically coherent subgroups or sub-communities of interests?
- To what extent are the social and semantic structure interdependent?
- Can socio-semantic blockmodelling reveal the basic structure of the forum communication in a meaningful way?

The paper is structured as follows: After this introduction Section 2 gives an overview on related work and current developments in the research on discussion forums in MOOCs. Section 3 describes the extraction of an information exchange network from Coursera forum data. An introduction to blockmodelling and a detailed description of the proposed extension is given in Section 4. The main findings are summarised in Section 5 and further discussed in Section 5, which concludes the paper.

2. BACKGROUND AND RELATED WORK

Discussion forums have been widely used in MOOCs to facilitate communication between participants and to scaffold collaboration. The collection of posts in a MOOC discussion forums accounts for the information flow between learners from various knowledge backgrounds and is an indicator for collaborative knowledge building between learners with diverse knowledge backgrounds [33]. Related research has shown that users engagement in MOOC discussion forum tend to differ. While many users are not active at all or use the forum purpose-specific (i.e. participants use forums either to ask for assignments' solutions or to get rapid and trustworthy response to specific questions) [28], MOOC forums are usually dominated by few, highly active users, who can influence other participants and stimulate and sustain the discussions [21, 35]. This diverse behaviour results in different roles of users that can be described in various aspects using different analysis techniques.

Techniques used for the analysis of MOOC discussion forums can be characterised as content related or communication related. Content related analyses aim to uncover the nature of forum contributions from the post content [31]. Cui and Wise [7] apply content analysis and machine learning to identify forum threads where participants discuss the course content which is important for the investigation of information exchange. Content analysis can also be used to characterise forum users based on the types of contributions they made [3, 24]. For communication related analyses, often social network analysis techniques are applied. Social networks between users based on common discussion threads can serve to investigate the coherence of the underlying social network [15], detection of communication patterns [14] as well as community support [26]. However, when it comes to role modelling based on social relations in discussion forums finer grained network modelling is required such that the concrete post/ reply communication between individuals are adequately reflected. In discussion forums with nested threads, these relations can be observed directly from the thread structure [29]. 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 appraoches such as discussion act tagging [3]. User roles can then be inferred based on communication behaviour which is reflected in the position of an actor in the social network. There are different possibilities to define user roles in social communication networks. Abnar et al. [1] use centrality measures in subcommunities to identify roles such as leaders and mediators in a forum communication network. In [20] users are characterised with respect to the number of help-giving and help-seeking posts giving higher values to those users who reach more others with their posts to those who repeatedly target a small set of communication partners.

In the work described in this paper, techniques of network and content analysis are combined to characterise 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). We have found a somewhat similar approach in the work of Yang et al. [36] 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. assume that there is an interplay between users' interests and social relations that is inherently encoded in the model. In our work, however, 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. Also, using block modelling approach, we do not assume that users with the same role have to form a cohesive subcommunity.

3. NETWORK EXTRACTION FROM FORUM DATA

The dataset comprises forum posts from the Coursera MOOC on "Introduction to Cooperate Finance" conducted in six weeks between 11/2013 and 12/2013. Overall there were 8336 posts in 870 different threads by 1540 different users. 1436 posts were made by anonymous users. Many of the discussion threads are 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 for the analysis 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 starting point of the analysis is the set of forum threads of the discussion forum. These threads contain a sequence of posts where for each post the unique identifier, post content, and the author's identity is available. 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. Thus, the initial task is to extract this network from the raw forum thread data and can be structured in three successive steps, namely (1) post classification, (2) post linking, (3) transformation to a social knowledge exchange network. An example of the procedure for the example discussion thread in Table 1 is shown in Figure 1.

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**

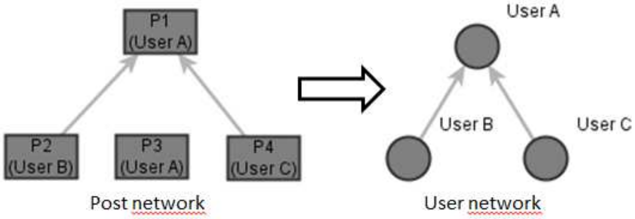

Figure 1: **Basic scheme of network extraction from forum posts.**

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 that can be considered as information-giving or information-seeking. In previous studies different types of posts in MOOC discussion forums are described [3, 16, 22, 24]. In this study the classification schemes for MOOC discussion forums described by Arguello and Shaffer [3] and the similar classification of Liu, Kidzinski and Dillenbourg [24] are generalised 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 this task an automated classification model was trained on 500 posts that were hand-classified by three experts. The validity of the classification scheme was ensured with a high interrater agreement among all three raters according to Fleiss-Kappa ($\kappa = .78, p < .005$).

The organisation of the course forum in sub forums is used to filter the dataset prior to the automatic post classification. In previous work [20], 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 using since the sub forum is a good predictor for those posts. In this study, however, information on the sub forum in which a discussion thread occurs is used to restrict the analysis to sub forums that are 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 [5] when bagging with 10 iterations was applied. Information seeking posts can be classified with high F1-score = 0.77. For information giving posts the F1-score is also moderately high F1-score = 0.66. However, posts of type “other” often lead to misclassifications as the confusion matrix in Table 2 shows.

Table 2: Confusion matrix for post classification.

	True inf. seeking	**True inf. giving**	**True other**	**Class precision**
Pred. inf. seeking	75	16	2	0.81
Pred. inf. giving	27	88	30	0.61
pred. others	0	17	36	0.68
class recall	0.74	0.73	0.53	

In order to reduce the effect of misclassified other posts, the final classification was improved using the iterative classification algorithm described by Duinn and Bridge [27]. This algorithm uses the results from the classifier described above to compute for each post the number of preceding posts of each class. Then an additional classifier is trained to incorporate this information updates the initial classification, which leads to improved results since misclassifications such as classification as information giving posts without a preceding information-seeking post can be avoided. This increases F1-scores for information-giving posts increases to 0.79 and for information-seeking posts to 0.71 based on evaluation on another 200 hand-classified posts.

3.2 Network Extraction

Based on the classified posts, we initialise the network of information seeking and related information giving posts.

As a first step, we remove 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. These users contributed 4096 posts in 502 threads that spread over 27 of the 40 sub forums. Out of these, 1523 posts were classified as “questions”, 1832 posts were classified as “answers” resulting in 1303 links between the users. 741 posts were classified as “other”, and thus, not reflected in the edges of the resulting network. On average, each user in the network made 4.34 posts (SD=7.246) in 2.61 threads (SD=3.608) over 1.71 forums (SD=1.281). From these posts, on average per user, 1.61 were classified as “questions” (SD=2.740), 1.94 were classified as “answers” (SD=4.042) and 0.78 were classified as “other” (SD=1.603). As it is shown from these distributions, users have a limited activity in the discussion forum throughout the course and they do not get involved or spread over many threads and forums.

In the following, all posts classified as “other” are filtered out from each thread such that only the “information seeking” and “information giving” posts remain. As an intermediate step before the social network between users can be created, a network of posts has to be built (see Figure 1). The basis for this is the observation that the users in Coursera discussion forum 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 further questions follow until an information giving post occurs in the sequence. After a sequence of information giving post 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 posts occurred in between. Sequences of comments attached to a parent posts can be seen as sub-threads that can contain both types of posts with the parent post as initial post. Consequently, a forum thread and the corresponding sub-threads based on comments 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. Next, each post node, labelled with the same author, is collapsed into a single node representing the user (Figure 1) resulting in the final knowledge exchange network between forum users, similar to the approach described in [19].

4. APPROACH: SOCIO-SEMANTIC BLOCKMODELLING

Blockmodelling [9] is a method to reduce a network to a macro structure by grouping actors groups based on their connection patterns and modelling relations between them. Those groups are commonly interpreted as roles or positions since it is assumed that similar connection patterns indicate the similar function. Figure 2 gives an example of a blockmodel with three roles and relations between them that reflects the hierarchical structure of the network. In this section the existing techniques for blockmodelling based on similarities of connection patterns of users are described first. 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 and do not necessarily have to be cohesive, in the sense that actors of the same role are densely interconnected among themselves. A blockmodel fitted to the network structure can be used to infer relations between those groups of actors. In generalized blockmodelling approach [9] 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. The most important types of relations for this work are depicted in Figure 3.

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 flow 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 as null relation.

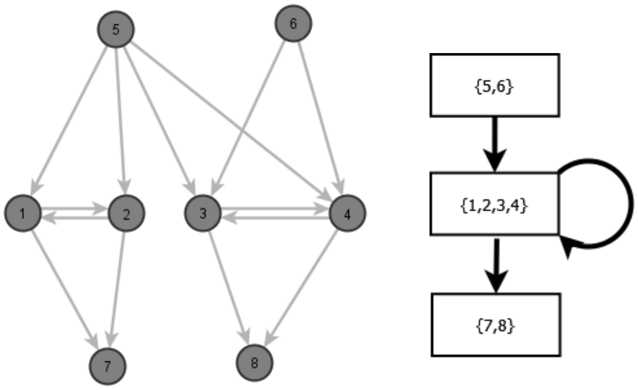

Figure 2: **Example network with regular and structural equivalences.**

It is important to note that in forum networks there is often no perfect fit of the relations between groups of users to 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. However, it can also not be considered as null-relation as it is defined. In cases were none of the described relations are applicable, the relation is chosen that can be applied with minimal modifications of the links between the actors in A and B. The total number of such modifications is referred as the blockmodel error.

An important fact that is often ignored is that blockmodelling can clearly be distinguished from the more common sub-community detection [13] in social network analysis. Even though, both, blockmodelling and sub-community detection group users in to clusters the objectives of these methods are quite different. Community detection methods aim to find densely connected substructures in the network by finding a clustering such that the number of connections within the cluster exceeds the number of connections between actors of different clusters as much as possible. Blockmodelling does not require any connections between actors of the same cluster at all, although they are not forbidden (see group B in Figure 2). Moreover, in a blockmodel users belong to the same group since they have similar connection patterns to users in other groups. Thus, a cluster can be interpreted as users with 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.

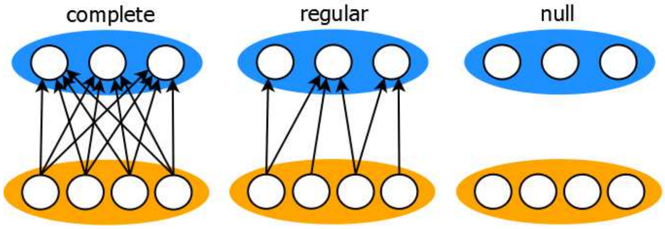

Figure 3: Relation types between two groups of actors.

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 to roles/positions such that the relations described before between groups of actors are inherently induced by the grouping of the actors. Graph-based similarity measures that are commonly applied for blockmodelling are structural and regular similarity.

4.2.1 Structural Similarity

Structural similarity [25] is related to the position of the actors within the network. Structural similarity can be assesses by 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 structural 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 structural 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 favor of regular similarity described next.

4.2.2 Regular Similarity

In contrast to structural similarity, regular similarity [34] between two actors 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. Thus, actors with a high regular similarity are considered to have the same role in the network. The problem then becomes assigning roles to actors such that actors within the same role are as similar as possible with respect to the roles of the actors they are connected to. If there is an assignment of actors to roles such that actors within a role are regular equivalent (maximum regular similarity), the fitted blockmodel has only regular and null blocks without any errors. For example in Figure 2 a perfect 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 [4] is applied.

4.3 Semantic Similarity

In contrast to graph based similarity described before, 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 blockmodels [32]. 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 4). 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.

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

The term $sim\left(t_{x,i}, T_y^P\right)$ corresponds to the similarities between the ith 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 $sim_{sem}^{S}\left(u_x, u_y\right)$ 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.

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

The distinction between information giving and information seeking interests is crucial for role 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 would not be made, the resulting blockmodel is likely to contain mostly relations from a certain role to the role itself and would hardly allow for a distinction between social and semantic roles since communication in one thematic area implies corresponding connections in the information exchange network.

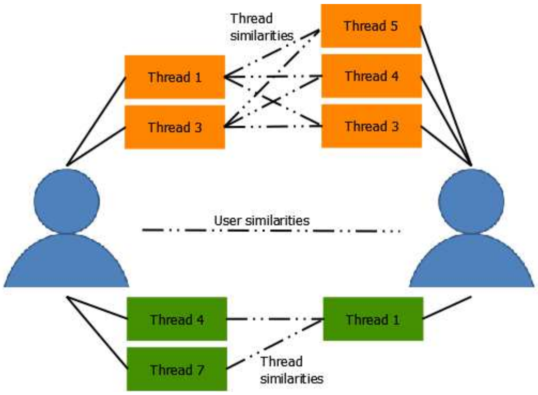

Figure 4: Semantic similarity of two users based on the similarity of threads in which they provide information (orange) and seek for information (green).

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. Then, one possibility would be to calculate their semantic similarities based on latent semantic indexing (LSI) [8], which is 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, which is 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 [17], which combines latent semantics 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[2]. It extracts concepts from textual documents by comparing the documents to to Wikipedia pages. This has several benefits compared to other approaches for keyword extraction. First, the concepts do not have to be exactly mentioned 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 inherent resolution of synonyms, polysemy, and disambiguation. This solves also the problem of short text and inexact language which is common in discussion forums. Additionally, this approach 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 Approach

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 is, given an allocation of users to roles, to identify regular relations between semantic coherent (but not necessarily socially coherent) roles in the knowledge exchange network extracted for 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 of at least one user in A (c.f. Section 4.2.2).

Semantic similarity, as described in Section 4.3, identifies semantic 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 of a role with many ingoing connections (information consumers). A semantic role can characterise users who have problems with topic X" or who have an expertise on topic Y. The combination of both can then be seen as a social role in 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 that is 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 then 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 a 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, can be subdivided into different types based on their connection patterns in the network discovered based on regular similarity. Combining user features with network structure [32], and finding the optimal blockmodel with respect to multiple objectives by optimizing role allocations is a hard problem [6, 18]. 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 5.

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

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 [11], a method that estimates the optimal number of clusters given distances/similarities of objects and a clustering function by minimising cluster instability.
3. Assign the role relations such that the blockmodel error is minimal described in Section 4.1.

The sparsity of the network is a problem since it biases the inference of relations 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 [37].

5. RESULTS

Regular similarity inherently assigns users to roles such that the relations between the roles are either (almost) regular or (almost) null/non-existent relations. The questions we aim to answer in the following Section 5.1 is to what extent the social and semantic structure of the community is interleaved. More concretely, how well does role assignment based on semantic similarity induces a blockmodel that has a small error according to regular relations between roles and whether roles deriving from regular similarity are also semantically coherent. In Section 5.2 the community in the MOOC discussion forum is analysed using the hybrid blockmodelling approach introduced in Section 4.4.

5.1 Semantic vs. Social Structuring

In the following, 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 is investigated.

Table 3: Correlations between different types of similarities

	structural	**regular**	**semantic**
structural	1	-0.19	-0.16
regular	-0.19	1	0.36
semantic	-0.16	0.36	1

First, we conducted a correlation analysis between the graph-based (social) similarities described in Section 4.2 and semantic similarity of users (Section 4.3). If social and semantic structure

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

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 3. All correlations are statistically significant (p << .05). There is a low 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 Cooperate 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 a different emphasis of regular (social) and semantic similarity by varying the paramters σ_{reg} and σ_{sem} (equation 3). For each blockmodel the normalized blockmodel error (*bm_err*) 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*.

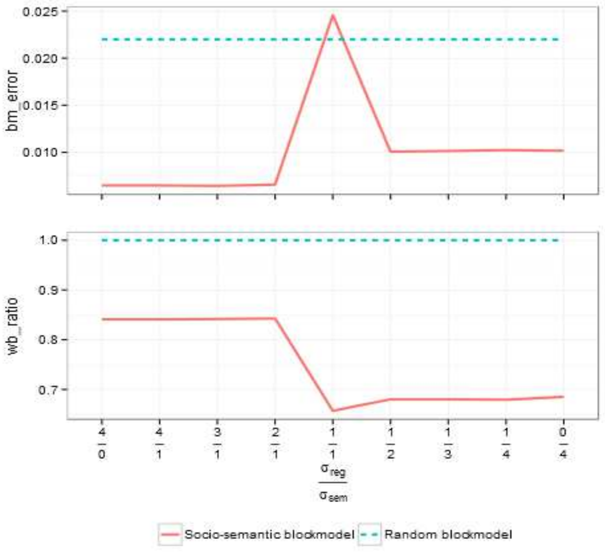

Figure 5: Blockmodel error (top) and ratio of average semantic distance within roles and between roles (bottom) for different ratios of σ_{reg} and σ_{sem}.

The results are presented in Figure 5. For both cases, *bw_ratio* and *bm_err*, there is a state transition between role assignments that emphasize more on social similarity and role assignments that emphasize on 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 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. The proposed hybrid blockmodelling approach described in Section 4.4 can be applied to map the information flow between different socio-semantic roles, as be described in Section 5.2.

5.2 Socio-semantic Blockmodelling

In the following, the socio-semantic structure of the forum communication is analysed based on a hybrid blockmodel. For our analysis we take into account the semantic coherence of roles as well as the blockmodel error in terms of regular relations. In order to do this, first a good level of emphasis of social and regular similarity according to equation 3 has to be found. Figure 6 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 $\sigma_{reg}=1$ and $\sigma_{sem}=2$.

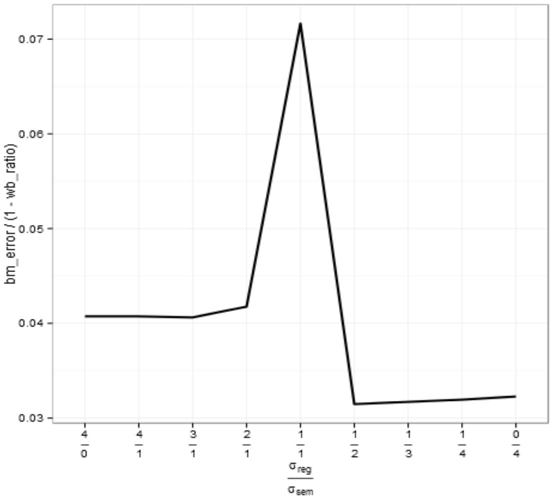

Figure 6: Ratio between blockmodel error and semantic coherence of the roles.

The resulting blockmodel is depicted in Figure 7. 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.

It is shown that there is one dominant role (role 1) comprising 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 there is a smaller set of users who can be characterised 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 4). As shown in [20], in- and outreach combine the post quantity of a user with the number of connections the user has to others. Users with a high inreach post many information- seeking posts and receive information from many different other users. Outreach is defined similar for outgoing information- giving relations. Thus, inreach corresponds to information- seeking behaviour and high outreach to information-giving behaviour. 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. Roles 2 and 3 can then be seen as users who are more specialised 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 4. For role 1, there is no clear semantic distinction between information giving and seeking interests which is also reflected by the self-loop in the blockmodel (Figure 7). The concept “Mathematical finance” is associated to every role since it is an important general concept that has been 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 characterise the particular roles.

The most frequent other concepts that were extracted from forum threads in which users of role 1 appear as information seekers and givers are “Investment”, “Depreciation”, and “Taxation”. These are 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.

The information seeking role 2 has no key concepts assigned to their information giving interests. These users seek for information especially in areas related to investments. They receive help form users in role 1 and role 3 on this topic. The nature of role 2 is further underlined by the fact that many of the threads they are active in are additionally annotated with the keyword “question”.

Role 3 can be interpreted as experts for the topics related to investment appraisal. However, despite from being a relatively small role in terms of number of users and the mean outreach is moderately high, users in this role could either be the ones who provide some information to a course topic they are good in and then stop participating in the forum or show a kind of “elder statesman” behaviour in the sense that they occasionally contribute to the information exchange in the forum as experts in topics that are of wide interest for the whole community.

In general, it can be said that three discovered socio-semantic roles structure reflects the general assumptions on MOOC discussion forums very well. There is a core community (role 1) that 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 (role 2 and 3) correspond to the users who participate in the forum communication occasionally and are either information givers or information seekers on certain topics.

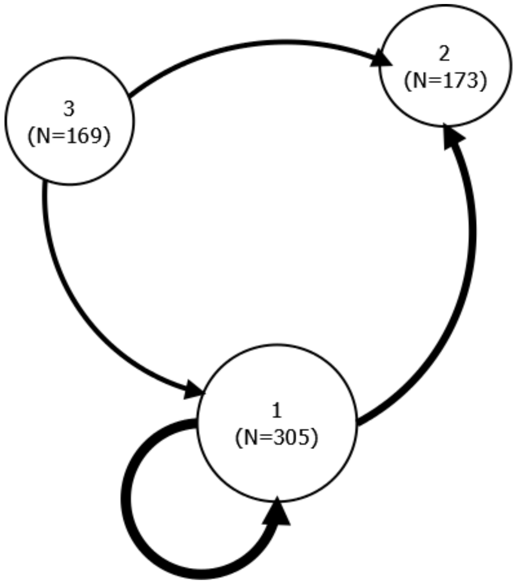

Figure 7: Blockmodel for the forum discussion in the MOOC discussion forum.

Table 4: Properties of the discovered roles.

Role	**Top inform. giving**	**Top inform. seeking**	**Mean in-reach**	**Mean out-reach**
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

6. CONCLUSION

We have analysed the social and semantic structure of a community of learners participating in a MOOC discussion forum with respect to user roles in social and semantic context. In the social dimension users were assigned to roles based on their regular similarity in the information exchange network of forum users. In the semantic dimension, roles were modelled based on the thematic areas in which users were active in by providing or seeking information. Those semantic roles can be also interpreted as expertise and information seeking for specific themes respectively.

We applied our approach on the dataset of a discussion forum that supported the online Coursera course "Introduction to Corporate Finance". Our research objective was three-fold: a) to define to what extent the forum communities of users are socially and semantically structured, b) to study to what extent the social and semantic structures interdependent and c) to explore whether socio-semantic blockmodelling can reveal meaningful information about the forum communication structures.

The results of our study showed that both social and semantic role structures are present in the discussion forum of the course (Section 5.1). The semantic coherence of user roles with respect to the semantic similarity of the 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 the users in terms their connection patterns in the information exchange network. Consequently, the community in the discussion forum did not evolve completely random as it might be suggested by the known differences of behaviour and engagement of participants in MOOC discussion forums.

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

For our dataset, three different roles were discovered based on the hybrid social-semantic blockmodelling approach. There was a majority of users who discuss the main course content. While for the other roles there was only occasional information exchange between users within the role, users of this majority role also had heavy communication with each other. Apart from that there were also users in the two smaller roles who could either be considered as users who contributed less to the forum communication where one of the roles contained more information providing users on specific course topics and the other comprise users who only seek for information on very concrete issues.

All these findings suggest that there is 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. 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 to find proper communication partners for certain thematic areas. This might enhance the engagement of learners in sustainable knowledge building dialogues and information exchange in the discussion forum.

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. Open issues are still to find out which external factors drive the evolution of the community and the emergence of different user roles and how users of different roles are engaged in other activities of the online course. Therefore, on the methodological level, it can be interesting to incorporate also user similarity based on resource access or engagement patterns into the role modelling.

7. REFERENCES

[1] Abnar, A., Takaffoli, M., Rabbany, R. and Zaïane, O. SSRM: structural social role mining for dynamic social networks. *Social Network Analysis and Mining*, 5, 1 (2015).

[2] Anderson, A., Huttenlocher, D., Kleinberg, J. and Leskovec, J. Engaging with Massive Online Courses. In *Proceedings of the 23rd International Conference on World Wide Web.* (Seoul, Korea), 2014, 687-698.

[3] Arguello, J. and Shaffer, K. Predicting Speech Acts in MOOC Forum Posts. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media.* (Oxford, UK), 2015.

[4] Borgatti, S. P. and Everett, M. G. Two algorithms for computing regular equivalence. *Social Networks*, 15, 4 (1993), 361-376.

[5] Breiman, L. Random Forests. *Machine Learning*, 45, 1 (2001), 5-32.

[6] Brusco, M., Doreian, P., Steinley, D. and Satornino, C. Multiobjective Blockmodeling for Social Network Analysis. *Psychometrika*, 78, 3 (2013), 498-525.

[7] Cui, Y. and Wise, A. F. Identifying Content-Related Threads in MOOC Discussion Forums. In *Proceedings of the Second ACM Conference on Learning @ Scale.* (Vancouver, BC, Canada). ACM, New York, NY, USA, 2015, 299-303.

[8] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), (1990) 391-407.

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

[10] Engle, D., Mankoff, C. and Carbrey, J. Coursera's introductory human physiology course: Factors that characterize successful completion of a MOOC. *The International Review of Research in Open and Distributed Learning,* 16, 2 (2015), 46-68.

[11] Fang, Y. and Wang, J. Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, 56, 3 (2012), 468-477.

[12] Ferschke, O., Howley, I., Tomar, G., Yang, D. and Ros\'e CP. Fostering Discussion across Communication Media in Massive Open Online Courses. In *Proceedings of the 11th International Conference on Computer Supported Collaborative Learning. (Gothenburgh, Sweden*), 2015, 459-466.

[13] Fortunato, S. Community detection in graphs. *Physics Reports*, 486, 3 (2010), 75-174.

[14] Gillani, N. and Eynon, R. Communication patterns in massively open online courses. *The Internet and Higher Education*, 23, 10 (2014), 18-26.

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

[16] Glenda S. Stump, Jennifer DeBoer, Jonathan Whittinghill, Lori Breslow. Development of a Framework to Classify MOOC Discussion Forum Posts: Methodology and Challenges. Available online: https://tll.mit.edu/sites/default/files/library/Coding_a_MOOC_Discussion_Forum.pdf. 02/04/2015.

[17] Han, L., Kashyap, A. L., Finin, T., Mayfield, J. and Weese, J. UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics,* Association for Computational Linguistics, 2013.

[18] Harrer, A. and Schmidt, A. An Approach for the Blockmodeling in Multi-Relational Networks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, (Istanbul, Turkey) *IEEE,* 2012, 591-598.

[19] Harrer, A., Zeini, S. and Sabrina Ziebarth. Visualisation of the Dynamics for Longitudinal Analysis of Computer-mediated Social Networks - Concept and Exemplary Cases. In *From Sociology to Computing in Social Networks. Theory, Foundations and Applications.* Springer, Vienna, 2010.

[20] Hecking, T,, Harrer, A., Hoppe, H.U. Uncovering the Structure of Knowledge Exchange in a MOOC Discussion Forum. In Anonymous *Proceedings of the International Conference of Advances in Social Network Analysis and Mining.* (Paris, France). IEEE, 2015, in press.

[21] Huang, J., Dasgupta, A., Ghosh, A., Manning, J. and Sanders, M. Superposter Behavior in MOOC Forums. In *Proceedings of the First ACM Conference on Learning @ Scale Conference.* (Atlanta, Georgia, USA). ACM, New York, NY, USA 2014, 117-126.

[22] Kim, S. N., Wang, L. and Baldwin, T. Tagging and Linking Web Forum Posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning.* (Uppsala, Sweden). Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, 192-202.

[23] Kizilcec, R. F., Schneider, E., Cohen, G. L. and McFarland, D. A. Encouraging Forum Participation in Online Courses with Collectivist, Individualist and Neutral Motivational Framings. *Proceedings of the European MOOCs Stakeholder Summit*, (Lausanne, Swizerland), 2014.

[24] Liu, W., Kidzinski, L. and Dillenbourg, P. Semi-automatic annotation of MOOC forum posts. In *Proceedings of the 2nd International Conference on Smart Learning Environments.* (Sinaia, Romania), 2015.

[25] Lorrain, F. and White, H. C. Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1, 1 (1971), 49-80.

[26] Malzahn, N., Harrer, A. and Zeini, S. The Fourth Man - Supporting self-organizing group formation in learning communities. In *Proceedings of the Computer Supported Collaborative Learning Conference 2007.* (New Brunswick, NJ, USA). ICLS, 2007, 547-550.

[27] Ó Duinn, P. and Bridge, D. Collective Classification of Posts to Internet Forums. In *Case-Based Reasoning Research and Development* LNCS 8765 (2014), 330-344.

[28] Onah, D. F., Sinclair, J., Boyatt, R. and Foss, J. G. Massive open online courses: learner participation. In *Proceeding of the 7th International Conference of Education, Research and Innovation.* (Seville, Spain). IATED Academy, 2014, 2348-2356.

[29] Rabbany, R., Takaffoli, M. and Zaiane, O. R. Analyzing participation of students in online courses using social network analysis techniques. In *Proceedings of educational data mining.* (Eindhoven, The Netherlands), 2011, 21-30.

[30] Rosé Carolyn P, Goldman, P., Zoltners Sherer, J. and Resnick, L. Supportive technologies for group discussion in MOOCs. *Current Issues in Emerging eLearning*, 2, 1 (2015), 5.

[31] Rossi, L. A. and Gnawali, O. Language independent analysis and classification of discussion threads in Coursera MOOC forums. In *Proceedings of the 15th International Conference on Information Reuse and Integration,* (Redwood City, CA, USA)*,* 2014, 654-661.

[32] Rossi, R. A. and Ahmed, N. K. Role Discovery in Networks. *CoRR*, abs/1405.7134 (2014).

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

[34] White, D. R. and Reitz, K. P. Graph and semigroup homomorphisms on networks of relations. *Social Networks*, 5, 2 (1983), 193-234.

[35] Wong, J., Pursel, B., Divinsky, A. and Jansen, B. An Analysis of MOOC Discussion Forum Interactions from the Most Active Users. In *Social Computing, Behavioral-Cultural Modeling, and Prediction* LNCS 9021, (2015), 452-457.

[36] Yang, D., Wen, M., Kumar, A., Xing, E. P. and Rose, C. P. 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 (2014).

[37] Ziberna, A. Generalized blockmodeling of sparse networks. Metodolozkizvezki, 10, (2013), 99-119.