

Characterising Learners in Online Communities Based on Actor-Artefact Relations

Von der Fakultät für Ingenieurwissenschaften,
Abteilung Informatik und Angewandte Kognitionswissenschaft
der Universität Duisburg-Essen

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte kumulative Dissertation

von

Tobias Ludger Hecking
aus
Arnsberg

1. Gutachter: Prof. Dr. Heinz Ulrich Hoppe
2. Gutachter: Prof. Dr. Pierre Dillenbourg

Datum der mündlichen Prüfung: 08.12.2016

Abstract

Online communities are of huge interest in terms of learning and knowledge creation because of the potential to distribute knowledge among possibly large audience independently from time and place. In this context, various forms of online learning have developed over time ranging from small learning groups to massive open online courses (MOOCs) with thousands of participants.

In order to support learning in those settings an increased understanding of specific characteristics of learners in online communities is necessary. Thus, dedicated means to gather valuable information from data produced in online learning environments have to be developed. This cumulative dissertation includes five publications aiming to make progress in this direction with a particular focus on the advancement of methods to analyse activity and interaction data of learners.

The methodological foundation of the work is (social) network analysis, which provides a well-grounded set of methods for structural analysis of relational data. Network analysis is especially suited since the collected data about actors (in this thesis mostly learners) who create and consume digital content (artefacts) can be modelled as actor-artefact networks. Those actor-artefact networks denote the starting point of all analyses presented in this dissertation, which target different aspects of learning in online communities, in particular the usage of learning resources, emergence of interest profiles, and information exchange between learners.

In the course of this work, stable artefacts that are not assumed to have changing content over time are distinguished from time-evolving dynamic artefacts (typically user generated content). In the case of stable artefacts, affiliations of learners to learning resources in online courses are analysed by identifying mixed clusters of learners and resources using network clustering algorithms. The evolution of these learner-resource clusters over time is investigated in detail leading to discoveries of typical resource access patterns that characterise learners regarding their interests in provided learning materials. The approach is further extended and combined with content analysis techniques to analyse thematic development in discussion forums.

Discussion forums are also the subject of two other studies investigating information exchange between learners in MOOCs. The evolving discussion threads are considered as dynamically evolving artefacts that are used to extract social networks reflecting information exchange between forum users. These networks are analysed to uncover different roles of forum users with respect to their positions in the network. For this task different approaches are described that are capable of modelling structural characteristics of the information exchange network over time and further take discussion topics as additional information into account.

List of Included Publications

Chapter 2

Hecking, T., Ziebarth S., & Hoppe H. U. (2014). Analysis of Dynamic Resource Access Patterns in Online Courses. *Journal of Learning Analytics (JLA)*. 1(3), 34-60.

Chapter 3

Hecking, T., Steinert L., Göhnert T., & Hoppe H. U. (2014). Incremental Clustering of Dynamic Bipartite Networks. In *Proceedings of the 1st European Network Intelligence Conference* (pp. 9-16), Wroclaw, Poland, IEEE. doi: 10.1109/ENIC.2014.15

Chapter 4

Hecking, T., Chounta I. - A., & Hoppe H. U. (2015). Analysis of User Roles and the Emergence of Themes in Discussion Forums. In *Proceedings of the 2nd European Network Intelligence Conference* (pp. 114-121), Karlskrona, Sweden, IEEE. doi: 10.1109/ENIC.2015.24

Chapter 5

Hecking, T., Hoppe H. U., Harrer, A. (2016) Discovery of Structural and Temporal Patterns in MOOC Discussion Forums. Preprint of chapter in J, Kawash, N. Agarwal, T. Özyer (Eds.) *Prediction and Inference from Social Networks and Social Media* (pp. 153-180), LNSN, Springer. doi: 10.1007/978-3-319-51049-1_8

Chapter 6

Hecking, T., Chounta I. - A., & Hoppe H. U. (2016). Investigating Social and Semantic User Roles in MOOC Discussion Forums. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge* (pp. 198–207), Edinburgh, UK, ACM. doi: 10.1145/2883851.2883924

Auxiliary Works

The publications listed below describe works that are relevant in the context of this thesis, but these are not included as dedicated chapters.

Hecking, T., Göhnert, T., Zeini, S., & Hoppe, H. U. (2013). Task and Time Aware Community Detection in Dynamically Evolving Social Networks. *Procedia Computer Science*, 18, 2066-2075.

Zeini, S., Göhnert, T., Hecking, T., Krempel, L., & Hoppe, H. U. (2014). The Impact of Measurement Time on Subgroup Detection in Online Communities. In F. Can, T. Özyer, & F. Polat (Eds.), *State of the Art Applications of Social Network Analysis* (pp. 249-268). LNSN, Springer, Cham.

Göhnert, T., Harrer, A., Hecking, T., & Hoppe, H. U. (2014). A Workbench for Visual Design of Executable and Re-usable Network Analysis Workflows. In J. Kawash (Eds.), *Online Social Media Analysis and Visualization* (pp. 181-199). LNSN, Springer, Cham.

Halatchliyski, I., Hecking, T., Goehner, T., & Hoppe, H. U. (2014). Analyzing the Path of Ideas and Activity of Contributors in an Open Learning Community. *Journal of Learning Analytics, JLA*, 1(2), 72-93.

Ziebarth, S., Neubaum, G., Kyewski, E., Krämer, N., Hoppe, H. U., Hecking, T., & Eimler, S. (2015). Resource Usage in Online Courses: Analyzing Learner's Active and Passive Participation Patterns. In *11th International Conference on Computer Supported Collaborative Learning*. (pp. 395-402) Gothenburg, Sweden. International Society of the Learning Sciences (ISLS).

Wichmann, A., Hecking, T., Elson, M., Christmann, N., Herrmann, T., & Hoppe, H. U. (2016). Group Formation for Small-Group Learning: Are Heterogeneous Groups More Productive? In *Proceedings of the International Symposium on Open Collaboration (OpenSym '16)*. Berlin, Germany.

Contents

Abstract.....	iii
List of Included Publications	iv
1 Introduction	1
1.1 Context and Research Objectives.....	2
1.1.1 Actors, Communities, and Artefacts	2
1.1.2 Characterising learners in online communities	6
1.1.3 Exploration of analytics methods	10
1.2 Synopsis of Included Publications	15
1.2.1 Relations of actors and stable artefacts	16
1.2.2 Social relations between actors and dynamic artefacts	20
1.2.3 Combination of semantic and social structures	22
2 Analysis of Dynamic Resource Access Patterns in Online Courses.....	25
3 Incremental Clustering of Dynamic Bipartite Networks.....	53
4 Analysis of User Roles and the Emergence of Themes in Discussion Forums.....	63
5 Discovery of Structural and Temporal Patterns in MOOC Discussion Forums.....	73
6 Investigating Social and Semantic User Roles in MOOC Discussion Forums.....	103
7 Summary and Future Perspectives.....	115
7.1 Summary.....	115
7.2 Future Perspectives	116
7.2.1 Advancing online learning environments	116
7.2.2 Widening the scope of applications.....	117
References	121

1 Introduction

With growing scalability of information technology in education and knowledge management, learning and knowledge acquisition have undergone a shift from physical into virtual environments. Various forms of knowledge building in large and heterogeneous online communities have emerged that rely on information creation and distribution in digital spaces. In addition to self-directed search in ubiquitously available information sources, online courses offer the opportunity to get a more guided access to knowledge provided and maintained by experts. In this context, a recent development are massive open online courses (MOOCs), which are university level courses that solely take place in online environments and target a broad audience of learners. Presence lectures are replaced by videos and digital reading material. Activities such as quizzes and peer-reviewing are used to assess the learning outcome of a possibly large amount of users. Apart from the extreme case of MOOCs, different types of online courses have been developed also on smaller scale focussing on different aspects of online learning. Some regular university courses have been enriched by additional online activities inspired by MOOCs (Volk, Reinhardt, & Osterwalder, 2014; Ziebarth & Hoppe, 2014). In these settings, the main process of knowledge acquisition can take place in an online learning environment accompanied by presence classes that can be used for discussions and assistance. These are typically referred to as “Flipped Classroom” models (Bishop & Verleger, 2013).

Common to all these forms of learning environments is the use of rich multimedia content accessible independently of time and place. The intention is to enable learners to acquire knowledge in a self-directed manner utilising various types of learning resources and tools. Despite the chances large scale and open online learning environments offer for learning and distribution of knowledge there are also several challenges. The openness of online platforms goes along with the need for adaptation to a heterogeneous and possibly large population of users in terms of personal backgrounds, interests, and goals. In the absence of immediate social interaction in a physical space, support for individuals is limited. Typical problems of collaborative learning in physical learning spaces such as motivational deficits, lack of individual accountability of learners, time management, and coordination issues have to be addressed differently in digital environments since interaction with learners is usually mediated and limited by technology.

Adaptation and personalisation of those learning environments is considered as a key concept to improve the learning experience (Brusilovsky & Millán, 2007). In order to achieve this, a better understanding of the specific characteristics of individuals as well as the function of learning related online communities as a whole is necessary. The processes of knowledge acquisition in online communities, and in particular open in online courses, are not yet fully understood and computational methods to gather valuable information are not exhaustively explored. This includes questions on how individuals access and consume information artefacts and how information exchange and diffusion among learners takes place. The work outlined in the following chapters aims at making progress in this direction. It can be located in the context of the emerging field of learning analytics that utilise the growing availability of machine readable data produced in different contexts to optimise learning environments (Ferguson, 2012). The focus of

the presented work, thereby, is not on the explicit development of dedicated support mechanisms that directly target learners or tutors. Instead, a primary goal is to gain knowledge about phenomena and properties of learning in online communities in a more general sense that can be used by practitioners and researchers for the redesign and development of online learning environments. In this aspect, the large amount of available data produced by online learning communities introduce new opportunities to study learning, knowledge creation, and knowledge diffusion processes on a large scale. Along with the emergence of knowledge discovery from large amounts of data using computational methods as a scientific paradigm (Hey, Tansley, & Tolle, 2009), large online courses can be considered as “massive research laboratories” (Diver & Martinez, 2015) that allow for conducting experiments and to observe different characteristics of online learning. However, in order to gain insights beyond obvious properties of those learning environments, adequate research instruments have yet to be developed. In this regard, the methodological objective of this thesis is the development and exploration of computational approaches, which are primarily rooted in the field of social network analysis (Brandes & Erlebach, 2005; Wasserman & Faust, 1994) in combination with analysis of time dependent data and produced content.

The following Section 1.1 provides the context of this research work and outlines the research objectives more concretely. Section 1.2 summarises the included publications (Chapters 2-6), highlights relations and differences among them, and positions them in a contextual framework. Chapter 7 summarises the main outcome of this thesis and gives an outlook on ongoing and future developments.

1.1 Context and Research Objectives

This section outlines preliminaries and research objectives of this thesis. It provides a framework to contextualise the following chapters and to locate them in relevant fields of research. First, the meaning of actors, communities, and artefacts for this work is outlined against the background of existing work. Second, important considerations and objectives regarding the investigations of online communities are described. The third subsection focuses on the envisaged methodological contribution to the fields of learning analytics and social network analysis.

1.1.1 Actors, Communities, and Artefacts

Actors. In this work, the term actor generally refers to all human beings who actively interact with and within a virtual learning environment. By the scope of this thesis the actors of interest are mostly learners, i.e. the actual users of an online learning environment. However, most of the methods described later on can be applied in a more general context including other individuals participating in a knowledge creation process such as tutors, scientists in scientific communities, or editors of publicly available information resources.

Online Communities. The notion of (online) community is used very differently in the literature (Porter, 2004). In the most general sense Porter (2004) defines an online (or virtual community) as:

“... an aggregation of individuals or business partners who interact around a shared interest, where the interaction is at least partially supported and/or mediated by technology and guided by some protocols or norms.”

Common notions of communities are “communities of interest” and “communities of practice”. Communities of interests (Cols) are gatherings of people with a common topic of interest, for example, news groups (Fischer, 2001). Communities of practice (CoPs) (Wenger, 1998) are communities of actors with similar professions who share a collective identity and constantly collaborate to collectively gain knowledge and improve professional practice.

While the definitions of Cols and CoPs are not necessarily coupled to learning Carlen and Jobring (2005) relate these concepts to the case of online learning communities (OLCs) and provide a typology incorporating different existing definitions. They derive three types of OLCs that are found in purely online but also blended settings where face-to-face meeting mix with online activities. Professional OLCs are related to CoPs and comprise of members who have a special interest in sharing knowledge about their work practices, and thus, can be seen as a form of advanced vocational training. In contrast, online interest communities do not necessarily share the same profession. This type of OLC especially emerges in forums and discussion boards on specific topics of the members’ interest. Apart from communities that are constituted from common professions or interests of its members the typology of Carlen and Jobring (2005) also introduces a notion of educational OLCs that is especially important for this work. Educational OLCs differ from the other types in the sense that they are established in a virtual environment that intentionally supports learning. Those communities exist within formal educational systems such as online courses that are created by educational managers for distance education and are structured around pre-defined tasks. Since educational OLCs are managed by instructors, they can be self-organised only to certain extent, and consequently, a data driven evaluation and development of support mechanisms can only be valuable if it explicitly takes into account the conditions set by educational design. The online courses investigated in this thesis can be seen as particular realisations of educational OLCs. However, regarding studies on knowledge sharing in MOOCs that will be described later on, properties of interest based communities are relevant as well. For example, many MOOC participants only use parts of the courses driven by a common interest to gain specific knowledge on certain subtopics or just out of curiosity, and thus, do not adhere to the originally intended use of such courses (Ferguson & Clow, 2015).

Rather than suggesting a strict categorisation, it is reasonable to highlight different aspects of communities that are important for the rationale of this thesis. In this sense, online communities can be placed in different contexts, i.e. production, socialising, and the aspect of common affiliations.

The production aspect is in the focus of communities in which people collaborate with the primary purpose of creating digital artefacts. Typical examples are groups of authors who

contribute to wiki articles (Bryant, Forte, & Bruckman, 2005) or open source software developer communities (Gacek & Arief, 2004). While the educational online communities investigated in this work are not primarily productive communities, the production aspect is still present with respect to learner generated content. The characterisation of actors regarding the creation of digital artefacts often requires computational methods for the analysis of the produced content, which will be described in Section 1.1.3.

In the view of socialising between actors, the primary purpose of online communities is interaction and exchange between its members. Dedicated social media platforms are used by actors to maintain social contacts that they have established in physical spaces (Ellison, Steinfield, & Lampe, 2007). This aspect is particularly important in the theory of social learning (Bandura & McClelland, 1977) that builds upon the basic assumption that learning always takes place in some sort of social context. In addition, it was shown that communities forming on online social media platforms have the potential to spread information very quickly among a large audience (Bakshy, Rosenn, Marlow, & Adamic, 2012; Romero, Meeder, & Kleinberg, 2011). In online courses, information spreads over interpersonal relations between actors, for example in chats, forums and discussion boards. Social network analysis provides a rich set of methods that are especially suited for analysing social aspects of online communities, which is also used extensively throughout this thesis.

An online community is not only denoted by socialising and collaborative content production but also by the affiliations of its members. Especially in interest based communities, the strength of social bonds can be low and the members identify themselves more with the topic of interest than with the community itself (Henri & Pudelko, 2003). In this sense, the typical activities, available resources, and the interests of the actors can be seen as the affiliation aspect of a community. Henri and Pudelko (2003) further point out that the emergence of intention in an online community is coupled with an increasing awareness of its existence, which is materialised by the online environment and provided resources. This view is especially important in educational online learning communities that are in first place shaped by the learning environment and the interest of its members. A large proportion of the work described in the following chapters deals especially with the affiliation aspect of online learning communities by investigating actors' affiliations to learning resources or themes utilising network analysis techniques.

Artefacts. In this work, an artefact can be any type of digital object that contains externalised information relevant for the purpose of the online (learning) environment. In digital environments the log protocols of actors' usage or modifications of artefacts constitute actor-artefact relations. Those log protocols are often the only data source that can be used for the purpose of analysis and characterisation of learners in online communities. In online learning communities artefacts play a major role as information containers, in communication processes, and for collaborative knowledge building for the externalisation of information (Belanger & Thornton, 2013; Cress & Kimmerle, 2008).

On the one hand there are stable artefacts, for example, learning resources such as videos and reading materials. These types of artefacts are stable in the sense that the content usually cannot be edited by learners, and thus, does not change over time. Stable artefacts are usually provided by a tutor or a platform manager. The main purpose of stable artefacts is to function as information sources providing relevant content for the learning objective. Resource based learning environments (RBLs) (Hill & Hannafin, 2001) were proposed as a design principle that utilises the high variety and availability of digital learning resources. In those environments learners have to be supported by tools and scaffolds that enable them to deal with the amount of available information depending on different contexts or learning goals. Tools incorporate facilities for effective search, processing, manipulation, for example, note taking and communication. Furthermore, scaffolding mechanisms have to be applied to assist learners in gathering dedicated knowledge from the available resources. Conceptual scaffolds provide guidance on how to use resources to achieve a specific goal, for example, how to write a survey on a specific topic. On the procedural level, scaffolds can be helping mechanisms for learners how to use specific resources. Other types of scaffolds are metacognitive scaffolds to support reflection of the learning process and strategic scaffolds outlining different ways to solve a task, for example, recommendations. More recently, the role of stable artefacts related to resource based learning has been emphasised in the context of online courses where various types of resources co-exist to support multiple learning styles (Grünewald, Meinel, Totschnig, & Willems, 2013). At the extreme end, connectivist approaches (Siemens, 2014) consider the entire learning process as building networks between different information sources, often in form of digital resources. This concept is adapted in connectivist MOOCs (cMOOCs) (Fini, 2009) and personal learning environments (PLEs) (Dabbagh & Kitsantas, 2012) that support learners and their learning processes in terms of management and retrieval of collections of relevant artefacts using social media and cloud computing technologies.

Apart from the view on artefacts as potentially stable information sources, there are also dynamic artefacts such as Wiki articles, blogs, and discussions threads in forums that play a role in communities of learners. The content is usually user generated and evolves over time. Cress and Kimmerle (2008) see collaborative editing of dynamic artefacts as an interleaved process in which externalised information and the knowledge state of contributors co-evolves. In this aspect, the collaborative creation of content is considered as an important concept for learning.

The distinction between stable and dynamic artefacts is crucial for the information one can acquire from the analysis of actor-artefact relations. Relations between actors and stable artefacts are typically derived from resource access protocols and can be used to infer semantic relations between actors, such as common interests or experiences. These hidden semantic relations can be used for community support mechanisms, for example, recommendations of social contacts and facilitation of thematic navigation in a virtual community (Hoppe et al., 2005). On the contrary collaboratively edited dynamic artefacts can be used to infer social relations between the actors since these processes typically require reactions to actions previously performed by others. In contrast to semantic relations, these relations are typically known to actors since the collaborative editing of an artefact usually creates awareness of the other person.

This leads to communication through artefacts which is particularly important in asynchronous settings since communication can only be established through some shared representations that can constantly modified by actors (Hoppe, 2009; Hoppe et al., 2005). Figure 1 depicts a schema of relations between actors derived from actor-artefact relations.

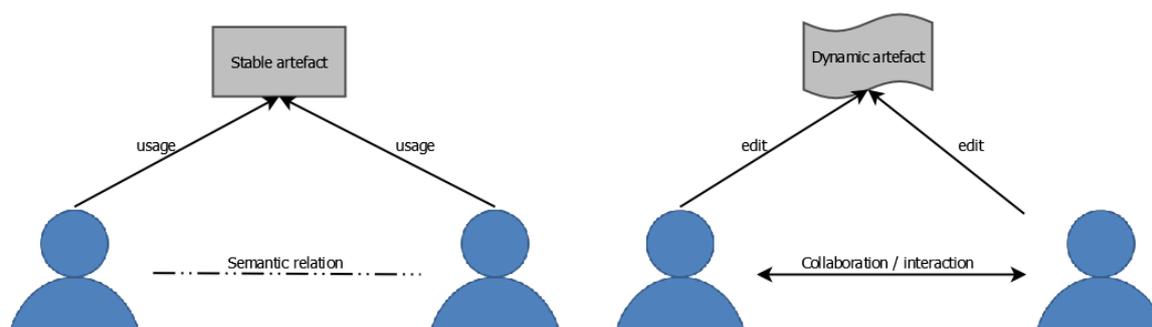


Figure 1 Inferring interpersonal relations from actor-artefact relations.

Technically, there are several ways to model relations between actors and artefacts and to transform these relations into social and semantic relations between actors. This includes semantic approaches that utilise the type of actor-artefact relations inducing directionality, for example, production and consumption (Reinhardt, Moi, & Varlemann, 2009; Suthers & Rosen, 2011). The combination of both perspectives – the social relations between actors and semantic relations based on common interests – has also been investigated in contexts not directly related to learning (Mika, 2007; Roth & Cointet, 2010). The potential of this combination in the learning and knowledge management domain, especially for recommendation of actor and artefact relations, has been outlined clearly by Harrer, Malzahn, Zeini, and Hoppe (2007). However, the integrated view on social and semantic relations is widely unexplored in the analysis of learners in online communities. Advances in this direction are among the contributions of this thesis that will be described explicitly later on.

1.1.2 Characterising learners in online communities

This subsection first provides the general background of the task of characterising learners in online communities by introducing a general scheme community analysis on different levels. After that, the expected outcome of this thesis with respect to decision support of online learning communities will be discussed.

Granularity of analysis. The function and specific characteristics of online learning communities and individual learners can be analysed on different levels of granularity. Primarily three layers can be distinguished that are not strictly separated, namely the individual-, the meso-, and the community level. Analytical models on one layer can also contribute to gain information on an adjacent layer via top down or bottom up inference. Figure 2 depicts this basic scheme and the information transfer between levels graphically, which will be described in more detail in the following.

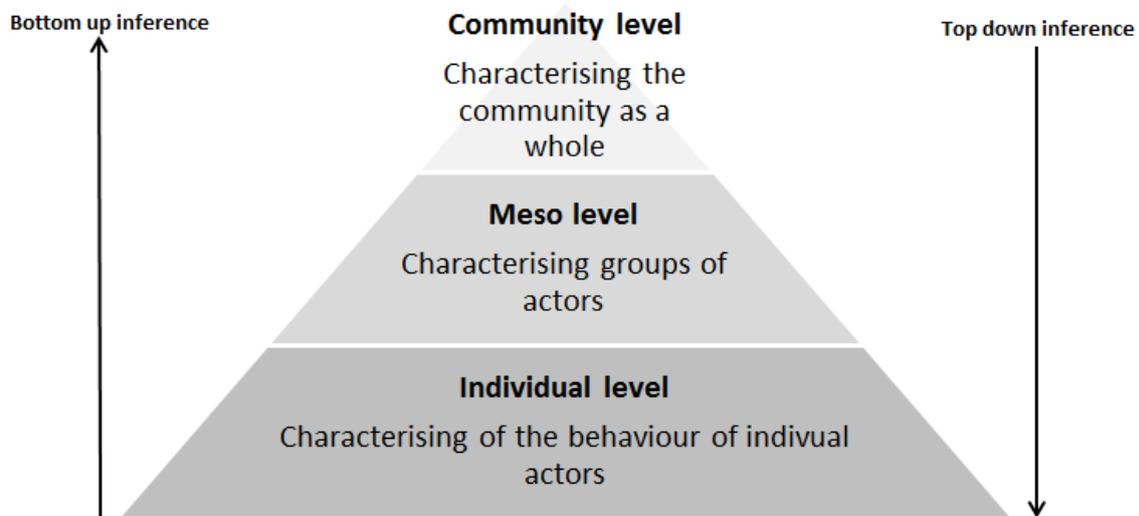


Figure 2 Levels of community analysis and types of inference.

The top level is referred to as the community level. Analyses on this level provide information about the structure and functionality of a community in a global view. This includes, for example, simple descriptive statistics on the number of actors or the social or semantic structure of the whole community. For example, in social network analysis a global measure is the clustering coefficient (Watts & Strogatz, 1998) which measures the extent of transitivity of network connections, i.e. the fraction of closed triangles (clique of three actors) and the number of open triangles. If this value is high, one can infer that the community is likely to bear subcommunities of densely connected regions in the network on a meso-level. In the context of learning communities it is then worth to identify such subcommunities and to characterise them in terms of social cohesion and information exchange.

Meso-level analyses often deal with the aggregation of individual actors into groups. Clustering is a typical approach to reduce the complex structure of a community to an interpretable macro-structure of the aforementioned subcommunities (Fortunato, 2010). The distributions and characteristics of groups of actors can then be transformed to statements about the whole community on the top level of analysis. For example, Palla, Barabasi, and Vicsek (2007) identify the evolution overlapping subcommunities of actors and generalise these meso-level observations to statements about the social dynamics in large collaboration and communication networks. However, clustering of actors is not the only task on the meso-level. There can also be meso-level representations of process data, such as process models that are discovered from atomic action logs of several individuals. These models reflect inherent strategies or workflows that are not directly visible from the individual activity traces (Reimann, Markauskaite, & Bannert, 2014). Bottom up inference can then be used to link the meso-level processes to more general rules and functions of the community as a whole. Apart from inferences from the meso-level to the community level, the assignment of actors to groups naturally leads to a characterisation on the level of individuals. Examples are role models that group actors based on certain notions of

similarity of their positions in a social network. These groups are often interpreted as sets of actors with similar roles (Doreian, Batagelj, Ferligoj, & Granovetter, 2004). A large proportion of the work described later on relies on community analysis on the meso-level, while the results are used to analyse the community itself, but also to characterise individual learners.

Community analysis on the level of individuals is mostly concerned with measures to describe properties and behaviour of particular actors. A proper characterisation of individual actors is important to make learning environments adaptive in the sense of personalisation and to create tailored support mechanisms. Typical characterisations for learners in online communities are quantifications of the learners' activities within a learning platform, like engagement patterns (Ferguson & Clow, 2015), measures of performance and affective states (Baradwaj & Pal, 2012), or demographics (Guo & Reinecke, 2014). Most prominent in social network analysis are centrality measures that are frequently used as indicators for the importance of actors in a networked community depending on their position in the social network (Koschützki et al., 2005). Analysis of a community on the individual level often precedes analyses on the meso-level since the grouping of actors, for example based on a notion of similarity, usually requires a previous characterisation individuals.

Enabling effective community support. A major objective of this thesis is to provide information to facilitate the evaluation and design of learning in online communities. Thereby, analytics and in particular learning analytics can contribute to community support in numerous ways. One can distinguish three interleaved areas, reflection and monitoring, predictions, and exploratory data analysis (Baker & Inventado, 2014). As mentioned in the beginning, the focus of this work is on inductive approaches and explorations for post-hoc evaluation of learning activities. This objective will be contrasted with related areas in the following.

Reflection support aims to help learners to become aware of their own learning processes as well as the learning processes of others, and thus, triggers reasoning about certain activities. In this sense, reflection mechanisms support the actors in a community by enabling them to adapt and to improve their learning processes on their own. This can be, for example, achieved by awareness tools that provide learners with easily interpretable representations, for example, concept clouds reflecting their coverage of a particular knowledge domain based on analysis of learner generated content (Manske & Hoppe, 2016). In contrast to self-reflection support for learners, monitoring tools enable tutors to keep track of the activities and learning progress of learners with the goal of more informed interventions (Lockyer, Heathcote, & Dawson, 2013). Reflection and monitoring techniques are especially concerned with data visualisations and understandable placement of information, which is usually established through the use of dashboards (Verbert et al., 2014).

In contrast to reflection and monitoring where the actors (learners and tutors) are provided with necessary information to actively improve individual and community learning processes, predictions (e.g. of course performance) can be used to create system interventions that support an online learning community without explicit involvement of human actors. Typical forms of system generated feedback are recommendations of resources, people, or activities (Verbert et

al., 2012). Approaches originating in intelligent tutoring systems denote an extreme case where knowledge estimation and hint generation is completely automatized (Rivers & Koedinger, 2015). Learning analytics for reflection, monitoring, and predictions explicitly targets students and tutors as stakeholders and the analysis results are directly fed back into the learning environment. This can be referred to as the short cycle of learning analytics (Clow, 2012). Apart from the intervention centric view on learning analytics there is also a need for inductive approaches that explore learning in online communities in order to acquire general insights into their special characteristics contributing to the development of theories and concepts (Diver & Martinez, 2015). A proper understanding of the roles of learners and the processes that take place within a community on all described levels of granularity is crucial for the development of novel concepts that improve learning in online environments in the long run. However, the rapid development of platforms and concepts for online learning such as MOOCs often precedes insights into the characteristics of the communities that emerge in this context. For this reason, one goal of this thesis is to contribute to a reduction of this information deficit of different stakeholders including researchers.

The importance of digital learning resources as static and dynamic artefacts for storing and distribution of knowledge was outlined previously in Section 1.1.1. Therefore, substantial parts of this work aim to acquire detailed knowledge about the usage of learning resources that goes beyond counting the number of accesses for each provided resource. Moreover, characterising learners based on their affiliations to available learning resources can help to adjust online learning environments to the needs and preferences of a heterogeneous audience. In the affiliation aspect of online communities it is also desirable to create an overview about the learners' general thematic interests instead of affiliations to concrete artefacts. This can be a valuable to gain knowledge about the general mindset of a community and to estimate competencies of learners in particular domains.

Furthermore, another important element in very large educational learning communities, such as the audience of MOOCs, is information exchange between peers since tutor assistance for individual learners is limited given the number participants. Thereby, it is of interest how the exchange between individuals takes place and whether the provided means of communication are sufficient to support collaborative knowledge construction. To this end, another objective that will be targeted in throughout this work is to gain insights into information brokerage between learners when communication only takes place asynchronously and mediated by artefacts.

Observations made on the previously mentioned levels of granularity are supposed to be brought into the scientific discourse to facilitate advances in the improvement of online learning concepts in terms of personalisation, provided content, and short cycle interventions. In this aspect, it is crucial to develop advanced analytics methods that help to acquire meaningful information from data and enable observations of non-obvious characteristics of learners in online communities.

1.1.3 Exploration of analytics methods

Ferguson (2012) outlined several challenges for learning analytics, particularly the need for advanced methodology to utilise increasingly challenging and complex datasets. While other fields such as business intelligence (Albright & Winston, 2014) and social computing (Macy & Willer, 2002) can build upon a rich and consolidated foundation of computational methods, in learning analytics a methodological discourse on how approaches from different can be utilised is about to emerge. To structure the range of computational methods for learning analytics a classification scheme is introduced by Hoppe (2016), which is depicted in Figure 3. In the following, this scheme is used to give an overview on different approaches used in learning analytics before the methodological objective of this thesis is explained in detail at the end of this section.

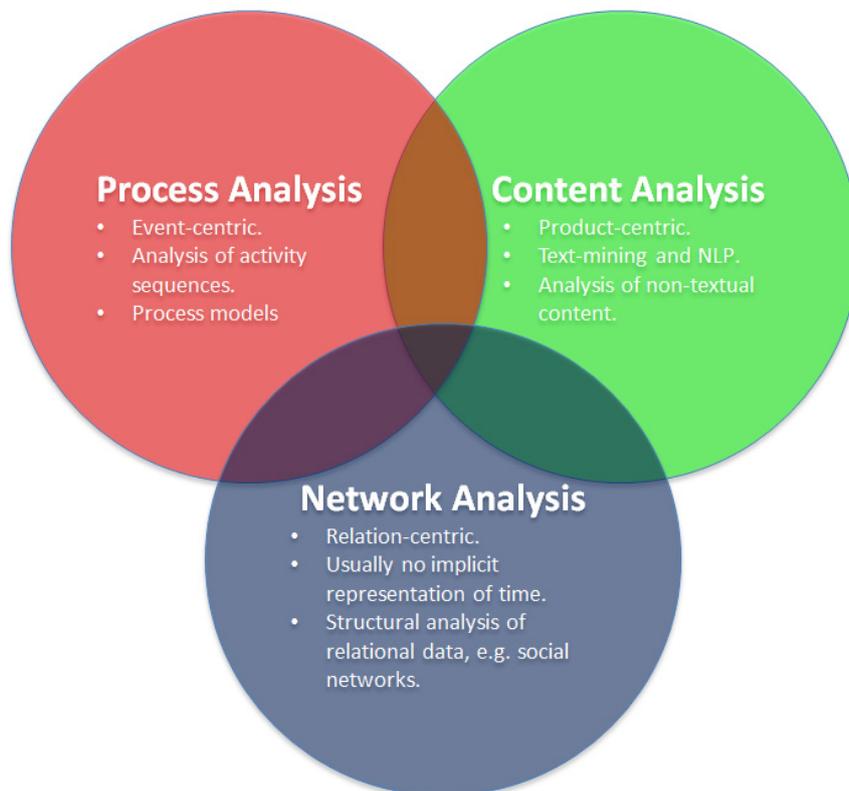


Figure 3 “Trinity” of computational methods for learning analytics.

Process Analysis. Techniques that can be used to investigate temporal activities of learners based on timestamped activity logs can be subsumed as “Process Analysis”. It has been argued that this event centric perspective is especially suited for the usually temporal data produced in collaborative learning (Reimann, 2009). In contrast to variable and variance based approaches, which are usually very restrictive on the type of data and further ignoring the ordering of events, process analysis can be used to link quantitative methods with qualitative process models to capture developments over time. Sequential pattern mining as a typical approach can be applied

to identify characteristic temporal patterns in the activity traces of actors recorded on web platforms (Srivastava, Cooley, Deshpande, & Tan, 2000), particularly in online learning environments (Perera, Kay, Koprinska, Yacef, & Zaïane, 2009). The main goal is to discover the typical order in which certain activities are performed to gain insights into learning processes and for comparison of actual with expected learner behaviour. A related approach is process mining (van der Aalst et al., 2007). In process mining, the goal is to generalise multiple event sequences, activity traces to a process model, i.e. a state transition system that can be represented as a Petri net or a probabilistic Markov chain model. Originally developed for process discovery and modelling in business applications, process mining has also been adapted in the educational domain to identify general learning processes from fine-grained log protocols or messages in group discussions (Bannert, Reimann, & Sonnenberg, 2014). While there are limited degrees of freedom in the actual model creation and sophisticated algorithms are available (van Dongen, de Medeiros, Verbeek, Weijters, & van der Aalst, 2005), a major challenge is to pre-process the input data, which is in the extreme case a detailed clickstream, such that the result can be related to the actual learning process. This requires the filtering and aggregation of atomic actions to higher-order tasks which is difficult since it requires detailed knowledge about the interaction patterns of learners with the system. This information, however, is necessarily incomplete since otherwise there would be no need for data analysis on the process level. Sequential pattern mining and process mining are not part of the work included in this thesis, however, temporal and process aspects are explicitly considered in conjunction with network analysis methods described later on.

Content analysis. As already stated in Section 1.1.1, dynamic and static artefacts can be considered as a building block of most virtual environments for learning and knowledge creation. Thus, another important methodological aspect is content analysis, especially of learner generated artefacts. While the range of available methods is very broad, in learning analytics content analysis mostly deals with the application of text-mining techniques to acquire information from textual data. Only a few studies consider other types of artefacts such as concept maps, for example the work of Clariana, Engelmann, and Yu (2013). An important application is discourse analysis, which is concerned with collaborative knowledge construction through participation in dialogues (De Wever, Schellens, Valcke, & Van Keer, 2006; Liddo, Shum, Quinto, Bachler, & Cannavacciuolo, 2011). Discourse analysis is a relevant subject for the work on analysing information sharing in discussion forums presented later on in Chapters 4, 5, and 6.

Another important goal of the analysis of textual data in online learning communities is to model thematic contributions of learners in collaborative writing tasks which can be used for role modelling of contributors and estimation of individual knowledge domains (Southavilay, Yacef, Reimann, & Calvo, 2013). The recently developed concept of meaningful purposive interaction analysis (Wild, 2016) combines latent semantic analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) and relations between actors and learning opportunities (e.g. online courses) to model competencies and conceptual development. Among other applications, this framework allows for comparison of the actors' competencies as well as combined visualisations of concepts and assessment of the actors' knowledge. Chapters 4 and 6 in this thesis

also deal with thematic modelling as a means to identify essential concepts from textual contributions of actors for the identification of interests and expertise.

Content analysis is further applied in text classification tasks to extract characteristic features from written language that can be used to train classification models. In the field of learning analytics text classification is often used to link textual contributions to theoretical constructs from learning sciences, such as taxonomies of cognitive learning levels (Wong, Pursel, Divinsky, & Jansen, 2015b) and collaborative knowledge construction activities (Rosé et al., 2008). Text classification is particularly important for the work described in Chapters 6 and 7 for the identification of meaningful discourse elements such as forum posts that are relevant for the exchange of information.

Apart from methods that consider documents as “bag of words” ignoring the position of words in a text, network text analysis aims to preserve the conceptual structure of texts. In those approaches words are mapped to concepts and connected based on the proximity of their positions in a text to establish a concept network. An application of network text analysis is the modelling and visual mapping of the conceptual structure of a knowledge domain of the authors of a text (Daems, Erkens, Malzahn, & Hoppe, 2014). This aspect of content analysis, however, is only peripherally covered in the following.

Network analysis. The third category of methods is (social) network analysis. These approaches explicitly deal with relational data by investigating networks of interconnected entities (represented as nodes). These entities are often actors in a social network. In this thesis, however, multipartite (also multi-mode) networks (Wasserman & Faust, 1994) with edges between nodes of different types are especially important since large proportions deal with bipartite actor-artefact networks. A basic property of these networks is that edges cannot exist of between nodes of the same type. This poses some restrictions on the length of cycles and paths between node types, which often creates the need to adapt methods designed for the unipartite case. This issue will be handled especially in Chapters 2-4. Network representations enable the application of methods rooted in the well-founded graph theory, and thus, network analysis is especially suited for structural analyses of networked communities on all three levels described in Section 1.1.2, cf. (Brandes & Erlebach, 2005). Related fields are graph data mining and statistical relational learning that have a stronger focus on data intensive classification and inference problems on networks such as link prediction and entity resolution (Getoor & Diehl, 2005).

Individual actors in a network are often characterised by centrality measures indicating their importance with respect to their connections to others. The up to day most comprehensive reviews of different notions of centrality can be found in (Friedl & Heidemann, 2010) and (Koschützki et al., 2005). It is widely assumed that the position an actor has in a social network has a strong influence on the ability to spread and receive information and to accumulate social capital (Borgatti, Mehra, Brass, & Labianca, 2009), and thus, centrality indices of special interest in learning contexts (de Laat, Lally, Lipponen, & Simons, 2007).

As mentioned in Section 1.1.2, social network analysis can be applied for community analysis on the meso-level by identifying densely connected groups of actors (Fortunato, 2010). Those,

densely connected regions in a social network are often called subcommunities or modules and play a crucial role in information spreading processes. Information can circulate very quickly within those modules since there are typically many connection paths between actors of the same subcommunity. However, the addition of new information to a subcommunity heavily relies on relations between actors of different subcommunities. These weak ties were described and studied intensively by Granovetter (1973). Subcommunity detection methods are not restricted to social networks in the original sense. The discovery of densely connected cohesive parts also matters in affiliation networks such as bipartite actor-artefact networks, as a special type of network clustering. These affiliation clusters can be used in various ways to characterise an educational online community and will be one of the main topics of the research described later in Chapters 2-4 of this thesis. In addition to clustering actors into cohesive subcommunities, actors can also be grouped based on the similarity of their connection patterns which does not necessarily imply cohesiveness of a cluster (Doreian et al., 2004). This type of network modelling is an important building block for role modelling in information exchange networks and will also be discussed explicitly in Chapters 5-6.

Apart from actor centric measures and grouping of actors, global properties of networked structures on the community level are also intensively studied in different ways. The aim is to find general rules explaining the emergence of various phenomena in networks such as short average path lengths which is known as the small-world property (Watts & Strogatz, 1998) or the evolution of highly connected nodes (hubs) in scale-free networks (Barabási & Albert, 1999). In the domain of knowledge related online communities an important task on the community level is to investigate the capability of different network configurations for spreading information among interconnected actors (Liben-Nowell & Kleinberg, 2008).

In the works reported later on temporal aspects of network evolution are of particular interest. Although real-world networks are often dynamic in the sense that nodes and edges are constantly added and deleted over time, network representations usually do not implicitly reflect the history these events. For example, when one creates a network of actors who communicate over a certain period, it is necessary to aggregate all communication events that occurred during a specified time window losing the temporal order of events in favour to map the structural relationships between the actors. However, the evolution of networked structures over time is an important aspect of network analysis. Thus, different techniques exist for handling this issue depending on the nature of the available data.

Most commonly, sequential approaches aggregate nodes and edges created in successive or overlapping time windows of specified size into a sequence networks. The choice of an appropriate time window size, thereby, has a huge effect on the result and is far from being trivial. It could be shown that the duration of typical production cycles of a community is a proper heuristic for making this choice (Zeini, Göhnert, Hecking, Krempel, & Hoppe, 2014), but the applied techniques and analysis goals have to be considered carefully as well (Hecking, Göhnert, Zeini, & Hoppe, 2013). In the case of online courses that will be discussed in the following

chapters, a “production cycle” is typically given by the frequency of course sessions (usually weeks).

In contrast to sequential modelling of dynamic networks, streaming approaches represent an evolving network as a stream of temporally ordered events of node/edge creation and deletion (McGregor, 2014). Starting with an initial state of the network, the structure is constantly modified by adding and removing elements when such events occur. This modelling technique keeps the order of node and edge creation events and no aggregation is necessary. Thus, properties of nodes or the network as a whole can be modelled as a continuous function over time. However, this method is only applicable if nodes and edges have a lifespan larger than the typical arrival time of events. This is, for example, the case in power grids or friendships in online social networks where nodes and edges once created are likely to persist over a longer time span. In the actor-artefact networks or communication networks relevant for this thesis, edges are typically observable only at the exact point in time when they are created, usually recorded as a timestamped event without duration. A streaming approach would boil down to a sequence of events (represented as networks with only one or very few edges) without being able to capture structure.

Concerning temporal dynamics, networks that model information flow over time denote a special case since these types of networks implicitly represent time. Assuming that nodes are activated by information passed between them over directed edges only forward in time, the edges that lead to new node activations induce a partial temporal order of the nodes with respect to their activation times. Citation networks of scientific publications are prototypical examples of such information flow networks. Since new publications can only take information from already existing publications they cannot contain cycles and the partial temporal order of publication dates is preserved in the partial topological order of nodes. Every directed network with time dependent edges can be transformed into a network with a partial temporal order of nodes if different versions of nodes are introduced such that the timestamp of ingoing edges of all nodes are always smaller than the timestamps of outgoing edges. For example, Halatchliyski, Hecking, Goehner, and Hoppe (2014) transform the typically cyclic hyperlink network between wiki articles into a directed acyclic graph by representing each article revision as a single node. In this way directed acyclic graphs between interlinked revisions are derived enabling to application of network based approaches to investigate the flow of ideas between articles over time.

Methodological objective. Apart from the research objective outlined in the previous Section 1.1.2 that is more oriented towards the expected outcome of investigations of online learning communities with respect to community support, this work also aims to expand the methodological spectrum of learning analytics and social network analysis in a more general sense. The methodological contribution of this thesis can be subsumed as development and exploration of analytics methods to characterise learners in online communities in terms of content production and usage, and information sharing.

Although the nature of the data produced by learners in online communities such as forum communication, collaborative content production, and artefact consumption is inherently of

relational nature, in many existing studies the full potential of the rich body of research in network analysis (sometimes referred to as network science) is not fully exploited. As pointed out by Hoppe, Harrer, Göhnert, and Hecking (2016) advanced network analysis methods applied to evaluate online learning scenarios can open up new perspectives in community analysis and support. From this point of view, a productive discourse on the exploration and application of computational methods beyond descriptive and inference statistics applied in the learning domain is considered as desirable for enabling more informed decisions in community support.

In order to proceed in this direction, adaptations of network analysis methods as well as new approaches are described that take into account the different perspectives described before. While the primary approaches in this work are based on network analysis, concepts from process and content analysis are incorporated, as well, leading to mixed method approaches. A special focus is on the application of network analysis techniques to temporal data. Explicit consideration of time allows for getting a more complete picture on dynamic processes of individual and community behaviour. This incorporates the process oriented perspective described above that considers timestamped events as the basic entity of analysis. As outlined before, the combination of the advantages of network analysis and temporal aspects to investigate structural dynamics is still a challenge in research in social network analysis that will be particularly addressed in the following chapters.

1.2 Synopsis of Included Publications

This section outlines the main contributions of the publications that constitute the following chapters. The chapters can be contextualised according to the investigated aspect of online communities and types of actor-artefact relations described in Section 1.1.1, the research objective according to different scopes and levels of community analysis as outlined in Section 1.1.2, and the methodological contributions with respect to the development and application of mixed method approaches discussed in Section 1.1.3.

As described in Section 1.1.1 different aspects of an online community can be analysed depending on the type of actor-artefact relations that are investigated. Particularly important is whether the artefacts are considered to be static (relations indicate actors' information consumption or interests), or dynamic (actors manipulate artefacts). Applying this scheme to the following chapters leads to the graphical outline depicted in Figure 4.

Relations between actors and stable artefacts are analysed by applying network analysis techniques for dynamic bipartite networks that are capable to discover meaningful patterns even if only limited data, i.e. resource access logs of actors are available (Chapters 2-4). The focus in these chapters is on the affiliation aspect and information consumption of actors in the investigated learning communities. As mentioned earlier the actors' relations to static artefacts can then be used to infer semantic relations between actors based on common affiliations.

Social actor-actor relations originating from co-editing dynamic artefacts are used to develop role models from forum communication between learners in large scale online courses (Chapter 5). Consequently, the social and content production aspect of communities is much more salient

than in Chapters 2-3. The combination of semantic relations between actors based on common interests and social communication relations among them will be discussed explicitly in Chapter 6 using mixed method approaches comprising of network and content analysis.

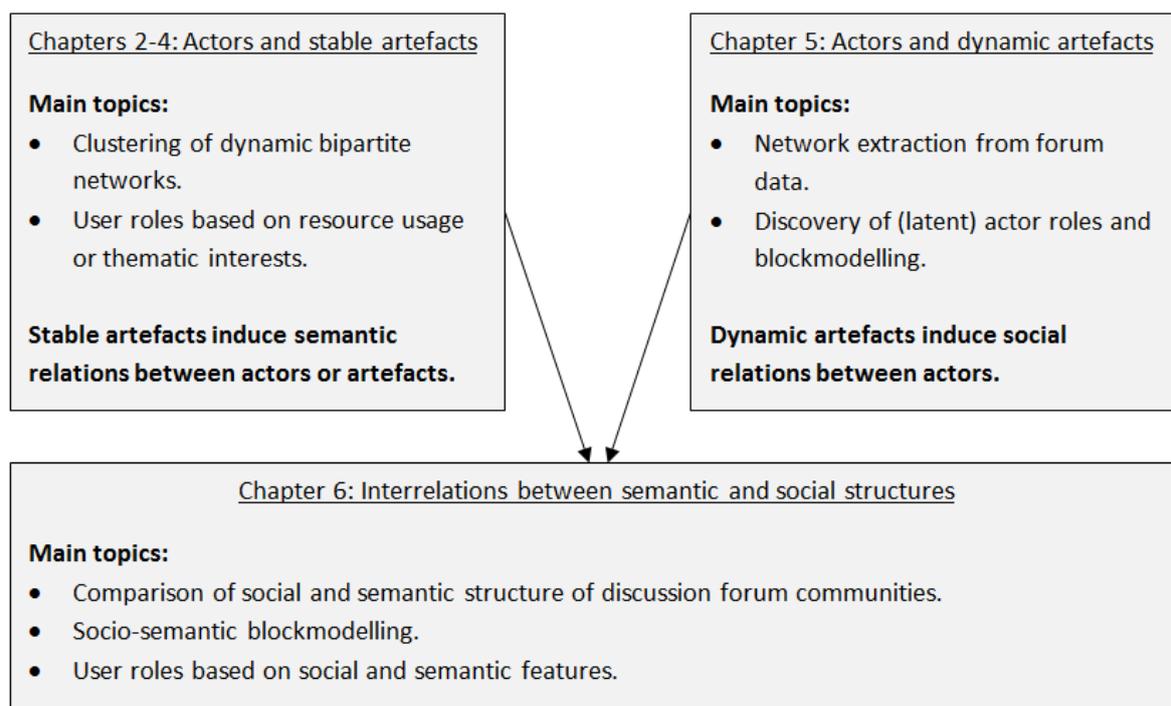


Figure 4 Graphical outline of content of selected papers and their interdependencies.

1.2.1 Relations of actors and stable artefacts

The chapters subsumed in the following report the development of methods for the analysis of bipartite networks of actors connected to artefacts. The described works serve the objectives of characterising learners according to resource usage and thematic interests outlined in Section 1.1.2 as decision support for community management. Furthermore, the challenge of combining network analysis with temporal aspects described in Section 1.1.3 is particularly addressed.

Chapter 2. The next chapter presents a study on learning resource usage of learners in online courses adapting methods from network analysis. In contrast to often used descriptive statistics on resource usage, the goal is to find more sophisticated patterns and regularities of the resource access of learners in such courses over time. The conducted analysis is solely based on the log data of resource accesses which is often the only available information about the learning processes of participants. Two studies were conducted; the first on a blended educational online learning community forming during a master-level university course with possible unobserved interactions outside the system, and the second on a course with sole online activities. Although the communities are productive to some extent since creating wiki articles as part of regular exercises, the focus is on the affiliation aspect and information consumption rather than on content production and social interactions. This poses the question whether there are emergent

patterns of resource usage in these courses even if direct social influence between the learners is limited. Learning resources are considered as stable artefacts that are used by the course participants as information sources. In this sense, actor-artefact relations are represented as a bipartite network of learners who are connected to learning resources they accessed according to activity log protocols. By applying “Biclique Communities” method for network clustering introduced by Lehmann, Schwartz, and Hansen (2008), overlapping clusters of learners and learning resources are derived. Each of these clusters is constituted by a group of learners who establish common relations to a set of learning resources while this set of learning resources is also more densely connected to the learner group than to learners outside the cluster. This duality allows for a differentiated interpretation. The learners in such clusters can be seen as a group of learners with similar interests (or resource usage behaviour). According to Section 1.1.1 a hidden or semantic relation between the learners can be assumed even if direct social relations are not present, which allows for a categorisation of the learners based on their association to such clusters. To this end, in an optimistic learner model even similar potential knowledge can be assumed for learners in the same cluster. In the same way, the artefacts that are associated with the same cluster can also be considered as related. The applied methodology relies on the notion of bicliques which are maximally connected subgraphs in bipartite networks. Consequently, sparsely connected actors and artefacts are inherently filtered by the clustering methods since they cannot be part of a biclique of a specified minimum size. The strict notion of bicliques further allows for discovery of structural patterns even in very densely connected networks of learners and learning resources.

In addition, a framework is developed for tracking the evolution of the discovered actor-artefact clusters over time, which has not been considered much in existing work on dynamic bipartite networks. The incorporation of time combines structural analysis with the process aspect described in Section 1.1.3 supporting an in-depth view on the changing roles of learners and resources during an online course. This enables the discovery of prior non-observable patterns of the evolution of relational structures of learners and learning resources. Groups of learners with similar learning processes manifested in their resource usage are likely to be part of the same bipartite clusters over time. On the other hand, learning resources that frequently occur in the same clusters because they are used by the same course participants have a potential overall importance for the particular subgroup. More complex patterns of diversification of resource usage in certain periods of the investigated courses, especially in the exam preparation phase, can be discovered as well.

According to the scheme introduced in Section 1.1.2, the approach can be seen as a characterisation of learners (actors) on the meso-level since it operates on subgroups of actors. However, the interpretation of the results can also characterise the community as a whole in terms of diversity of resource usage and importance of different types of material. The presented research of the evolution of bipartite clusters over time has been taken up by Ziebarth et al. (2015) to further characterise individual participants of an online course. Based on the observation that a majority of course participants occur in large clusters over time that contain the main learning resources, they can be classified according to the number and sizes of clusters

they were associated with to during the course. Participants who mainly occur in large clusters with the main course material can be seen as “mainstreamers” whereas others can be identified who show more diverse resource access behaviour. Triangulation of resource access pattern discovery with statistical evaluations further revealed relations between mainstreaming behaviour and exam preparation strategies.

Chapter 3. The analysis methods for dynamic bipartite networks used in Chapter 2 are well suited for very dense networks since the notion of bicliques poses strict requirements on the formation of bipartite clusters. It accounts for possible multiple memberships of actors or artefacts in more than one cluster by allowing overlaps between the clusters. However, there can be application scenarios in which a partitioning of a bipartite actor-artefact network into non-overlapping groups of nodes is more desirable, especially in sparse networks in which bicliques only rarely exist. Chapter 3 focuses on algorithmic aspects of network clustering and describes a new method for partitioning bipartite networks into densely connected components that accomplishes the possibilities of clustering dynamic actor-artefact networks described in Chapter 2.

The notion of modularity (Newman, 2006) for networks defines a measure that can be used to quantify how separated different modules in a partitioned network are. In general, modularity measures for each partition (or module) the difference of the edges that connect nodes of the partition and the number of such connections that are statistically expected under the given degree distribution of the network. Consequently, modularity can be used as an optimisation criterion for network clustering algorithms. Finding an optimal partitioning of a network with respect to maximum modularity is an NP-complete problem (Brandes et al., 2008) but there exist various modularity optimisation methods for networks in the general case that approximate the optimal solution (Fortunato, 2010). However, methods that are tailored for bipartite networks are not well explored. Since the number of expected edges within a partition of a bipartite network is different from the expected number of edges in the general case, existing modularity optimisation methods do not necessarily produce a good partitioning in the bipartite case. For this reason, the developed method adapts the Louvain algorithm introduced by Blondel, Guillaume, Lambiotte, and Lefebvre (2008) to the bipartite case. Thus, it is named “bipartite Louvain”. It produces better results than general modularity optimisation and other clustering methods for bipartite networks on different datasets. Bipartite Louvain adapts local updates of the overall modularity similar to the original method if nodes are moved from one cluster to another which avoids costly global re-calculations of the modularity. Therefore, the algorithm is capable of clustering larger networks than methods that rely on bicliques.

With respect to dynamic bipartite networks, Chapter 3 also introduces a method to compute bipartite clusters of successive time slices of dynamic networks in an incremental fashion. In particular, the clustering of a time slice of an evolving bipartite network is used to compute the clustering of the next time slice. This does not necessarily produce the best clustering for each time slice as it would be the case if the partitioning of each time slice of a network were built independently. However, the transitions between evolving clusters are much smoother using incremental cluster updates since clusterings of successive time slices are more similar than in

sequential clustering, if the changes in the network structure are not too big. This makes it easier to track the evolution of the clusters over time that do not change much and further takes into account the assumption that actors remember former affiliations to different artefacts for a certain amount of time.

In general, Chapter 2 and 3 provide complementary methods that can be used to analyse dynamic bipartite actor-artefact networks on the meso-level. Which method is eventually applied very much depends on the use case and the information need. The Biclique Communities method used in Chapter 2 using sequential clustering of time slices is especially suited for dense networks as they are derived from resource access protocols where heavy structural changes are expected due to the agenda of an online course. The strict definition of bicliques as building blocks of bipartite clusters allows for identifying meaningful substructures even in those networks and accounts for the potentially large overlaps of clusters. In contrast, the bipartite Louvaine method with incremental cluster updates across time slices of an evolving actor-artefact network is better suited for large and sparse networks where a partitioning into cohesive components that can be further investigated often is the better choice.

Chapter 4. This chapter widens the scope of application of the bipartite clustering approach in dynamic actor-artefact networks. In contrast to Chapters 2 and 3 the analysed community comprise of participants of a publicly available MOOC who contribute to the discussion forum of the course. Thus, the focus is much more on self-organised information acquisition and sharing in a loosely connected forum community. To this end, the goal is not to characterise actors in terms of their usage of learning resources but to map the coevolution of discussions themes and the interests of actors. Thus, the production aspect of online communities is more present than in Chapters 2 and 3. The approach can be used for, both, getting a better understanding of the use of discussion forums in large online courses, as well as the development of adaptive community support mechanisms such as recommendations. The basic artefacts are the discussion threads. Although discussion threads are dynamic artefacts that change over time, they are not used to infer social relations between the actors, as stated in Section 1.1.1. Instead, the discussion threads are replaced by keywords extracted from the text content yielding networks between actors and keywords extracted from the discussions they participated in. This accounts for a mapping of forum users and their thematic interests. The combination of network analysis with content analysis allows for the identification of emergent structural patterns of thematic affiliations of users.

A strong focus of this chapter is to use the clustering method from Chapter 2 as a typical meso-level analysis to infer characteristics of particular actors on the individual level. Since a cluster is considered as a set of actors with common interests connected to a set of related keywords, the overlaps between those clusters are of particular interest. Actors who are part of multiple actor - keyword clusters in certain time periods during the course can be considered as having more diverse interests and bridge between different thematic areas. Those actors have a higher potential to spread information across discussions on different topics than actors who are active only in a little number of thematic areas. It will be shown that the group of actors who occur in

overlaps between clusters have different characteristics regarding their posting activities and in extreme cases behave like instructors even if they are regular participants.

On the community level, the approach is further used to characterise the evolution of themes. The chapter, thereby, takes a user centric view on the emergence of themes. In particular, a topic is established as a set of keywords that is of common interest for set of actors as a result of bipartite clustering applied to actor-keyword networks. This covers a different aspect of thematic modelling of actors' contributions in collaborative learning settings (cf. Section 1.1.3 "Content Analysis") and can be contrasted to pure content related topic modelling such as Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003). From this point of view, the identification of evolution patterns of bipartite clusters over time as described in Chapter 2 and 3 can be used to map the coevolution of the actors' interest and themes simultaneously in an interpretable framework.

In the outline of this thesis, Chapter 4 can be considered as a transition towards a second line of research that investigates the direct communication relations between actors inferred from dynamic artefacts, namely forum discussions as depicted on the right site of Figure 4. This will be described in the following.

1.2.2 Social relations between actors and dynamic artefacts

In Section 1.1.1 the important role of constantly modified dynamic artefacts as mediating objects for asynchronous collaboration and communication in online communities was described. As stated in Section 1.1.2, one goal is to contribute to a better understanding of the mechanisms behind artefact mediated information exchange between peers. To that end, the chapters outlined in the following analyse social relations between actors manifesting from their contributions to forum discussions (considered as dynamic artefacts) in large scale online courses with respect to the information exchange, as well as collective knowledge building. The developed methods and the outcome of the conducted studies are considered to be valuable for the design of community support mechanisms for information exchange that takes into account different roles of actors in the social aspect of online learning communities.

Chapter 5. This chapter addresses several problems of modelling roles of actors with respect to information exchange with a focus on methodological aspects. As in Chapter 4, the work is in the context are MOOC discussion forums, but here the particular discussion threads are considered as dynamic artefacts that establish social relations between actors. While in Chapter 4 the view was more on the affiliation aspect of the forum community, Chapter 5 focuses on the interpersonal aspect of content production and information exchange through social relations. While it is known that many course participants use discussion forums passively as information source (Anderson, Huttenlocher, Kleinberg, & Leskovec, 2014), the reported studies focus on those who actively contribute in discussions and knowledge sharing to contribute to a better understanding collaboration in large scale online courses.

First the problem of transforming the relations between forum users and discussion threads into social relations between the users is addressed. This task is not trivial since the applied transformation has a fundamental impact on the information that can be gathered by following

analysis steps. In the context of the research objective of modelling roles of actors in information exchange, the goal is to model a network that adequately reflects who provides information to whom. For this reason, supervised classifiers are built based on structural and content features of forum posts to classify information seeking posts, information giving posts, and posts that do not relate to information exchange. The classification is used to infer directed relations between pairs of forum users where one is the information giver and the other the information seeker. This approach filters for communication events related to information exchange, and thus, maps the information flow much better than existing approaches where social relations between users are established based on activities in common threads.

The extracted information exchange network is used to perform analyses on the level of individuals, as well as on the meso-level. On the individual level, an important task is to characterise users according to their posting behaviour with respect to information giving and seeking. This can, for example, help to identify experts and users who raise important questions in the community. Simply counting the number of posts that request information and posts that react to information requests lead to an incomplete picture since high values might result from temporary high communication activity with a small set of individuals. On the other hand counting the number of communication partners does not take into account the intensity of communication. In order to deal with these issues, new measures for the information giving (“outreach”) and information seeking behaviour (“inreach”) of individual forum users are introduced that combines the diversity of the communication in terms of the number of communication partners and the quantity of forum posts of an user. Individual measures of influence in communication networks can lead to wrong conclusions if time is ignored. Therefore, by the aggregation of communication events over time high centrality values of actors that result from constant activity cannot be distinguished from high values resulting from only few short episodes of high activity. This problem has been addressed systematically by Braha and Bar-Yam (2009). For this reason, the measures for individual behaviour are considered in successive short periods which lead to individual trajectories (or fingerprints) of the posting behaviour of forum users. These trajectories can be used to characterise individual actors but the results can also be transferred to the meso-level of analysis and summarised into characteristic activity patterns of posting behaviour for different types of actors.

Besides the individual characteristics of users in information exchange, it is also desirable to characterise the structure of information exchange on the meso-level. The particular goal is to identify groups of users having similar structural connection patterns in the information exchange network. A common approach for this is blockmodelling (Doreian et al., 2004) that can be used to reduce a complex social network into a macro-structure that reflects the structural dependencies between different components of the network. This is especially suited to answer questions regarding the cohesion of the communication and structural roles of actors. These models, however, do not incorporate structural changes over time. Therefore another class of role models based on tensor decompositions (Kolda & Bader, 2009) is considered as well. These models can be used to group actors who have similar connection patterns over time. From the methodological point of view, Chapter 6 combines two yet separated lines of research on role modelling in

communication networks (the more traditional blockmodelling approaches which are more rooted in mathematical sociology and tensor decomposition models with origins in statistical relational learning) by systematically exploring different combinations of both approaches. While blockmodels have mainly been applied to small and medium sized static networks, tensor decomposition methods are designed for large and dynamic networks but are less interpretable than the more graph theoretically and sociologically grounded blockmodels.

In general the results show common patterns across two investigated MOOC discussion forums that are in line with existing research reporting a separation of MOOC forum communities into active core users and a majority of peripheral less active users (Wong, Pursel, Divinsky, & Jansen, 2015a), which empirically validates the developed approaches. Additionally further insights into the development of actor roles over time and interdependencies between these roles, as well as specific characteristics on the community level can be shown. In particular, the applied network based approaches do not only reflect the activity of individual users but also the evolution of the social structure that emerge from the forum communication.

1.2.3 Combination of semantic and social structures

The combination of the semantic and social perspective of actor-artefact relations is based on the basic assumption that individual characteristics of actors and the topology of the underlying social network adapt to each other. Thus, in a broad sense the work summarised in the following can be located in the field of adaptive networks. Theoretical models of those adaptive network where co-evolution of individual node properties and network topology takes place are intensively studied in statistical mechanics and complex systems research (Gross & Blasius, 2008). Evidence for the interdependence of social and semantic structures is also given by observations in empirical data, such as biological networks or social media (Roth & Cointet, 2010). However, concrete forms socio-semantic network analysis are not well explored in the area of online learning and knowledge creating communities, although its potential for the design of community support mechanisms such as social recommendation (Harrer et al., 2007).

Chapter 6. The work combines the social relations derived from artefact mediated communication and semantic relations based on the interests of actors that were considered separately in Chapters 2-5. In particular, it extends the research on knowledge exchange in MOOC discussion forums reported in Chapter 5, where role models of forum users were created solely based on social connections, by incorporating content analysis of the users' textual contributions. This allows for studying possible interdependencies between the social relations based on information exchange and the thematic orientation of forum users in an integrated framework. Similar to Chapter 4, the thematic affiliations of forum users are derived by automatic extraction of the main concepts from the produced discussion content. Thus, the chapter deals with all the three aspects of online communities described in Section 1.1.1, namely the social aspect, the production aspect, and the affiliation aspect.

Especially with the advent of large scale collaboration in online courses there is a need to understand the potential complex social and semantic mechanisms of knowledge exchange in

online courses especially with respect to the design of collaborative elements (cf. Section 1.1.2 “Enabling effective community support”). Simultaneously, more sophisticated datasets became available that allow for studying forum communication on large scale and to develop novel methodological approaches. In this context, the work described in Chapter 6 is part of an emerging line of research in learning analytics that is especially concerned with the question how and to what extent learners in MOOCs engage in content related knowledge exchange, for example, by using text classification for the identification of content related discussions based on indicator phrases (Wise, Cui, & Vytasek, 2016). In contrast to those solely content based approaches, the connection patterns between forum contributors are additionally taken into account in this chapter.

In the sense of coherent and self-organised knowledge exchange between peers, it is desirable that the semantic structure of a community i.e. similarity of users in terms of their topics of expertise and topics of lacking experience is reflected by the structure of communication relations between forum users. For this reason, the approach outlined in Chapter 6 determines the social role of forum users by identifying groups of users with similar positions in the information exchange network using graph theoretic notions of similarity as in Chapter 5. In addition, the semantic similarity between two forum users is determined based on the thematic overlap of topics in which they provide information to others and topics in which they ask for information, which further reflects the thematic context of information exchange.

In this way, social and semantic similarity between users in a forum can be correlated as a type of community level analysis assessing to which extent the social and semantic structure of the community is interrelated. The results show that in the investigated discussion forum the semantic and social structures are only moderately correlated indicating an increased need for external discussion support mechanisms. This finding is also supported by recent work of Rosé and Ferschke (2016).

However, since joint evolution of the social and semantic structure is not completely absent. The similarities of users are thus transformed into socio-semantic blockmodels that group actors with similar positions in the social and semantic space yielding an interpretable model to characterise the information exchange between different components of the communication network of forum users. By top down inference, this meso-level group structures can be interpreted as different roles that can be assigned to individual forum users in addition to the models described in Chapters 2-5.

Apart from contributions to the discourse on information exchange between peers in online learning environments in learning analytics, the chapter also makes methodological contributions to multi-objective blockmodelling (Žiberna, 2014) by describing a flexible approach that is applicable to large datasets in contrast to optimisation approaches (Brusco, Doreian, Steinley, & Saturnino, 2013).

2 Analysis of Dynamic Resource Access Patterns in Online Courses

A first version of this paper was presented at the 4th Conference on Learning Analytics and Knowledge (LAK 2014)¹. As one of the best rated full paper contributions (13 out of 44 submissions), it was selected by the program chairs of the conference for an extended version in the Journal of Learning Analytics. This journal has been established by the Society of Learning Analytics² (SoLAR) as an interdisciplinary network of leading scientists in the field of learning analytics as the first journal that explicitly target the field. The journal version, which constitutes this chapter, was accepted in 09/2014 and extends the original conference paper significantly by applying the analysis methodology to an additional dataset and a more extensive evaluation.

Hecking, T., Ziebarth S., & Hoppe H. U. (2014). Analysis of Dynamic Resource Access Patterns in Online Courses. *Journal of Learning Analytics (JLA)*. 1(3), 34-60. (<http://www.learning-analytics.info/journals/index.php/JLA/issue/view/336>). ISSN 1929-7750 (online).

The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

Author	Contribution	%
Tobias Hecking	<ul style="list-style-type: none">- Conceptualisation of the approach.- Data collection and cleaning.- Implementation of algorithms.- Evaluation.	65%
Sabrina Ziebarth	<ul style="list-style-type: none">- Statistical evaluation.- Interpretation of results as a tutor and co-designer of the blended learning course.	20%
H. Ulrich Hoppe	<ul style="list-style-type: none">- Supervision and advice in the conceptualisation phase and contextualisation.- Interpretation of results as lecturer of the blended learning course.	15%

¹ Hecking, T., Ziebarth, S., & Hoppe, H. U. (2014). Analysis of dynamic resource access patterns in a blended learning course. In *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (pp. 173-182), Indianapolis, IN, USA, ACM.

² <https://solaresearch.org/>

Analysis of Dynamic Resource Access Patterns in Online Courses

Tobias Hecking, Sabrina Ziebarth, H. Ulrich Hoppe

University of Duisburg-Essen, Germany

hecking@collide.info

ABSTRACT: This paper presents an analysis of resource access patterns in two recently conducted online courses. One of these has been a master level university lecture taught as a blended learning course with a wide range of online learning activities and materials, including collaborative wikis, self-tests, and thematic videos. The other course has been offered in the form of a MOOC. As a specialty of this course, master level students from two different universities could participate as a regular university class and receive credits for successful completion. In both courses, online learning resources such as videos, scientific literature, and wikis played a central role. In this context, the motivation for our research was to investigate characteristic patterns of resource usage of the learners. In order to gain deeper insights into the usage of learning materials, we have adapted methods from social network analysis and applied them to dynamic bipartite student-resource networks built from event logs of the students' resource access. In particular, we describe the clustering of students and resources in such networks and propose a method to identify patterns of the cluster evolution over time.

KEYWORDS: Learning resources, MOOCs, SPOCs, social network analysis

1 INTRODUCTION

Learning materials play an important role in online courses. When there is no or very limited direct communication between learners and a knowledge conveying person (such as a teacher or tutor), learners have to be provided with a set of learning resources that enable self-directed acquisition of knowledge. To achieve this, traditional reading materials and lecture videos can be accompanied by quizzes for self-assessment and by wikis or forums to support collaborative writing, peer reviewing, and discussing the material (Ziebarth & Hoppe, 2014). The combination of the aforementioned resource types and activities provides the typical setting of what we call a “resource-intensive online course.”

This paper presents a study on the student-resource interactions in two of such online courses, both using Moodle¹ as a learning management system (LMS). Network analysis methods were used to investigate the dynamics of relations between students and resources in order to identify characteristic patterns of the courses. The first case study focuses on a resource-intensive blended master-level university course. While in previous instances of these lecture the LMS was only used to distribute lecture slides and exercise assignments, this course followed a new approach inspired by the concept of small private online courses (Fox, 2013) that combine elements from massive open online courses (MOOCs) with traditional university classes. The course took up elements of both xMOOCs and connectivist MOOCs (cMOOCs) (cf. Rodriguez, 2013), including the supply of lecture videos, reading materials, weekly online exercises, quizzes, and discussion forums as well as collaborative activities such as the production of wiki articles based on group writing and peer reviewing activities. These

¹ <https://moodle.org/>

(2014). Analysis of Dynamic Resource Access Patterns in Online Courses. *Journal of Learning Analytics*, 1(3), 34–60.

collaborative online activities completely replaced the previous presence-based and primarily teacher-centric exercise. The course was not opened to the outside, thus there was no “massification” beyond the expected audience of about 40 students from four M.Sc. programs, all related to computer science and interactive media. In a second case study, the same method was applied to data gathered from an online course designed as a relatively small MOOC with 173 participants on computer-mediated communication (CMC). The focus of resource usage in this course was more on videos and reading material, which led to different resource access behaviour in comparison to the first case study.

The contributions of this paper are twofold: first, a new method for learning analytics is introduced that allows identifying and tracing cohesive clusters of students and resources over time; second this method is applied to the two aforementioned case studies.

Based on the log protocols of resource access of the students, relations between students and learning resources are modelled as bipartite networks. In such networks, a link between a student and a learning resource can be interpreted as interest of the student in this resource. This interest either can be intrinsically motivated — for example, by curiosity — or externally stimulated by assignments and course design. Cohesive clusters of students and resources indicate a concentration of interests and activities around certain materials by specific subgroup of students in a given time segment. This does not necessarily imply a direct interaction between students, although this may be the case especially for content created by groups of learners (such as wiki articles). The tracking of the evolution of such mixed clusters of students and learning resources further describes how the affiliations between students and resources change over time. This information can be used to characterize different online courses according to the specific evolution of resource usage, e.g. courses with homogeneous or diverse resource usage.

The proposed approach was applied to two case studies to gain insights into the resource usage in different kinds of online courses. The main research questions are as follows:

- Do students use content produced by peers, or do they prefer the materials provided by teachers?
- Which resources are important for which groups of students?
- How do the relations between students and learning resources evolve over time?
- Are there differences regarding the resource access patterns of students of different study programmes participating in the same course?
- How can the evolution of resource access over time be characterized in different online courses?

Based on the answers to these questions, we hope to improve the design of future instances of the courses.

The rest of the paper is structured as follows: Section 2 describes the theoretical background concerning the role of learning resources and activities in general and more specific in the context of online learning. In section 3, we introduce our analysis methods before we explain the analysis of the mentioned case studies in section 4. Section 5 concludes the main results and gives an outline of possible future research directions.

2 BACKGROUND

2.1 The Role of Learning Resources in Online Courses

Online learning environments and their ability to facilitate distributed learning activities enable a wide variety of course designs, especially in academic education. More than a decade ago, Hill and Hannafin (2001) introduced the concept of resource-based learning environments (RBLEs), where learners are provided with various static and dynamic learning resources in conjunction with tools for manipulation and search in order to satisfy different learning needs. Today, massive open online courses (MOOCs) constitute a current trend in practical applications of technology-enhanced learning. Because of the large number of participants in such online courses, the personal supervision of learners by teachers is not possible. Therefore, lessons are given as video lectures. Exercises often comprise online quizzes/tests that can be automatically evaluated as well as other tasks reviewed by peers (e.g., see Belanger & Thornton, 2013). While in xMOOCs students mainly act as consumers of predefined learning contents, connectivist MOOCs (cMOOCs), as introduced by G. Siemens and S. Downs (c.f. Fini, 2009), focus on the production and sharing of knowledge artefacts like blogs or wiki articles by and between learners (Rodriguez, 2013). Apart from targeting a broad audience, the shift of learning activities into online environments can be of advantage even on smaller scale (Bruff et al., 2013; Harrer et al., 2007). Thus, “blended learning courses” provide hybrid approaches that combine the benefits of online learning and face-to-face sessions. Such course designs enable learners to consume learning resources and share and discuss results asynchronously — independent of time and place — while still having classes/events with physical attendance for community building and supervision (Garrison & Kanuka, 2004). In general, learning resources of various types play an important role in course design regarding the support of autonomous learning. It has also been argued that due to the variety of media and delivery channels online courses are particularly suited to address the needs of different learner types (Grünwald et al., 2013).

2.2 Analysis Methods for Learning Resource Usage

By investigating the relations between learners and learning resources, we expect to gain deeper insight into the function of the learning community. In general, analysis methods for learning resource usage can be characterized in two dimensions, namely structural and sequence analysis.

Modelling relations between learners and the resources they employ can be used to reflect structural characteristics of the resource usage of learners. Nachmias and Segev (2003) showed that a large proportion of the resources provided in a web-based learning environment is used by the students. However, by comparing the resource usage on the level of individuals, large differences regarding the quantity of different resources used could be identified. Hoppe et al. (2005) introduced the concept of social-thematic navigation through the sharing of learner-created “emerging learning objects.” In this view, relations between groups of learners induced by thematically related learning objects indicate learners with a common interest. The learners could interact indirectly mediated by those objects without necessarily having person-to-person communication. Social network analysis methods have also been applied to networks of people and artefacts to support the evolution of a knowledge-creating community in terms of the identification of people with common interests as well as trend analysis (Harrer et al., 2007). The work presented in Romero, Ventura, and García (2008) evaluates data-mining methods for resource usage of students in LMS including statistics, clustering, and classification.

The dimension of sequential analysis is mostly concerned with the behaviour of learners over time. There are, especially, data mining methods used to identify typical sequences of resource access over time (Perera et al., 2009). This also enables the identification of different learning paths for different learners that can be compared to the intended learning path by the teacher (Pahl & Donnellan, 2002; Romero et al., 2008). More recently, Kizilcec, Piech, and Schneider (2013) discovered engagement patterns of learners in MOOCs based on their access to video lectures and assessments in order to classify the users into subpopulations.

The method described in this paper enables structural analysis of networks between students and learning resources. While apart from sequential analysis most methods rely on static snapshots of the state of a learning course, and thus are not able to capture the dynamics, this work explicitly considers the variation of the student-resource access patterns over time. Further, the thematic classification of learning resources used together by groups of students are taken into account, which will be described in section 3.5.

3 ANALYSIS METHODS

3.1 Building Student-Resource Networks from Students Action Logs

Actor-resource networks are bipartite networks. Bipartite networks (Wasserman & Faust, 1994), also called “two-mode” networks or “affiliation networks,” contain two distinct types of vertices. Edges can only occur between vertices of different types.

In the case of bipartite actor-resource networks, one set of vertices represents actors who have connections to a second set of vertices that represent resources. In our study, we investigate bipartite networks of students and learning resources over time. The networks are based on the students’ activities within the learning management system Moodle where the relations of students and resources can be observed by event logs of resource access. Whenever a student accesses a resource (opens a lecture video, for example), an event is generated and stored in the Moodle database. These event logs contain the identification of the student, the name, and identification of the resource, as well as the timestamp of the event. Based on these log files, a time series of bipartite student-resource networks can be extracted by sliding a specified time window over the stream of resource access events. Each network of the series (time slice) contains all relations between students and resources that occurred within the corresponding time window.

3.2 Clustering of Student-Resource Networks

Actors in evolving social networks usually develop a community structure. This means that certain subsets of actors tend to be more densely connected to each other than to outsiders (Backstrom et al., 2006; Watts & Strogatz, 1998). The detection of such cohesive subgroups, or subcommunities of actors in a social network, is a common task in social network analysis (Fortunato, 2010). Subcommunity detection methods can partition a graph into disjointed sets of nodes while maximizing the modularity between clusters (Girvan & Newman, 2002). Other methods do not cluster all nodes exhaustively and allow overlaps between clusters so that nodes cannot necessarily be assigned to a subgroup uniquely. A

prominent example for methods that detect overlapping subgroups is the Clique Percolation Method (CPM) (Palla et al., 2005).

In this study, however, we investigate bipartite graphs where students are connected to learning resources they accessed during a certain period. In such actor-resource networks, it is also possible to identify densely connected substructures. Thus, a bipartite cohesive subgroup, also called bipartite cluster, is a bipartite subgraph where a set of actors is more densely connected to a set of resources than to resources outside the cluster. This can help researchers to understand which students are interested in which resources, and how such affiliations change over time.

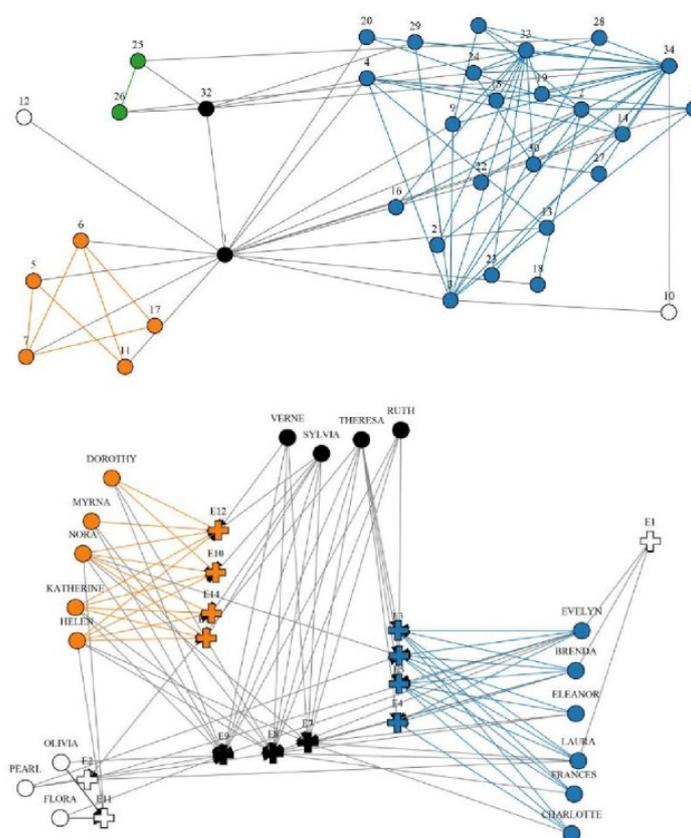


Figure 1: Comparison of clustering of a one-mode social network and a bipartite actor-resource network. The node shapes indicate different node types. Black nodes are in more than one cluster.

In comparison to clusters in one-mode networks, a bipartite cluster $BC = \{A, R\}$ consists of two parts, namely a set of actors (students: $A = \{a_1, a_2, \dots, a_n\}$) and a set of resources ($R = \{r_1, r_2, \dots, r_m\}$). Figure 1 shows the difference between clusters in one-mode and two-mode networks. For example, the orange bipartite cluster in the bottom part of the figure contains the actor group $A = \{Dorothy, Myrana, Nora, Cathrin, Helen\}$ and the resource group $R = \{E10, E11, E14, E15\}$. In contrast to subgroup detection in one-mode social networks, where social ties between actors are directly observable, the actors within a bipartite actor-resource cluster are not necessarily connected by social relations. Moreover, the actors share a common interest in or awareness of the resources of the bipartite cluster. In the following, the set of actors in a bipartite cluster is called “actor group of the

cluster” and the set of resources is called “resource group of the cluster.”

The methods for subgroup detection in one-mode networks mentioned above cannot be applied directly to identify bipartite subgroups. One modification of a subgroup detection method that works for bipartite graphs is based on the Clique Percolation method (Palla et al., 2005). The Biclique Communities method (Lehmann, Schwartz, & Hansen, 2008) relies on the definition of a $K_{a,b}$ biclique. This is a maximal connected bipartite subgraph with a nodes of the first mode and b nodes of the second mode. Thus, if a set of a actors all are connected to each of the b resources, they form a $K_{a,b}$ biclique. A bipartite subgroup also called biclique community in Lehmann, Schwartz, and Hansen (2008) is defined as the union of a series of adjacent $K_{a,b}$ bicliques. Two $K_{a,b}$ bicliques are adjacent if they share at least $a - 1$ nodes of the first mode and $b - 1$ nodes of the second mode. Figure 2 depicts an example of two adjacent $K_{2,2}$ bicliques. Since only nodes that are part of a $K_{a,b}$ biclique are assigned to clusters, the method produces non-exhaustive clustering. Further, it is possible that nodes are assigned to more than one cluster, which results in overlapping clusters.

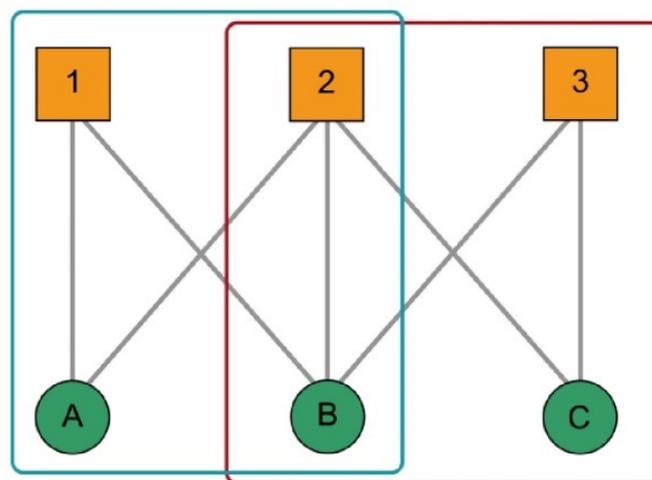


Figure 2: Example of two adjacent $K_{2,2}$ bicliques. First clique: {A, B, 1, 2}, second clique: {B, C, 2, 3}.

3.3 Detection of Evolution Patterns in Student-Resource Networks

The analysis of a static snapshot of a social network or actor-resource network can give various insights into their structures (Wasserman & Faust, 1994). However, since most networks evolve over time, the community structure also changes. In section 3.1 the transformation of event logs of resource access into subsequent snapshots (time slices) of an evolving bipartite network has been described. It is likely that a subgroup detection method will uncover different subgroups in each of the time slices. Since subgroups do not appear and disappear at random, there can be similar subgroups in consecutive time slices. This phenomenon has recently been studied in social network analysis (Leskovec, Kleinberg, & Faloutsos, 2005; Palla, Barabasi, & Vicsek, 2007). In social networks, a group of people may stay in contact for a certain period. New members may join the group and others may leave at certain points but a stable majority is considered as the same subgroup over several points in time. In such situations, one can speak of the same subcommunity with different instances in different time slices or snapshots of a dynamic network.

Palla, Barabasi, and Vicsek (2007) define six basic lifecycle events for subcommunities in one-mode social networks. These events are:

- Birth: A subcommunity is identified the first time.
- Growth: A subcommunity acquires new members but its core stays the same.
- Contraction: A subcommunity loses members over time.
- Merge: The members of distinct subgroups merge to one subgroup at a later point in time.
- Split: One subcommunity splits into two new sub-communities.
- Death: A subcommunity disappears over time.

In bipartite networks of students and learning resources, one may also find patterns of the evolution of student-resource clusters. The difference to one-mode networks is that in bipartite networks the previously described events can happen to the actor and resource groups of the bipartite clusters separately. By applying a bipartite subgroup detection method to different slices of an evolving student-resource network, one can identify traces of dynamically evolving actor groups and resource groups. Figure 3 depicts an example of such traces. The circle- and the square-shaped nodes represent actor and resource groups respectively. Each group of actors forms a bipartite cluster with a group of resources at a certain point in time, indicated by a dotted connection. According to the change of affiliation of students to resources, the detected bipartite clusters can change over time. However, similar actor groups (student groups) can be detected over time as part of different bipartite clusters. Thus, one can identify relations between similar actor or resource groups of subsequent time slices. These relations are depicted as arrows in Figure 3.

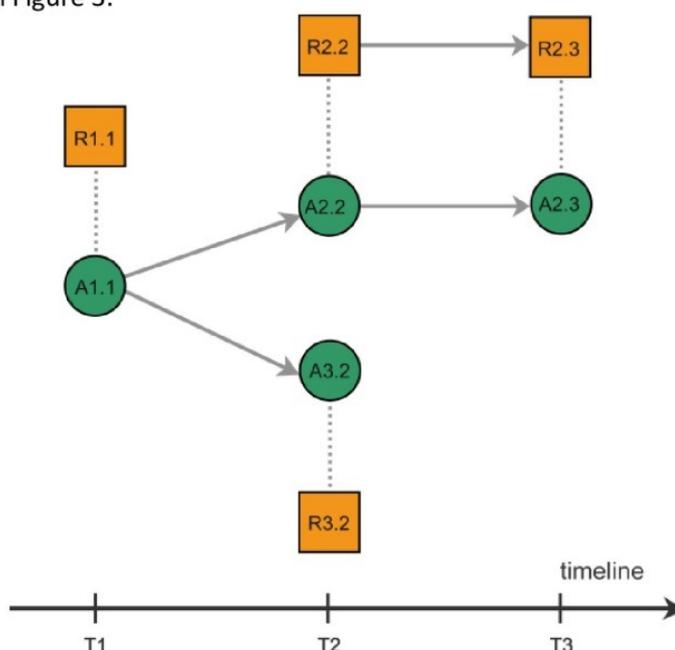


Figure 3: Traces of dynamic actor (green circles) and resource groups (orange squares) of bipartite clusters over time.

Table 1 shows those relations that allow the tracing of actor and resource groups over time, which results in a timeline as in Figure 3.

3.4 Tracing the Evolution of Dynamic Actor and Resource Groups

Given a list of subsequent time slices of k evolving actor-resource networks (G_1, G_2, \dots, G_k), the first step is to apply the Biclique Communities method, described in section 3.2, to each of these graphs. The result is an ordered list of k sets of bipartite clusters (or “biclique communities”) for the k time slices ($\Theta_1 = \{BC_{1,1}, BC_{1,2}, \dots\}, \Theta_2 = \{BC_{2,1}, BC_{2,2}, \dots\}, \dots, \Theta_k = \{BC_{k,1}, BC_{k,2}, \dots\}$). As explained in section 3.2, each of the bipartite clusters $BC_{i,j} = \{A_{i,j}, R_{i,j}\}$ contains an actor group and a resource group. In the context of the tracing algorithm, the actor groups and resource groups of particular time slices are called “step-actor groups” and “step-resource groups” respectively.

Similar step-actor/resource groups occurring in subsequent time slices are assumed to be instances of the same dynamic group. Thus, a “dynamic actor group” is a timeline of similar actor groups and a “dynamic resource group” is a timeline of similar resource groups. Figure 3 depicts the traces of a dynamic actor and a resource group with step-groups as particular instances. By tracing step-groups of actors and resources over several time slices, different lifecycle events, or evolution patterns, of bipartite clusters can be identified (see Table 1).

To identify the described events, a method is needed that matches the step-groups of the time slices with step-groups of previous time slices. Although there are methods for tracing the evolution of dynamic subcommunities in one-mode social networks (Asur, Parthasarathy, & Ucar, 2009; Greene, Doyle, & Cunningham, 2010; Palla, Barabasi, & Vicsek, 2007), to our knowledge there are no such methods for bipartite networks that allow the identification of the evolutionary events in Table 1. In the following, an adaption of the method of Greene, Doyle, & Cunningham (2010) is introduced, which was originally designed for one-mode networks to trace the subgroup structure in bipartite networks like learner resource networks:

Initialization:

The method maintains two sets of not matched groups, one for not matched step-actor groups (NMA) and one for not matched step-resource groups (NMR). At the beginning of the process, the set NMA is the set of step-actor groups of the first time slice — $T_1(NMA = \{A_{1,1}, A_{1,2}, \dots\})$ — and NMR contains all step-resource groups — ($NMR = \{R_{1,1}, R_{1,2}, \dots\}$).

Matching phase:

After the initialization, the algorithm proceeds with the step-groups of the next time slice, T_i . The next step is to compute the similarity of the step-actor groups of T_i with all step-actor groups in NMA and the similarity of the step-resource groups in T_i with the step-resource groups in NMR . As in Greene, Doyle, and Cunningham (2010) and Palla et al. (2007) the Jaccard coefficient is used as measure of similarity between two groups:

$$sim(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Here, X and Y are either two step-actor groups or two step-resource groups. If $sim(X, Y)$ exceeds a certain threshold, the groups will be matched.

There are four different situations that can occur when matching the current step-groups in T_i to the not

matched step-groups in *NMA* and *NMR* in T_{i-1} .

1. One-to-one match between a step-group of the current time slice and a previously not matched group in *NMA* or *NMR*. In this case, the two groups belong to the same timeline of a dynamic group. In Figure 3 this is the case for the dynamic resource group R_2 , which occurs in T_2 and T_3 .
2. One-to-many match: One not previously matched group in T_{i-1} matches to more than one group of the current time slice T_i . This indicates that the not previously matched group splits up into the matched step-groups of T_i . The matched step-groups of T_i are then considered as the first occurrences of a new dynamic group. In Figure 3, the actor group $A_{1,1}$ in timeline 1 matches with the two actor groups $A_{2,2}$ and $A_{3,2}$, which will then be considered instances of new dynamic actor groups A_2 and A_3 .
3. Many-to-one match: More than one not previously matched group in *NMA* or *NMR* match to the same step-group of T_i . This means that these groups merge to the matching group of T_i . The fourth column of Table 1 is an example of such a situation.
4. A step-actor or step-resource group of T_i cannot be matched to any of the groups in *NMA* or *NMR*. In this case, such a group will be the first instance of a new dynamic group. An example for this is the resource group $R_{2,2}$ in Figure 3.

After all step-groups of the currently investigated time slice T_i were checked, all groups from *NMA* and *NMR* that could be matched with one or more step-groups of the currently investigated time slice T_i will be removed from these sets. All step-groups of the current time slice T_i are then added to *NMA* and *NMR* respectively.

Then the algorithm moves on to the set of groups in the next time slice. The algorithm stops when the last time slice is processed. The result can then be visualized as a swim lane graph, as in Figure 3. In this visualization, a student-resource cluster is represented by two nodes. The square nodes represent the resource groups of the bipartite clusters while the circle nodes represent the student (or actor) groups. A student group node is always connected to a resource group node by a vertical line. This indicates that the particular student group forms a bipartite cluster together with the resource group in a corresponding time slice. As described above, the different groups found in one time slice of the evolving student-resource network can be re-identified in future time slices by the described similarity-based matching condition. Visually, groups matched across subsequent time steps are connected by a horizontal arrow. Hence, these arrows always point forward in time and allow the depiction of merge, split, and continuing student and resource groups.

Table 1: Possible events in bipartite cluster evolution over time.

	<p>A group of actors and their affiliations to resources stays stable over subsequent steps in time.</p>
	<p>A group of actors shift their interest to new resources.</p>
	<p>A new group of actors (A2) forms around an existing set of resources.</p>
	<p>Groups A1 and A2 merge and share their affiliations to resources.</p>
	<p>A group of actors (A1) splits into two new groups and no longer shares their affiliations to resources.</p>

3.5 Fostering Cluster Interpretability through Content Analysis

The method described above can be used to track the general resource access behaviour of students over time by clustering students and learning resources simultaneously. In order to interpret the results, one must investigate the content of the clusters. This step can also be supported by computational methods. Cluster labelling (Treeratpituk & Callan, 2006) is one technique commonly used in document clustering to assign keywords, which helps to interpret the content of the clusters. This idea can be adapted to label the resource groups of student-resource clusters by utilizing the metadata of the resources. The cluster labelling in this work follows a semi-automatic approach. Most of the learning resources in the online courses under investigation have assigned metadata such as keywords. Videos can be annotated with tags. Most scientific literature is also annotated with keywords. However, quizzes, lecture slides, and wiki articles are not necessarily described by keyword lists. In these cases, the course managers need to annotate these resources manually.

The resource groups of the student-resource clusters can then be labelled using a simple frequency-based approach. A keyword is assigned to a resource group if it appears in the metadata of at least 70 percent of the resources in the group. In a case where no keyword has support above the threshold of 70 percent, the most frequent keywords are used instead.

Given an example of a student-resource cluster $BC = \{A, R\}$ with a resource group $R = \{r1, r2, r6, r7\}$ from the blended learning course, which will be described in more detail in section 4.1, each resource in the cluster can be interpreted as a list of keywords (see Table 2). Keywords can either be automatically extracted from the resource metadata or added by hand by the course manager. The result for the given example would be the cluster label “ITS” because three of the four resources have this keyword in common. If needed, an analyst can refine this initial labelling by hand.

Table 2: Example of resources and allocated keywords. Italic keywords indicate manually assigned keywords. Other keywords could be directly extracted from resource metadata.

Resource ID	Resource name	Keywords
r1	ITS slides	ITS, <i>Pseudo Tutors</i>
r2	Constrained based tutors (wiki)	<i>Cognitive models, student modelling</i>
r6	Intelligent Tutoring Systems (part 3)	ITS, constraint based tutors
r7	Ragnemalm — “Student diagnosis in practice; Bridging a gap”	Student diagnosis, student modelling, ITS

This semi-automatic cluster labelling usually has the advantage of being more accurate than fully automatic labelling. The keyword table can be initially filled in by the extracted keywords from the resource metadata. A teacher or course manager knows the resources used in the course very well and thus can refine the keyword tables by adding keywords and unifying synonyms like “ITS” and “Intelligent Tutoring Systems.”

4 CASE STUDIES

4.1 The GILLS Blended Learning Course

GILLS (“Gestaltung interaktiver Lehr- / Lernsysteme,” or “Design of interactive teaching and learning systems”) is a master level course in the Department of Computer Science and Applied Cognitive Science at the University of Duisburg-Essen (Germany) that focuses on user/task modelling and models of interactive and collaborative learning environments. It targets students from the three Master’s programmes of “Applied Computer Science,” “Interactive Media and Cognition,” and “Computer Engineering” and also higher education teachers/students of computer science. The course is taught once per year with about forty participants. It consists of a weekly three-hour lecture accompanied by a one-hour exercise with weekly assignments.

Probably due to the given schedule of the course, attendance at the exercise was typically lower than at the lecture and the willingness for active participation (i.e., presentation of results on the part of students) was also low. Thus, in summer semester of 2013, the exercise was “virtualized” when the presence-based activity was replaced by a combination of individual and group assignments with reporting back incorporated into the Moodle platform (Ziebarth & Hoppe, 2014). The plenary face-to-face setting of the lecture was used as a forum to take up and discuss results from the virtual exercise activities. The role of the lecture was redefined in the sense of providing orientation knowledge (“advance organizer”) and core definitions with a limited number of characteristic examples, leaving specific extensions to the virtual exercise activities. These lecture elements were provided as smaller thematic video clips, available on the platform. Evidently, this new approach was inspired by the concept of MOOCs and takes up elements of both xMOOCs and cMOOCs while being neither massive nor open.

The Moodle interface contained one forum for announcements and one for questions and discussion for the participants. The students were asked to post questions relevant to the whole course on the forum rather than to write emails to the supervisors. Additionally, the forum was used to support group exercises. Furthermore, a wiki was introduced to build a glossary for repetition and exam preparation. A typical exercise regarding a wiki article consisted of one week of creating the initial version in a small group, one week of individually reading the articles of other groups and writing peer feedback, and one week of revising the article in the original small group. Each group handled a different topic, enabling students to specialize regarding their interests. Thus, groups were formed by topic using Moodle’s choice activity, which was set up to ensure groups of similar size. Furthermore, students could see the choices of their peers to enable group formation and communication.

Overall, the online exercises included (per student):

- Creating and revising four wiki articles in small groups
- Two tasks regarding constructive problem solving and modelling (partially using external tools) in small groups
- Writing peer reviews (as wiki comments) regarding the group results above (individual task)

(2014). Analysis of Dynamic Resource Access Patterns in Online Courses. *Journal of Learning Analytics*, 1(3), 34–60.

- Creating and extending a concept map regarding the topics of the first chapter using external tools (group task)
- Three quizzes regarding the three main topics for self-testing, which could be completed several times (individual task)
- Creating questions and sample solutions in a wiki regarding the lecture topics (individual task)
- Playing the serious game Matchballs (Ziebarth, Malzahn, & Hoppe, 2012) for training relations between important concepts of the lecture (individual task).

To motivate students for participating in the online exercises, bonus points were introduced.

The Moodle course contained the following types of resources:

- Lecture videos
- Lecture slides
- Additional reading material like scientific papers
- Assignments including self-test quizzes
- Wiki articles
- Forum discussions

4.1.1 General analysis

Forty-eight students visited the first lecture of the GILLS course; of the 44 students doing the first exercise, 40 regularly did their exercises until the end of the course and 36 did the oral exam.

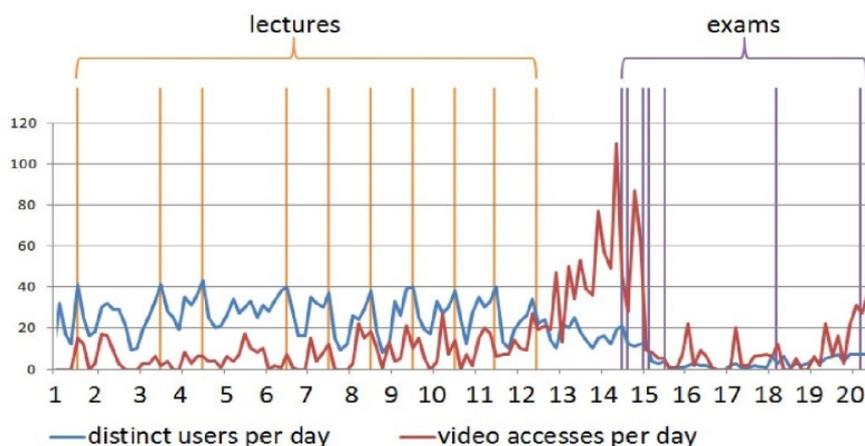


Figure 4: Moodle usage over time.

The Moodle course was frequently visited: There were 24 distinct students per day and 45 per week on average (see also Figure 4). The highest number of distinct users per day was on lecture days, the lowest was on weekends. The deadline for and assignment of new exercises was typically on lecture days, which may explain the high number of accesses on these days.

Figure 5 shows the relations of students in the course based on group choices. The edge width indicates

how often students were in the same group; the darker the edge, the stronger the relation.

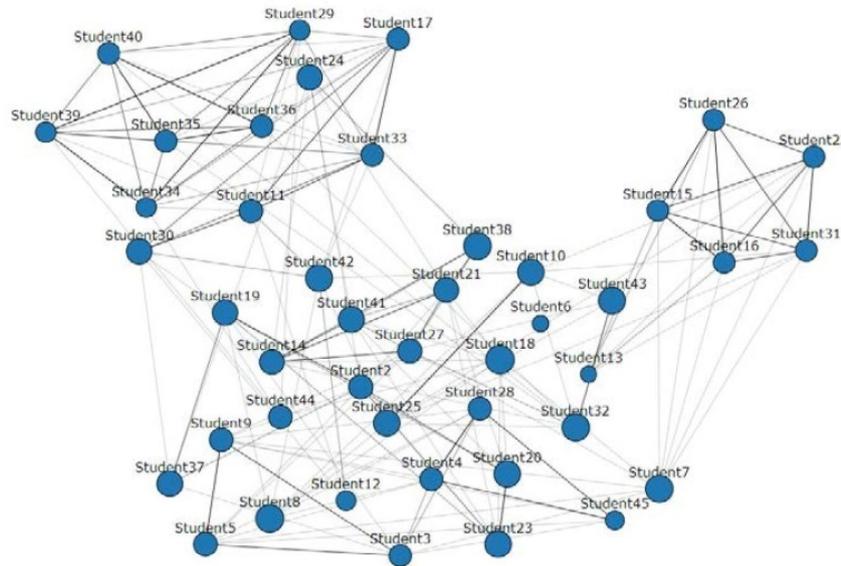


Figure 5: Student relations based on group choices.

The node size represents the centrality of the students; the bigger a node, the more relations to other students. On the one hand, the network shows some groups of students often working together in the same constellation (in the periphery of the graph). These groups often consist of students of the same study programme. On the other hand, there are students without a stable affiliation who work with a high number of different students (in the centre of the graph). These might be the ones choosing an exercise/group by topic rather than for social reasons.

4.1.2 Overlapping clusters of student and learning resources

By applying the Biclique Community method (see section 3.2) to the student resource networks generated from the resource access logs of the students, overlapping bipartite student-resource clusters were retrieved. Learning resources that frequently appear in overlaps between bipartite clusters may be of particular importance because students of different groups have relations to them.

As a prototypical example for the bipartite clustering of students and resources, Figure 6 depicts the subgroup structure of the student–resource network in the 10th week of the lecture period, where the topics were “Intelligent Tutoring Systems” and “Cognitive User Modelling.” The exercise assignment for this week was to create wiki articles on these topics. It is worth mentioning that the applied Biclique Communities method for clustering does not cluster the nodes in a network exhaustively. Nodes that do not belong to a $K_{a,b}$ clique (see section 3.2) cannot belong to a cluster. Students who cannot be assigned to bipartite clusters uniquely either have a broader interest in learning resources than other students or do not access enough resources to be part of a $K_{a,b}$ biclique. In this case, $K_{3,5}$ bicliques were used to construct the clusters. Hence, Figure 6 only shows those students and their relations to resources that are part of detected bipartite clusters. The node sizes in Figure 6 correspond to the number of connections of a student or a resource respectively.

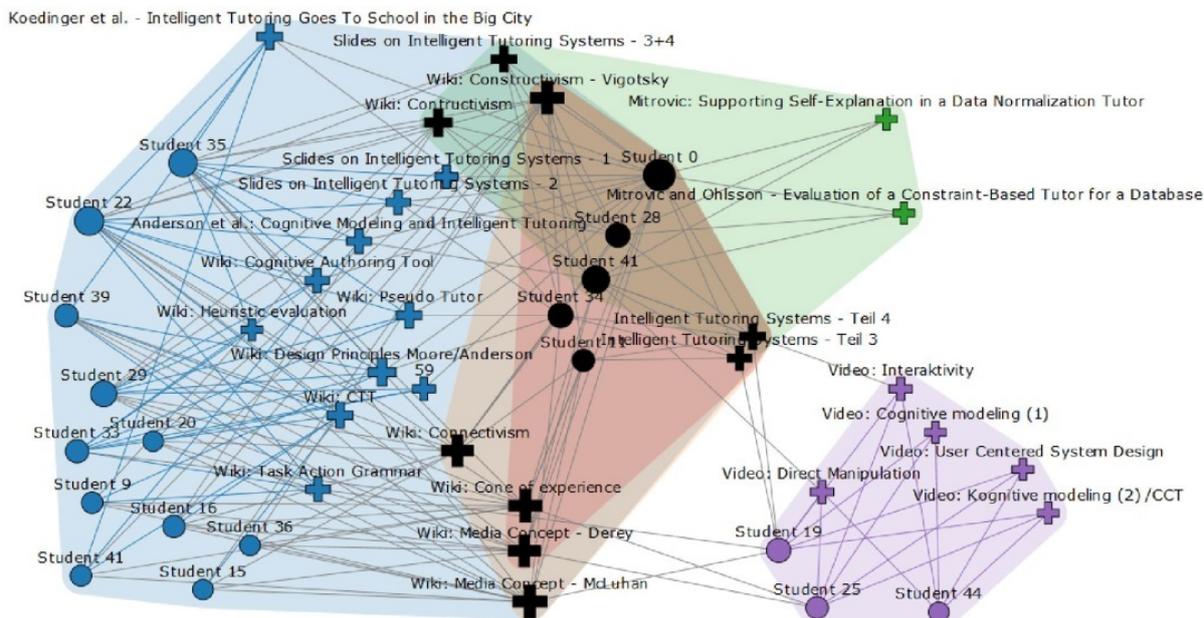


Figure 6: Overlapping bipartite subgroups in the 10th week of the lecture period. Black nodes are in multiple subgroups.

The light red cluster mainly contains the wiki articles created in the previous week, thus, the included students can be considered to have accessed them for creating peer reviews. This cluster overlaps completely with other clusters, since writing peer reviews was only part of the exercise. The second part was writing wiki articles on ITS (two-week task). There were two articles on “Constraint Based Tutoring” (CBT) and two on “Model Tracing Tutors” (one on “Cognitive Tutors” and one on “Pseudo Tutors”). The green cluster focuses on resources (papers, slides, and videos) regarding CBT. The blue cluster, on the one hand, contains resources regarding Model Tracing Tutors (papers, slides, videos) as well as some of the wiki articles that should be created (Cognitive Authoring Tool, Pseudo Tutor). On the other hand, it contains some of the wiki articles (CTT, Heuristic Evaluation, Task Action Grammar) that had to be corrected in the previous week. Since the deadline of the old exercises and the assignment of new ones are on the same day, these old assignments are also considered in the data.

The only resources in the purple cluster are videos of the first weeks of the lecture that are not directly connected to the topics of the considered time interval (week 10). This cluster does not correspond to any *a priori* expectation. Since video resources play an important role for exam preparation, it might indicate an early activity of this type.

4.1.3 Resource access during the lecture period

The resource access of the students is modelled as subsequent time slices of an evolving bipartite network. The size of the time slices is one week so that the resource access of the students during one lecture week is aggregated in one time slice. The first four weeks of the lecture period were skipped, because at this time there are not many resources accessible in the course. To identify bipartite subgroups in the time slices, the Biclique Communities method described in section 3.2 was applied with the parameters $a = 3$ and $b = 5$. This means that an uncovered subgroup contains at least a students

(2014). Analysis of Dynamic Resource Access Patterns in Online Courses. *Journal of Learning Analytics*, 1(3), 34–60.

and b resources. This choice is reasonable, since most students share some common resources (e.g., lecture slides) and consequently smaller values lead to one big cluster in most of the time slices. Larger values for a and b have the effect that most nodes cannot be clustered because they do not belong to a corresponding $K_{a,b}$ biclique.

The tracing algorithm described in section 3.4 requires a similarity threshold for the matching of subsequent step-groups. This threshold was set to 0.3. This threshold is recommended in Greene, Doyle, and Cunningham (2010) for tracing one-mode clusters, but is also suitable for our method because it allows the detection of merges and splits of dynamic groups but is not too low that matching between step-groups occurs at random.

The results for tracing the bipartite clusters during the lecture period are depicted in Figure 7. For subgroup traces, the swim lane visualization introduced in Halatchliyski et al. (2013) was applied. The horizontal lines connect step-actor and step-resource groups that match between two time steps respectively. The vertical lines connect a step-actor group (circle) with a step-resource group (square), if they belong to the same bipartite cluster in the corresponding week. The node sizes correspond to the contained number of resources or students respectively.

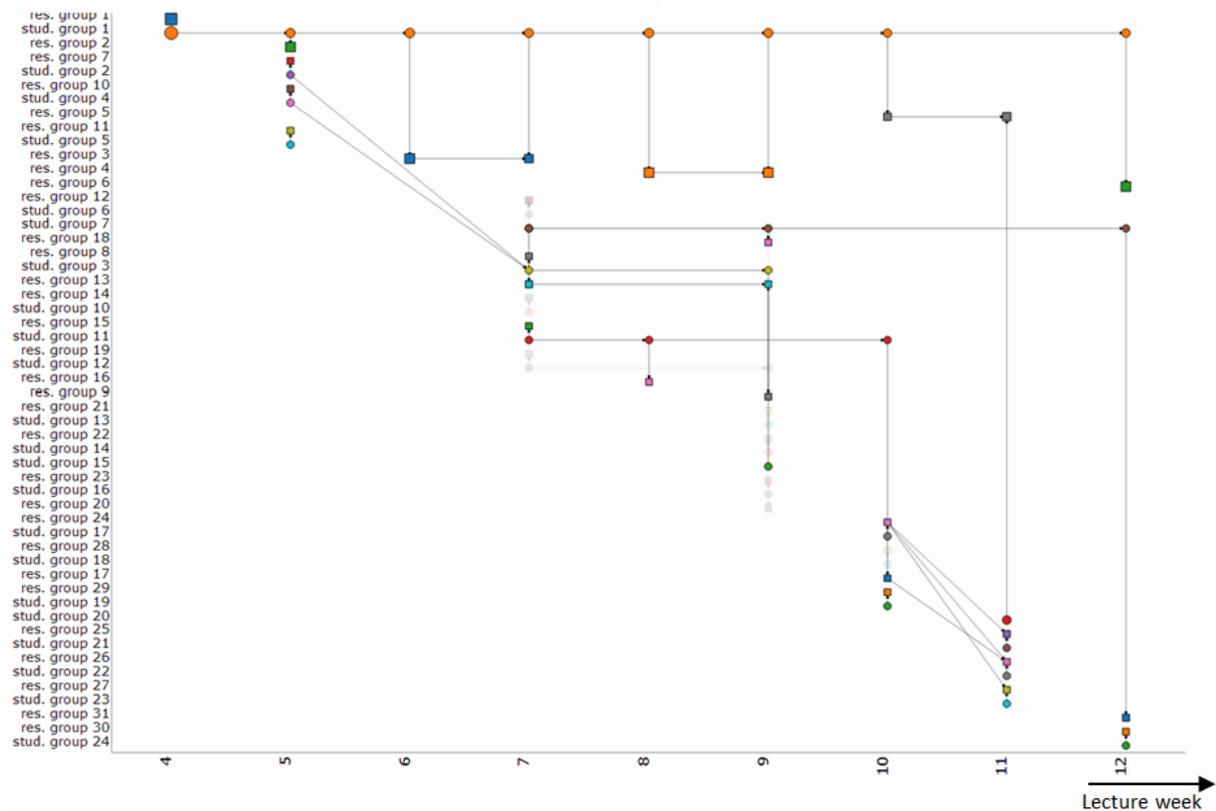


Figure 7: Visualization of the subgroup evolution of the last 9 weeks of the lecture period with interesting traces highlighted.

During the lecture period, there is one large, stable actor group (stud. group 1). This actor group forms different bipartite clusters every week with resource groups containing the core lecture material for the particular week as well as the wiki pages created by the exercise groups. This indicates that the students

(2014). Analysis of Dynamic Resource Access Patterns in Online Courses. *Journal of Learning Analytics*, 1(3), 34–60.

of this group are not only affiliated with the core material and the wiki content created by their own group, but also access the content created by others. This can be seen in Figure 6, where the big blue bipartite cluster depicts student group 1 in week 10 and their affiliations to course resources. Many students of the major student group 1 are also members of smaller actor groups at the same time. Not all of the smaller student groups correspond to the exercise groups that collaboratively edit wiki pages. Moreover, the majority of students use the wiki pages across exercise groups, most likely an effect of the peer review exercises. The smaller groups mostly contain students with close relations to the additional learning material, as well as lecture videos. The patterns described in Table 1 correspond to the “stable group” pattern, described in the first row of Table 1. Student group 1 accesses similar sets of resources in weeks 6 and 7 and also in weeks 8 and 9. In these cases, learning resources of the previous week are still of importance. This shows that the peer reviewing and wiki article revising exercises indeed encourage many students to occupy themselves with the course material of previous weeks. However, the patterns of simultaneous merging or splitting of the actor groups and resource groups of bipartite clusters, described in rows 4 and 5 of Table 1, could not be found.

4.1.4 Resource access during the exam preparation phase

In contrast to the lecture period, exam preparation is self-directed. Thus, in this period, students choose the resources they need for their exam preparation based on their own schedule and assessment. Relevant questions regarding exam preparation are:

- Which resources are often used by certain groups of students?
- What are typical traces of actor groups and resource groups in such informal settings?

Our method can help to answer these questions in order to provide students with additional resources before the exam and to identify those combinations of resources considered relevant by the students. Figure 8 depicts the traces of bipartite student resource clusters during the exam phase. The parameter settings for the clustering and the tracing algorithm are the same as in section 4.1.3. However, the size of the time slices is only four days instead of one week as in 4.1.3. This is due to the more dynamic resource access during the exam preparation phase.

There is one big cluster of learning resources, namely resource group 1. This dynamic resource group is relatively stable, which indicates a set of learning resources of particular importance to students for exam preparation. The different step-resource groups of this dynamic group contain primarily lecture slides, especially on the topics of CSCL, CSCW, and Intelligent Tutoring Systems, which are the last lecture topics relevant for the exam. The majority of these step-groups also contain self-test quizzes. A relatively stable actor group with step-actor groups often forms a bipartite cluster with particular instances of resource group 1. This indicates that a large group of students considers the resources within this group most relevant in the days and weeks before the exam.

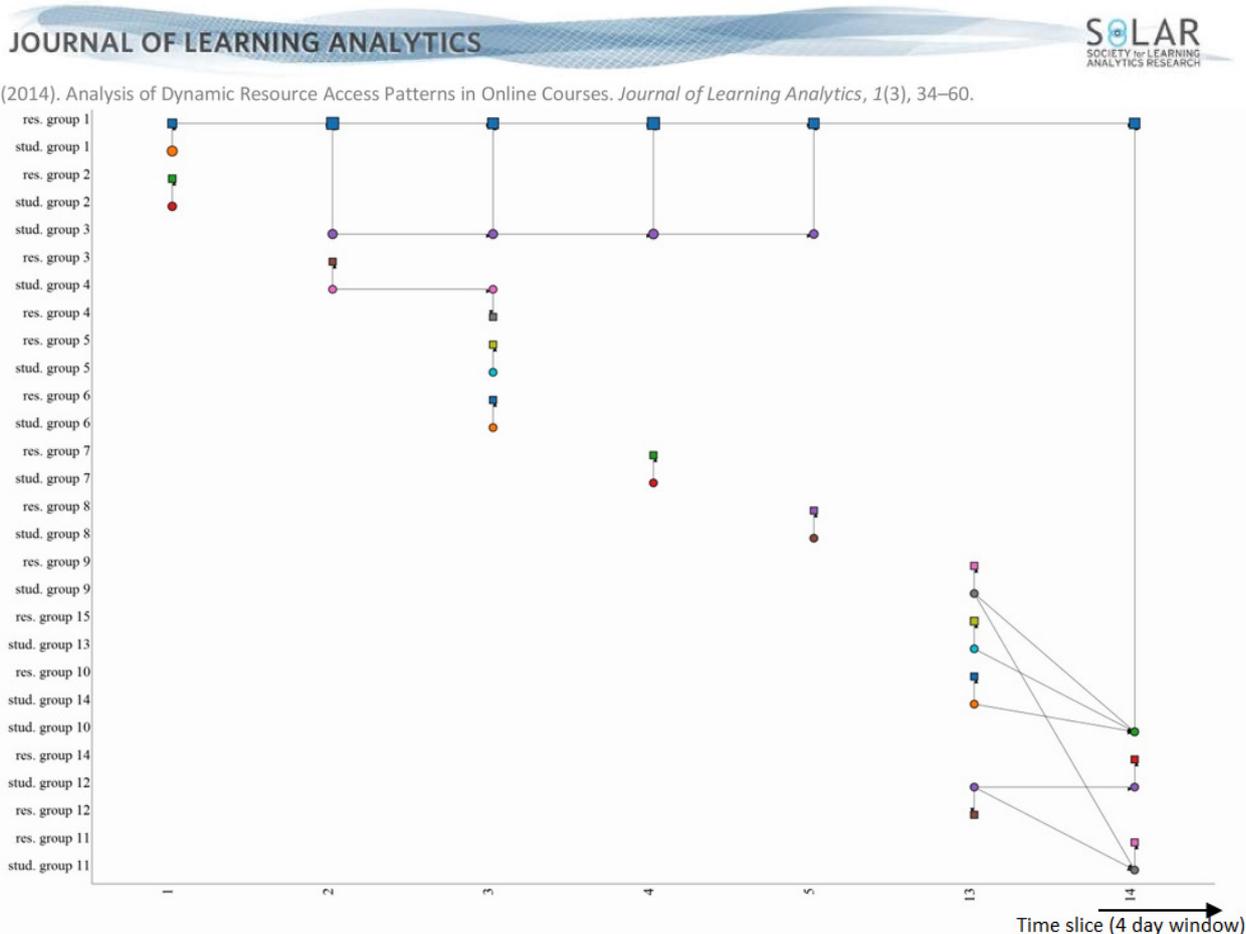


Figure 8: The evolution of bipartite subgroups during the oral examination phase.

As in the lecture period, the found clusters overlap pervasively. Learning resources that occur more than two times in more than one cluster include the following:

- Lecture slides “CSCL”
- Lecture slides “Knowledge Diagnosis”
- Wiki article “Methods and Models for Human Computer Interaction”
- Wiki article “Fitts’ Law”

These resources were used by students irrespectively of the other resources they used for exam preparation.

Between points in time 6 and 12, no clusters can be found, meaning that single students only accessed resources sporadically. This can also be observed in Figure 4 between week 15 and 20, which can be explained by the scheduling of oral exams. While most students had their oral exams in the four weeks after the lecture period, a smaller group in another study programme had exams after eight weeks. Students in the second study programme seem to begin exam preparations eight days before the exam at the earliest, which is much less time than students prepare in the first exam period. Regarding exam results, students from the first group tended to get better grades (1.3 on average) than students from the second group (2.1 on average). Furthermore, the resources accessed by the second group were different at the beginning of their active phase. Instead of a dominating group of students who mainly accessed the lecture slides for exam preparation, as did the students in the beginning of the exam phase, the active students in the late phase of the exam period can be clustered into small overlapping

groups affiliated either with wiki articles or lecture videos (see Figure 9). In the last time slice, most of the students merge to a single group, but instead of forming a bipartite cluster with mostly wiki and video resources, which would correspond to the merging pattern of the 4th column of Table 1, they refocus their learning interest on the “core” preparation material (res. group 1), as can be seen in Figure 9.

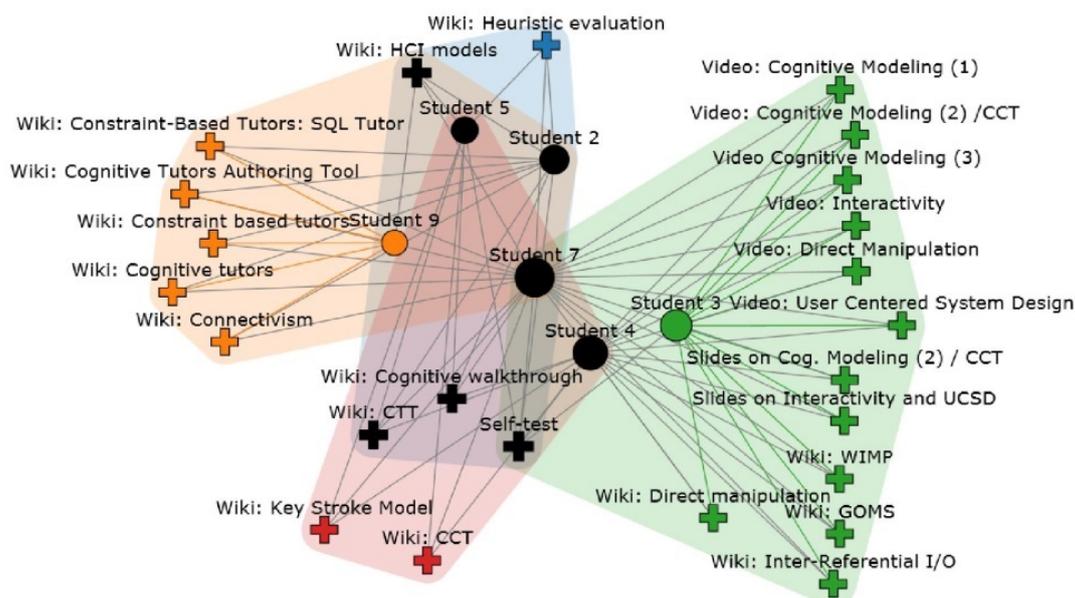


Figure 9: The student–resource clusters in time slice 13 of the exam preparation phase.

4.2 The Computer-Mediated Communication MOOC

The course “Computer-mediated communication” (CMC) was originally designed as a joint blended learning seminar for students from the courses “Applied Communications and Media Science” at the University of Duisburg-Essen and “Pedagogics” at the University of Bochum. In the winter term 2013/14, the obligatory presence sessions were replaced and the course on computer-mediated communication was redesigned as a MOOC, which was also open to participants outside the two universities. With 173 participants in total and a maximal number of 134 active participants per week, the course was not really “massive,” but it had other characteristics of a MOOC, such as video based lectures and extensive online activities. As in the GILLS course, Moodle was the chosen platform for online activities. A distinctive feature of this course was that regular Master’s level students from “Applied Communications and Media Science” at the University of Duisburg-Essen and “Pedagogics” at the University of Bochum had the opportunity to receive regular credits for successful completion of the course. Videos were one important type of learning resource with video messages introducing the main course objectives for each week as well as videos of expert talks and interviews. Additionally, weekly reading assignments included a variety of scientific literature. There were also self-tests and three workshops. For workshop assignments, students had to solve tasks in small groups. In contrast to the GILLS course described in section 4.1, students were not supposed to build groups themselves. Instead, they were grouped automatically by the system.

(2014). Analysis of Dynamic Resource Access Patterns in Online Courses. *Journal of Learning Analytics*, 1(3), 34–60.

Figure 10 depicts the number of active participants per week. The Christmas gap in course activity occurs between weeks 10 and 12. Course activity shows a decrease over time during the progress of the course. This effect is commonly observed in MOOCs (Clow, 2013); however, in this course the decline of active users is not as steep as in large MOOCs since only 14 days of inactivity were tolerated in order to be allowed to participate in the final exam and to receive course credit.

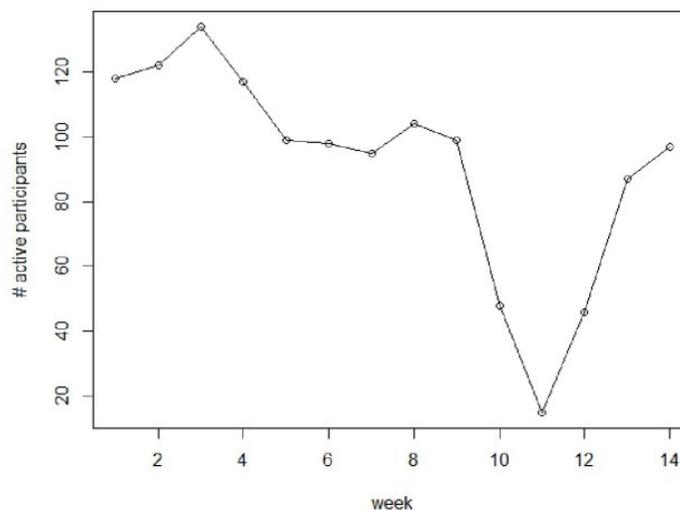


Figure 10: Number of active participants in the CMC MOOC.

4.2.1 Resource access during the course period

As in the analysis of the lecture period of the GILLS course (section 4.1), the bipartite student-resource network was sampled into time slices corresponding to each week of the lecture. Since the CMC dataset contains by far more students but fewer learning resources compared to the GILLS course analysed in section 4.1, the biclique percolation algorithm was applied to the CMC dataset with a different parameterization. The clusters were constructed from $K_{5,4}$ bicliques. This means that only students who used at least four different resources and resources used by at least five different students can appear in a bipartite cluster.

Figure 11 depicts the trace of the identified clusters during the time of the course. Similar to the GILLS course, there is a large relatively stable group of students who accessed a common set of resources over a longer period of time (see student group 1). However, especially in the beginning, many fewer clusters can be found in comparison to the blended learning course. In the first three weeks of the course, the resource access behaviour of active students was very stable such that only one student-resource cluster can be identified. In addition, the set of resources accessed during the first three weeks is very stable. The content analysis of the resources in resource group 1 in the first three weeks of the lecture period reveals that the students in this phase are mainly focused on videos and reading material on general theories of computer-mediated communication. This corresponds to the continuous interest pattern described in the first row of Table 1. The fact that a large group of students used similar resources over several weeks, which leads to the re-identification of a resource group by the proposed algorithm, repeats also in preceding weeks. Thus, a characteristic of the course is that the participants

(2014). Analysis of Dynamic Resource Access Patterns in Online Courses. *Journal of Learning Analytics*, 1(3), 34–60.

deal with the learning material from previous weeks instead of focusing exclusively on the resources of the current week.

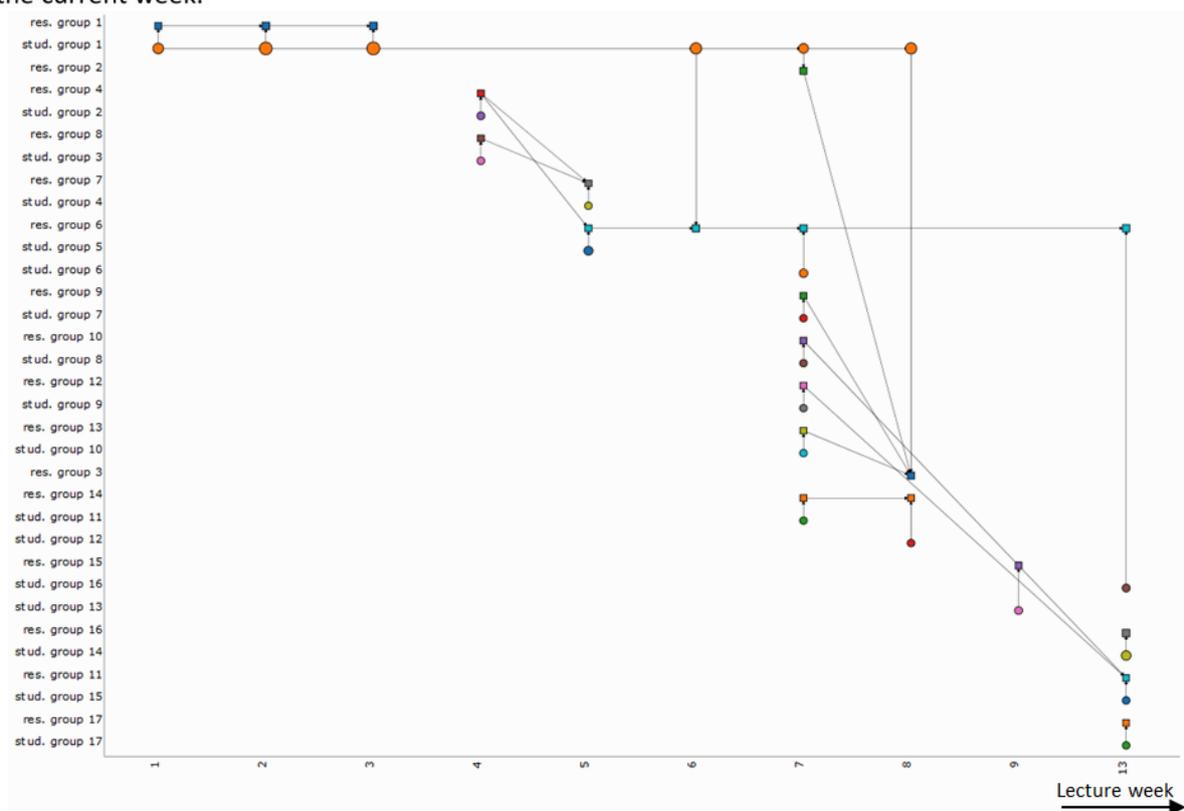


Figure 11: Visualization of the subgroup evolution CMC MOOC.

As shown in Figure 10, there is a continuous decline in the number of active students between weeks 2 and 7, which is also reflected in the swim lane diagram in Figure 11. In weeks 4 and 5, the number of resources used by each student is very low. Many of the students cannot be assigned to clusters because they do not access enough resources. This circumstance will be discussed in more detail in the next section, 4.2.2. In this phase, the large student group 1 disappears. In week 4, only two relatively small and overlapping student-resource clusters can be identified. The topic of the week was about trust and group work. The first cluster, comprising student group 2 and resource group 4, can be labelled with the keywords “distributed groups,” “trust,” and “CMC.” The other cluster includes students and mainly resources on general topics in CMC. Resource group 4 splits into two other resource groups in week 5, namely resource groups 6 and 7. Resource group 7 is labelled again with the keyword “CMC” and resource group 6 with “distributed groups” and “trust.”

Whereas in the first 6 weeks the evolution of the student-resource clusters could be characterized as stable and not very diverse, in week 7 the number of identified clusters increased and the clusters became much more diverse in terms of the resource groups. This may result from the “assignment of the week” where students were asked to create their own exam questions and therefore needed to review the material of the previous weeks. There is also a correspondence to the number of active participants depicted in Figure 10. Beginning with week 7, course activity rose slightly again. This was exactly the point in time where the clustering algorithm detected a higher number of clusters and, at the

same time, less stable clusters. Apart from the special course assignment, another reason for this is that in this late phase of the course, some participants were already preparing for the final exam and therefore accessed additional learning resources apart from the course material of the corresponding week. This can be seen especially in week 13, the last week of the course.

4.2.2 Further analysis of the clusters

As in the results for the GILLS course, the biclique percolation community-detection-method identifies highly overlapping clusters of students and resources. Figure 12 depicts a snapshot of the student-resource clusters of the 7th week of the lecture period where resource access was at its most diverse. It is worth mentioning that this is only a subset of all active students in this week. Since the clusters were built from $K_{5,4}$ bicliques, the clusters contain only those students who were active enough to appear in at least one student-resource cluster. In this case, clustered students have connections to a minimum of four resources and clustered resources to at least five students. Nodes that belong to more than one cluster are colored in black. Hence, even if six clusters were detected, only four can be visually distinct by node colour because two are completely overlapping with one or more of the others. Hence, the resource access behaviour of students in different clusters is not completely distinctive. Moreover, a set of core learning resources are of interest to the majority of the learners in the course.

The size of a node corresponds to the number of its connections, which is a useful visual representation to combine connection statistics with the clustering approach. This shows clearly that the students who access many different resources and the resources that attract many students can be found in the overlaps between two or more clusters (black nodes). Nearly every active student is interested in the quiz on expert–layman communication, the video message of week 7, and the corresponding reading material independently from the cluster allocation. An overlapping subset of students has additional affiliations to given video messages.

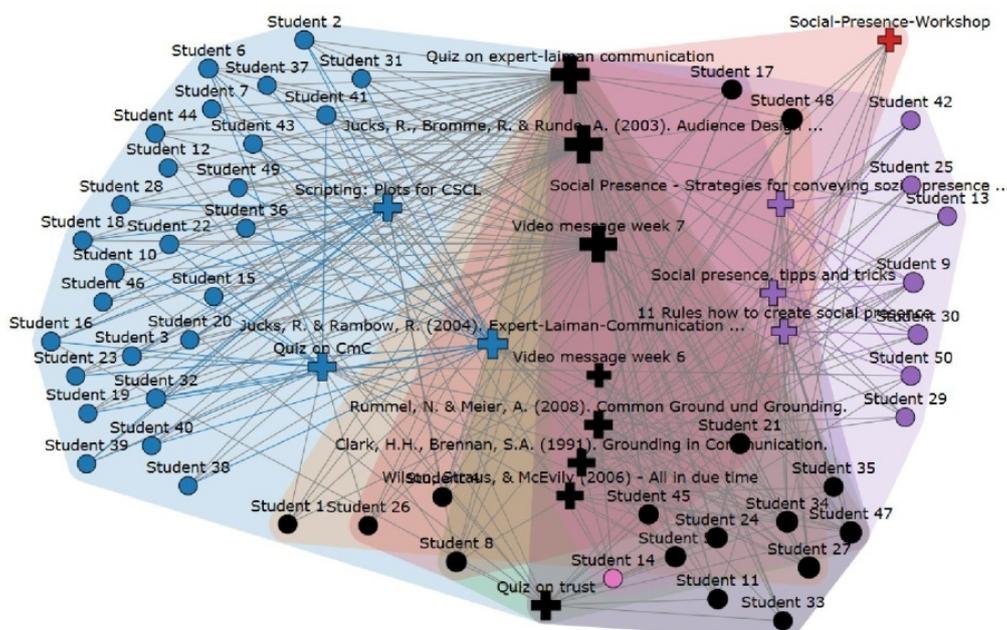


Figure 12: Student-resource clusters in the 7th week of the course.

In order to analyze the fraction of students and resources that could be assigned to bipartite clusters and the fraction of nodes that appear in the overlap between clusters, Table 3 lists basic statistics of the student-resource cluster of each week.

Table 3: Statistics of the clusters in each week of the CMC course.

Week	# Active students	# Students in clusters	# Students in overlaps	# Resources used	# Resources in clusters	# Resources in overlaps
1	118	52	0	12	8	0
2	122	92	0	17	14	0
3	134	97	0	18	14	0
4	117	16	2	20	6	3
5	99	27	2	17	6	3
6	98	64	0	23	11	0
7	95	50	16	28	15	8
8	104	67	3	26	11	2
9	99	10	0	20	4	0
10	48	0	0	10	0	0
11	15	0	0	9	0	0
12	46	0	0	34	0	0
13	87	39	10	41	20	9

This shows that whenever more than one cluster could be detected, there were always a large number of students and resources in the overlap of different clusters. Although large overlaps of the clusters are found in particular weeks of the course and the students often reuse material from previous weeks (section 4.2.1), only a few learning resources occur in cluster overlaps in multiple weeks. These are mainly workshop resources and scientific literature on trust in computer-mediated group work.

Regarding the number of students and resources that could be assigned to clusters, it is interesting to see that even if there were around 100 different learners visiting the course in weeks 4 and 5, only a minority were part of a $K_{5,4}$ biclique. That means that the majority of students accessed only a few resources in this phase. By looking at the swim lane diagram in Figure 11 in conjunction with Table 3, one can see that in weeks where a large percentage of students were accessing enough resources to be allocated to a cluster, the algorithm detects only one big student-resource cluster. This means that when the quantity of different resources used became more diverse, the concrete resource usage of the students also became more diverse, such that the algorithm detected more clusters.

5 CONCLUSION

The analysis of the two online courses using the proposed method has provided some useful insights concerning the patterns of resource usage by students. The courses, namely the GILLS course and the MOOC on computer-mediated communication (CMC), not only differ in openness and number of participants but also in course design. The GILLS course was conducted as a blended learning course with presence lectures but extensive online activities oriented towards content production in the form of wiki articles peer reviewed by other participants. In the CMC MOOC, reading assignments and self-tests were the main tasks for participants. Since, in the CMC course, there was no break between the

(2014). Analysis of Dynamic Resource Access Patterns in Online Courses. *Journal of Learning Analytics*, 1(3), 34–60.

lecture and the exams, exam preparation started in parallel with the lecture. This also led to different resource access behaviour in the two courses.

In the GILLS course, the majority group accessed the lecture slides assigned to the topic of each particular week and the wiki articles written by other students. The continuous interest in wiki articles created by other students indicates that the peer review concept does work appropriately.

As described in section 4.1, students could earn bonus points for the exam when they did the group exercises, namely writing and reviewing wiki articles. Our findings that the articles of the glossary wiki were used across exercise groups and in addition to the lecture slides for reworking the lecture indicates that the self-generated content is important to students. Some groups of students who often overlap with the major student group are also interested in additionally provided scientific articles or the lecture videos. Lecture videos are often found in small bipartite student-resource clusters as in Figure 6. Thus, they seem to be used on demand by small groups of students as additional learning resources or to rework a presence lecture they had not attended.

During the exam preparation phase, we discovered a set of core learning resources accessed by a more or less stable group of students. By investigating the overlaps between the student–resource clusters, it is possible to identify a small set of learning material also used by more than one group of students, see section 4.1.4. For future instances of the GILLS course, the course designers should further investigate the role of these learning materials, particularly for exam preparation.

As described in section 4.1.1, students tend to work with others in their own study programme. In section 4.1.4, the analysis shows that students whose oral exams were scheduled later than those of other programmes also had different patterns of resource access during the exam preparation phase. They began preparation late and focused more on lecture videos and wiki articles than other material. In future, we can tailor the course more specifically for those distinct learner types.

In comparison to the GILLS course, the student-resource clustering in the CMC MOOC yielded fewer clusters. Since the MOOC had not such a strong emphasis on group work, the students seemed to focus more on the same resources. Another interesting insight gained by the tracing of bipartite clusters during the time of the CMC course was that the continuous interest pattern (first row of Table 1) was dominant in the first weeks of the course. This clearly indicates that the resources provided were important to students over several weeks during the lecture period. It is interesting to see that from week 7 on the resource access of course participants became more diverse as the course activity started to rise again after weeks of decline. By a deeper investigation of the clusters, it could be shown that there were periods in the course when students tended to access only a few resources and hence could not be allocated to a cluster. This is a good example of the ability of the proposed method to focus on the most active users within the MOOC. In more active periods, large numbers of students and resources are allocated to more than one cluster.

From a methodological point of view, this paper presented a modification of the subcommunity tracing method for one-mode social networks described in Greene, Doyle, and Cunningham (2010), so that it can be used to trace clusters in bipartite networks. The utility of this kind of analysis for the discovery of characteristic properties of the relations between students and learning resources was demonstrated. A major advantage of the proposed analysis in contrast to purely descriptive analysis of resource access is that one can identify not only popular and hence potentially useful resources, but also who is affiliated

(2014). Analysis of Dynamic Resource Access Patterns in Online Courses. *Journal of Learning Analytics*, 1(3), 34–60.

with the resources, and how such affiliations change over time. This allows for the identification of common student interest in resources.

One can think of several other applications of the method in different learning environments: In cMOOC-like environments or open learning platforms such as Khan Academy² where users are more or less free in the resources they choose, the tracing of dynamic learner-resource networks can help to identify learners with similar learning traces and similar interests. The results could be used for peer recommendation and the formation of learning groups (cf. Harrer et al., 2007). In blended learning scenarios, where learning activities take place in real-world as well as virtual environments, student groups based on resource access that are stable over time can be an indicator for group work in the real world. Research on this can help to develop tools for better course management in such scenarios. The described method not only provides structural information about learners and their relation to learning resources, it also reflects the underlying learning processes. In environments with predefined resource access, such as sequential lectures or scripted inquiry learning settings, the method can help to verify if students use resources as intended. Unintended usage can be detected and may trigger corresponding adjustments on the part of instructors.

ACKNOWLEDGEMENTS

The authors thank the members of the research group on *Social Psychology: Media and Communication* at the Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, for providing the data of the CMC MOOC and Alfredo Ramos for the implementation of the swim-lane diagram visualization.

REFERENCES

- Asur, S., Parthasarathy, S., & Ucar, D. (2009). An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data*, 3(4), 16:1–16:36.
- Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks: Membership, growth, and evolution. *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining* (pp. 44–54), Philadelphia, PA: ACM.
- Belanger, Y., & Thornton, J. (2013). Bioelectricity: A quantitative approach Duke University's first MOOC. Technical Report, Duke University, NC (2013).
- Bruff, D. O., Fisher, D. H., McEwen, K. E., & Smith, B. E. (2013). Wrapping a MOOC: Student perceptions of an experiment in blended learning. *Journal of Online Learning & Teaching*, 9(2), 187–199.
- Clow, D. (2013). MOOCs and the funnel of participation. *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 185–189), Leuven, Belgium: ACM.
- Fini, A. (2009). The technological dimension of a massive open online course: The case of the CCK08 course tools. *The International Review of Research in Open and Distance Learning*, 10(5). Retrieved from <http://www.irrodl.org/index.php/irrodl/article/view/643/1402>.
- Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3), 75–174.

² <https://www.khanacademy.org/>

(2014). Analysis of Dynamic Resource Access Patterns in Online Courses. *Journal of Learning Analytics*, 1(3), 34–60.

- Fox, A. (2013). From MOOCs to SPOCs. *Communications of the ACM*, 56(12), 38–40.
- Garrison, D. R., & Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *The Internet and Higher Education*, 7(2), 95–105.
- Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12), 7821–7826.
- Greene, D., Doyle, D., & Cunningham, P. (2010). Tracking the evolution of communities in dynamic social networks. *Proceedings of the International Conference on Advances in Social Network Analysis and Mining* (pp. 176–183), Odense, Denmark: IEEE.
- Grünewald, F., Meinel, C., Totschnig, M., & Willems, C. (2013). Designing MOOCs for the support of multiple learning styles. In D. Hernández-Leo, T. Ley, R. Klamka, & A. Harrer (Eds.), *Scaling up Learning for Sustained Impact, Lecture Notes in Computer Science 8095* (pp. 371–382). Springer Berlin Heidelberg.
- Halatchliyski, I., Hecking, T., Göhnert, T., & Hoppe, H. U. (2013). Analyzing the flow of ideas and profiles of contributors in an open learning community. *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 66–74), Leuven, Belgium: ACM.
- Harrer, A., Malzahn, N., Zeini, S., & Hoppe, H. U. (2007). Combining social network analysis with semantic relations to support the evolution of a scientific community. *Proceedings of the 8th International Conference on Computer Supported Collaborative Learning* (pp. 270–279), New Brunswick, NJ: ISLS.
- Hill, J., & Hannafin, M. (2001). Teaching and learning in digital environments: The resurgence of resource-based learning. *Educational Technology Research and Development*, 49(3), 37–52.
- Hoppe, U., Pinkwart, N., Oelinger, M., Zeini, S., Verdejo, F., Barros, B., & Mayorga, J. I. (2005). Building bridges within learning communities through ontologies and thematic objects. *Proceedings of the Conference on Computer Support for Collaborative— Learning* (pp. 211–220), Taipei, Taiwan: ISLS.
- Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 170–179), Leuven, Belgium: ACM.
- Lehmann, S., Schwartz, M., & Hansen, L. K. (2008). Biclique communities. *Physical Review E*, 78(1), 016108.
- Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. *Proceedings of the Eleventh International Conference on Knowledge Discovery in Data Mining* (pp. 177–187), Chicago, IL: ACM.
- Nachmias, R., & Segev, L. (2003). Students' use of content in web-supported academic courses. *The Internet and Higher Education*, 6(2), 145–157.
- Pahl, C., & Donnellan, D. (2002). Data mining technology for the evaluation of web-based teaching and learning systems. *Seventh International Conference on E-Learning in Business, Government and Higher Education* (pp. 1–7), Montreal, Canada.
- Palla, G., Barabasi, A. L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446, 664–667.
- Palla, G., Derenyi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435, 814–818.
- Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaiane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6), 759–772.
- Rodriguez, O. (2013). The concept of openness behind c and x-MOOCs (massive open online courses). *Open Praxis*, 5(1), 67–73.

(2014). Analysis of Dynamic Resource Access Patterns in Online Courses. *Journal of Learning Analytics*, 1(3), 34–60.

- Romero, C., Gutiérrez, S., Freire, M., & Ventura, S. (2008). Mining and visualizing visited trails in web-based educational systems. *First International Conference on Educational Data Mining* (pp. 182–186), Montréal, Canada.
- Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1), 368–384.
- Treeratpituk, P., & Callan, J. (2006). Automatically labeling hierarchical clusters. *Proceedings of the 2006 International Conference on Digital Government Research*, San Diego, CA. 167–176.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (1st ed.) Cambridge: Cambridge University Press.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393, 440–442.
- Ziebarth, S., Malzahn, N., & Hoppe, H. U. (2012). Matchballs: A multi-agent-system for ontology-based collaborative learning games. *Proceedings of the 18th International Conference on Collaboration and Technology* (pp. 208–222), Raesfeld, Germany: Springer.
- Ziebarth, S., & Hoppe, H. U. (2014). Moodle4SPOC: A resource-intensive blended learning course. *Proceedings of the 9th European Conference on Technology Enhanced Learning* (pp. 259–372), Graz, Austria: Springer.

3 Incremental Clustering of Dynamic Bipartite Networks

The paper was presented at the 1st European Network Intelligence Conference (ENIC 2014). The ENIC conference series emerged from the ENGINE project³ funded by the European Union as part of the 7th framework program for research, technological development, and demonstration. The intention is to establish a platform for exchange and cooperation between European researchers from the areas of social network analysis, recommender systems, and data mining. The scope of the conference series ranges from novel network based algorithms to case studies in various areas of application.

© 2014 IEEE. Reprinted, with permission, from Tobias Hecking, Laura Steinert, Tilman Göhnert, H. Ulrich Hoppe, *Incremental Clustering of Dynamic Bipartite Networks*, *Proceedings of the 1st European Network Intelligence Conference (ENIC)*, 09/2014 (pp. 9-16).
<http://dx.doi.org/10.1109/ENIC.2014.15>

Author	Contribution	%
Tobias Hecking	<ul style="list-style-type: none"> - Conceptualisation of the approach. - Data collection and cleaning. - Implementation of algorithms. - Design and accomplishment of the evaluation. 	60%
Laura Steinert	<ul style="list-style-type: none"> - Implementation of artificial network generators for evaluation. 	15%
Tilman Göhnert	<ul style="list-style-type: none"> - Data collection from publication databases and pre-processing 	15%
H. Ulrich Hoppe	<ul style="list-style-type: none"> - Supervision and advice during conceptualisation. 	10%

³ <http://engine.pwr.wroc.pl>.

Incremental Clustering of Dynamic Bipartite Networks

Tobias Hecking, Laura Steinert, Tilman Göhnert,
H. Ulrich Hoppe

University of Duisburg-Essen
Dept. of Computer Science and Applied Cognitive Science
{hecking,steinert,goehnert,hoppe}@collide.info

Abstract— This paper deals with the problem of identifying clusters in evolving bipartite networks over time. In bipartite networks there exist two types of nodes while ties can only occur between nodes of different types. Hence, a cluster in a bipartite network consists of two node sets for the two node types each. A major challenge regarding the evolution of those clusters over time is that the two parts of a bipartite cluster may evolve independently. While there is already an increasing amount of research on the identification of clusters in dynamic unipartite networks, the bipartite case is still underrepresented. After a clear motivation of the problem, an adaptation of an existing method for optimising modularity in unipartite networks is extended to dynamic bipartite networks. The method is evaluated on computer generated as well as real world networks.

Keywords—Bipartite networks; Community detection; Modularity optimisation

I. INTRODUCTION

The identification of densely connected subgroups, most commonly named community detection in networks, has attracted much attention in network analysis research [1]. Most of the developed methods have been applied to social networks between entities of the same type, e.g. friendship networks in online communities and co-authorship networks between scientists. However, in many real world settings affiliations between entities are of a bipartite nature. That means that there are two types of nodes and connections only occur between nodes of different types. These bipartite affiliation networks can be observed in various domains.

Examples for bipartite networks are networks showing connections between customers and the products they bought, Wiki articles and their editors, forum users and topics they posted in or scientists and the papers they wrote.

Often unipartite (one-mode) networks between entities of the same type are derived from those bipartite (or two-mode) networks based on mutual affiliations of nodes of one type to nodes of the other type. This is reasonable since mutual connections often imply connections between the nodes of one type itself. For example, a common paper of two scientists constitutes a co-authorship connection between the authors themselves. Deriving a co-purchasing network from a bipartite customer - product network gives a network of customers with similar preferences. Performing community detection on such

networks can give valuable insights into the interconnection between persons mediated by certain objects.

On the other hand, the projection of two-mode networks into one-mode networks always loses the information about one type of nodes. When community detection is performed on the bipartite network itself, the resulting clusters comprise nodes of both types. This allows for identifying persons with similar affiliations to certain objects by keeping the objects itself as part of the cluster and vice versa. This approach has been applied, for example, in [2], where students were clustered around learning resources they used in an online course. This has led to insights on which groups of students are densely connected to which groups of resources and which groups of resources are used by which groups of students.

Another aspect which is of increasing interest in community detection is to analyse the evolution of the community structure over time in evolving networks [3, 4]. Given a sequence of time slices of an evolving network, communities may persist over several time steps. Since bipartite clusters comprise nodes of two types such clusters can be subdivided into two parts. When looking at evolution of bipartite clusters over time, the two node sets can be considered separately. While a bipartite cluster of a previous time slice might not be completely present in a subsequent time slice, it might be the case that one or both parts can be observed as part of another bipartite cluster.

This paper describes an approach for the incremental identification of evolving bipartite clusters. Incremental means that the clustering of a time slice can be partially derived from the clustering of previous time slices. The advantages are twofold. First, computation time can be saved when historical cluster information is used to build clusterings. Second, incremental detection of bipartite clusters can produce smoother transitions between subsequent clusterings. This means that subsequent clusterings are more similar if only little changes in the network occur. This makes the re-identification of cluster parts across time slices and the interpretation of the results easier.

The paper is structured as follows: After this introduction, section II provides a clear formulation of the problem of identifying and tracking bipartite clusters over time. After a review of existing work on bipartite clustering and dynamic

community detection in section III, the method described in [5] which was originally designed for one-mode networks is adapted in order to detect clusters in evolving bipartite networks in an incremental fashion (section IV). In section V the methods are evaluated on real world as well as computer generated datasets. Section VI concludes the findings and gives an outlook on possible improvements.

II. PRELIMINARIES

A. Clusters in bipartite networks

In contrast to unipartite (one-mode) networks $G=(N,E)$ where all nodes are of the same type and edges can occur between any pair of nodes, bipartite (two-mode) networks $G=(N_a, N_b, E)$ comprise two set of nodes N_a and N_b . In the following these sets are called modes. Edges $e \in E$ can only occur between nodes of different modes. This induces some special properties of bipartite networks one needs to be aware of. Paths of even length can only occur between nodes of the same mode and paths of odd length only between nodes of different modes. Consequently, cycles of odd length cannot exist. Thus, clustering methods that rely on cliques such as Clique Percolation [6] or on the basic definition of the clustering coefficient [7] cannot be used without adaption to identify clusters in bipartite networks because triangles cannot exist.

As already stated in the introduction, a bipartite cluster $c=(N_{c,a}, N_{c,b})$ consists of two parts, $N_{c,a}$ and $N_{c,b}$. One part are the nodes that belong to the first mode N_a and the other part are the nodes belonging to the second mode N_b . Figure 1 depicts two examples of clusters in a bipartite network. Considering the network on the left side of figure 1 there exist two clusters of nodes indicated by the grey boxes. Each of the two clusters comprises nodes of both types, indicated by vertex shapes. In the upper left cluster the first part comprises nodes $N_{c,a}=(1,2)$ and the second part comprises the nodes $N_{c,b}=(A,B,C)$.

B. Evolutionary events

When performing community detection in evolving unipartite networks, six different kinds of life cycles can be detected for communities in two successive time slices [3]:

- Birth:** A cluster is identified the first time.
- Growth:** A cluster acquires new members but its core stays the same.
- Contraction:** A cluster loses members over time.
- Merge:** The members of distinct subgroups merge to form one subgroup.
- Split:** One cluster splits into two new born sub-communities.
- Death:** A cluster disappears over time.

In a static time slice the two parts $N_{c,a}$ and $N_{c,b}$ of a bipartite cluster c are treated as one cluster. However, in the transition from one time slice to another the events described above can

occur to the parts of a bipartite cluster separately [2]. This leads to some additional challenges in tracking and identifying evolving bipartite clusters over time. The events *merge*, *split*, *growth* and *contraction* can occur to both parts of a bipartite cluster simultaneously. However, it is also possible that the events occur independently in the two parts of a bipartite cluster as shown in figure 1.

Example: A network of users and forum threads where a group of users forms a bipartite cluster with a set of forum threads is given. The same group of users might be identified as part of a cluster with another set of forum threads in the following time slice of the network because they abandon the old threads and collectively refocus their interests to another set of forum threads. This leads to a pattern as in figure 1 where the user part (circle nodes) and the forum thread part (square nodes) of the clusters at time $t-1$ persists at time t but form a bipartite cluster with other users and threads respectively.

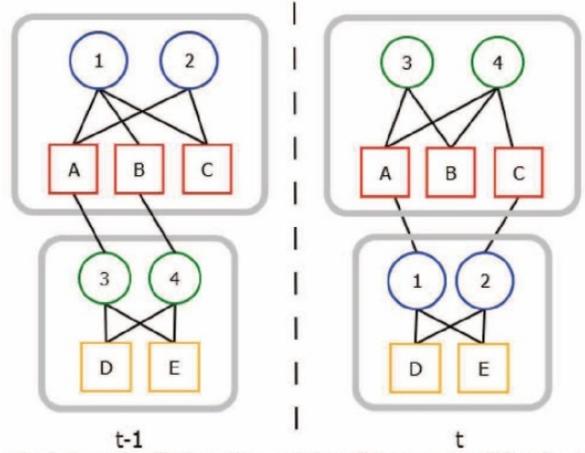


Fig. 1. Example of independent evolution of the two parts of bipartite clusters. The particular parts of the clusters in $t-1$ persist over time but in different clusters.

III. BACKGROUND AND RELATED WORK

A. Clustering methods for bipartite networks

Like unipartite networks, bipartite networks often tend to develop densely connected substructures (clusters) [8, 9] where nodes within a cluster have more connections to nodes of the own cluster than to nodes outside the cluster. Given a clustering of a network the modularity measure defined by Newman [10] is a quality measure to measure how separated the clusters are from each other in terms of connections. In general the modularity sums the fraction of edges of a graph within the particular clusters subtracted by the expected number of edges within the clusters when the edges of the network are randomly rewired by preserving the original degree distribution. The expected number of inner cluster edges in bipartite networks differs from the expected number of edges in unipartite networks because edges cannot occur between all pairs of nodes. For this reason Barber [11] defined

the bipartite modularity measure as given by equation 1. Note c_i and c_j are the cluster ids of nodes i and j . $\delta(c_i, c_j)$ is the Kronecker-Delta.

$$Q(G) = \frac{1}{|E|} \sum_i \sum_j \left(A_{i,j} - \frac{d(i) * d(j)}{|E|} \right) \cdot \delta(c_i, c_j) \quad (1)$$

The BRIM algorithm [11] optimizes the bipartite modularity based on a partial clustering of the nodes of the first type. The nodes of the second type are assigned to the existing clusters in a way that gives the maximal gain in bipartite modularity. After that the procedure starts with the derived clustering of the nodes of the second type and reassigns the node of the first type. The method proceeds in this way until a local maximum of the bipartite modularity is reached.

The Biclique Percolation method [12] adapts the Clique Percolation Method (CPM) described in [6] to bipartite networks which produces a non-exhaustive clustering of a unipartite network with overlapping clusters. The original CPM method relies on the identification of k -cliques (complete graphs of k nodes). As already described in section II.A cliques do not exist in bipartite networks in the original sense because of the absence of cycles of odd length. Hence, a percolation method based on the notion of a $K_{a,b}$ biclique is introduced in [12]. A $K_{a,b}$ biclique is a maximal connected bipartite subgraph comprising of a nodes of the first type and b nodes of the second type. As in CPM clusters can overlap and not all nodes are necessarily assigned to clusters.

A similar approach is described in [13]. This method, however, requires no input parameters. It first identifies maximal bicliques and merges them according to a closeness function.

Other methods rely on the redefinition of the clustering coefficient for bipartite networks [8]. Zhang et al. [14] use the bipartite edge clustering coefficient to decompose a bipartite network into separated modules by successively removing edges with the highest edge clustering coefficient.

Apart from the aforementioned methods that have been especially designed for bipartite networks, there are methods developed for unipartite networks that do not rely on the presence of odd cycles in graphs. Therefore, these methods can be applied to both unipartite and bipartite networks without adaption. This includes modularity optimization methods like the Louvain method [5], the GN (Girvan, Newman) method [15] and the label propagation method [16]. The algorithm described in this paper builds upon the Louvain method but optimizes the bipartite modularity instead of the unipartite modularity and will be described in more detail in section IV.

B. Clustering of evolving networks

As already described in section II.B, in evolving networks clusters might be identified completely or in parts over subsequent time slices of the dynamic network. Clustering algorithms for evolving networks can in general be divided into two classes [17]. Sequential methods, also called offline methods, start the clustering procedure in each time slice from scratch. In a second phase the identified clusters of the most

recent time slice are matched with clusters identified in the previous time slice based on a matching rule [18]. In contrast to that, incremental methods (online methods) use the cluster information of previous time slices to build the clustering of the most recent time slice. In general incremental cluster identification comprises two tasks, namely identification and update. The identification is done clustering algorithms. Based on a clustering of a time slice at time $t-1$ in the update phase some nodes are assigned to clusters at time t based on their previous clusters before the identification starts again. For the sequential approach the focus lies on the matching procedures between clusters of subsequent time slices. Matching can, for example, be based on Jaccard similarity [18] or measures of the inclusion of one cluster into another [19]. There are also hybrid measures as in [20] where inclusion is combined with the social position of the cluster members in social networks. The main motivation for incremental methods is to save computation time because the cluster identification does not have to be calculated completely new in every time slice of the evolving network. Apart from that incremental clustering leads to more smooth transitions between subsequent clusterings [21]. Transition smoothness is measured according to the similarity of subsequent clusterings. While sequential approaches might detect many of the events described in section II-C based on small changes in the network, incremental detection algorithms that produce smoother cluster transitions can be of benefit for the interpretation because evolutionary events are only detected if there is a significant change in the cluster structure. Some incremental approaches for unipartite networks already exist. Hartman et al. [17] give a good overview. Those methods can be based on modularity optimisation [21-23], agglomerative methods based on local rules [24] and label propagation [25]. This work describes incremental methods applied to evolving bipartite networks.

To our best knowledge there is only one incremental method that explicitly targets the clustering of evolving bipartite networks. Greene et al. [26] use a spectral clustering method based on singular value decomposition to identify clusters in evolving bipartite networks. This method however requires prior knowledge about the expected number of clusters. Since this is often not the case, the methods described in the next section of this paper do not require any input parameters.

IV. APPROACH

This section outlines an adaption of the modularity optimisation method in [5] to the bipartite case. The original method is sometimes referred to as Louvain method. Since the target function of [5] aims at optimising the modularity according to its definition for unipartite networks the result is not necessarily a locally optimal result in the sense of the bipartite modularity formulation (see eq. 1). Hence, the adapted version of the algorithm for the bipartite case incorporates a greedy optimisation strategy with the bipartite modularity as target function. Further, a strategy how clusterings of evolving bipartite networks can be updated

incrementally by using information of previous clusterings is outlined.

A. Original modularity optimisation method

The original version of the Louvain method starts with each node assigned to a single-node cluster. Then the method reassigns the nodes iteratively to the cluster - including its own - which yields the highest positive gain of modularity. The result is an initial clustering as depicted in the left part of figure 2.

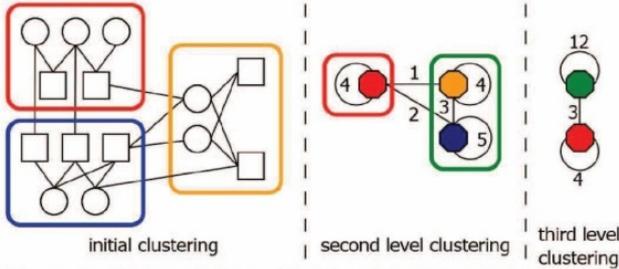


Fig. 2. Application of the modularity optimisation method to bipartite networks.

Further improvements can be achieved if the initial clusters are aggregated into single vertices. Edge weights between these new vertices correspond to the sum of weights of the links between the clusters. Then the modularity optimisation starts again and produces a higher level clustering. The algorithm stops when no more modularity improvement can be achieved by the optimisation and merge steps. Figure 2 outlines the process graphically. The Louvain method is in general applicable to bipartite networks. However, since the original version optimises the modularity according to the definition for unipartite networks, the result might only be an approximation of a locally optimal solution in the sense of the bipartite formulation of modularity (equation 1). Thus, in the following an adapted optimisation procedure is proposed that optimises the bipartite modularity.

B. Optimisation of the bipartite modularity

Calculating the change in modularity when moving a single node from one bipartite cluster to another can be split into two parts, namely removing the node from its original cluster and adding it to a new one. According to equation 1 for the bipartite modularity a cluster with nodes of only one type – which is possible but not desirable in bipartite clusterings – contributes to the overall modularity with 0. Hence, removing a node i from its cluster c and putting it as a single-node cluster causes the modularity to change according to equation 2.

$$\Delta Q_{c-i} = \left(\frac{|E_{c,in}| - |E_{c,in_i}|}{|E|} - \frac{D_c - D_{c,i}}{|E|^2} \right) - \left(\frac{|E_{c,in}|}{|E|} - \frac{D_c}{|E|^2} \right) \quad (2)$$

$|E_{c,in}|$ is the number of edges within cluster c . $|E_{c,in_i}|$ is the number of edges within cluster c incident to node i . $D_c = \sum_{j \in N_{c,a}} \sum_{k \in N_{c,b}} d(j)d(k)$ is the sum of the pairwise products of the degrees of the nodes of each type in cluster c .

$D_{c \setminus i} = d(i) \sum_{j \in N_{c,a}} d(j)$, if node i is of type B and $D_{c \setminus i} = d(i) \sum_{j \in N_{c,b}} d(j)$, otherwise. Merging a single-node cluster with node i as its only member into another cluster d yields the modularity gain of eq. 3.

$$\Delta Q_{d+i} = \left(\frac{|E_{d,in}| + |E_{d,in_i}|}{|E|} - \frac{D_d + D_{d,i}}{|E|^2} \right) - \left(\frac{|E_{d,in}|}{|E|} - \frac{D_d}{|E|^2} \right) \quad (3)$$

Consequently, the modularity gain when moving node i from cluster c to cluster d can be efficiently calculated on a local basis by equation 4.

$$\Delta Q_{c-i,d+i} = \Delta Q_{c-i} + \Delta Q_{d+i} \quad (4)$$

Given a bipartite clustering of a network G , a greedy algorithm that moves nodes iteratively from one bipartite cluster into another cluster such that the bipartite modularity gain is maximal, produces a clustering of G which is a local optimum of bipartite modularity.

As in the original Louvain method in the initial step each node forms a cluster on its own. However, in the adapted version the initial clustering is generated by iteratively reassigning the nodes to clusters such that the bipartite modularity gain according to equation 4 is maximal. In the next step each cluster found in the first step is aggregated into a single node. The resulting graph is not a bipartite graph anymore, as can be seen in figure 2. However, the bipartite modularity can be further improved by clustering the aggregated graph again, using the original version of the Louvain method.

The merge and optimise steps proceed until there is no further improvement of the modularity. Since from the second level clustering on, both parts of the bipartite clusters are merged simultaneously by optimising the unipartite modularity, the result is only an approximation of a locally optimal bipartite clustering regarding bipartite modularity. Thus, the bipartite greedy optimisation is applied again and the result is the final bipartite clustering. A summary of this procedure can be seen in lines 4-8 of listing 1.

C. Incremental detection of evolving bipartite clusters

In the first step of the method described before each node is assigned to a single-node cluster. In incremental clustering, this step is only necessary for the first network of the sequence. For the following time slices, an initial cluster allocation can be derived from the previous clustering according to an update rule. One general strategy which can be applied to derive an initial clustering of a time slice at time t based on a clustered time slice at time $t-1$ is to initialise nodes that are affected by network changes as single-node clusters while all other nodes keep their cluster allocation from the previous time slice [21]. The same is done with nodes that are new at time t . This general strategy is also applied in this work.

Edges that newly appear within a cluster or edges disappearing between clusters strengthen the modularity [23], which is also true for bipartite networks. If a node of type A gains many

connections to nodes of type B from foreign clusters this will likely cause the optimisation method described in section IV.B to change this node to another cluster. Consequently, nodes that are affected by these changes can keep their cluster allocations of the previous time slice.

More difficult to handle is the case when intra-cluster edges disappear or nodes are removed, because this might cause a division of one or both parts of the cluster. In the unipartite case, Nguyen et al. [23] propose to reassign nodes to clusters based on a clique analysis of the affected cluster. However, in the bipartite case there are no cliques in the original sense (see section II.A). The adaption of this rule to the bipartite case by using bicliques [12] instead of cliques has two drawbacks. First, the identification of bicliques is complex and may counterpoise the advantage in computation time for the incremental method compared to the sequential method. Second, in section II it has been explained that the evolution of both parts of a bipartite cluster has to be tracked separately, but bicliques can only be used to detect simultaneous splits of both parts of a bipartite cluster. Thus, in this work a simpler rule is used. The connectedness of a node i of type A within cluster c can be calculated as the fraction of connections to nodes of type B in cluster c and the total number of nodes of type B in c (equation 5).

$$con(i, c) = \frac{|E_{c, in_i}|}{|N_{c, b}|} \quad (5)$$

Each node that loses connectedness to its original cluster at time t compared to $t-1$, becomes re-initialised as a single-node cluster. Other nodes keep their cluster allocation from the previous time slice.

For a sequence of input networks Γ the procedure can be summarised by the following pseudocode in listing 1.

```

1 For  $G_i$  in  $\Gamma$ 
2   IF ( $G_i$  is first graph in  $\Gamma$ )
3     Assign the nodes of  $G_i$  to
       single-node clusters.
4   ELSE
5     Build initial clusters of  $G_i$  based
       on clusters of  $G_{i-1}$  by update rule.
4   1st greedy optimisation of bip. mod.
5   Merge to second level network
6   Apply Louvain method
8   2nd greedy optimisation of bip. mod.

```

Listing 1: Pseudocode of the incremental bipartite modularity optimisation.

V. EVALUATION

This section provides the evaluation results for the incremental bipartite modularity optimisation algorithm described in section IV on artificial as well as real-world networks. For comparison the method was also applied sequentially, i.e. without the incremental updates of the clusters between successive time slices as described in section IV.C. The artificial networks have the advantage that they bear a ground truth clustering structure, with predefined evolutionary events.

Hence, this data is used to evaluate the event detection capabilities of the method. The real-world datasets are used to evaluate the bipartite modularity over time and the transition smoothness between clusterings of subsequent time slices. The transition smoothness is evaluated using the normalised mutual information (NMI), as it was applied in [4], between two clusterings. Since, the node set might not be stable across successive time slices the NMI was calculated only considering nodes that are present in both time slices. However, smoothness cannot be a quality measure on its own because algorithms that produce too smooth cluster transitions may not be able to reflect the actual cluster structure of the evolving network. All results are compared to the LP_BRIM method [27], which combines label propagation [16] with the BRIM method described in [11].

A. Real world networks

The evaluation on real world datasets was performed on a dynamic bipartite network of users and forums as well as a network of scientists and venues in which they have published.

1) Forum dataset

The forum dataset [9] was collected from a Facebook-like online community of students at the University of California, Irvine, in 2004. Students are related to forum topics they were interested in. For this study a subset of this dataset has been used. From the first 30 days, which was the most active period, a sequence of bipartite networks has been sampled into 3-day time slices. Figure 3 depicts the number of users, the number of topics and the number of links between them in this period. An interesting feature of this dataset is that after 15 days (time slice 5), the number of edges decreases sharply. It is worth to study how the dynamic clustering algorithms react to this change in the network structure.

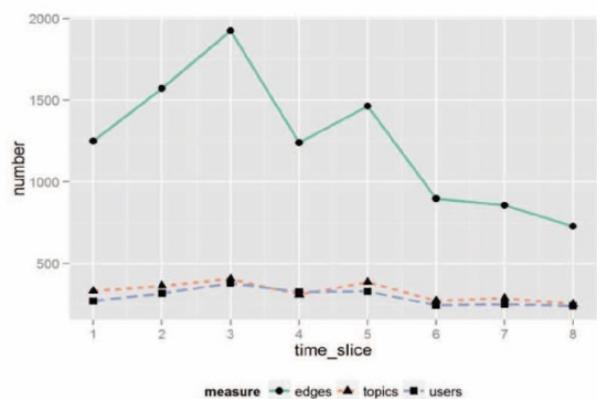


Fig. 3. Statistics of the time slices of the evolving user – forum topic network.

Figure 4 depicts the results for the bipartite modularity over time for the bipartite modularity optimisation method described in section IV and LP_BRIM.

As can be seen in figure 4, the sequential bipartite modularity optimisation performs best on the forum dataset. This is

reasonable because the incremental version has less degree of freedom in creating a clustering than the sequential counterpart that starts from scratch in every time slice.

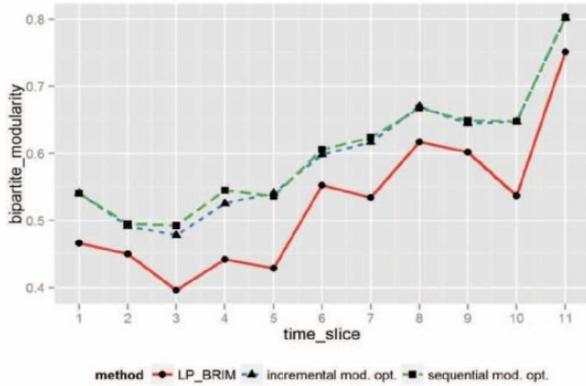


Fig. 4. Bipartite modularity over time for the user - forum topic network.

The results for transition smoothness of the clusterings from one time slice to the next in terms of normalised mutual information in figure 5 show that the incremental version of the bipartite modularity optimisation method produces slightly smoother clustering transitions than the sequential counterpart in the first 5 comparison steps.

LP_BRIM also produces a very smooth clustering, however, because of the trade-off between a good bipartite clustering and the bipartite modularity optimisation the incremental modularity optimisation method can be considered as superior. Interesting to see is that the NMI between successive clusterings for the incremental version of the bipartite modularity optimisation falls slightly behind the sequential version. This is the time slice from which on the number of edges in the network starts to drop (see figure 3). This indicates that the incremental algorithm produces smooth cluster transitions when less network changes occur, but is also able to react to sudden heavy changes in the network structure.

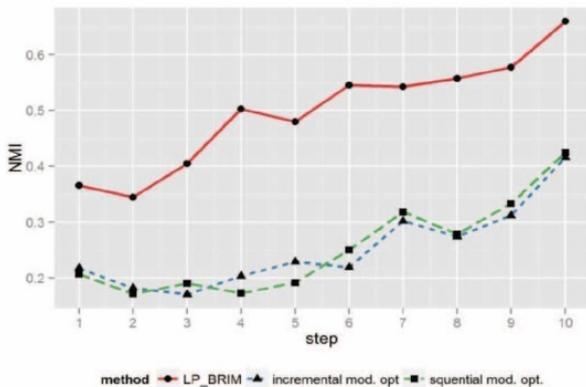


Fig. 5. NMI of the clusterings of two subsequent time slices of the user - forum topic network.

2) Author - venue network dataset

The author-venue network was extracted from the Faceted DBLP¹ database information, based on the dump taken at 2014-03-01. The extraction process started with extracting the set of all publications by authors who ever published at the CRIWG conference. The next step was the extraction of all authors involved in these publications. This author set is the one contained in the network. The venue-nodes in the network correspond to all venues at which at least one of the items in the publication set was published. The links were established based on the author-publication relations. Whenever an author published at least once at a certain venue, a link exists in the network. The extraction process considers all types of publications available in the DBLP database but Thesis, Proceedings (meaning the conference proceedings themselves, not papers in conference proceedings), and WWW resources. Although the CRIWG conference itself is restricted to the years 1995 - 2013, the extracted network is not restricted in time as all publications by the seed authors are included. It is assumed that there is a more or less stable set of authors and venues in recent years. Hence, for evaluation purposes of the dynamic clustering methods, in this study the data was sampled into a sequence of six networks representing yearly time slices of the author-venue network between 2008 and 2013. The smallest time slice (2008) comprises 998 authors and 227 venues and the largest time slice from 2011 has 1298 authors publishing on 327 venues.

The results on this larger dataset show a similar pattern to the previous results of the forum dataset regarding the bipartite modularity over time (figure 6). When the bipartite modularity optimisation is applied sequentially to the network time slices, the bipartite modularity of the resulting clustering is highest. The best transition smoothness can be achieved again by applying LP_BRIM. However, LP_BRIM falls behind regarding modularity. The incremental version of the bipartite modularity optimisation method provides a highly modular clustering as well as smooth transitions (figure 7).

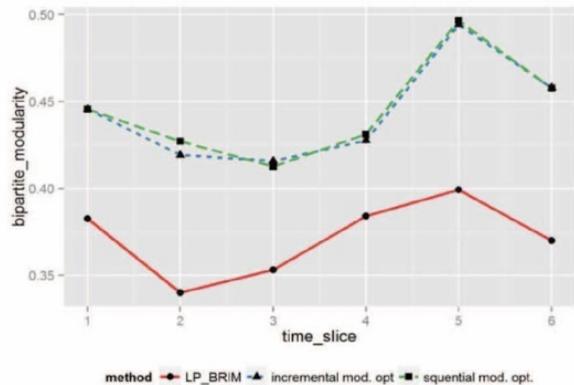


Fig. 6. Modularity over time for the author - venue network.

¹ <http://dblp.l3s.de/dblp++.php>

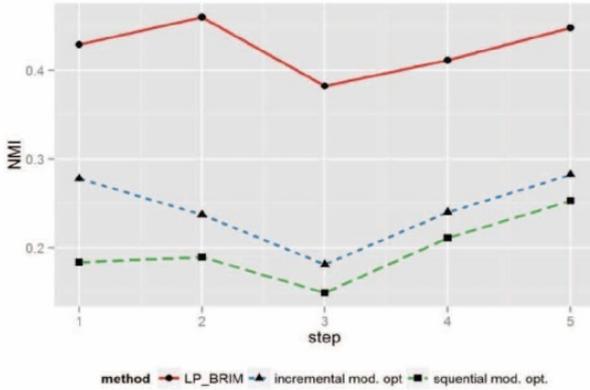


Fig. 7. NMI of the clusterings of two subsequent time slices of the author - venue network.

B. Artificial networks

In order to conduct benchmark tests, artificial bipartite dynamic graphs with predefined clusterings were used. Many algorithms for graph generation exist, e.g. [28]. Yet very few consider dynamic clustered graphs. Among these none met all required criteria. Therefore, a new generator was developed. The artificially generated network has a predefined clustering structure as well as a predefined set of the evolutionary events described in section II.B that occur to the clusters. Hence, it is on the one hand useful to compare the produced clustering of the bipartite modularity optimization method to the predefined clustering. On the other hand it allows evaluating to what extent the predefined cluster events can be detected by matching the clusters of subsequent time slices.

The network generation process can be controlled by a set of six parameters as follows:

- Number of initial nodes $N \in \mathbb{N}$
- Intra-cluster edge probability $\rho_{intra} \in [0; 1]$
- Inter-cluster edge probability $\rho_{inter} \in [0; 1]$

However, these parameters do not control the cluster dynamics. In the few graph generators for dynamic clustered graphs mentioned in the literature this is often achieved indirectly by specifying the number of times each event shall occur, e.g. in [18]. But directly controlling the cluster dynamics has the advantage of being able to easily test the algorithm's event detection.

Therefore, the proposed graph generator receives a metagraph G_t as input, which explicitly gives the cluster dynamics (see figure 8). In this graph every node stands for a part of a bipartite cluster of the graph to be generated. Reconsider that each bipartite cluster consists of two parts, one for the two node types each. Directed edges between nodes of the same type connect cluster parts of different time slices whereas dashed edges between nodes of different types connect nodes that represent parts of the same bipartite cluster in the same time slice.

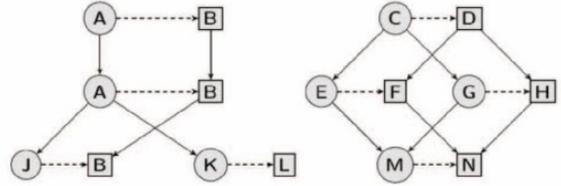


Fig. 8. An example of an input graph G_t for the graph generator.

The generator varies the cluster assignment of each node in such a way that the cluster structure evolves as specified by the metagraph G_r .

By altering the edge probabilities ρ_{inter} and ρ_{intra} it can be tested how cliquelike and isolated the communities need to be to be correctly identified by the algorithm. For this evaluation computer generated dynamic graphs were produced with $\rho_{intra} \in \{1, 0.9, 0.8, 0.7, 0.6, 0.5\}$ and $\rho_{inter} = 1 - \rho_{intra}$. Figure 9 depicts the results for the incremental version of the bipartite modularity optimization applied to a generated dynamic network corresponding to the metagraph in figure 8. The similarity between the identified and the predefined clustering was measured in terms of normalised mutual information.

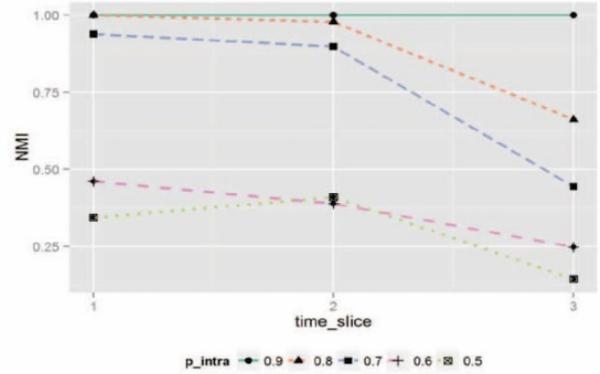


Fig. 9. NMI between predefined and identified clustering for different values for ρ_{intra} .

Furthermore, the correct number of merges, splits and persisting parts of bipartite clusters were correctly detected up to an inter-cluster edge probability of $\rho_{intra} = 0.7$. To identify the events a measure of inclusion was used.

Two parts $N_{c,a}$ and $N_{d,a}$ of two subsequent bipartite clusters C and D at time $t-1$ and time t are considered as related if the nodes of smaller part of C and D is included at least to an extend of 80% in the larger one. This allows the detection of merges, splits and continuing parts of bipartite clusters.

VI. CONCLUSION

This paper framed the problem of the identification of dynamic bipartite clusters over time. Although there are many possible application scenarios ranging from economics to social media the problem has not been tackled very much in existing work. As an attempt to incrementally identify clusters in evolving bipartite networks, the modularity based clustering

method of [5] has first been extended such that it optimises the bipartite modularity instead of the unipartite modularity. Furthermore, the new method has been put in a framework that allows incremental updates of the clustering structure in evolving bipartite networks. The method was evaluated on real world as well as artificial networks. As the results in section V show the method is able to detect clusters in dynamic bipartite network with high quality. Moreover, the incremental version of the method can also increase the smoothness of the transitions between subsequent clusterings. This property is useful for the interpretation of the results because the clustering is not as much affected by small changes in the network as in sequential identification. However, a too stable and coarse grained clustering as produced by the LP_BRIM method are of disadvantage because the identified cluster structure might not reflect the real cluster structure in the network. Because of the lack of existing approaches for dynamic bipartite clustering, more methods can be developed in future works. This can include the extension of existing methods for dynamic unipartite networks to the bipartite case as in this paper, as well as the invention of methods explicitly tailored to bipartite networks. One challenge in the dynamic clustering of dynamic bipartite networks is that the two parts of a bipartite cluster can evolve independently as described in section II.B. This puts further complexity on the task of identifying the cluster parts over time. In the future this aspect could be more emphasised.

REFERENCES

- [1] M. A. Porter, J. Onnela and P. J. Mucha, "Communities in networks," *Notices of the AMS*, vol. 56, pp. 1082-1097, 2009.
- [2] T. Hecking, S. Ziebarth and H. U. Hoppe, "Analysis of dynamic resource access patterns in a blended learning course," in *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, Indianapolis, Indiana, 2014, pp. 173-182.
- [3] G. Palla, A. L. Barabasi and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, pp. 664-667, 2007.
- [4] Y. Lin, Y. Chi, S. Zhu, H. Sundaram and B. L. Tseng, "Analyzing Communities and Their Evolutions in Dynamic Social Networks," *ACM Trans. Knowl. Discov. Data*, vol. 3, pp. 8:1-8:31, apr, 2009.
- [5] V. D. Blondel, J. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, pp. P10008, 2008.
- [6] G. Palla, I. Derenyi, I. Farkas and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814-818, 06/09, 2005.
- [7] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 2658-2663, 2004.
- [8] P. G. Lind, M. C. Gonzalez and H. J. Herrmann, "Cycles and clustering in bipartite networks," *Phys Rev E*, vol. 72, pp. 056127, Nov, 2005.
- [9] T. Opsahl, "Triadic closure in two-mode networks: Redefining the global and local clustering coefficients," *Social Networks*, vol. 35, pp. 159-167, 2013.
- [10] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 8577-8582, 2006.
- [11] M. J. Barber, "Modularity and community detection in bipartite networks," *Physical Review E*, vol. 76, pp. 066102, 2007.
- [12] S. Lehmann, M. Schwartz and L. K. Hansen, "Biclique communities," *Phys Rev E*, vol. 78, pp. 016108, Jul, 2008.
- [13] N. Du, B. Wang, B. Wu and Y. Wang, "Overlapping community detection in bipartite networks," in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, 2008, pp. 176-179.
- [14] P. Zhang, J. Wang, X. Li, M. Li, Z. Di and Y. Fan, "Clustering coefficient and community structure of bipartite networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, pp. 6869-6875, 2008.
- [15] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 7821-7826, 2002.
- [16] U. N. Raghavan, R. Albert and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E*, vol. 76, pp. 036106, 2007.
- [17] T. Hartmann, A. Kappes and D. Wagner, "Clustering Evolving Networks," *CoRR*, vol. abs/1401.3516, 2014.
- [18] D. Greene, D. Doyle and P. Cunningham, "Tracking the Evolution of Communities in Dynamic Social Networks," *Social Network Analysis and Mining. International Conference on Advances In*, vol. 0, pp. 176-183, 2010.
- [19] M. Takaffoli, F. Sangi, J. Fagnan and O. R. Zaiane, "Community evolution mining in dynamic social networks," *Procedia-Social and Behavioral Sciences*, vol. 22, pp. 49-58, 2011.
- [20] P. Bródka, S. Saganowski and P. Kazienko, "GED: the method for group evolution discovery in social networks," *Social Network Analysis and Mining*, vol. 3, pp. 1-14, 2013.
- [21] R. Görke, P. Maillard, A. Schumm, C. Staudt and D. Wagner, "Dynamic graph clustering combining modularity and smoothness," *Journal of Experimental Algorithmics (JEA)*, vol. 18, pp. 1-5, 2013.
- [22] T. Aynaud and J. Guillaume, "Static community detection algorithms for evolving networks," in *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium On*, 2010, pp. 513-519.
- [23] N. P. Nguyen, T. N. Dinh, Ying Xuan and M. T. Thai, "Adaptive algorithms for detecting community structure in dynamic social networks," in *INFOCOM, 2011 Proceedings IEEE*, 2011, pp. 2282-2290.
- [24] M. Takaffoli, R. Rabbany and O. R. Zaiane, "Incremental local community identification in dynamic social networks," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2013, pp. 90-94.
- [25] J. Xie, M. Chen and B. K. Szymanski, "LabelRankT: Incremental community detection in dynamic networks via label propagation," in *Proceedings of the Workshop on Dynamic Networks Management and Mining*, New York, New York, 2013, pp. 25-32.
- [26] D. Greene and P. Cunningham, "Spectral co-clustering for dynamic bipartite graphs," in *Pensa, RG Et AI (Eds.). DyNaK 2010: Proceedings of the 1st Workshop on Dynamic Networks and Knowledge Discovery Barcelona, Spain, September 24, 2010, CEUR Workshop Proceedings, Vol. 655*, 2010, .
- [27] X. Liu and T. Murata, "Community detection in large-scale bipartite networks," in *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences On*, 2009, pp. 50-57.
- [28] A. Lancichinetti and S. Fortunato, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," *Physical Review E*, vol. 80, pp. 016118, 2009.

4 Analysis of User Roles and the Emergence of Themes in Discussion Forums

The paper was presented at the 2nd European Network Intelligence Conference (ENIC 2015). As in the previous year, the majority of contributions could be located in the field of social network analysis with a particular focus on applications related to social media, knowledge management, and learning.

© 2015 IEEE. Reprinted, with permission, from Tobias Hecking, Irene Angelica Chounta, H. Ulrich Hoppe, *Analysis of User Roles and the Emergence of Themes in Discussion Forums, Proceedings of the 2nd European Network Intelligence Conference (ENIC), 09/2015 (pp. 114-121)*. <http://dx.doi.org/10.1109/ENIC.2015.24>

Author	Contribution	%
Tobias Hecking	<ul style="list-style-type: none">- Conceptualisation of the approach.- Data collection and cleaning.- Design and accomplishment of the evaluation.	70%
Irene Angelica Chounta	<ul style="list-style-type: none">- Support in contextualisation and conceptualisation.- Accomplishment of statistical evaluations.	20%
H. Ulrich Hoppe	<ul style="list-style-type: none">- Supervision and advice.	10%

Analysis of User Roles and the Emergence of Themes in Discussion Forums

Tobias Hecking, Irene-Angelica Chounta, H. Ulrich Hoppe

University of Duisburg-Essen
Department of Computer Science
and Applied Cognitive Science
Duisburg, Germany
{hecking,chounta,hoppe}@collide.info

Abstract— This work explores network analysis methods for the analysis of emergent themes as well as types of users in discussion forums. The paper provides both, a description of the analysis approach and its application as a case study. To that end, keywords are extracted from forum threads and then linked to the forum users resulting in a bipartite network based on their activity in discussion threads. Applying bipartite clustering on those networks, groups of users with common interest in themes can be identified as well as groups of related keywords based on their common relations to users. As a case study, the approach is applied to a discussion forum of a Coursera MOOC. The results reveal some interesting patterns and phenomena of thematic development that take place in such large scale learning courses.

Keywords—*Bipartite Networks; Community Detection; Thematic development; Discussion Forums; MOOCs*

I. INTRODUCTION

Discussion forums are a common means for enabling asynchronous information exchange on the web. The content of those forums can grow very large and can be of different nature, ranging from question-answer discussions to coordination and socializing among users. Especially in the case of the recently emerged, massive open online courses (MOOCs), discussion forums play an important role. In such courses, the discussion forums are often the only possible channel for communication and knowledge exchange between participants as well as between participants and the course staff. The discussion topics range from technical issues support to peer help and social conversations [1].

In order to understand the processes and underlying mechanisms in such online discussion forums, information has to be extracted on different levels. On the level of individuals, the challenge is to model roles of contributors and to identify the important users who are indispensable for the cohesiveness of the community while, on a more abstract level, the content of the forum discussions and the patterns of user contribution is of particular interest. Since online discussions can be very dynamic, methods for analysing thematic development are needed. Probabilistic topic models like Latent Dirichlet Allocation [2] can be used to model topics in texts (in our case forum threads). However, further insight can be acquired if relations among keywords and interest user groups as well as their evolution over time are extracted from the data. To that

end, we present an approach for clustering bipartite networks of forum users and keywords extracted from the forum threads. A bipartite cluster of users and keywords can be interpreted as a group of users with common interests and a group of keywords that are related since they are densely connected to a common set of users. The evolution of those clusters can then be tracked over time, which enables the identification of evolutionary events like merging, splitting, and continuing of interest groups and keyword groups. Furthermore, it is shown that the proposed framework should not be considered as a substitute of probabilistic topic models but rather as an additional analysis method for thematic development. All results are presented along a case study on an anonymised forum dataset of the Coursera course titled “Global Warming: The Science and Modeling of Climate Change” provided by the University of Chicago (See [1] for data description).

The paper is structured as follows: Section II gives an overview of the related work in discussion forum analysis in and methods for analysing emergent themes from textual data that are relevant for this work in general. Section III outlines the methodological foundation of this work. The case study on the mentioned Coursera MOOC forum is presented in Section IV. A more formal evaluation and comparison with existing methods follows in Section V. Finally, Section VI concludes the main findings of this work and gives an outlook of possible further work.

II. RELATED WORK

The analysis of thematic development in online communities has been an active research topic in the recent years. One of the first studies on online mass communication by Whittaker et al. [3] described dependencies between different properties of Usenet groups such as thread depth, message length, and demographics. Later the roles of users and knowledge diffusion processes were investigated in more detail [4, 5]. However, these kinds of studies did not yet incorporate text mining of the content of user contributions. This has been done later to solve tasks like post classification, and discussion disentanglement. Especially in the case of MOOC discussion forums it is of huge interest to identify content related threads in which exchange of knowledge between participants takes place [1, 6] as well as the estimation of discussion quality [7].

Probabilistic models of thematic development or topics in online discussions rely on the principle of Latent Dirichlet Allocation (LDA) [2]. Dynamics is especially taken into account in dynamic topic models (DTM) [8] and authors' relations to topics in author-topic-models (ATM) [9]. The combination of both in one model is presented by Xu et al. [10] which comes closest to the idea of this paper. However, it still restricted to a fixed number of topics and interest groups and does not allow tracing the full history of thematically related keywords and interest groups of users simultaneously.

Apart from probabilistic topic analysis, network based methods have as well been applied for discourse analysis, for example, in the context of communication in organisations or scientometrics [11, 12]. Network analysis has found to be appropriate for modelling thematic dynamics in online discussions as well. Introne and Drescher [13] applied the clique percolation community detection method [14] to networks of words extracted from chat messages of users who collaboratively solve a fictive criminal case. The found sub-communities of words are interpreted as topics. By applying community tracking [15, 16] - the re-identification of sub-communities across time slices, they were able to trace the life time of a topic including evolutionary events like topic splits and merges of topics (word clusters). This approach has some similarities with ours presented in Section III. However, in instead of extracting word-to-word networks from the forum posts, our approach links thread keywords to users who are active posters in the thread. The clustering of those bipartite networks does not only allow finding groups of related keywords but also groups of users with a common interest as well as their evolution over time. This approach of modelling artefacts related to the users who use them rather than the direct induction of relations between keywords or users has successfully been applied in social media analytics [17] and scientometrics [18]. A previous step into this direction was presented by Lipizzi et al. [19]. In their work bipartite networks of users linked to keywords extracted from their Twitter tweets are modelled based on tweets on certain products. Different to the work presented in this paper, for further analysis they project the user-keyword relations into a unipartite network of keywords which loses the information about the users. These concept networks are then analysed in terms of cohesiveness and sentiment.

III. METHODOLOGY

A. Network Extraction

For the purpose of the study, we build time slices of the evolving network of users and keywords corresponding to the forum activity in each week of the course. In each time slice there is a set of active users U posting to a subset of threads T . The first step is to extract a set of keywords K from the active threads in the corresponding period. This is done by aggregating the posts within one thread to a single document and application of the keyword extraction algorithm provided by Alchemy API¹. This algorithm computes a ranked list of keywords with relevance values ranging between 0 and 1. In this work we used keywords with relevance above 0.8

(experimentally determined value). The result can be modelled as a bipartite network where threads are represented as nodes of one type linked to keywords as nodes of the other type. Its adjacency matrix A_{TK} is of the following form:

$$\begin{bmatrix} 0 & B_{TK} \\ B_{TK}^T & 0 \end{bmatrix} \quad B_{TK} \in \{0,1\}^{|T| \times |K|} \quad (1)$$

In the same manner, another bipartite network of users and forum threads can be build where users are linked to threads they posted in during the course week. The resulting adjacency matrix A_{UT} has the form:

$$\begin{bmatrix} 0 & B_{UT} \\ B_{UT}^T & 0 \end{bmatrix} \quad B_{UT} \in \{0,1\}^{|U| \times |T|} \quad (2)$$

If the rows of B_{TK} and the columns of B_{UT} are in the same order, the adjacency matrix A_{UK} of the final user – keyword network can be calculated as:

$$A_{UK} = \begin{bmatrix} 0 & B_{UK} \\ B_{UK}^T & 0 \end{bmatrix}, \quad B_{UK} = B_{UT} \times B_{TK} \quad (3)$$

B. Network Clustering

The next step is to discover bipartite clusters of closely related users and keywords. A cluster of users and keywords can be seen as a group of users with a common thematic interest. In case of discussion forums users who post in the same thread automatically form such an interest group since they form a biclique with all the keywords of the thread, provided that the thread has enough keywords. However, defining interest groups only on the thread level might give an incomplete picture. Therefore, the biclique percolation method [20] is appropriate to discover interest groups of people in multiple threads based on common keywords. Figure 1 gives an example of a bipartite cluster which is not restricted to authors and keywords from only one forum thread. The biclique percolation method has two further advantages for the task of clustering users and keywords simultaneously. First, the method allows overlaps between the clusters. This property is essential since it is likely that a keyword can be used in different contexts. The same is true for users. A user can be part of more than one interest group. Second, the method only assigns users and keywords to clusters if they are part of a biclique with a nodes of the one mode and b nodes of the second type. This can be considered as an inherent filtering procedure that helps to focus on the important parts of the network. Keywords and users who are not well connected within the network are not part of the final clustering result.

The principle of biclique percolation adapts the well-known clique percolation method to bipartite networks where no cliques in the original sense are present. The method relies on the definition of a $K_{a,b}$ biclique. This is a maximal connected bipartite subgraph with a nodes of the first mode and b nodes of the second mode. Thus, if a set of a actors all are connected to each of b resources, they form a $K_{a,b}$ biclique. A bipartite subgroup also called biclique community is defined as the union of a series of adjacent $K_{a,b}$ bicliques. Two $K_{a,b}$ cliques are considered adjacent if they share at least $a-1$ nodes of the one type and $b-1$ nodes of the other (see Figure 1).

¹ <http://www.alchemyapi.com/>

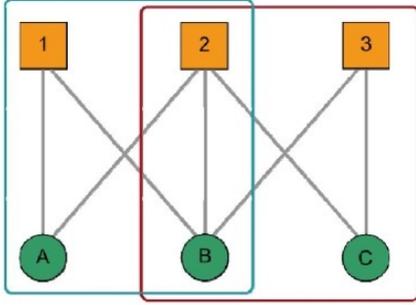


Fig. 1. Example of two adjacent $K_{2,2}$ bicliques. First clique: $\{A, B, 1, 2\}$, second clique: $\{B, C, 2, 3\}$ (Source [24]).

C. Thematic dynamics

As described in [21], in a static snapshot of an evolving bipartite network both parts of a cluster can be considered at once. However, regarding the evolution of the network these two parts should be considered separately. It is possible to re-identify the keyword group or the user group of a bipartite cluster across time slices. An example is given in Figure 2.

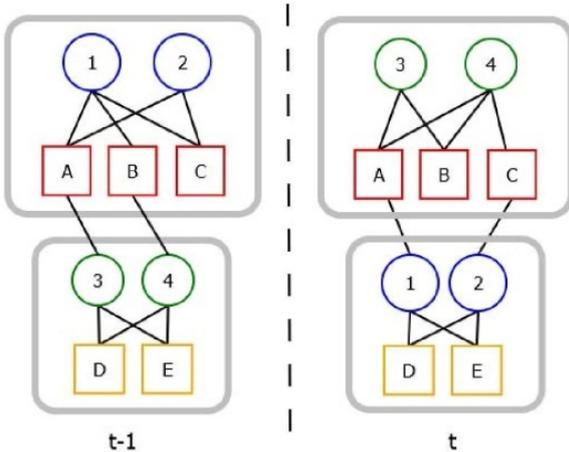


Fig. 2. Example of independent evolution of the two parts of bipartite clusters. The particular parts of the clusters in $t-1$ persist over time but in different clusters. (Source: [21])

This analytical approach can also be applied to the user-keyword networks that are extracted from successive time slices, as described before. Thus, it can provide insight regarding the analysis of thematic dynamics in discussion forums. The assumption is that keywords which occur in one bipartite cluster can be interpreted as semantically related, since they are densely related to a subset of the users in the network. Furthermore, the users of such a cluster can be considered as an interest group with common relations to a subset of concepts. Such a mixed cluster of users and concepts not necessarily correspond to a single thread where all users in a thread are linked to all keywords assigned to the thread, as explained in Section III.A, but may contain users and keywords of different threads. This allows finding keywords that belong to a certain thematic area and subgroups of users who are interested in this area, but also to trace the whole history of

thematic areas and interest groups by applying community matching [22, 23]. A group of keywords in a bipartite cluster at a given time slice t can sometimes be matched to groups of keywords in a time slice $t-x$ by different measures of similarity. The same is true for groups of users. Since there is not always a one-to-one matching between two groups in two different time slices of the network, it becomes also possible to identify splitting and merging topics (groups of keywords) or interest groups (groups of users) solely based on the changes in the user-keyword network. The advantage of the bipartite clustering approach to other dynamic topic modelling approaches is that the model completely maintains the relation of users and keywords in bipartite user-keyword clusters. This enables to track the shifts of interests of certain groups of users simultaneously.

In the case of bipartite clusters, the re-identification of parts of clusters is more complicated as in unipartite networks since the two groups of nodes in such clusters can evolve independently (Figure 2), and therefore, the matching procedure has to be applied to both parts of a bipartite cluster separately. There are different approaches to match groups of nodes in a dynamic network over time. The work of Bródka et al. [22] gives a good overview and evaluation. In this work we chose the inclusion measure. Its value indicates to what extent a smaller group is contained in a larger group and can be calculated as:

$$\text{sim}(g1, g2) = \frac{|g1 \cap g2|}{\min(|g1|, |g2|)} \quad (4)$$

The inclusion measure is most suitable to identify merges and splits of groups. According to the approach presented by Greene and Doyle [23] and its adaptation to bipartite clusters [24] the matching of groups across time slices two sets of not matched groups are maintained. A group (keyword group or user group) is considered as not matched if there is no group with similarity above a certain threshold. The method proceeds as follows:

The procedure subsequently examines the clusters found in each time slice. As stated before, a bipartite cluster contains two groups, one keyword group and one user group. For all groups in a particular time slice t , the similarity to all the not matched groups in previous time slices is computed. If the similarity to one or more not matched group is above the threshold, a relation between the groups is introduced. Consequently, the matched groups from previous time slices are deleted from the set of not matched groups and all the groups of the current time slice are added to this set. This has to be done for the keyword groups and the user groups of the bipartite clusters separately in order to identify evolutionary events of these groups separately. The matching threshold has been fixed to 0.75 in this study.

IV. CASE STUDY

A. Dataset description

For this paper, we study a Coursera MOOC titled “Global Warming: The Science and Modeling of Climate Change”. A discussion forum was used to support the communication between the users and to promote social interaction and

information exchange. The activity in the discussion forum was recorded in log files from October 21, 2013 to January 11, 2014 [1]. The discussion forum consisted of forums and sub-forums, further divided in threads consisting of posts and comments to posts. The structure of the discussion forum also reflects different thematic areas that are facilitated, such as general discussion, help on assignments, feedback on course organizational issues, etc.

Users were able to start their own threads, post in different threads and forums and also comment directly on existing posts. The MOOC participants could post or comment in the discussion forum either using a personalized user account or anonymously. Overall, we identified four different user types: students (1000 users), staff (4 users), and instructors (1 user). Additionally anonymous posts (5.9% of all posts) were allocated to a single fictive anonymous user. In the present study, this anonymous user was removed from the analysis since we cannot gain insight with respect to their user status (students or staff) or to personalize user activity. The dataset consisted of 1005 participants, who created 5336 posts in 2027 threads distributed over 1874 sub-forums. In Figure 3, we present the distribution of user activity, with respect to posting, over the threads and forums of the discussion board. On average, each user posts on 3 threads (mean=2.94, $\sigma=6.9$) and 2 forums (mean=1.86, $\sigma=2.08$). As it is shown from the distribution, users do not get involved or spread over many threads and forums. This indicates limited activity that focuses on particular thematic areas without further expanding to others.

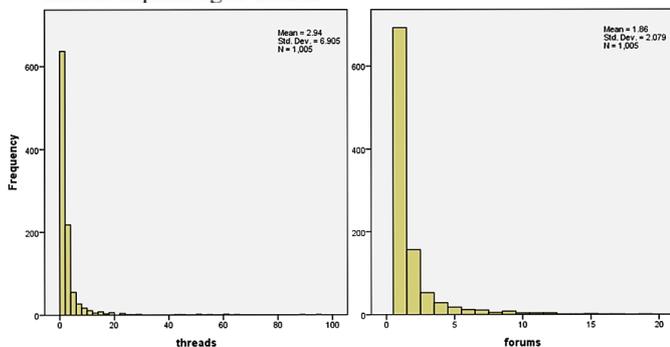


Fig. 3. Distribution of user posts over threads and forums of the discussion board for the whole duration of the MOOC.

In Figure 4, we present the distribution of posts per individual user for the whole duration of the course. On average, each user created 5 posts (mean=5.31, $\sigma=18.86$) while it was shown that the majority of users posted less than 50 posts in the discussion forum. The users of the discussion forum could vote for the posts that were created by other users by adding a +/-1. This way, the users themselves provided a rating of the quality and usefulness of posts. Overall, 3844 posts (72% of the total number of posts) received no votes, while the rest posts were rated with 0.47 votes on average. The number of posts per user, was found to correlate highly with the number of threads ($\rho=0.876$, $p<0.01$) and the number of forums

($\rho=0.819$, $p<0.01$) that the particular user had been posting. Additionally the number of posts correlated statistically significantly with the average number of votes per user ($\rho=0.234$, $p<0.01$). This suggests, that the users who have a high posting activity, also contribute to different thematic areas and are acknowledged as influential by the other users of the discussion forum.

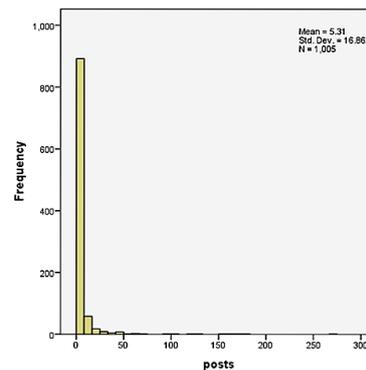


Fig. 4. Distribution of user posts over the whole duration of the MOOC course.

B. Bipartite Clustering of Users and Keywords

The users and keywords were clustered using the described biclique percolation method. The parameters were set to $a=b=3$ based on the observation that relevant forum threads have usually 3 or more keywords and 3 or more users who post in this thread. Other threads do not affect the clustering since a biclique must contain at least 3 keywords and 3 users. A typical result is depicted in Figure 5.

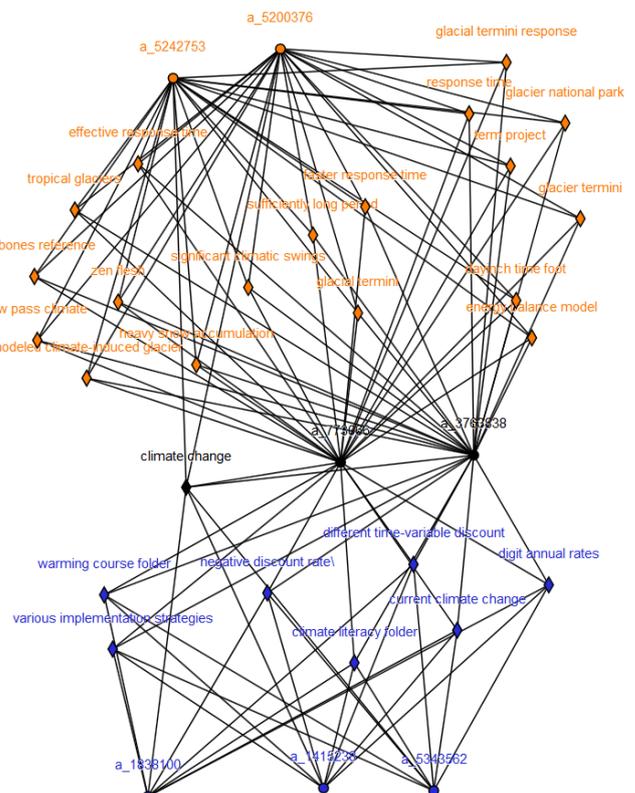


Fig. 5. Bipartite clusters of users and keywords

The depiction is an excerpt of the resulting clusters in the seventh week of the course and shows two user-keyword clusters that share a common user. The orange cluster mainly contains keywords related to glaciers and the blue cluster relates to technical terms.

C. Bridging concepts and bridging users

Since users can be active in different threads, and hence, can have broader thematic interests, it may occur that those users cannot be assigned to a cluster uniquely. As one would expect this is often the case for the course administration. However, there are also regular course participants who bridge between different thematic areas as the example in Figure 5 shows. On the other hand, there are also concepts that are of general importance. These concepts frequently appear in the overlap of bipartite clusters as well. Consequently, the proposed approach can be used to identify, both, concepts of general importance for the community and users who have the potential to spread information between different thematic areas. Table I depicts the concepts that occur in the overlaps of the user-keyword clusters over time most often. In addition, the third column of Table I subsumes the time slices in which the corresponding node occurs. Concepts that often occur in more than one cluster at a time are of different types. For example, “Climate change”, which is central to the course topic “Global warming” in general, occurs in overlap of clusters in 6 of 11 time slices. This is expected for these kinds of concepts. Furthermore, there are concepts like “Environmental Science Master” that are not directly related to the course content. Those can frequently be found in the overlap of bipartite clusters, as well. An explanation could be that topics like potential further activities after the course are discussed in many different threads and in different contexts.

TABLE I. CONCEPTS MOST FREQUENTLY IN OVERLAPS

Keyword	Present in time slices	Times in overlap
Climate change	1,2,3,5,6,7,8,9	6
Sustainable resources	3,4,6,7,8	3
Environmental Science Master	6,7,8	3

Since the retrieved cluster can overlap in both modes, namely concepts and users, Table II subsumes the users who occur in overlaps most often. These users can be considered to have a broader interest, and thus, are closely connected to more than one user-keyword cluster. Since MOOC discussion forums are partly moderated by course staff who answer important questions regarding different topics, it is reasonable to find course staff people most often in the overlap. However, it is interesting that we could find such people also among the regular users. These users have a broader interest and participate in more threads than others, which can as well result from superposting behaviour (c.f. Huang et al. [25]).

TABLE II. USERS MOST FREQUENTLY IN OVERLAPS

User	User type	Present in threads	Times in overlap
a_4977322	Course staff	1,2,3,4,5,6,9,10	6
a_5219940	Regular user	1,4,5,6,7	6

a_1947440	Regular user	1,2,3,6,7,8	5
-----------	--------------	-------------	---

For the purpose of this study, we identified the users who appear in overlaps (Group 1) and further studied their practice with respect to their posting activity during the course. Additionally we compared them to users who don't appear in overlaps (Group 2). The results for the two groups are presented in Table III.

TABLE III. USER ACTIVITY STATISTICS FOR GROUP 1 AND GROUP 2

Groups	#threads	#forums	#posts	#votes per user
Group 1	5.55	2.74	10.9	0.91
Group 2	1.7	1.45	2.65	0.26

Overall, 324 users were found to appear in the overlaps (Group 1), 321 students, 2 staff members and 1 instructor. These users posted on average 11 posts throughout the duration of the course, in 6 threads (mean=5.55) and 3 forums (mean=2.74). Moreover, the posts they created were rated on average with 0.91 votes. On the contrary, the users who did not appear in overlaps (Group 2 - 681 users) created fewer posts which were distributed in fewer threads and over fewer forums in comparison to the activity of Group 1. Furthermore, the users of Group 2 were voted lower for their posts. From this data, it can be concluded, using the Mann-Whitney U Test, that Group 1 had statistically significantly higher posting activity, more distributed over thematic areas and was rated higher than the activity of Group 2 ($p < 0.01$). The distribution of user activity per Group with respect to posts and threads is presented in Figure 6.

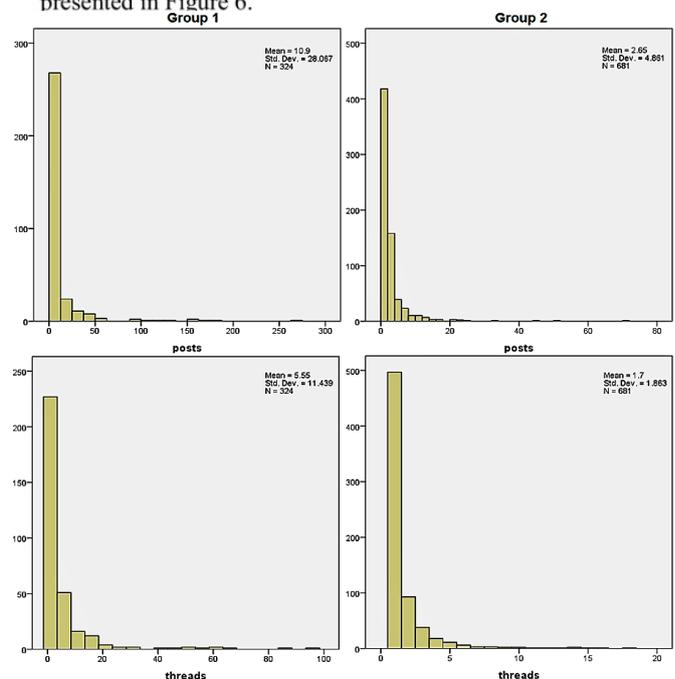


Fig. 6. Distribution of number of posts and threads for the users of Group 1 and Group 2.

D. Evolution User-Keyword Clusters

Since user parts and keyword parts of the clusters can continue, merge, and split independently, there is a wide variety of possible evolutionary events. Some could be identified in our dataset. The most frequent events are listed in Figure 7. Event 1 in the depiction occurs most often. In these cases a topic is taken up by a different group of users. This happens often when forum threads remain active over several weeks or become active again after a period of inactivity. In this case study this is, for example, the case for threads on general discussions or keywords related to technical or organizational issues of the course.

On the contrary, interest groups of users could be identified across several time slices but with changing connections to thread keywords (event 2). This pattern is very interesting since it shows that groups of users sometimes change their interests simultaneously without necessarily having direct communication in the forum. As a prototypical example, a group of students with common connections to keywords of the area “carbon emissions” in week 4 of the course could be re-identified in week 7. However, in this week they talk about number crunching techniques for measuring climate change. An explanation could be that they first discuss more general about factors of climate change and afterwards show more interest into concrete measuring techniques.

Splitting groups of keywords (event 3 in Figure 7) could be observed occasionally. In the beginning of the course when a long thread for the general discussion of the topic “climate change” was taking place that evolve a large amount of the users. The resulting keyword group splits afterwards into three smaller keyword groups resulting from more focused discussions on special areas.

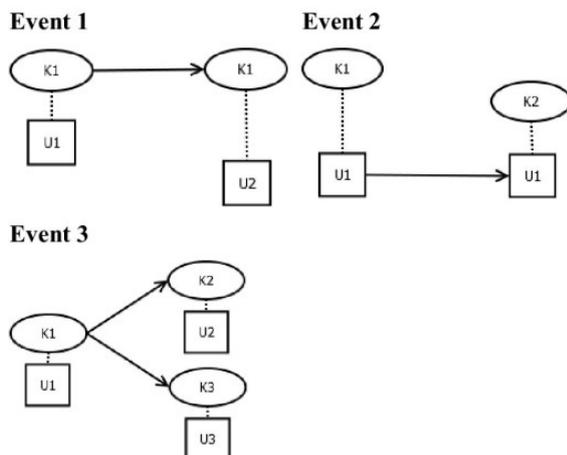


Fig. 7. Evolutionary events for discovered for the bipartite user-keyword clusters.

V. COMPARISON WITH TRADITIONAL TOPIC MODELLING

While not completely comparable, the presented approach has some similarities with topic modelling. To some extent, the keyword part of a user-keyword cluster can be interpreted as a topic since their common relations to a set of users introduce some semantic relatedness. It is very difficult to compare these two approaches quantitatively. However, for the sake of completeness Table IV depicts the similarity of topics found by LDA and keyword groups of the user-keyword clusters identified in each time slice. LDA requires the number of topics as a parameter. For comparison this parameter is set to the number of user-keyword clusters found in the corresponding time slice. The input of LDA is a document-term-matrix which elements are the number of occurrences of each term in the document. In our case a document is defined to be a single discussion thread in the forum. The term vocabulary corresponds to the keywords extracted with the Alchemy API as for the clustering approach. However, the keywords for each thread were not restricted to have a relevance value above a certain threshold as for the clustering approach. Since in LDA each topic is represented as a probability distribution of words, it is necessary to transform this distribution into a bag of words by specifying a minimum rank for words to be representative for a topic. For this evaluation this value was set to 10. The resulting bag of words and the keyword groups introduced by the clustering approach were compared in terms of inclusion similarity. It measures to what extent the smaller group of words is contained in the larger one similar to equation 4. For each keyword cluster the similarity with the best matching bag of words induced by LDA was calculated and average similarity of all pairs calculated. The problem is that a restriction to a fixed number of keywords is not possible for the clustering approach since there is no ranking of the keywords. Thus, the comparison with LDA is only limited.

The results in Table IV clearly show that the results of the user-keyword clusters are very different to the topics modelled by LDA. This huge differences result from the different interpretation of a topic. While LDA models topics as a probability distribution of words based on their occurrences in forum threads, the keywords in a bipartite cluster are the result of common relations to users.

TABLE IV. COMPARISON OF LDA AND BIPARTITE CLUSTERING.

Time slice	Number of topics	Avg. incl. similarity (LDA, clusters)
1	30	0.19
2	21	0.175
3	20	0.135
4	9	0.01
5	12	0.08
6	11	0.2
7	10	0.15
8	11	0.29
9	11	0.29
10	7	0
11	0	1

VI. CONCLUSION

In this paper we explored bipartite clustering of users and keywords for the analysis of user roles and thematic

development in online discussion forums. We demonstrated the potential of the approach with the application of the method to a MOOC discussion forum. The interesting feature of co-clustering of users and keywords in a bipartite user-keyword network is that users and keywords extracted from forum threads are grouped simultaneously into mixed clusters. In such clusters, the group of users can be interpreted as a group with a common interest and introduces semantic relations among the keywords. Since the biclique percolation method [20] was applied, overlapping clusters are possible. Overlaps in the user dimension indicate users with broader interests and overlaps in the keyword dimension help to discover keywords that are important for more than one interest group.

In order to gain further insight, we applied the proposed methodology to a three-month MOOC course. In general, the participants of the course appeared to have low posting activity that focused on certain thematic areas. However, posting activity correlated to the number of threads and number of forums users were active as well as to the number of votes they received. After applying bipartite clustering of users and keywords, we were able to identify two groups of MOOC's users: those who appear in overlaps of the discovered clusters and those who do not. From the analysis of user activity of these groups, it was shown that users who appear in overlaps create more posts that also spread in more threads and forums, and thus, over various thematic areas, acting as disseminators of knowledge. Additionally, these users received more votes, i.e. get higher ratings from other users of the discussion forum, indicating that they are identified as influential and important nodes. Regarding the emergence of themes in the discussion forums the identified clusters were tracked over successive time slices. Different evolutionary events could be detected for, both, the keyword groups and the user groups of the bipartite clusters. One common event discovered was the persistence of a group of keywords across time slices. However, these groups of keywords were not necessarily connected to the same group of users over time. This indicates that themes are sometimes taken up by different groups of users. On the contrary the same pattern could be identified for persisting groups of users who build bipartite clusters with different groups of keywords over time. Splitting groups of keywords could be identified as well.

There are some commonalities to existing topic modelling methods. However, the presented approach is not directly comparable to these methods and should not be considered as a substitute of traditional topic modelling approaches. Most existing approaches like LDA and derivatives use probabilistic modelling. They do not allow for the tracking of topic evolution considering a topic as the result of evolutionary events like "merge", "split" of previous topics. Some methods model user relations to topics but not the dynamics of interest groups of users over time. Our approach allows for both, tracking of the history of emerging groups of related keywords as well as maintaining the relations to users. Simultaneously tracking of changing interests of user groups can also be identified as presented in section IV.D. The results in Section V have shown that the results by the bipartite clustering approach are very different compared to the results of LDA topic modelling. Hence, one should be careful with the interpretation of a cluster of users and keywords. While in

probabilistic modelling a topic is a probability distribution of words based on the word's occurrences in documents, a group of words in a bipartite cluster is solely based on dense relations to a group of users with similar interests. This is a completely different view on thematic development in online discussions and the clustering approach yields different insights into the thematic dynamics.

In future work the presented approach should be applied to a wider range of online communication, like chats and Twitter. This would be a further step to understand thematic development in these areas and the identification of important users and their interests.

REFERENCES

- [1] L. A. Rossi and O. Gnawali, "Language independent analysis and classification of discussion threads in coursera MOOC forums," in *Proceedings of the 15th {IEEE} International Conference on Information Reuse and Integration, {IRI} 2014, Redwood City, CA, USA, August 13-15, 2014*, 2014, pp. 654-661.
- [2] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," *J.Mach.Learn.Res.*, vol. 3, pp. 993-1022, mar, 2003.
- [3] S. Whittaker, L. Terveen, W. Hill and L. Chery, "The dynamics of mass interaction," in *From Usenet to CoWebs* Anonymous Springer, 2003, pp. 79-91.
- [4] J. Zhang, M. S. Ackerman and L. Adamic, "Expertise networks in online communities: Structure and algorithms," in *Proceedings of the 16th International Conference on World Wide Web*, Banff, Alberta, Canada, 2007, pp. 221-230.
- [5] L. A. Adamic, J. Zhang, E. Bakshy and M. S. Ackerman, "Knowledge sharing and yahoo answers: Everyone knows something," in *Proceedings of the 17th International Conference on World Wide Web*, 2008, pp. 665-674.
- [6] Y. Cui and A. F. Wise, "Identifying content-related threads in MOOC discussion forums," in *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, Vancouver, BC, Canada, 2015, pp. 299-303.
- [7] J. Kim, E. Shaw, D. Feng, C. Beal and E. Hovy, "Modeling and assessing student activities in on-line discussions," in *Proc. of the AAAI Workshop on Educational Data Mining*, 2006, .
- [8] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 113-120.
- [9] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth and M. Steyvers, "Learning author-topic models from text corpora," *ACM Transactions on Information Systems (TOIS)*, vol. 28, pp. 4, 2010.
- [10] S. Xu, Q. Shi, X. Qiao, L. Zhu, H. Jung, S. Lee and S. Choi, "Author-Topic over Time (AToT): A Dynamic Users' Interest Model," vol. 274, pp. 239-245, 2014.
- [11] J. Diesner and K. M. Carley, "Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis," *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations*, pp. 81-108, 2005.
- [12] L. Leydesdorff and I. Hellsten, "Measuring the meaning of words in contexts: An automated analysis of controversies about 'Monarch butterflies,' 'Frankenfoods,' and 'stem cells'," *Scientometrics*, vol. 67, pp. 231-258, 2006.
- [13] J. E. Introne and M. Drescher, "Analyzing the flow of knowledge in computer mediated teams," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, San Antonio, Texas, USA, 2013, pp. 341-356.
- [14] G. Palla, I. Derenyi, I. Farkas and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814-818, 06/09, 2005.

- [15] G. Palla, A. L. Barabasi and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, pp. 664-667, 2007.
- [16] D. Greene, D. Doyle and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference On*, 2010, pp. 176-183.
- [17] K. El-Arini, M. Xu, E. B. Fox and C. Guestrin, "Representing documents through their readers," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illinois, USA, 2013, pp. 14-22.
- [18] E. Yan, Y. Ding and E. Jacob, "Overlaying communities and topics: an analysis on publication networks," *Scientometrics*, vol. 90, pp. 499-513, 2012.
- [19] C. Lipizzi, L. Iandoli and J. E. R. Marquez, "Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers' reactions to the launch of new products using Twitter streams," *Int. J. Inf. Manage.*, vol. 35, pp. 490, 2015.
- [20] S. Lehmann, M. Schwartz and L. K. Hansen, "Biclique communities," *Phys Rev E.*, vol. 78, pp. 016108, Jul, 2008.
- [21] T. Hecking, L. Steinert, T. Gohnert and H. U. Hoppe, "Incremental clustering of dynamic bipartite networks," in *Network Intelligence Conference (ENIC), 2014 European*, 2014, pp. 9-16.
- [22] P. Bródka, S. Saganowski and P. Kazienko, "GED: the method for group evolution discovery in social networks," *Social Network Analysis and Mining*, vol. 3, pp. 1-14, 2013.
- [23] D. Greene, D. Doyle and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference On*, 2010, pp. 176-183.
- [24] T. Hecking, S. Ziebarth and H. U. Hoop, "Analysis of Dynamic Resource Access Patterns in Online Courses," *Journal of Learning Analytics, JLA*, vol. 1, pp. 34-60, 2014.
- [25] J. Huang, A. Dasgupta, A. Ghosh, J. Manning and M. Sanders, "Superposter behavior in MOOC forums," in *Proceedings of the First ACM Conference on Learning @ Scale Conference*, Atlanta, Georgia, USA, 2014, pp. 117-126.

5 Discovery of Structural and Temporal Patterns in MOOC Discussion Forums

An early version of the work described in this paper was presented as a poster paper at the 7th Conference on Advances in Social Network Analysis and Mining (ASONAM 2015)⁴. The conference series is highly competitive (18% full paper acceptance rate) and has a strong focus on computational aspects of social network analysis. Based on the presentation at the conference the work was selected for publication as chapter of a book of the Lecture Notes in Social Networks series. This chapter constitutes an extensive revision of the original work and extends the first version especially in methodological aspects. It has been accepted for publication in 05/2016.

Authors' preprint version of Hecking, T., Hoppe H. U., Harrer, A. (2016) Discovery of Structural and Temporal Patterns in MOOC Discussion Forums. In J, Kawash, N. Agarwal, T. Özyer (Eds.), Prediction and Inference from Social Networks and Social Media (pp. 153-180). Lecture Notes in Social Networks, Springer, http://dx.doi.org/10.1007/978-3-319-51049-1_8

Author	Contribution	%
Tobias Hecking	<ul style="list-style-type: none"> - Conceptualisation of the approach. - Exploration of methods. - Design and implementation of algorithms. - Design and accomplishment of the evaluation. 	75%
Andreas Harrer	<ul style="list-style-type: none"> - Support in conceptualisation and algorithmic aspects as author of relevant previous work. - Advice in presentation and contextualisation. 	15%
H. Ulrich Hoppe	<ul style="list-style-type: none"> - Supervision and advice in conceptualisation and contextualisation. 	10%

⁴ Hecking, T., Hoppe, H.U. & Harrer, A. 2015. Uncovering the Structure of Knowledge Exchange in a MOOC Discussion Forum. In *Proceedings of the 7th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 1614-1615), Paris, France, ACM.

Discovery of Structural and Temporal Patterns in MOOC Discussion Forums

Tobias Hecking, Andreas Harrer, H. Ulrich Hoppe

Abstract This work aims to explore methods to investigate the structure of knowledge exchange in discussion forums in massive open online courses (MOOCs) explicitly taking into account changing patterns over time. The paper covers three different aspects of forum analysis combining different methods. First, an approach for the extraction of dynamic communication networks from forum data based on the classification of forum posts is presented that takes into account the information exchange relations between forum users. Second, measures that characterise users according to information seeking and information giving behaviour are introduced and the development of individual actors is analysed. Third, blockmodelling and tensor decomposition approaches for reducing a dynamic network to an interpretable macro-structure reflecting knowledge exchange between clusters of actors over time are evaluated. This allows for the analysis of the communication structure related to information exchange between participants of large scale online courses in different aspects. The utility of the analytics framework is demonstrated along two case studies on forum discussions in two MOOCs offered on the Coursera platform.

1 Introduction

Discussion forums are a common element in massive open online courses (MOOCs). Since online courses with a large amount of participants cannot be supported individually by a tutor, discussion forums are often used for information seeking and information giving among the participants themselves. In previous studies it was

Tobias Hecking
University of Duisburg-Essen, e-mail: hecking@collide.info

Andreas Harrer
Univeristiy of Applied Sciences Dortmund, e-mail: andreas.harrer@fh-dortmund.de

H. Ulrich Hoppe
Univeristiy of Duisburg-Essen, e-mail: hoppe@collide.info

shown that only a small fraction of all course participants actively take part in forum discussions. However, those who frequently participate in activities and possibly complete the course, which is also only a small fraction of all participants [6], are much more likely to be active in the forum [2]. Recent studies also relate the forum activity of course participants to influence in the community of learners [35]. This leads to the assumption that discussion forums are a communication channel for knowledge exchange between the core cluster of participants who are really willing to finish the course. This and the fact that course forums are more restricted in the topics that can be discussed makes these types of discussion forums an interesting case to study phenomena of knowledge exchange in electronic communities.

The goal of this work is to uncover the structure of knowledge exchange in MOOC discussion forums adapting mixed methods. The work is a significant extension of our previous work [17], incorporating new datasets and refining methodology. There exist several studies on forums for knowledge in different research areas including network analysis. However, most of the existing approaches analyse a static snapshot of a forum. This paper extends this body of research by investigating individual behaviour as well as interpersonal relationships between actors with a special focus on temporal dynamics, i.e. changes of behaviour and roles of actors over time. It aims at exploring the full process of structural analysis of discussion forums including network extraction, characterisation of individuals, and uncovering the latent meso-level structure of the communication networks. Thus, our approach covers a range of analytical methods combining text analysis and social network analysis techniques. This process includes the following three steps:

(Step 1) The first crucial step in applying network analysis methods to forum data is the extraction and modelling of a social network from forum threads that reflects directed knowledge exchange relations between actors. This is not a trivial task since the fact that two actors are active in the same discussion thread does not imply that these two also share knowledge in the sense that one of both replies to an information request of the other. Thus, the first step is to identify information seeking posts and related information giving posts in the discussion threads of the forums based on textual and structural features. This leads to a directed network of forum posts where each post that provides information points to one or more information seeking posts. As a second step, this network is transformed into a directed network of actors where the edges represent information giving relations. All edges carry timestamps, which allows for splitting the network into time slices that represent the information seeking/information giving structure of the forum communication in a particular period of time.

(Step 2) The extracted networks are analysed on the level of individuals in terms of trajectories of information seeking and information giving behaviour. Clustering actors according to similar trajectories yields interesting insights into the development of forum actors during the online course. Therefore, we define measures that properly characterise the information seeking and information giving behaviour of forum users taking into account, both, the number of connections they have and their post quantity.

(Step 3) In discussion networks where potentially everyone can talk to everyone else there is no obvious structure or network topology. Especially in networks of the size of the ones under consideration and larger, it is almost impossible to make statements on the possible latent organisation of the network structure. For this reason different approaches can be applied to reduce the network to a macro-structure that captures the interaction pattern between components of the network, and thus, allow for a better interpretability of the latent organisation of the network. In particular, blockmodelling and tensor decomposition methods are modified and evaluated to perform this task for dynamic networks. Eventually, the discovered macro-structures are mapped to a knowledge exchange graph that depicts information flow between clusters of actors over time.

The methods are evaluated with and applied to anonymised datasets of two discussion forums of MOOCs offered on the Coursera platform¹ which are described in more detail in [29]. The first course is on “Introduction to Cooperate Finance” conducted during 11/2013 and 12/2013. Overall there were 8336 posts in 870 different threads by 1540 different actors. The second course is on “Global Warming: The Science and Modeling of Climate Change” with a discussion forum comprising 1007 actors with 5546 posts in 1020 different threads. For both datasets the forum activity peaked in the beginning and decreased afterwards until the end of the course which is typical for a MOOC discussion forum. In the following we refer to this dataset as “Corporate Finance” or “Global Warming” respectively.

The paper is structured as follows: After this introduction, Section 2 gives an over-view of the related work of discussion forum analysis and the methodological back-ground of this paper. The mentioned analysis steps are discussed separately in Sections 3-5. Results of the application of the developed methods to the two datasets are presented in Section 6. Section 7 concludes the paper and gives an outlook on possible directions of further research.

2 Background

2.1 Analysis of Discussion Forums

Research on discussion forums can be classified into content related and structure related analysis. Content related analysis deals with the content of the forum posts. Typical tasks are post classification, and discussion disentanglement. Especially in the case of MOOC discussion forums it is of huge interest to identify content related threads in which exchange of knowledge between participants takes place [29, 8] as well as the estimation of discussion quality [20].

In contrast to the content related analysis where the individual posts or threads are the main object of inquiry, structural analysis aims to model relations between entities in discussion forums in order to answer questions on knowledge diffusion

¹ <https://www.coursera.org/>

Tobias Hecking, Andreas Harrer, H. Ulrich Hoppe

through forum communication. One of the first studies on online mass communication by Whittaker et al. described dependencies between different properties of Usenet clusters such as thread depth, message length, and demographics [33].

Based on the information who gives information to whom in a question answer forum, one task is the identification of expert actors. Measures of expertise can be based on the quantity of questions and answers actors post to a forum or their position in the Q/A communication network [37]. Adamic et al. [1] draw upon this research by investigating forum data from Yahoo answers. They took a deeper look into the structural properties of networks between information givers and information seekers as well as their interest in topics. The gained knowledge could be applied to distinguish different types of discussions and best answer prediction. Most recently, Gillani et al. [14] compared the structure of social networks extracted from different MOOC forums according to vulnerability and information diffusion. They could show that the cohesiveness of the networks depend on very few actors for most of the investigated forums.

An initial challenge that usually arises when network analysis methods are applied to forum data is the modelling of the underlying network. Especially in forums that have a non-nested thread structure, it is a challenge to establish relations between posts and consequently also between posters. While in forums like Yahoo answers, the relations between questions and answers are directly observable by the thread structure; this is usually not the case in MOOC forums. This problem is well known in language processing [21], however, there are only a few studies in network analysis research that tackle this problem explicitly, for example, [14, 28]. Thus, the study in this paper combines both content related and structure related properties of discussion forums to model the underlying social network.

2.2 Identification of Roles in Communication Networks

Role modelling is of particular interest to characterise actors, for example, peripheral participants or “lurker” or active advice givers in the community [11]. The identification of the fundamental structures and topologies of networks is a means to understand the nature of interaction and actors’ behaviour in complex networks. Role models cluster actors based on their position and connection patterns in the network. Thus, a cluster can be interpreted as users with similar role in the network. In contrast to community detection actors in the same cluster do not necessarily have to be densely connected within their cluster but only sparsely to outsiders. Moreover, role models do not require any connections between actors of the same cluster at all, although they are not forbidden. The goal is to identify connection patterns between clusters, which reduces the complex network structure to an interpretable macro-structure that reflects latent connection patterns between different roles of actors in the network.

2.2.1 Blockmodelling

Analyses that group with similar actors are subsumed as “positional analysis” [31]. One of the prominent approaches in that strand is “blockmodeling” [9] that assigns actors to clusters of similar relations with other actors and reduces the actor network to a more coarse grained network of clusters that represents the original network structure on a higher level (represented by an image matrix or block matrix). Those patterns can also be detected when analysing the algebraic composition of multiple relations [27], which is especially well suited for understanding the fundamental structure of a network when combined with a positional network that already reflects the basic structure of the network.

Connection patterns and role structures can be derived from different notions of equivalence and similarity between nodes [23]. Two important notions of equivalence of actors in a social network are structural and regular equivalence [31]. Two actors are structural equivalent if they have exactly the same neighbours in the network and consequently are not separated by more than two steps. A partitioning of the actors of a network based on structural equivalence results in a blockmodel with only complete or null relations between clusters (Figure 1). This definition is usually too strict for typical sparsely connected information exchange networks. Thus, a more appropriate notion of equivalence is regular equivalence. Two actors are regular equivalent if they have the same connections from and to equivalent actors [32]. If there exists a regular relation from a cluster c_i to a cluster c_j , all actors in c_i point to at least one actor in c_j and all actors in c_j have at least one ingoing relation from actors in c_i . Consequently, regular equivalence blockmodels only contain regular or null relations (Figure 1). Note that complete relations are a special case of regular relations. The different types of relations between clusters can be expressed by a $|C| \times |C|$ image matrix, where $|C|$ is the number of clusters. Each element of the image matrix gives the type of relation found between each pair of clusters.

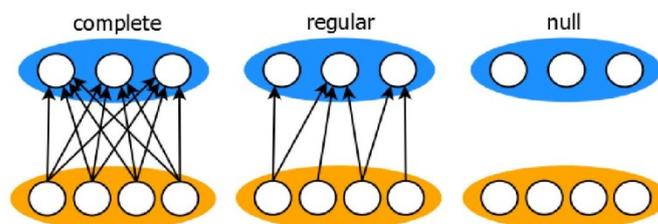


Fig. 1 Different types of blocks (relations between clusters).

Using a notion of equivalence to partition the network is often too strict and can lead to unrealistic results. Finding a partition of a network into clusters of regular equivalent nodes likely can result in an inappropriate high number of small clusters comprising only one or two nodes or only one cluster comprising all nodes (if the network does not contain sources and sinks). Thus, different relaxations can be

applied [34]. On the one hand, one can relax the assignment of nodes to clusters such that each node has assigned a weight for each cluster that reflects to what extent the node belongs to this cluster [34]. A more common way is to relax the equivalence requirement for the nodes in the same cluster to similarities such that nodes within the same cluster should be almost structural or regular equivalent [9]. This is also the line followed in this paper. To measure the goodness-of-fit of such approximate blockmodels one has to compare the derived block structure with an ideal model that perfectly matches the desired relations, e.g. regular equivalence. The error of such approximate blockmodels is measured according to the minimum number of modifications (adding and deletion of edges) to be made such that the blockmodel is ideal.

Finding a nontrivial partition of nodes into k clusters such that they resemble an ideal blockmodel as good as possible is an NP-hard problem [34]. Different approaches exist to approximate an optimal solution [9]. The direct approach aims to optimise an initial partition by iteratively moving nodes from one partition to another or switching the clusters of two nodes such that the blockmodel error is continuously minimised. This procedure is computationally expensive and not feasible for large networks. Alternatively, one can compute the extent of regular equivalence of nodes (regular similarity) beforehand and cluster the nodes into clusters using arbitrary clustering algorithms. This indirect approach is much faster than direct optimisation but does not guarantee to find a local minimum of the blockmodel error.

2.2.2 Tensor decomposition for role modelling

Apart from the complexity of the direct optimisation of a block structure, there are other drawbacks of blockmodels. First, traditional blockmodelling approaches tend to focus very strongly on the typical core- periphery structure of forum networks comprising a large cluster of active actors and many small clusters of sparsely connected actors. This is, on the one hand, a reasonable macro-structure but might not reflect the possibly latent interactions between different clusters of actors. Alternative approaches for mapping the network structure to a higher-order structure are based on tensor decomposition methods. These approaches are well suited for dynamic networks since the adjacency matrices of successive time slices of the network can be stacked to a third-order tensor (see left side of Fig. 2). These approaches are not only successfully used in relational learning tasks such as link prediction [10] and community mining [12] but are also applicable for role modelling [22, 25]. In contrast to blockmodelling, these methods do not optimise the assignment of nodes to clusters/roles towards fitting a target tensor that reflects an ideal block structure. Moreover, these methods optimise the partition towards the tensor representation of the evolving network itself.

Two related methods that are suitable for modelling dynamic and asymmetric relations between latent clusters are RESCAL [25] and DEDICOM [3]. The adjacency matrices for each time slice of an evolving network are modelled as a third-order tensor $X \in \mathbb{R}^{(|Act| \times |Act| \times T)}$, where Act is the set of actors in the network and T the

number of time slices. The identification of actor roles and relations between them is performed simultaneously by finding a good fitting decomposition of X such that the following equations (12) are minimised:

$$\min_{A,R,D} \sum_{k=1}^T \|X_k - A \times D_k \times R \times D_k \times A^T\|_F \quad (1)$$

for DEDICOM and for RESCAL:

$$\min_{A,R} \sum_{k=1}^T \|X_k - A \times R_k \times A^T\|_F \quad (2)$$

RESCAL can be considered as a relaxation of the DEDICOM model since it allows for varying relations among the latent roles over time. Since it is expected that relations between roles in discussion forums are also dynamic, in the following, this work is restricted to the RESCAL decomposition of the adjacency tensor. Graphically this decomposition can be depicted as in Figure 2. The decomposition of the tensor can be efficiently computed using algorithms based on alternating least squares that update the matrices A , R_k , D_k in alternating fashion by minimising the objective functions in Equations 1 and 2. For more details of the concrete procedure we refer to [25] and [3].

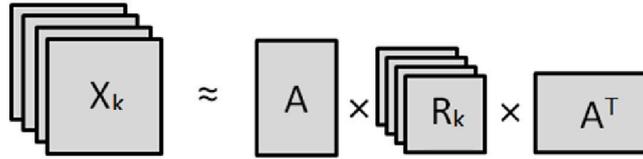


Fig. 2 Graphical depiction of the RESCAL decomposition

The rows of the matrix $A \in \mathbb{R}^{|A_c| \times |C|}$ correspond to the actors in the network and the columns to the clusters. Each element of $a_{i,j}$ of A indicates a loading for actor i for a latent factor (cluster) j . The matrices $R_k \in \mathbb{R}^{|C| \times |C|}$ can be interpreted as the latent relations between the clusters $c_j \in C$ for each time slice. Consequently, the matrices can be seen as a specific notion of an image matrix as used in blockmodelling. If two nodes have similar loadings for the latent factors according to the matrix A , they have similar relations to other clusters or roles over time according to the relation tensor R . This bears some similarity with blockmodelling based on regular similarity in dynamic networks. The combination of both will be described in more detail in Section 5.

2.2.3 Estimating the number of clusters

All the investigated methods require a pre-specified number of clusters (roles). There is no general rule for good choices of this parameter and it has even been stated that this decision is “more an art than a science” [3]. The blockmodel error tends to decrease with growing number of clusters. The reason is that the more and smaller the clusters are the easier it is to establish nearly regular relations with a small number of relations between two clusters. If in an extreme case the number of clusters is equal to the number of nodes (each cluster contains one node), the block-model error is 0 since the set of nodes induce a perfect (but not desirable) regular equivalence partitioning. Moreover, the parameter specification should be reasonable for the goal of the analysis. Since the identified macro-structure models should be interpretable, typically one uses domain knowledge and hypotheses to decide on the number of clusters. This leads to a pre-specification of the structure (types of relations, number of roles) one aims to identify. This can also be extended to a deductive model fitting approach ([9] pp. 233-244).

3 Network Extraction from Forum Posts

Starting with the lists of posts for each discussion thread, the goal of the network extraction step is to model a directed knowledge exchange network between the forum actors that reflects the information giving relations between them as good as possible. Each post entry contains information of the discussion thread it belongs to, the sub-forum, the (anonymised) actor, and the post content. There are two types of posts, namely regular posts and comments. Regular posts occur in linear sequence without sub-threading. In Coursera forums, actors can comment on regular posts but it is not possible to comment on a comment. The result is a flat thread structure in which regular posts are arranged in linear fashion, and comments always have a single parent post. While it is easy to relate a comment to its parent post, it is more complicated to relate a regular post to previous regular posts. In the following the network extraction steps are described in detail.

3.1 Forum post Classification

Since not every forum post can be related to knowledge exchange, in a first step, information seeking and information giving forum posts have to be identified. There exist several tag sets for classes of forum posts which are often very fine grained, e.g. differentiating between initial questions and repeated questions and different types of answers [21]. However, since we are only interested in persons who have provided information or asked for some information in the forum, no further distinction of different types of information seeking and information giving posts is made.

The result is a simplified tag set comprising “information seeking”, “information giving”, “social posts”, and “others”. Information seeking posts are course content related questions and problem descriptions. Information giving posts subsume all posts that provide some content related information. Apart from that, the forum is also extensively used by people who search for study clusters and general discussions. Those posts are labelled as “social” posts. All other posts that do not fit in any of the other categories are classified as “others”. This classification also goes along with the observations made in [30]. Information seeking and information giving posts are the only relevant categories for the following studies. Social posts and others are only used for the purpose of filtering. Since discussion forums in Coursera courses are organised in sub-forums, it is easy to filter “social posts” since they usually occur in dedicated sub-forums that are not used for information exchange. The content related discussions take place in sub-forums that especially target content related issues regarding lectures and assignments. Thus, automatic classifiers are trained only to detect “information seeking”, “information giving”, and “others” posts from these types of sub-forums. Each forum post is represented by a vector of features adapted from [21]. In addition more content related features and a flag that indicates whether a post is a regular post or a comment were added. Content related features are different variations of indicator phrases that suggest different post types. While phrases mapped to the “help_me” indicator phrase such as “need help” clearly indicate an information seeking post, other phrases such as “thanks” indicate a post that can be labelled as “others” such as “Thank you for your help”. Table 1 gives an overview of the features used to encode each forum post. This list results from a larger list of structural and content features. A good selection and combination of features was derived by applying a genetic algorithm [24]. Starting with 30 random combinations 10-fold cross validation was used as fitness function to assess the quality of a combination. A new generation of subsets of features are created out of the best 25% of previous combinations performing tournament selection [24]. The result is the best combination found in 30 generations (Higher number of generations does not lead to further improvements).

A bagged random forest classifier [5] yielded the best results. The classification model was trained on 500 posts that were hand-classified by three experts by taking the majority of manual assigned classes for each post (interrater agreement according to Fleiss-Kappa $\kappa = .78, p < .005$). The classification has been evaluated using 10-fold cross validation. The F1-score for the classification of information seeking posts is good (F1-score = 0.77) and acceptable for information giving posts (F1-score = 0.66). However, posts of type “other” often lead to misclassifications as Table 2 shows. For this reason, the final classification was done by an iterative classification procedure (c.f. [26]). This algorithm uses an additional classifier trained on a dataset where the types of preceding posts of each posts are known. This makes the classification of information giving posts easier since those posts must have at least one preceding information seeking post. An initial classification is retrieved as described before. Then the additional classifier is applied to the data with the initially assigned class labels. This increases F1-scores for information giving posts

Table 1 Feature Set for Post Classification

Feature	Description
Forum ID	Not used for training but to restrict the automatic classification to sub-forums dedicated for course content related exchange.
Lexical similarity with initial post	Initial posts are often questions. High cosine similarity of the word vectors of a post and the initial post after stopword removal is an indicator for an information giving post.
Votes	Number of votes for each post. In Coursera actors can rate posts of other actors. Usually only content related posts receive votes whereas "other" posts like "Thank you for your answer" are unlikely to receive many positive votes.
Order	Position in the thread. The first post has position 0. The second 1. The position of a comment is its position in the comment chain + the position of its parent post. Information seeking posts often appear at the beginning of a discussion thread followed by information giving posts.
Is comment	Is the post a regular post or a comment on a previous post?
Length	Number of words of the post after stopword removal.
Question mark	Question marks are a good indicator for information seeking posts.
Exclamation mark	Exclamation marks occur often in information seeking and information giving posts, but not in "other" posts.
FiveWoneH	Number of occurrences of "Why", "What", "Where", "When", and "How". A high number indicates information seeking posts.
Special phrases	Phrases that indicate specific post types, i.e. variations of: help_me (e.g. "need help"), help_you, thank, did not, similar (e.g. "same problem"), wrong

increases to 0.71 and for information seeking posts to 0.79 based on evaluation on another set of 200 hand-classified posts.

Table 2 Precision and Recall of the Post Classification based on 10-fold cross validation.

	Precision	Recall
Information seeking	0.81	0.74
Information giving	0.61	0.73
Other	0.68	0.53

3.2 Network Extraction

After post classification, the next task is to extract the knowledge exchange network. First, all posts that are not classified as either information giving or information seeking are removed from each thread leaving only question answering threads. The threads in Coursera discussion forums are not nested, and thus, it is not directly

visible from the thread structure which post refers to which previous posts. However, it is possible to attach comments to particular posts. Comments are often used to refer to older posts in the thread when the discussion leading to some kind of sub-thread. Thus, a network of posts is built according to the rules described in the following:

- **Information seeking/giving sequence:** A sequence of information seeking post is usually succeeded by a sequence of information giving posts. After the information giving sequence sometimes a new information seeking sequence starts followed by another information giving sequence. Thus, a thread can be decomposed into a set of such information seekinggiving sequences. The first rule applied is to link each of the information giving post to the information seeking posts in each information seeking/giving sequence in a thread.
- **Comments as sub-threads:** Comments are attached to a parent post. Comments attached to a parent post are treated as regular posts of a sub-thread with the parent post as initial post. Then the sequence rule described before is applied to this sub-thread.

The comparison of the extracted relations by the rules and the relations identified by the human classifiers yields high accuracy (F1-score = 0.89). The performance achieved by the rules described above indicates that actors in the investigated MOOC forum usually maintain the structure of a thread themselves, using comments instead of regular posts if they refer to much earlier posts. Consequently the discussion structure in threads is usually not very entangled and the simple rules are very effective for network extraction. We also tried to incorporate the lexical overlap between the posts as it was done for the case of chat disentanglement [18], but this did not improve the results. A reason can be that discussion threads structured according to discussion topics already induce a lexical overlap of many of the contained posts such that these features do not provide new information.

In the post network resulting from the previous step each post node is annotated with the author of the post and a timestamp. Furthermore, the network is highly disconnected since links are based on individual forum threads. Next, each post node labelled with the same author is collapsed into a single node representing the actor resulting in the final knowledge exchange network between forum actors, similar to [16]. Figure 3 gives an example:

The actors who post neither information giving nor information seeking posts do not appear in the extracted network since “social” and “other” posts were ignored. After deletion of isolated actors and the actor representing all anonymous posts, the resulting network for the “Cooperate Finance” course comprises of 647 actors who actively participated in content related forum discussions by asking for information or providing information to information seekers. The network is very sparse with only 1303 edges. For the smaller course, “Global Warming”, there remain 348 actors as nodes in the extracted network. The actors in this network are more densely connected (1291 edges) than in the previous one.

Tobias Hecking, Andreas Harrer, H. Ulrich Hoppe

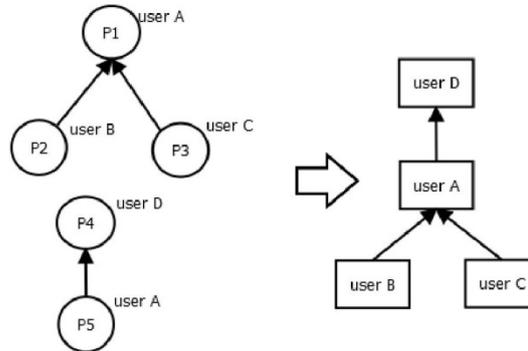


Fig. 3 Mapping a post network to a network of actors.

4 Individual Development - Behavioural Roles over Time

In the following we present the analysis approach for identifying patterns of individual development of forum actors over time. The goal is to characterise actors according to their information seeking and information giving behaviour. The question is whether there exist actors who are important in the sense that they either provide much information (information giving) or ask important questions or raise problems that stimulate the community to respond (information seeking). Simply counting the number of information giving and information seeking posts of each actor can be too simple. Someone who has a high number of information giving posts might not necessarily be a real information giver in the community since the high number of posts can result from a long discussion with one single actor [37]. On the other hand taking the out-degree or in-degree centrality as a characteristic measure can also lead to unreliable results since a high degree can result from one single post. Consequently, an information giver should have many information giving posts reaching many different information seekers and an information seeker should have many information seeking posts that are replied by many other actors. Thus, our solution is to use a combined measure of the number of posts and the diversity of connections resulting in two measures we refer to as outreach and inreach that will be defined below.

4.1 Definition of Inreach and Outreach

Given a single node of a weighted and directed network, the diversity of its in and outgoing relations can be characterised by a measure of entropy. Equations 3 and 4 calculate the diversity of outgoing and ingoing relations for a node i , where $w(e_{i,j})$ is the weight/multiplicity of an edge from i to j and $od(i)$ is the out-degree and $id(i)$ the in-degree of i (taking into account edge weights, which is equal to the number of its information giving posts or information seeking posts, respectively, except in

rare cases where one information giving post matches multiple information seeking posts).

$$H_{out}(i) = \frac{-1}{od(i)} \sum_{j \in outneigh(i)} w(e_{i,j}) * \log\left(\frac{w(e_{i,j})}{od(i)}\right) \quad (3)$$

$$H_{in}(i) = \frac{-1}{id(i)} \sum_{j \in inneigh(i)} w(e_{j,i}) * \log\left(\frac{w(e_{j,i})}{id(i)}\right) \quad (4)$$

The value for the connection entropy of node i reaches its maximum, if all posts of i address different nodes and its minimum 0 on the other extreme. In order to combine diversity and posting activity the number of corresponding posts of node i ($= od(i)$ for information giving, $= id(i)$ for information seeking) can be multiplied with $(H_{out}(i) + 1)$ or $(H_{in}(i) + 1)$, respectively, resulting in equations 5 and 6 for the outreach and the inreach.

$$outreach(i) = od(i) - \sum_{j \in outneigh(i)} w(e_{i,j}) * \log\left(\frac{w(e_{i,j})}{od(i)}\right) \quad (5)$$

$$inreach(i) = id(i) - \sum_{j \in inneigh(i)} w(e_{j,i}) * \log\left(\frac{w(e_{j,i})}{id(i)}\right) \quad (6)$$

As the result, the outreach of actor i is at minimum the number of its information giving posts, if all posts address the same actor. The statement is similar for the inreach of an actor but with respect to its ingoing (information receiving) relations. Consequently the two described measures are helpful to detect active actors that reach or are reached by many other actors in the network. The comparison of the outreach and inreach of actors allows for a characterisation of the actors with respect to their behaviour, i.e. information seeking and information giving.

4.2 In- and Outreach over time - Identification of Characteristic Actor Trajectories

The behaviour of an actor in a course forum usually will change over time. To cope with this, the next step after the definition of the in- and outreach measures is to uncover typical trajectories of information giving and information seeking behaviour of actors over time. Since the edges of the evolving knowledge exchange network carry timestamps it is possible to calculate the in- and outreach of the actors at different times by splitting the dynamic network into successive time slices. However, calculating these measures for each time slice independently does not account for a possible long term effect of actor activities. Consequently the calculation of an in- and outreach trajectory of an actor has to take into account the history of the actors behaviour as well. In time slice t of the network older edges should not be weighted as high as recent edges but can still have an effect. In order to achieve this

Tobias Hecking, Andreas Harrer, H. Ulrich Hoppe

we apply a growing window approach to model the evolution of the knowledge exchange network at different points in time. The approach includes a linear forgetting function which weights the edges according to their recentness as suggested in [7]. The parameter Θ can be used to control the extent of decline of edge weight. Other weighting functions, for example, exponential decline of edge weights are also possible. The weighted adjacency matrix $wAdj_t$ of the t th time slice is then calculated as a weighted sum of the unweighted adjacency matrices Adj_i of the network in the time slices up to t as in equation 7.

$$wAdj_t = \sum_{i=1}^t Adj_i * \Theta^{t-i}, \Theta \in (0, 1] \quad (7)$$

The calculation of in- and outreach on each of the resulting weighted networks derived by this growing window approach results in an inreach sequence and outreach sequence for each actor. The growing window approach for modelling evolving networks is an additive procedure and the information of earlier time slices is not completely lost in later time slices. An actor who stops communicating in the forum has still a value greater than 0 for in- or outreach lowered by the forgetting function in each slice. Thus, the slices should be kept rather short to capture the dynamics of the changes of individual in- and outreach. In this work this size is fixed to 3 days since it can be assumed that a forum post receives most of the reactions within this time window [35]. The goal is to discover characteristic patterns in the actor behaviour over time based on the similarity of those sequences. A naive approach would consider the sequences as numerical feature vectors and apply e.g. k-means clustering to reduce the set of sequences to a small number of clusters of similar sequences. However, using this approach, both, the inreach and outreach sequences have to be considered simultaneously resulting in a high number of dimensions of the feature vector which can be problematic for traditional clustering methods. Thus, k-medoids clustering [19] is applied in an alternating fashion. First, clusters are derived by partitioning the actors according to their inreach sequences. In a second step the medoids of the found clusters are used to initialise the k-medoids clustering according to the outreach sequences. This procedure alternates until the clusters have stabilised.

The medoids of each cluster can be considered as its prototypical representative. Thus, this method is appropriate to reduce the vast amount of individual sequences to uncover an interpretable set of typical trajectories actors can have in the knowledge exchange forum. Results in Section 6.1 will show the utility of this approach to identify the emergence of different behavioural roles and their changes over time.

5 Macro-structure of Evolving Knowledge Exchange Networks

In both of the dynamic knowledge exchange network communication links repeat very rarely over several time slices. More than 80% of the communication links be-

tween two actors occur only once. This leads to the assumption that in the forum there exist no stable cohesive subcommunities of actors over time. However, there can be actors who behave similar over time with regarding their connection patterns to others without necessarily having a direct connection. Thus, reducing the network to a macro-structure that reflects the information flow between different clusters of actors according to connection patterns can lead to interesting insights into the overall structure of the knowledge exchange in discussion forums. As described in Section 2.2, for the discovery of those macro-structures in evolving networks there exist different approaches, blockmodelling and tensor decomposition. Both have certain advantages and disadvantages for the task. In the following the two approaches are compared regarding the utility for the analysis of knowledge exchange networks. We will also propose adaptations of existing methods that better fit the needs for uncovering macro-structures of knowledge exchange.

In particular we aim to find regular relations between clusters. On the one hand, complete relations based on structural equivalence, as described in Section 2.2, are too strict for our sparse forum discussion network. On the other hand, regular relations reflect information flow between clusters. For information giving relations between actors, regular relations between clusters can be interpreted as existing information flow from cluster c_i to cluster c_j . This does not require that everyone in cluster c_i has to talk to everyone in cluster c_j .

5.1 Dynamic Blockmodelling

The task of finding a well-fitting block structure of a network as described in Section 2.2 becomes even more complex in dynamic networks. An ideal blockmodel for dynamic networks defines a partitioning of the nodes into k clusters such that the nodes within the same cluster are almost regular equivalent in each time slice. For a given dynamic network sampled in T time slices, the goal is then to find a partition of the nodes that is as close as possible to an ideal dynamic blockmodel in each time slice.

A simple approach for indirect blockmodelling in dynamic networks would be to calculate the extent of regular equivalence (or regular similarity) $REGESim_t(i, j)$ for a pair of nodes i and j in each time slice t and take the average as in equation 8.

$$dynamicREGESim_{i,j} = \sum_{t=1}^T \frac{REGESim_t(i, j)}{T} \quad (8)$$

For computing the regular similarity of all node pairs the REGE algorithm [4] is applied. The resulting overall similarity can then be used as input for a clustering algorithm that assigns the nodes to clusters. Here hierarchical clustering is used. In the following this approach is referred to as SIDBM (Sequential Indirect Dynamic Blockmodelling). However, focusing on each time slice separately has the disadvantage that if the similarity of node pairs varies heavily over time, a high similarity

Tobias Hecking, Andreas Harrer, H. Ulrich Hoppe

between two nodes in only one time slice can have a huge effect on the outcome even if the nodes are very dissimilar in other time slices.

To reduce this problem, the incremental blockmodelling method for multi-relational networks proposed by Harrer and Schmidt [15] can be utilised. Optimising partitions across time slices of evolving networks can be considered as a special case of multi-relational blockmodelling when the edges in each time slice are considered as edges of one particular relation. The approach identifies a blockmodel across multiple relations as follows:

1. Find a blockmodel for each relation (time slice) starting from a random partition.
2. Select the blockmodel that yields the smallest average error for each partition.
3. Find other blockmodels by optimising the partition found in step 2 for each time slice.
4. Repeat steps 2 and 3 for n iterations and select the blockmodel that yields the smallest error.

This approach guarantees to find a partition of the nodes according to a given equivalence relation with a local minimum of the average blockmodel error across time slices. However, the third step incorporates direct optimisation of a given network partition to fit a certain blockmodel which is computational expensive especially for many time slices, and thus, it is not feasible for large networks. We propose an adaptation of the method that keeps the original procedure of incremental updates of the node partition but uses the indirect (similarity based) clustering instead. In the first step the extent of regular equivalence the REGE algorithm [4] is applied to compute the extent of regular equivalence for each node pair in each time slice t . Instead of directly optimising the partitions in step 3, the similarity of the nodes in time slice t $dynamicREGESim_t(i, j)$ is computed as the average regular similarity of the nodes in the time slice t and the similarity of the nodes of the time slice s which yields the best fitting blockmodel across all time slices (see equation 9).

$$dynamicREGESim_t(i, j) = \frac{REGESim_t(i, j) + dynamicREGESim_s(i, j)}{2} \quad (9)$$

In the next iteration again the similarity $dynamicREGESim_s$ that yields the best fitting blockmodel for all time slices is chosen to re-compute the similarities of node pairs according to equation 9. After a defined number of iterations (in this work 25) the best blockmodel will be returned. In the following, this method is referred to as IIDBM (Incremental Indirect Dynamic Blockmodelling).

5.2 Role Modelling based on Tensor Decomposition

While the more traditional blockmodelling approaches optimise the partitioning of the nodes of an evolving network towards ideal image matrices for the time slices (or

target tensor), RESCAL (see Section 2.2) aims to approximate the adjacency tensor itself by inferring the loadings of the nodes to latent clusters. Thus the results of link based clustering can differ much from an ideal regular equivalence blockmodel. In this aspect, the results are not as easy to interpret as traditional blockmodels. For example, the values of the relation matrices R_k (equation 2) do not have clear semantics in terms of certain types of relations and node equivalences such as the typical image matrices produced by blockmodels. To overcome this problem we introduce a slight modification of the RESCAL approach for link based clustering that biases the clustering of the rows of the matrix A towards a given type of node equivalence. Instead of clustering the rows of the resulting matrix of loadings A directly, A is modified to A' by the following matrix multiplication (equation 10):

$$A' = S \times A \quad (10)$$

$S \in \mathbb{R}^{|Act| \times |Act|}$ can be any similarity matrix between the nodes (actors) in the network. With this simple modification, the elements $a_{i,r}$ of the role matrix $A' \in \mathbb{R}^{|Act| \times |C|}$ contain high values if actor i has a high loading for cluster c_r and also a high similarity with other actors who also have high loading for cluster c_r . Consequently this approach biases the assignment of nodes to clusters towards a certain type of relations. Thus, in the following this adaptation of link clustering based on RESCAL is referred to as biased RESCAL and the matrix S is considered as the average regular similarity of node pairs as given by equation 8 in the section before. In order to allocate nodes to clusters uniquely any partitioning clustering (in this work hierarchical clustering) approach can be applied to the rows of the role matrix A' .

5.3 Formal Evaluation

Since there is no ground truth data for evaluation (i.e. no predefined classification of actors as in Bader et al. [3]), the described macro-structure models can only be verified by comparing the model to the actual data. One obvious measure is the blockmodel error that results by comparing the node partitioning to an ideal regular equivalence blockmodel comprising either regular or null blocks described in Section 2.2. Second, the discovered regular relations between clusters are evaluated with respect to the density of links between the related clusters.

As mentioned in Section 2.2 the number of clusters has to be decided based on empirical observations and hypotheses about the network structure. Based on previous studies it is reasonable to assume a core-periphery structure with a core of actors who are well connected and active in the forum [35, 13]. It can also be assumed that there are many peripheral nodes that appear either as information seeker or information giver only a very few times. Further, it is likely that there exist a (semi-peripheral) role comprising mostly information givers and a role comprising of mostly information seekers.

Tobias Hecking, Andreas Harrer, H. Ulrich Hoppe

Thus, only results where the number of clusters was fixed to 5 are reported. Higher numbers scale the resulting blockmodel error but do not lead to changes in the ranking of the models. Further, the resulting models are more difficult to interpret in application scenarios. The size of a time slice was set to 2 weeks of forum communication resulting in 3 slices for the “Corporate Finance” forum and 4 slices for the “Global Warming” forum. This choice has been made since the most MOOCs are organised in thematic blocks of one or two weeks and it is likely that the forum communication is oriented towards this pace. It was shown that the size of time slices have a systematic effect on clustering outcomes and a good resolution can be achieved if the time slices are long enough to capture the typical duration of production cycles of the studied communities [36].

5.3.1 Fitting an ideal regular block structure

As already mentioned regular block structures are especially suited for the mapping of information exchange as evidence for existing information flow within one or between two clusters. Table 3 depicts the sum of blockmodel errors of the time slices with respect to an ideal regular equivalence blockmodel of the previously described algorithms.

Table 3 Error of regular equivalence blockmodels produces by different methods.

	RESCAL	Biased RESCAL	SIDBM	IIDBM	Random partition- ing
Corporate Finance	828.9	646.39	397.15	321.96	1027.71
Global Warming	872.71	437.01	504.56	447.02	1143.11

It can be seen from Table 3 that the incremental indirect dynamic blockmodelling approach (IIDBM) yields better results than the simple sequential approach (SIDBM) for both datasets. For the “Cooperate Finance” discussion forum, this approach also gives the best fitting blockmodel. The tensor decomposition method RESCAL falls behind which is expectable since it does not explicitly optimise the portioning of the nodes according to regular equivalence. In contrast to that, the biased version of RESCAL leads to slightly better results than IIDBM for the “Global warming” dataset.

5.3.2 Density patterns

Evaluating macro-structures according to the fit to an optimal regular equivalence does not allow for statements about density patterns between clusters of nodes since a regular relation between two clusters c_i and c_j is fulfilled if all nodes in c_i have at least one outgoing relation to one node in c_j and all nodes in c_j have one ingoing relation to one node in c_i . In the following the number of links between the cluster pairs for which a regular relation was discovered by the described macro-structure modelling methods is investigated in more detail. The density of links ρ_{c_i, c_j} between two clusters c_i and c_j is simply the fraction of links actually pointing from nodes in c_i to nodes in c_j , and the number of links that could exist between the two clusters.

Let \mathfrak{R}_{reg} be the set of ordered pairs of clusters $\langle c_i, c_j \rangle$ for which a regular relation exists from cluster c_i to cluster c_j . The average relation density of the regular relations ρ_C for a given clustering C is then given by equation 11 where Adj denotes the adjacency matrix of the network.

$$\rho_C = \frac{1}{|\mathfrak{R}_{reg}|} \sum_{c_i, c_j \in C: \langle c_i, c_j \rangle \in \mathfrak{R}_{reg}} \frac{1}{|c_i| * |c_j|} \sum_{l \in c_i, m \in c_j} Adj_{l,m} \quad (11)$$

The average relation density is evaluated for all cluster pairs for which a relation is closer to be regular than to being non-existent. Results for the different macro-structure modelling methods are given in Table 4 for the time slices of the ‘‘Cooperate Finance’’ network and Table 5 for the ‘‘Global warming’’ network. The number of clusters was fixed to 5, as above. The numbers in brackets denote the number of role relations classified as regular.

Table 4 Average relation density ρ_C of different macro-structure models for the ‘‘Corporate Finance’’ discussion forum (Number of relations classified as regular in brackets).

Time slice	RESCAL	Biased RESCAL	SIDBM	IIDBM
1	.01 (17)	.004 (13)	.004 (9)	.005 (6)
2	.0003 (15)	.002 (4)	.002 (4)	.0001 (8)
3	.0003 (16)	.0002 (8)	.002 (16)	.005 (4)

It can be seen from Tables 4 and 5, that the average density of the relations induced by the RESCAL model is higher on average compared to the models that aim to optimise a regular equivalence blockmodel. Further, the number of relations that are considered approximately regular is much higher. The high blockmodel error for RESCAL reported in Table 3 indicates that the discovered regular relations are much more approximate than for other models. Overall the relations are denser for the ‘‘Global Warming’’ course which results from the higher network density (see Section 3.2). However, the values are overall very close to 0. An explanation is the sparsity of the networks extracted from the discussion forums when those networks are restricted to information giving relations.

Tobias Hecking, Andreas Harrer, H. Ulrich Hoppe

Table 5 Relation density ρ_C of different macro-structure models for the “Global Warming” discussion forum (Number of relations classified as regular in brackets).

Time slice	RESCAL	Biased RESCAL	SIDBM	IIDBM
1	.031 (15)	.022 (9)	.011 (10)	.014 (8)
2	.038 (17)	.018 (9)	.016 (8)	.014 (9)
3	.037 (12)	.013 (7)	.013 (6)	.010 (8)
4	.059 (15)	.015 (9)	.018 (6)	.028 (6)

5.3.3 Assessment of methods

Previously it was shown that blockmodelling and tensor decomposition approaches can be used to uncover different types of macro-structure in knowledge exchange networks. Blockmodelling approaches that cluster nodes based on the extent of regular equivalence between pairs are a useful means to identify parts in the network with specific function. This type of modelling is more suited to identify functional parts of the network like core-periphery structures, clusters of experts or information seekers, etc. In contrast to that, the original RESCAL tensor decomposition model is more capable of reflecting the number of outgoing links that point from c_i into c_j while failing to partition the network according to regular relation and null relation patterns in the network. In general, it can be seen that blockmodelling methods perform much better on the task of fitting a regular relation structure between roles than tensor decomposition on the bigger and much more sparsely connected “Corporate Finance” network. On the other hand, in the smaller and more densely connected “Global Warming” network the biased version of RESCAL and IIDBM perform similar with respect to the blockmodel error and slightly better regarding density patterns. Especially in those networks with denser interrole relations the combination of RESCAL and regular similarity can be considered as a good alternative to the more traditional blockmodelling approaches since it combines, both, a good fitting blockmodel and density of relations. Consequently in the following IIDBM is used for the “Corporate Finance” and biased RESCAL for “Global Warming” dataset.

6 Applications

In the following, the utility of the described methods for modelling behavioural roles (Section 4) and structural roles (Section 5) are demonstrated by applying them to the networks extracted from the two online courses.

6.1 Trajectories of Behavioural Roles

Based on the k-medoids clustering of in- and outreach sequences of the actors described in Section 4.2 different behavioural roles for both of the investigated discussion forums can be identified. The medoids of the clusters are taken as prototypical examples for a set of actors with similar behaviour over time. The networks were sampled in short time slices of 3 days for the reasons explained in Section 4.2. The damping factor Θ for equation 7 for creating in- and outreach sequences was fixed to 0.9 leading to a slight linear decline of the values over time if an actor stops posting to the forum. A proper value for the number of clusters was estimated by optimising the separation of clusters according to the average silhouette width. For the “Corporate Finance” forum 18 clusters and for the “Global Warming” forum 15 clusters were determined. Fig. 4 and Fig. 5 depict the trajectories of some of the medoids representing different types of in- and outreach patterns.

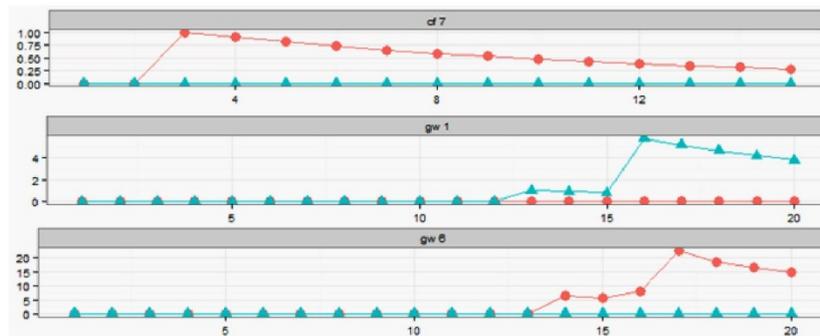


Fig. 4 Post and leave pattern (top). Late starter pattern (bottom). (Red line: inreach, Cyan line: outreach)

The depicted trajectories reflect typical actor behaviour in the discussion forums of the online courses. A pattern that is frequent in both discussion forums is the “post and leave” pattern. 76% in the “Corporate Finance” MOOC and 78% in the “Global Warming” MOOC are clustered around medoids that represent this pattern. Representative trajectories are given by Fig. 4. Those actors post only once very little to the discussion forum and then become inactive. A special case of this behaviour can be considered as “late starting” as the example (gw6) from the “Global warming” forum shows. Those actors are only active in a very late phase of the course and have higher values for inreach or outreach than in the post and leave pattern. Clusters around those medoids are rather small in the “Corporate Finance” course so that only 7% of actors are in clusters that can be classified as late starter cluster. Late starting is much more frequent in the “Global Warming” course where 55% of the actors are clustered around a medoid that can be considered as late starter. Note that a cluster considered as late starter cluster is also a post and leave cluster. An

Tobias Hecking, Andreas Harrer, H. Ulrich Hoppe

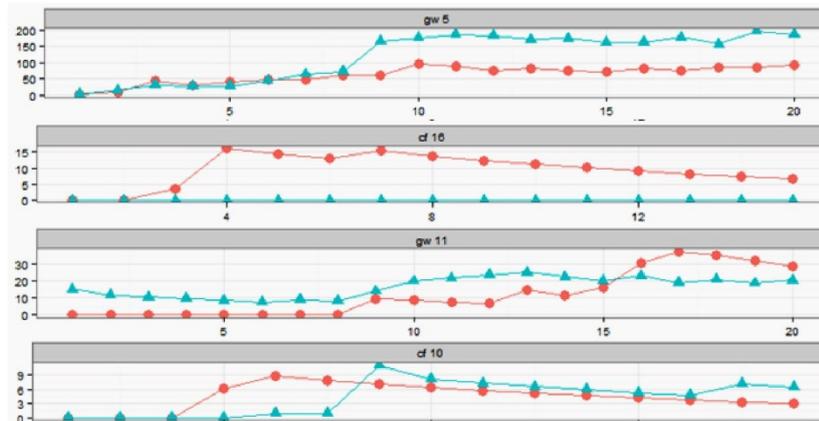


Fig. 5 Patterns of actors who are active during a longer period. (Red line: inreach, Cyan line: outreach)

explanation could be that some actors use the forum to discuss certain issues close to the final exam. On the contrary, there is a smaller set of actors who are active over longer periods and have a high outreach over time indicating expertise or a high inreach indicating that they use the discussion forum mainly for gathering information (“All time information giving/seeking”). Typical examples drawn from the discovered cluster medoids are depicted in Fig. 5. The trajectory at the top of Fig. 5 (medoid of cluster gw 5) depicts an actor who is an all time information giver in the “Global Warming” forum. This actor has an extremely high outreach over time while also receiving lots of information from other actors during the course. This would be typical for course staff, but in this case, it is a regular actor who behaves like a tutor in the course constantly providing information to many others. On the contrary, the second example from the top of Fig. 5 (cluster cf 16) shows a typical cluster representative for the “Corporate Finance” course who has a much higher inreach than outreach over the entire period of observation. Those actors use the discussion forum mainly for gathering information. Very interesting patterns that are quite seldom in both datasets (7% in “Corporate Finance” and 17% in “Global Warming”) are clusters comprising actors who develop from information seekers with higher inreach than outreach to information givers with higher outreach than inreach or vice versa as depicted by the two examples at the bottom of Fig. 5. The medoid of cluster gw 11 of the Global Warming course starts as someone who actively gives information to peers. In the last third of the course the actor develops into an active information seeker indicated by the increase of inreach. On the contrary, in the example at the bottom (cluster cf 10) the actor starts as an information seeker but develops into an information giver. These cases are of particular interest since it reflects personal development through active forum participation.

In general, the temporal pattern of individual development regarding clusters of in- and outreach sequences of the actors reflects the often reported observation that the majority of actors use the forum only occasionally and consequently do not produce many posts. On the other hand, there are some actors who are more active, and thus, often have a more diverse development over time.

6.2 Macro-structures of Knowledge Exchange

In the following the actor roles and relations between them in the two discussion forums under investigation are modelled according to the methods described in Section 5 to derive blockmodels that map the macro-structures of knowledge exchange between clusters of actors that can be interpreted as roles. In Section 5.3 it was shown that it is difficult to find clusters of actors so that actors of one cluster have dense outgoing relations to actors of the same or another cluster. On the other hand, regular equivalence relations could be discovered to some extent. As stated before, those regular relations are considered as important for mapping knowledge flow between roles in the course forums. Thus, the blockmodels derived by the IIDBM method for the “Corporate Finance” and by the biased RESCAL method for the “Global Warming” course were used since these have the best fit for the two datasets in terms of a low blockmodel error and moderate relation density, respectively, (see Section 5.3). Based on the considerations outlined in the beginning of Section 5.3 and the following evaluation the number of clusters roles was fixed to 5 and the networks were sampled in time slices of two weeks. This results in 3 time slices for the “Corporate Finance” and 4 time slices for the “Global Warming” dataset.

The knowledge exchange graphs of the particular time slices merged into a single representation are depicted in Fig. 6. Each cluster is considered twice per time slice, as sending and receiving cluster (depicted as pills). This allows for representing the structure of knowledge exchange between clusters across time slices (T1, T2, T3, T4) as a directed acyclic graph where the edges have a partial temporal order, which is beneficial for depicting information flow over time. An edge states that the relation between two clusters is almost a regular relation (imposed by regular similarity of the actors of the two clusters, respectively).

The communication in both discussion forums is dominated by one core cluster (Cluster 1 (23% of all actors) in the “Cooperate Finance” forum and Cluster C (10.5% of all actors) in the “Global Warming” forum. These clusters not only comprise of actors who are active in each time slice and have many information giving relations to the other clusters but also receive information others. Interesting to see is that they have regular relations with themselves in most time slices. In time slices when peripheral clusters of actors have regular relations with other clusters they tend to connect with the core cluster. This is typical for core-periphery structures in social networks ([9] p.p. 235-236). The concentration of the communication around a core is even more pronounced in the smaller and more densely connected “Global Warming” forum. There exists also a semi-periphery (cluster B and D). These clusters are

Tobias Hecking, Andreas Harrer, H. Ulrich Hoppe

active during the entire period but build connections mostly with the core cluster C. In this aspect, it can be assumed that in MOOC discussion forums information constantly circulates within a core cluster and occasionally alternates between this core cluster and the more peripheral clusters. Thus, the role of the core actors can be considered as especially important for the coherence of knowledge exchange.

Actors who are only connected to other clusters in particular time slices are primarily in clusters 3 and 5 in the “Corporate Finance” course and clusters A and E in the “Global Warming” course. This reflects the role of actors who show a “post and leave” behaviour and appear as information seekers or information givers only in particular time slices. A possible explanation is that those actors become active in the forum during course topics that are of interest or for which they encounter concrete problems. This actor role is much more salient in the “Cooperate Finance” course.

The knowledge exchange structures in both discussion forums further have in common that the most relations between roles exist in the first two time slices. This reflects the typical decrease in activity over time in MOOC discussion forums due to dropouts of the course and the “post and leave” behaviour reported in Section 6.1. This is especially the case in the “Corporate Finance” course. In the “Global Warming” forum, however, there exist relations from the core cluster to almost all other clusters in the last time slice. This is also reflected by the high fraction of “late starters” reported in Section 6.1 for this course. In the “Corporate Finance” course late starting can be associated with some actors in the cluster 4 which is the only active role in the last time slice except the core cluster 1.

In general, it can be said that the assignment of actors to clusters or different roles, respectively, by the IIDBM or biased RESCAL method mostly reflects the heterogeneous activity patterns of the actors. The flow graphs in Fig. 6 reveal that only a limited number of actors are available to establish (the desired) sustainable information exchange over time.

7 Conclusion and Further Work

In order to tackle the problem of identification and analysis of the structure of knowledge exchange in discussion forums one has to face several challenges. This paper presented a holistic approach incorporating network extraction and modelling from unstructured forum data, analysis of individual development, as well as an analysis of connection patterns in the forums of two Coursera MOOCs. A crucial step is the extraction of the underlying communication network from forum data taking into account the semantic of links. Although, the extraction of a directed network of actors who give information to each other using machine learning methods for forum post classification is not perfect, we expect that the resulting network reflects the true knowledge exchange structure much better compared to naïve approaches, for example, simply linking actors who appear in the same thread. In contrast to the most existing studies on network analysis on forum data time is explicitly taken into

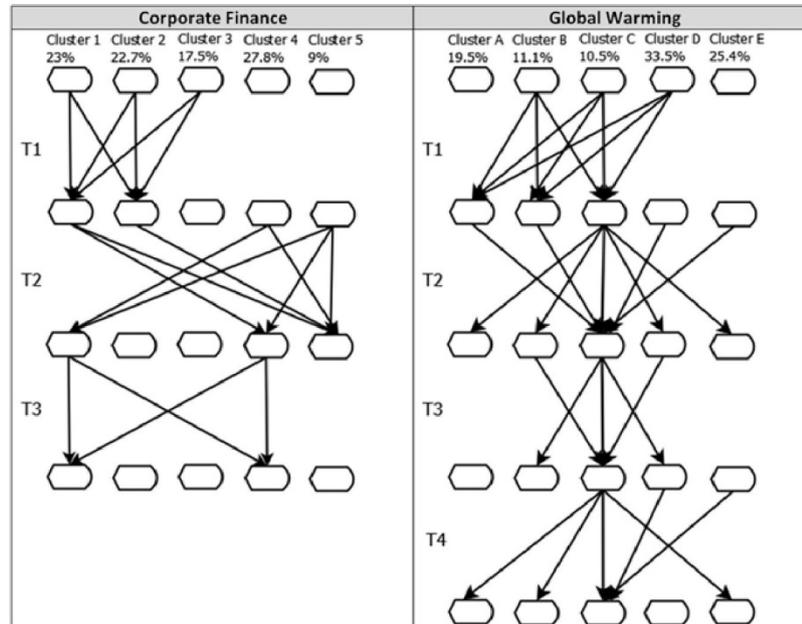


Fig. 6 Relations between clusters/roles over time. Percentage of actors is given for each cluster.

account enabling to investigate changes in the structure of the forum communication.

Based on the definition of the in- and outreach measures (Section 4), it is possible to characterise the posting behaviour of individuals, i.e. the role of actors in terms of information giving and information seeking. These measures combine the volume of posts with the diversification of connections in the extracted communication network. Analyses of the in- and outreach of actors over time in the presented case studies revealed different temporal patterns of posting behaviour. While the most actors are only partially active in the forum “post and leave”, the more active users can be divided into persons who constantly act as information givers or information seekers or actors who change their behaviour over time. These later cases are of particular interest for the research on knowledge building in large scale online courses since these patterns are indicators for individual development.

On a meso-level, different methods for clustering actors based on the similarity of their connection patterns were developed (Section 5) to uncover meaningful structures from the extracted communication networks enhancing the interpretability of the knowledge exchange structure. The clusters found are not necessarily cohesive but reflect the global position of actors in the knowledge exchange network. Thus, they can be interpreted as roles. The connection patterns between these clusters change over time, which introduces further challenges to the role modelling.

Tobias Hecking, Andreas Harrer, H. Ulrich Hoppe

It was shown that the modification of multi-relational blockmodelling methods and role modelling based on the RESCAL tensor decomposition are useful to detect underlying structures in the knowledge exchange network of forum users. In both courses, knowledge exchange is typically dominated by a small number of actors who actively participate in the forum discussion during the course, which is in line with previous findings such as [35, 14]. Other actors are not active over the entire period but appear as information seekers or information givers in particular time slices where they establish connections with the core cluster. Nonetheless, these actors can be temporally important with respect to the coherence of the discussion since they might provide important information input into the forums and initiate longer discussions.

In general, the results of the analysis of individual development (Section 6.1) and the macro-structure of knowledge exchange (Section 6.2) yield similar conclusions but highlight different aspects of forum communication in the investigated discussion forums. While Section 6.1 focused on the individual behaviour of actors over time, the results of Section 6.2 provided insights into the overall structural characteristics of the communication network. We believe that such a combination of different methods can contribute to a better and more substantial understanding of the complex process of knowledge exchange in discussion forums of large scale online courses.

Similar patterns related to forum activity were identified for the networks of both courses even if they differ in size and density. This also supports conclusions made in related studies, for example [38], that the emergence of a sustainable and constructive knowledge exchange as well as individual development has to be much better supported. This can, for example, be achieved by providing better communication channels such as Q/A systems or by using techniques of social recommendation to support participants in finding proper communication partners. To improve those support mechanisms, in the future the nature of the knowledge exchange between clusters of actors with similar roles in the network should be further explored. It is also worth to investigate the impact of role relations as described in Section 6.2 more deeply on the content level. This could, for example, be done by incorporating the discussion topics in the analysis. This would yield further insights into content related communication in MOOC discussion forums and the diffusion of information within the community of course participants. Furthermore, the results can be combined with analyses of other aspects of online courses. For example, it would be interesting to see whether a participant who develops from an information seeker to an information giver or vice versa in the discussion forum also changes behaviour regarding other course activities, or whether there are possible effects on course success.

References

1. Adamic, L.A., Zhang, J., Bakshy, E., Ackerman, M.S.: Knowledge sharing and yahoo answers: everyone knows something. In: Proceedings of the 17th international conference on World Wide Web, pp. 665–674. ACM (2008)
2. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J.: Engaging with massive online courses. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 687–698. ACM (2014). DOI 10.1145/2566486.2568042
3. Bader, B.W., Harshman, R.A., Kolda, T.G.: Temporal analysis of semantic graphs using asalsan. In: 7th IEEE International Conference on Data Mining. ICDM 2007., pp. 33–42 (2007). DOI 10.1109/ICDM.2007.54
4. Borgatti, S.P., Everett, M.G.: Two algorithms for computing regular equivalence. *Social Networks* **15**(4), 361–376 (1993). DOI 10.1016/0378-8733(93)90012-A
5. Breiman, L.: Random forests. *Mach.Learn.* **45**(1), 5–32 (2001). DOI 10.1023/A:1010933404324
6. Clow, D.: Moocs and the funnel of participation. In: Proceedings of the 3rd International Conference on Learning Analytics and Knowledge, LAK '13, pp. 185–189. ACM (2013). DOI 10.1145/2460296.2460332
7. Cortes, C., Pregibon, D., Volinsky, C.: Communities of interest. In: Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis (IDA, pp. 105–114 (2001)
8. Cui, Y., Wise, A.F.: Identifying content-related threads in mooc discussion forums. In: Proceedings of the 2nd ACM Conference on Learning @ Scale, L@S '15, pp. 299–303. ACM (2015). DOI 10.1145/2724660.2728679
9. Doreian, P., Batagelj, V., Ferligoj, A., Granovetter, M.: *Generalized Blockmodeling (Structural Analysis in the Social Sciences)*. Cambridge University Press, New York, NY, USA (2004)
10. Dunlavy, D.M., Kolda, T.G., Acar, E.: Temporal link prediction using matrix and tensor factorizations. *ACM Trans.Knowl.Discov.Data* **5**(2), 10:1–10:27 (2011). DOI 10.1145/1921632.1921636
11. Furtado, A., Andrade, N., Oliveira, N., Brasileiro, F.: Contributor profiles, their dynamics, and their importance in five q&a sites. In: Proceedings of the 2013 conference on Computer supported cooperative work, pp. 1237–1252. ACM (2013)
12. Gauvin, L., Panisson, A., Cattuto, C.: Detecting the community structure and activity patterns of temporal networks: A non-negative tensor factorization approach. *PLoS ONE* **9**(1), e86,028 (2014). DOI 10.1371/journal.pone.0086028
13. Gillani, N., Eynon, R.: Communication patterns in massively open online courses. *The Internet and Higher Education* **23**, 18–26. DOI 10.1016/j.iheduc.2014.05.004
14. Gillani, N., Yasseri, T., Eynon, R., Hjorth, I.: Structural limitations of learning in a crowd: communication vulnerability and information diffusion in moocs. *Scientific Reports* **4**, 6447 (2014). DOI 10.1038/srep06447
15. Harrer, A., Schmidt, A.: Blockmodelling and role analysis in multi-relational networks. *Social Network Analysis and Mining* **3**(3), 701–719 (2013). DOI 10.1007/s13278-013-0116-x
16. Harrer, A., Zeini, S., Ziebarth, S.: Visualisation of the dynamics for longitudinal analysis of computer-mediated social networks-concept and exemplary cases. In: N. Memon, R. Alhajj (eds.) *From Sociology to Computing in Social Networks: Theory, Foundations and Applications*, pp. 119–134. Springer, Vienna (2010). DOI 10.1007/978-3-7091-0294-7_7
17. Hecking, T., Hoppe, H.U., Harrer, A.: Uncovering the structure of knowledge exchange in a mooc discussion forum. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, ASONAM '15, pp. 1614–1615. ACM (2015). DOI 10.1145/2808797.2809359
18. Hoppe, H.U., Göhnert, T., Steinert, L., Charles, C.: A web-based tool for communication flow analysis of online chats. CEUR. URL http://ceur-ws.org/Vol-1137/LAK14CLA_submission_6.pdf
19. Kaufman, L., Rousseeuw, P.: Clustering by means of medoids. In: Y. Dodge (ed.) *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pp. 406–416. North-Hoand, Amsterdam (1987)

Tobias Hecking, Andreas Harrer, H. Ulrich Hoppe

20. Kim, J., Shaw, E., Feng, D., Beal, C., Hovy, E.: Modeling and assessing student activities in on-line discussions. In: Proceedings of the AAAI Workshop on Educational Data Mining (2006)
21. Kim, S.N., Wang, L., Baldwin, T.: Tagging and linking web forum posts. In: Proceedings of the 14th Conference on Computational Natural Language Learning, CoNLL '10, pp. 192–202. Association for Computational Linguistics (2010)
22. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM Review* **51**(3), 455–500 (2009). DOI 10.1137/07070111X
23. Lerner, J.: Role assignments. In: Br, U. es, T. Erlebach (eds.) *Network Analysis, Lecture Notes in Computer Science*, vol. 3418, pp. 216–252. Springer, Berlin / Heidelberg (2005). DOI 10.1007/978-3-540-31955-9_9
24. Mitchell, M.: An introduction to genetic algorithms. MIT press (1998)
25. Nickel, M., Tresp, V., Kriegel, H.P.: A three-way model for collective learning on multi-relational data. In: Proceedings of the 28th international conference on machine learning, pp. 809–816 (2011)
26. Ó Duinn, P., Bridge, D.: Collective classification of posts to internet forums. *Case-Based Reasoning Research and Development* **8765**, 330–344 (2014). DOI 10.1007/978-3-319-11209-1_24
27. Pattison, P.: Algebraic models for social networks. *Encyclopedia of Complexity and Systems Science* pp. 8291–8306 (2009). DOI 10.1007/978-0-387-30440-3_492
28. Ramesh, A., Goldwasser, D., Huang, B., Daume Iii, H., Getoor, L.: Understanding mooc discussion forums using seeded lda. In: Proceedings of the 9th ACL Workshop on Innovative Use of NLP for Building Educational Applications. ACL (2014)
29. Rossi, L.A., Gnawali, O.: Language independent analysis and classification of discussion threads in coursera mooc forums. In: Proceedings of the 15th IEEE International Conference on Information Reuse and Integration, pp. 654–661 (2014). DOI 10.1109/IRI.2014.7051952
30. Stump, G., DeBoer, J., Whittinghill, J., Breslow, L.: Development of a framework to classify mooc discussion forum posts: Methodology and challenges. Report (available online (26/09/2016)). URL https://tll.mit.edu/sites/default/files/library/Coding_a_MOOC_Discussion_Forum.pdf
31. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications, *Structural analysis in the social sciences*, vol. 1. Cambridge University Press (1994)
32. White, D.R., Reitz, K.P.: Graph and semigroup homomorphisms on networks of relations. *Social Networks* **5**(2), 193–234 (1983)
33. Whittaker, S., Terveen, L., Hill, W., Cherny, L.: The dynamics of mass interaction. In: From Usenet to CoWebs, pp. 79–91. Springer, London (2003)
34. Wiesberg, S., Reinelt, G.: Relaxations in practical clustering and blockmodeling. *Informatica* **39**(3), 249–256 (2015)
35. Wong, J.S., Pursel, B., Divinsky, A., Jansen, B.J.: An analysis of mooc discussion forum interactions from the most active users. *Social Computing, Behavioral-Cultural Modeling, and Prediction* **9021**, 452–457 (2015). DOI 10.1007/978-3-319-16268-3_58
36. Zeini, S., Göhnert, T., Hecking, T., Krempel, L., Hoppe, H.U.: The impact of measurement time on subgroup detection in online communities. In: F. Can, T. Özyer, F. Polat (eds.) *State of the Art Applications of Social Network Analysis*, pp. 249–268. Springer, Cham (2014). DOI 10.1007/978-3-319-05912-9_12
37. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th international conference on World Wide Web, WWW '07, pp. 221–230. ACM (2007). DOI 10.1145/1242572.1242603

6 Investigating Social and Semantic User Roles in MOOC Discussion Forums

This paper was presented at the 6th Conference on Learning Analytics and Knowledge (LAK 2016). This instance of the conference received a record number of full paper submissions (116) with an acceptance rate of 31%. As in previous instances of the conference, it provided a platform for discussion of new developments in the analysis of large online courses combining computational methods and application perspectives. Based upon selection by the program chairs of the conference, the paper was invited to appear as extended version in the Journal of Learning Analytics. This chapter includes the conference version.

© 2016 ACM, Inc. Reprinted with permission from Hecking, T., Chounta I. - A., & Hoppe H. U. Investigating Social and Semantic User Roles in MOOC Discussion Forums. *Proceedings of the 6th International Conference on Learning Analytics & Knowledge, 03/2016* (pp. 198–207). <http://doi.acm.org/10.1145/2883851.2883924>

Author	Contribution	%
Tobias Hecking	<ul style="list-style-type: none">- Conceptualisation of the approach.- Data collection and cleaning.- Design and accomplishment of the evaluation.	70%
Irene Angelica Chounta	<ul style="list-style-type: none">- Support in contextualisation and conceptualisation.- Statistical evaluations.	15%
H. Ulrich Hoppe	<ul style="list-style-type: none">- Supervision and advice during conceptualisation.- Support in contextualisation and planning.	15%

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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-4190-5/16/04\$15.00

DOI: <http://dx.doi.org/10.1145/2883851.2883924>

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 Coursera¹ 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

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 approaches 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

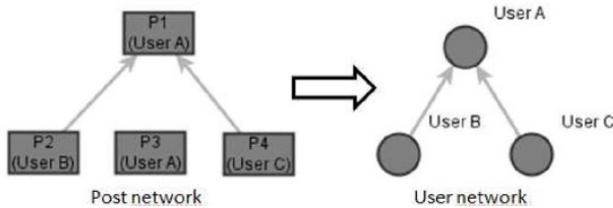


Figure 1: Basic scheme of network extraction from forum posts.

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.

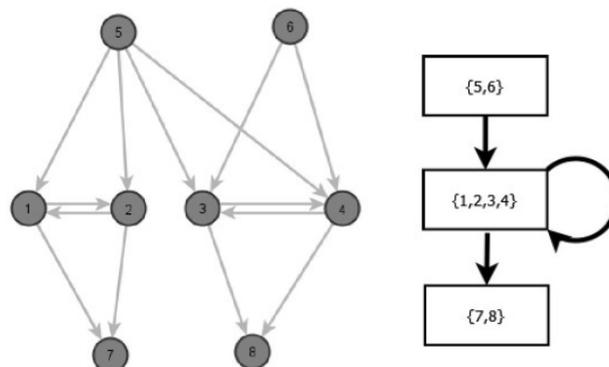


Figure 2: Example network with regular and structural equivalences.

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 where 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.

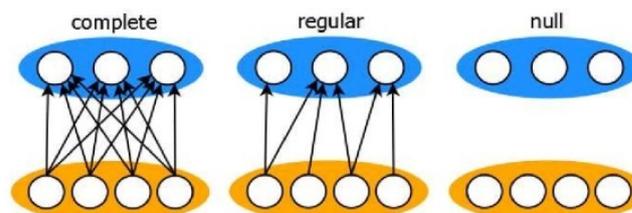


Figure 3: Relation types between two groups of actors.

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 assessed 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.

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

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.

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

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.

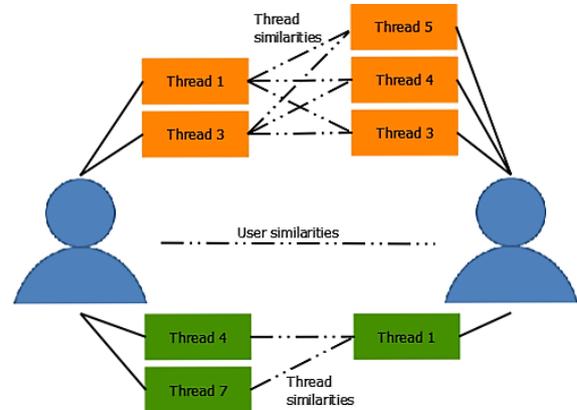


Figure 4: Semantic similarity of two users based on the similarity of threads in which they provide information (orange) and seek for information (green).

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 Reuters' Open Calais². 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 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 X³ 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.

² <http://www.opencalais.com/>

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.

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

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 [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

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 ($p \ll .05$). 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 parameters 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 .

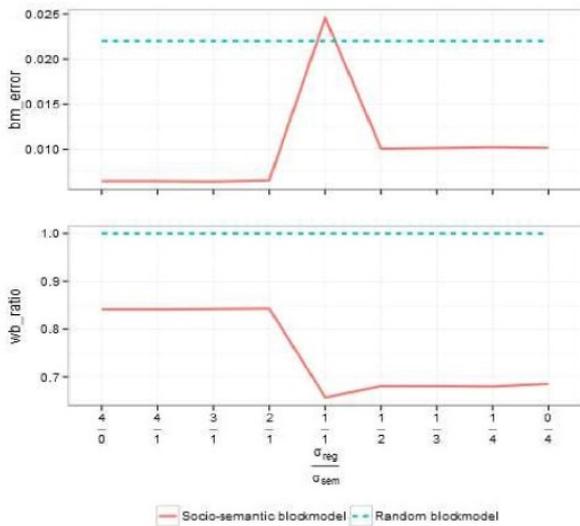


Figure 5: Blockmodel error (top) and ratio of average semantic distance within roles and between roles (bottom) for different ratios of σ_{reg} and σ_{sem} .

The results are presented in Figure 5. For both cases, wb_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 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 $\sigma_{sem}=2$.

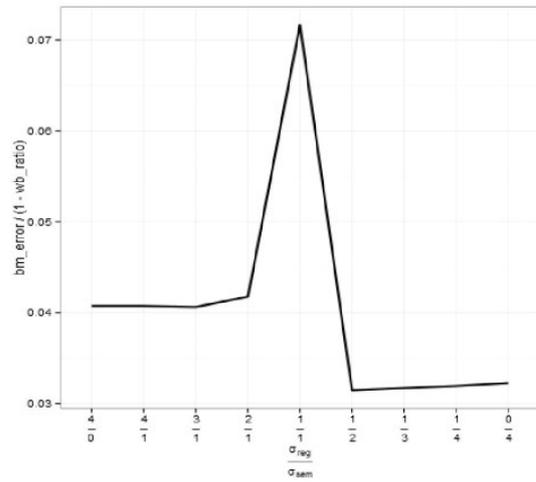


Figure 6: Ratio between blockmodel error and semantic coherence of the roles.

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 from 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.

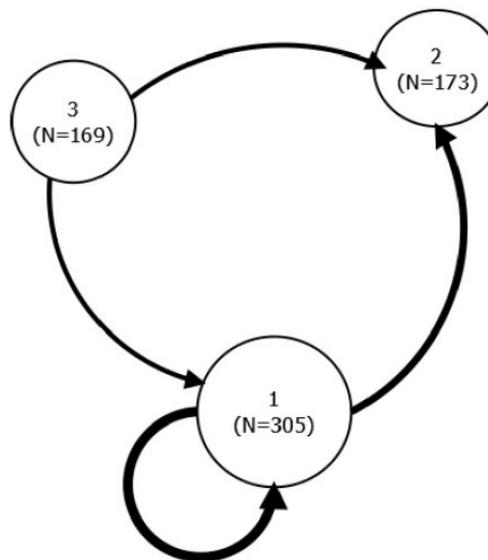


Figure 7: Blockmodel for the forum discussion in the MOOC discussion forum.

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. <i>Question</i>	3.58	0.43
3	1. Mathematical finance 2. Investment 3. Rate of return 4. Net present value	1. <i>Ambiguity</i> 2. <i>Decision theory</i>	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 Zaiane, 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 Blockmodelling 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 Blockmodelling (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 Rosve 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, Switzerland), 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. *Metodološkizvezki*, 10, (2013), 99-119.

7 Summary and Future Perspectives

7.1 Summary

This thesis dealt with the phenomena of learning and knowledge creation in online communities with a focus on the advancement of methods to analyse activity and interactions of actors in course environments. Analyses were conducted on different levels of granularity (community, meso-, and individual level (cf. Section 1.1.2) and the results were transformed into statements about different roles or characteristics of actors as well as properties of the community as a whole. The results contribute to a better understanding of learning in online communities, and thus, support future improvement of the design of dedicated learning scenarios. Furthermore, the outcome of this thesis enriches the methodological foundation of the field of learning analytics and exemplifies the potential of taking up approaches originating in social network analysis and related fields of data science, such as content and process analysis, for the acquisition of meaningful information from data produced in virtual learning environments.

By modelling actor-artefact relations as bipartite networks, it was possible to detect patterns in the usage of learning resources that are invisible to pure descriptive analyses (Chapters 2-3). This includes the identification of subsets of actors with different resource access strategies especially during exam preparations compared to the majority of course participants. The qualitative model of a clustered bipartite network keeps as much information about the original activity log data as possible. While it is not only possible to count the number of actors who are interested in a certain artefact or the number of artefacts accessed by an actor, the model also captures structure, i.e. relationships between individual actors and artefacts. This can be used to induce semantic relations between actors or between artefacts based on common interests or co-usage respectively. The dynamics of such relations over time was explicitly taken into account and different methods for tracing the evolution of bipartite actor-artefact clusters were developed. These approaches were also extended by incorporating content analysis of user generated content to replace the artefacts by meta-information i.e. keywords. Applied to discussion content in online courses different roles of contributors with respect to their interests, as well as the co-evolution of actors' interest and themes could be revealed (Chapter 4).

Chapters 5-6 shifted the focus from affiliation aspects of online communities to interpersonal relations derived from dynamic artefacts namely forum discussions. Therefore, social network analysis techniques, in particular centrality assessment and blockmodelling, were extended to incorporate time and discussion content. It could be shown that an in-depth analysis of connection patterns of actors in social networks derived from communication relations is a useful means for mapping the overall structure of information exchange, and further allows for a characterisation of actors with respect to their individual information giving and information seeking behaviour over time. While most of the users are active only in short periods in time, there are also exceptional cases where actors change their communication behaviour during the course. Furthermore, by reducing the forum interaction network to a macro-structure (blockmodel) that captures the information sharing between parts of the network, Chapters 5 and

6 clearly showed the limitations of MOOC discussion forums revealing the partitioning of forum communities into core and peripheral actors, while only the core actors build a cohesive network based information exchange relations.

7.2 Future Perspectives

7.2.1 Advancing online learning environments

This thesis explored different characteristics of learners in online communities with respect to their learning preferences and their roles in information sharing activities. A major prospect is to use the information gathered from the various conducted analyses for decision support in redesigning learning environments.

As mentioned in Section 1.2.1, methods for identifying resource access patterns have already been taken up in evaluation studies of online courses to gain insights into the role of different types of learning resources, and further for the identification of characteristic participation patterns in relation to course success (Ziebarth et al., 2015). The goal is to improve future instances of these courses with respect to the provided learning resources. It could be shown that affiliations of learners to course material can differ, especially affiliations to learner generated content such as wiki articles. These affiliation patterns are, in addition, not necessarily stable over time. To support learners with different preferences, online courses should incorporate a variety of resources to foster self-directed knowledge acquisition for different types of learners. However, in order to avoid overloading courses with learning material, participants have to be supported in structuring their learning process and in managing information sources. This can be achieved by personalisation of learning environments, for example, using recommendations and guidance mechanisms for learners to maintain their personal portfolio of resources. Especially online platforms that agglomerate different external tools and resources can benefit from the described approaches. A recent example of those platforms is ProSolo (Rosé et al., 2015), as an open environment for informal learning that incorporates various types of social media content to enable users to follow their learning goals in a self-directed manner. For adaptation of those services to users with heterogeneous interests in learning resources, goals, and backgrounds, resource access patterns based on dynamic actor-artefact relations can be a valuable source of information.

Another development in large scale online courses, such as MOOCs, is to support collaborative learning in groups (Staubitz, Pfeiffer, Renz, Willems, & Meinel, 2015; Wen, Yang, & Rosé, 2015). These concepts aim to exploit the heterogeneity of background knowledge and point of views that emerge from the large audience to facilitate knowledge exchange and critical discourse between participants. Moreover, there is evidence that the experiences of working together with other course participants counteract the problem of attrition in large scale online courses (Tomar, Sankaranarayanan, & Rosé, 2016). Thereby, it is known that heterogeneity of knowledge of group members is beneficial for collaborative learning (Webb, Nemer, & Zuniga, 2002). Thus, in computer supported collaborative learning, analytical approaches are used for automatic group formation based on data-driven learner modelling (Hoppe, 1995; Manske, Hecking, Chounta,

Werneburg, & Hoppe, 2015). However, especially in large online courses without face-to-face interaction a major challenge is to establish productive group learning since asynchronous communication, different motivation levels, and coordination issues between participants can complicate effective group work. Initial experiences provide evidence that in online courses the composition of groups according to previous activity levels of members can have a positive influence on the future group productivity, and consequently, informed group formation based on previous data analysis can be one building block to achieve a sufficient level of collaboration (Wichmann et al., 2016). In this aspect, characterising learners based on their resource access patterns or thematic interests using the approaches outlined in Chapters 2- 4 can account for, both, activity levels of learners as well as their possible knowledge. Low activity in the sense of a low coverage of learning material is directly expressed in the student-resource clusters introduced in Chapter 2. Resource access patterns can even be used as “optimistic learner models” assuming that the information contained in the accessed learning resources approximate the learners’ actual knowledge. Thus, the mentioned approaches are a promising means for enabling informed group formation in large scale online courses.

Apart from small-group learning, the existing collaboration tools, which are typically discussion forums, can be improved as well. The general outcome of the work described in Chapters 5 and 6 motivates the design of interaction support mechanisms that are more tailored to different roles of forum contributors. The “core” users who are active over longer periods are responsible for the cohesion of information exchange by establishing connections with many others. This group of forum users has the potential to form a community with stronger social bonds than the typical loose communication relations emerged from forum discussions. This is generally desirable since it has been argued that a sense of belonging to a community is promotive for collaborative learning (Wegerif, 1998). Therefore, novel tools that help those actors to maintain their social contacts and to support them in forming interest groups are needed. Models of the actors’ expertise and problem areas used in Chapter 6 can assist to assemble those groups. Furthermore, centralised discussion forums could be accompanied by tools allowing for ad-hoc discussions, for example, a question and answer board attached to a lecture video so that collaboration can take place in parallel to the actual learning activity. This can reduce barriers to engage in collaborative activities and create learning opportunities through discussions also for actors in peripheral roles. Similar tools are about to emerge but have not been applied at large scale, yet (Rosé & Ferschke, 2016).

7.2.2 Widening the scope of applications

All studies reported in this thesis were conducted in the context of online courses ranging from small blended learning courses to large scale MOOCs. However, the outlined approaches can be applied to other scenarios as well. This includes online communities forming in social media platforms, mass collaboration in wikis, and scientific communities. A lot of effort has been made in these contexts to understand complex phenomena of knowledge acquisition, information sharing, and evolution of interests or opinions, utilising computational methods (Lazer et al., 2009). The techniques developed in this thesis have been integrated as components into the

“Analytics Workbench” (Göhnert, Harrer, Hecking, & Hoppe, 2014) as a tool for assembling complex analysis workflows of data processors and visualisations. This enables the combination of the described approaches with other analysis methods and datasets, which facilitates the usage in other contexts from a technical perspective.

Investigating actor-artefact relations using dynamic network models can be of particular importance in online communities that primarily rely upon sharing and consumption of resources, like video platforms, file sharing communities, or production oriented communities (e.g. wikis, open source software developers). The artefacts can either be encoded explicitly in the network as in Chapter 2 or replaced by meta-information, such as topics derived from content analysis, as in Chapter 3. Tracing the evolution of clusters of actors and resources / topics does not only allow for the identification of actors with similar interest or associations between resources, which is a typical prerequisite for recommender systems, but also for capturing affiliation changes over time. This can, for example, contribute to gain insights about how collective behaviour or interests emerge in online communities, even if direct interaction between actors is limited.

With respect to the social and production aspect of online communities understanding the mechanisms behind the spread of information through social networks is of huge interest. Typical tasks are the identification of trends, influential actors, as well as explanation and prediction of infection processes (Guille, Hacid, Favre, & Zighed, 2013). To that end, the presented research in Chapter 5 and 6 can be extended from the identification of patterns of information sharing between peers to models of cascading information diffusion in social networks. From a pure network science perspective, particularly if the social network through which information spreads is considered as a closed world (i.e. ignoring external information sources), the information an actor can have depends solely on the position of this actor in the communication network. This assumption is clearly unrealistic but it can be used to develop models that approximate real world situations, especially if the produced content is taken into account, as in Chapter 6. Positional analysis, such as blockmodelling, can then be used to group actors with similar positions into clusters of actors for which similar knowledge can be assumed. The expected results can be helpful to investigate the roles of actors regarding the emergence of opinions and trends or to detect important information brokers. In particular, our previous work on the identification of “main paths” of information flow in directed acyclic hyperlink networks between collaboratively edited articles (artefacts) in wiki environments (Halatchliyski et al., 2014) can be taken up for this purpose. Using a similar approach to model time respecting acyclic networks between actors based on identification of interpersonal information spillovers, the described blockmodelling methods can be used to reduce the resulting network to a macro-structure mapping the main path of information diffusion between groups of actors and to discover different roles in the information spreading process.

In conclusion, the outcome of this thesis constitutes a step towards a better understanding of learning in online communities for the sake of improving dedicated concepts and environments. Furthermore, the developed methods can be utilised for research in various areas in the landscape of collaborative knowledge acquisition in digital spaces, and hence, also open up a variety of opportunities for future innovations and developments.

References

- Albright, S., & Winston, W. (2014). *Business Analytics: Data Analysis & Decision Making* (5 Ed.). Delmar Learning.
- Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). Engaging with Massive Online Courses. In *Proceedings of the 23rd International Conference on World Wide Web*. (pp. 687-698) Seoul, Korea. ACM. doi:10.1145/2566486.2568042
- Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In A. J. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 61-75). Springer, New York.
- Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*. (pp. 519-528). Montreal, Québec, Canada. ACM.
- Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning*, 9(2), 161-185. doi:10.1007/s11409-013-9107-6
- Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439), 509-512. doi:10.1126/science.286.5439.509
- Baradwaj, B. K., & Pal, S. (2012). Mining Educational Data to Analyze Students' Performance. *CoRR*, abs/1201.3417.
- Belanger, Y., & Thornton, J. (2013). Bioelectricity: A Quantitative Approach Duke University's First MOOC. *Duke Space Library*. Duke University.
- Bishop, J. L., & Verleger, M. A. (2013). The Flipped Classroom: A Survey of the Research. In *Proceedings of the 120th American Society for Engineering Education Annual Conference and Exposition*. (pp. 23.1200.1201 - 1223.1200.1218) Atlanta, GA, USA. ASEE.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.
- Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, 323(5916), 892-895.
- Braha, D., & Bar-Yam, Y. (2009). Time-Dependent Complex Networks: Dynamic Centrality, Dynamic Motifs, and Cycles of Social Interactions. In T. Gross & H. Sayama (Eds.), *Adaptive Networks: Theory, Models and Applications* (pp. 39-50). Springer, Berlin, Heidelberg.
- Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., & Wagner, D. (2008). On Modularity Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2), 172-188. doi:10.1109/TKDE.2007.190689
- Brandes, U., & Erlebach, T. (2005). *Network Analysis: Methodological Foundations*. Springer, New York.
- Brusco, M., Doreian, P., Steinley, D., & Saturnino, C. (2013). Multiobjective Blockmodeling for Social Network Analysis. *Psychometrika*, 78(3), 498-525. doi:10.1007/s11336-012-9313-1
- Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web*. (pp. 3-53). Springer.
- Bryant, S. L., Forte, A., & Bruckman, A. (2005). Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting group work*. (pp. 1-10). ACM.

- Carlen, U., & Jobring, O. (2005). The rationale of online learning communities. *International Journal of Web Based Communities*, 1(3), 272-295.
- Clariana, R., Engelmann, T., & Yu, W. (2013). Using centrality of concept maps as a measure of problem space states in computer-supported collaborative problem solving. *Educational Technology Research and Development*, 61(3), 423-442. doi:10.1007/s11423-013-9293-6
- Clow, D. (2012). The Learning Analytics Cycle: Closing the Loop Effectively. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. (pp. 134-138) Vancouver, British Columbia, Canada. ACM. doi:10.1145/2330601.2330636
- Cress, U., & Kimmerle, J. (2008). A systemic and cognitive view on collaborative knowledge building with wikis. *International Journal of Computer-Supported Collaborative Learning*, 3(2), 105-122.
- Dabbagh, N., & Kitsantas, A. (2012). Personal Learning Environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning. *The Internet and Higher Education*, 15(1), 3-8.
- Daems, O., Erkens, M., Malzahn, N., & Hoppe, H. U. (2014). Using content analysis and domain ontologies to check learners' understanding of science concepts. *Journal of Computers in Education*, 1(2), 113-131. doi:10.1007/s40692-014-0013-y
- de Laat, M., Lally, V., Lipponen, L., & Simons, R.-J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. *International Journal of Computer-Supported Collaborative Learning*, 2(1), 87-103. doi:10.1007/s11412-007-9006-4
- De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education*, 46(1), 6-28.
- 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.
- Diver, P., & Martinez, I. (2015). MOOCs as a massive research laboratory: opportunities and challenges. *Distance Education*, 36(1), 5-25.
- Doreian, P., Batagelj, V., Ferligoj, A., & Granovetter, M. (2004). *Generalized Blockmodeling (Structural Analysis in the Social Sciences)*. Cambridge University Press, New York, NY, USA.
- Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, 12(4), 1143-1168. doi:10.1111/j.1083-6101.2007.00367.x
- Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5), 304-317.
- Ferguson, R., & Clow, D. (2015). Examining Engagement: Analysing Learner Subpopulations in Massive Open Online Courses (MOOCs). In *Proceedings of the 5th International Conference on Learning Analytics And Knowledge*. (pp. 51-58) Poughkeepsie, New York, USA. ACM. doi:10.1145/2723576.2723606
- Fini, A. (2009). The technological dimension of a massive open online course: The case of the CCK08 course tools. *The International Review Of Research In Open And Distance Learning*, 10(5).
- Fischer, G. (2001). Communities of interest: Learning through the interaction of multiple knowledge systems. In *Proceedings of the 24th Information Systems Research Conference in Scandinavia*. Bergen, Norway. Department of Information Science, Bergen.

References

- Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3), 75-174.
- Friedl, D.-M. B., & Heidemann, J. (2010). A critical review of centrality measures in social networks. *Business & Information Systems Engineering*, 2(6), 371-385.
- Gacek, C., & Arief, B. (2004). The many meanings of open source. *IEEE software*, 21(1), 34-40.
- Getoor, L., & Diehl, C. P. (2005). Link mining: a survey. *SIGKDD Explor.Newsl.*, 7(2), 3-12. doi:10.1145/1117454.1117456
- Göhnert, T., Harrer, A., Hecking, T., & Hoppe, H. U. (2014). A Workbench for Visual Design of Executable and Re-usable Network Analysis Workflows. In J. Kawash (Ed.), *Online Social Media Analysis and Visualization* (pp. 181-199). LNSN, Springer, Cham.
- Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6), 1360-1380.
- Gross, T., & Blasius, B. (2008). Adaptive coevolutionary networks: a review. *Journal of the Royal Society Interface*, 5(20), 259-271.
- Grünwald, F., Meinel, C., Totschnig, M., & Willems, C. (2013). Designing MOOCs for the Support of Multiple Learning Styles. In Hernández-Leo, D., Ley, T., Klamma, R., Harrer, A. (Eds.) EC-TEL 2013. LNCS, vol. 8095 *Scaling up Learning for Sustained Impact*, (pp. 371-382), Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-40814-4_29
- Guille, A., Hacid, H., Favre, C., & Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(2), 17-28.
- Guo, P. J., & Reinecke, K. (2014). Demographic differences in how students navigate through MOOCs. In *Proceedings of the 1st ACM Conference on Learning@ Scale*. (pp. 21-30). ACM.
- Halatchliyski, I., Hecking, T., Goehnert, T., & Hoppe, H. U. (2014). Analyzing the Path of Ideas and Activity of Contributors in an Open Learning Community. *Journal of Learning Analytics*, JLA, 1(2), 72-93.
- Harrer, A., Malzahn, N., Zeini, S., & Hoppe, H. U. (2007). Combining Social Network Analysis with Semantic Relations to Support the Evolution of a Scientific Community. In *Proceedings of the 8th International Conference on Computer Supported Collaborative Learning*. (pp. 270-279) New Brunswick, New Jersey, USA. International Society of the Learning Sciences (ISLS).
- Hecking, T., Göhnert, T., Zeini, S., & Hoppe, U. (2013). Task and Time Aware Community Detection in Dynamically Evolving Social Networks. *Procedia Computer Science*, 18, 2066-2075. doi:10.1016/j.procs.2013.05.376
- Henri, F., & Pudelko, B. e. (2003). Understanding and analysing activity and learning in virtual communities. *Journal of Computer Assisted Learning*, 19(4), 474-487.
- Hey, T., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Microsoft research Redmond, WA, USA.
- Hill, J., & Hannafin, M. (2001). Teaching and learning in digital environments: The resurgence of resource-based learning. *Educational Technology Research and Development*, 49(3), 37-52. doi:10.1007/BF02504914
- Hoppe, H. U. (1995). The Use of Multiple Student Modeling to Parameterize Group Learning. In *Proceedings of the 7th World Conference on Artificial Intelligence in Education*. (pp. 234-241) Charlottesville, VA, USA.
- Hoppe, H. U. (2009). Integrating Learning Processes Across Boundaries of Media, Time and Group Scale. In M. Redondo, C. Bravo, & M. Ortega (Eds.), *Engineering the User Interface: From Research to Practice* (pp. 1-18). Springer, London.

- Hoppe, H. U. (2016). Computational Methods and Approaches for the Analysis of Learning and Knowledge-building Communities. In G. Siemens & C. Lang (Eds.), *Handbook of Learning Analytics and Educational Data Mining*. (to appear).
- Hoppe, H. U., Harrer, A., Göhnert, T., & Hecking, T. (2016). Applying network models and network analysis techniques to the study of online communities *Mass Collaboration and Education* (pp. 347-366). Springer.
- Hoppe, H. U., Pinkwart, N., Oelinger, M., Zeini, S., Verdejo, F., Barros, B., & Mayorga, J. I. (2005). Building Bridges within Learning Communities through Ontologies and "Thematic Objects". In *Proceedings of the 7th International Conference on Computer Supported Collaborative Learning*. (pp. 211-220) Taipei, Taiwan. International Society of the Learning Sciences (ISLS).
- Kolda, T. G., & Bader, B. W. (2009). Tensor Decompositions and Applications. *SIAM Review*, 51(3), 455-500. doi:10.1137/07070111X
- Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., & Zlotowski, O. (2005). Centrality Indices. In U. Brandes & T. Erlebach (Eds.), *Network Analysis: Methodological Foundations* (pp. 16-61). Springer, Berlin, Heidelberg.
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., . . . Van Alstyne, M. (2009). Computational Social Science. *Science*, 323(5915), 721-723. doi:10.1126/science.1167742
- Lehmann, S., Schwartz, M., & Hansen, L. K. (2008). Biclique communities. *Phys.Rev.E*, 78(1), 016108. doi:10.1103/PhysRevE.78.016108
- Liben-Nowell, D., & Kleinberg, J. (2008). Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12), 4633-4638.
- Liddo, A. D., Shum, S. B., Quinto, I., Bachler, M., & Cannavacciuolo, L. (2011). Discourse-centric learning analytics. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. (pp. 23-33) Banff, Alberta, Canada. ACM. doi:10.1145/2090116.2090120
- Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action: Aligning learning analytics with learning design. *American Behavioral Scientist*, 57(10), 1439-1459. doi:10.1177/0002764213479367
- Macy, M. W., & Willer, R. (2002). FROM FACTORS TO ACTORS: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology*, 28(1), 143-166. doi:10.1146/annurev.soc.28.110601.141117
- Manske, S., Hecking, T., Chounta, I. A., Werneburg, S., & Hoppe, H. U. (2015). Using Differences to Make a Difference: A Study on Heterogeneity of Learning Groups. In *Proceedings of the 11th International Conference on Computer Supported Collaborative Learning*. (pp. 182-189) Gothenburg, Sweden. International Society of the Learning Sciences (ISLS).
- Manske, S., & Hoppe, H. U. (2016). The "Concept Cloud": Supporting Collaborative Knowledge Construction based on Semantic Extraction from Learner-generated Artefacts. In *Proceedings of the 6th International Conference on Advanced Learning Technologies*. (vol. 1, pp. 302-306) Austin, TX, USA. IEEE. doi:10.1109/ICALT.2016.123
- McGregor, A. (2014). Graph stream algorithms: a survey. *ACM SIGMOD Record*, 43(1), 9-20.
- Mika, P. (2007). *Social Networks and the Semantic Web* (Vol. Semantic Web and Beyond, Vol 5.). Springer, New York.
- Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577-8582. doi:10.1073/pnas.0601602103
- Palla, G., Barabasi, A. L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136), 664-667.

References

- Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6), 759-772.
- Porter, C. E. (2004). A Typology of Virtual Communities: A Multi-Disciplinary Foundation for Future Research. *Journal of Computer-Mediated Communication*, 10(1). doi:10.1111/j.1083-6101.2004.tb00228.x
- Reimann, P. (2009). Time is precious: Variable-and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning*, 4(3), 239-257.
- Reimann, P., Markauskaite, L., & Bannert, M. (2014). e-Research and learning theory: What do sequence and process mining methods contribute? *British Journal of Educational Technology*, 45(3), 528-540. doi:10.1111/bjet.12146
- Reinhardt, W., Moi, M., & Varlemann, T. (2009). Artefact-Actor-Networks as tie between social networks and artefact networks. In *Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing, (CollaborateCom 2009)*. (pp. 1-10). doi:10.4108/ICST.COLLABORATECOM2009.8308
- Rivers, K., & Koedinger, K. R. (2015). Data-Driven Hint Generation in Vast Solution Spaces: a Self-Improving Python Programming Tutor. *International Journal of Artificial Intelligence in Education*, 1-28. doi:10.1007/s40593-015-0070-z
- Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*. (pp. 695-704) Hyderabad, India. ACM. doi:10.1145/1963405.1963503
- Rosé, C., & Ferschke, O. (2016). Technology Support for Discussion Based Learning: From Computer Supported Collaborative Learning to the Future of Massive Open Online Courses. *International Journal of Artificial Intelligence in Education*, 26(2), 660-678.
- Rosé, C., Ferschke, O., Tomar, G., Yang, D., Howley, I., Alevin, V., . . . Baker, R. (2015). Challenges and opportunities of dual-layer MOOCs: Reflections from an edX deployment study. In *Proceedings of the 11th International Conference on Computer Supported Collaborative Learning*. (vol. 2, pp. 459-467). International Society of the Learning Sciences (ISLS).
- Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 3(3), 237-271.
- Roth, C., & Cointet, J.-P. (2010). Social and semantic coevolution in knowledge networks. *Social Networks*, 32(1), 16-29.
- Siemens, G. (2014). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning*, 2(1). Retrieved from http://www.itdl.org/journal/jan_05/article01.htm.
- Southavilay, V., Yacef, K., Reimann, P., & Calvo, R. A. (2013). Analysis of collaborative writing processes using revision maps and probabilistic topic models. In *3rd Conference on Learning Analytics and Knowledge*. (pp. 38-47).
- Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, 1(2), 12-23.

- Staubitz, T., Pfeiffer, J., Renz, C., Willems, C., & Meinel, C. (2015). Collaborative Learning in a MOOC Environment. In *Proceedings of the 8th International Conference of Education, Research and Innovation*. (pp. 8237-8246) Seville, Spain. IATED.
- Suthers, D., & Rosen, D. (2011). A unified framework for multi-level analysis of distributed learning. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. (pp. 64-74). ACM.
- Tomar, G. S., Sankaranarayanan, S., & Rosé, C. P. (2016). Intelligent Conversational Agents as Facilitators and Coordinators for Group Work in Distributed Learning Environments (MOOCs). In *AAAI Spring Symposium Series*. Retrieved from <http://www.aaai.org/ocs/index.php/SSS/SSS16/paper/view/12728>.
- van der Aalst, W. M., Reijers, H. A., Weijters, A. J., van Dongen, B. F., De Medeiros, A. A., Song, M., & Verbeek, H. (2007). Business process mining: An industrial application. *Information Systems*, 32(5), 713-732.
- van Dongen, B. F., de Medeiros, A. K. A., Verbeek, H. M. W., Weijters, H. M. W., & van der Aalst, W. M. P. (2005). The ProM Framework: A New Era in Process Mining Tool Support. In *Proceedings of the 26th International Conference on Applications and Theory of Petri Nets*. (pp. 444-454) Miami. Springer. doi:10.1007/11494744_25
- Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Van Assche, F., Parra, G., & Klerkx, J. (2014). Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing*, 18(6), 1499-1514. doi:10.1007/s00779-013-0751-2
- Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., & Duval, E. (2012). Context-aware recommender systems for learning: a survey and future challenges. *IEEE Transactions on Learning Technologies*, 5(4), 318-335.
- Volk, B., Reinhardt, A., & Osterwalder, K. (2014). TORQUES: Riding the MOOC wave to benefit on-campus courses. In *European MOOC Stakeholder Summit*. (pp. 189-193) Lausanne, Switzerland.
- Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications* (Vol. 1). Cambridge University Press.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440-442.
- Webb, N. M., Nemer, K. M., & Zuniga, S. (2002). Short circuits or superconductors? Effects of group composition on high-achieving students' science assessment performance. *American Educational Research Journal*, 39(4), 943-989.
- Wegerif, R. (1998). The social dimension of asynchronous learning networks. *Journal of asynchronous learning networks*, 2(1), 34-49.
- Wen, M., Yang, D., & Rosé, C. P. (2015). Virtual Teams in Massive Open Online Courses. In *Proceedings of the 17th International Conference on Artificial Intelligence in Education*. (pp. 820-824) Madrid, Spain. Springer. doi:10.1007/978-3-319-19773-9_124
- Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.
- Wichmann, A., Hecking, T., Elson, M., Christmann, N., Herrmann, T., & Hoppe, H. U. (2016). Group Formation for Small-Group Learning: Are Heterogeneous Groups More Productive? In *Proceedings of the International Symposium on Open Collaboration (OpenSym '16)*. Berlin, Germany.
- Wild, F. (2016). Meaningful, Purposive Interaction Analysis *Learning Analytics in R with SNA, LSA, and MPIA*. (pp. 107-131). Springer, Cham.

References

- Wise, A. F., Cui, Y., & Vytasek, J. (2016). Bringing order to chaos in MOOC discussion forums with content-related thread identification. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*. (pp. 188-197) Edinburgh, United Kingdom. ACM. doi:10.1145/2883851.2883916
- Wong, J.-S., Pursel, B., Divinsky, A., & Jansen, B. J. (2015a). An Analysis of MOOC Discussion Forum Interactions from the Most Active Users. *Social Computing, Behavioral-Cultural Modeling, and Prediction*, 9021, 452-457. doi:10.1007/978-3-319-16268-3_58
- Wong, J.-S., Pursel, B., Divinsky, A., & Jansen, B. J. (2015b). Analyzing MOOC discussion forum messages to identify cognitive learning information exchanges. *Proceedings of the Association for Information Science and Technology*, 52(1), 1-10.
- Zeini, S., Göhnert, T., Hecking, T., Krempel, L., & Hoppe, H. U. (2014). The Impact of Measurement Time on Subgroup Detection in Online Communities. In F. Can, T. Özyer, & F. Polat (Eds.), *State of the Art Applications of Social Network Analysis* (pp. 249-268). LNSN, Springer, Cham.
- Žiberna, A. (2014). Blockmodeling of multilevel networks. *Social Networks*, 39, 46-61. doi:10.1016/j.socnet.2014.04.002
- Ziebarth, S., & Hoppe, H. U. (2014). Moodle4SPOC: A Resource-Intensive Blended Learning Course. In (pp. 359-372) Graz, Austria. Springer. doi:10.1007/978-3-319-11200-8_27
- Ziebarth, S., Neubaum, G., Kyewski, E., Krämer, N., Hoppe, H. U., Hecking, T., & Eimler, S. (2015). Resource Usage in Online Courses: Analyzing Learner's Active and Passive Participation Patterns. In *Proceedings of the 11th International Conference on Computer Supported Collaborative Learning*. (pp. 395 - 402) Gothenburg, Sweden. International Society of the Learning Sciences (ISLS).