Finding Community in the Crowd: The Importance of Tie Definition and Networking Partitioning in Examining Social Learning in MOOCs

Alyssa Friend Wise, Yi Cui & Wan Qi Jin

ACKNOWLEDGMENTS

An earlier version of this work was published in the ACM Proceedings of LAK'17, the 7th International Conference of Learning Analytics and Knowledge (doi=3027385.3027446). We thank Stanford University and the MOOCPosts Dataset team for their assistance in accessing and working with the data.

ABSTRACT

When applying SNA methods to study social interactions in MOOC discussion forums, it is important to construct the networks in ways that take into account the complexity and characteristic of those activities. This study examines the importance of content-based network partitioning and tie definition on social network structures and interpretation for MOOC discussion forums. Using dynamic interrelated post and thread categorization (Cui et al. in press) based on a previously developed natural language model (Wise et al. 2017), 817 threads containing 3124 discussion posts from 567 learners in a MOOC on statistics in medicine were characterized as either being related to the learning of course content or not. Content-related, non-content, and unpartitioned interaction networks were constructed based on five different tie definitions: Direct Reply, Star, Direct Reply+Star, Limited Copresence, and Total Copresence. Results showed the content-related and non-content networks to have distinct characteristics at the network, community, and individual node levels, with the unpartitioned network more closely resembling the non-content network. Network properties were less sensitive to differences in tie definition with the exception of Total Copresence, which showed distinct characteristics useful for detecting inflated social status due to "superthread" initiation. Networks partitioned based on content-relatedness are distinct in terms of participating members, the complexity of activities, participant's behavioral patterns, and their interaction techniques.

Keywords

Massive open online courses; social network analysis; discussion forum; network partitioning; tie extraction

1 INTRODUCTION

Massive open online courses (MOOCs) present many exciting opportunities for expanding learning by opening up accessibility to college-style courses while at the same time bringing together self-selected 'students' from all over the world. However, these online learning environments have faced substantial challenges including low completion rates and less than satisfactory learning experiences in many cases (Hew and Cheung 2014; Khalil and Ebner 2013). One commonly cited shortcoming contributing to these problems is the lack of social interaction in MOOCs (Rosé and Ferschke 2016). Interaction is an important element of quality in online learning generally (Trentin 2000) and of particular importance for engagement in MOOCs (Khalil and Ebner 2013), therefore increasing interaction is a promising route for addressing the well documented challenges in completion and satisfaction.

While interaction is a worthy goal, the tremendous numbers of students involved and the diversity in learner backgrounds, needs and intents (Jacobsen 2017) effectively prohibits sufficient interaction of the conventional student-to-instructor form. Many MOOCs therefore rely on peer-to-peer communication as the primary vehicle for interaction (Kellogg et al. 2014), with online discussion forums serving as the central medium. Despite the potential for peer interaction to improve student experiences and learning, the actual benefits of MOOC discussion forums reaped thusfar are questionable. First, MOOC discussions are often plagued by a host of problems that prevent them from meeting their full potential. These problems include low levels of participation (Breslow et al. 2013), overwhelming quantity and disorganization of posts (McGuire 2013), and a lack of responsivity between learners (Agrawal et al. 2015). Second, examinations of the relationship between MOOC forum participation and learning outcomes have yielded mixed and contradictory findings (Jiang et al. 2014; Santos et al. 2014). These two issues have prompted intense interest in investigating the interactions occurring in MOOC forums and their relationship to learning (e.g. Gillani and Eynon 2014; Gillani et al. 2014; Jiang et al. 2014; Kellogg et al. 2014). One common tool used in such studies is social network analysis (SNA). SNA is a useful method to investigate interaction in online discussion because of its focus on the connections between actors (Cho et al. 2007; Yusof and Rahman 2009). However, MOOC forum discussions differ from those in conventional online learning environments in several important ways. Specifically there is a dramatically greater number of learners interacting in perpetually different configurations in a relatively unstructured activity for a broad range of purposes (Kizilcec et al. 2013). Thus there is a need to apply SNA methods in ways that take into account the complexity and distinct characteristics of MOOC discussion activities.

This study addresses two underexplored areas important for the application of SNA to MOOCs: network partitioning and tie definition. In the following section, we review prior SNA research on discussion forums in online learning environments, and specifically in MOOCs, focusing on network partitioning approaches and the conceptualization and operationalization of social ties. We then describe and justify the methods used in the current study to partition discussion forum activities, construct ties, and investigate interactions in different networks. Based on this, we select a network construction method that best reflect our theory about interactions in discussion forum and produced interpretable results. We then analyze, both quantitatively and qualitatively, the characteristics of interactions in the partitioned networks constructed using this method to understand the relationships and activities in content-related and non-content networks. The ultimate goal is to provide both new perspectives and empirical evidence regarding social network construction and network partitioning for studying interactions in MOOC discussion forums.

2 LITERATURE REVIEW

SNA has been a useful tool with which to study online discussion forums because of its ability to extract patterns of connections between learners. Often used in combination with other methods, such as discourse analysis (e.g. Oshima et al. 2012), clustering and prediction modelling (e.g. Romero et al. 2013), SNA has been helpful for improving understanding about the general characteristics of social interactions and relationships in discussion forums and exploring their relationship with learning outcome in formal online learning contexts (Dawson, 2010; Rabbany et al. 2011). However, there is a need for caution in generalizing findings and methods from SNA in formal online learning environments to MOOCs due to the distinction between activities in the two types of contexts. For instance, discussion forums in formal online learning are usually dedicated to learning interactions closely related to the course content and participated in by a relatively small number of learners, who may know each other outside of the forums, and to whom the curriculum and course requirements are important factors influencing their behaviors (e.g. Romero et al. 2013). This stands in contrast to the large size and great diversity of participants and purposes of MOOC forums described above. These distinct characteristics need to be taken into account in considering how SNA can be used to usefully examine interactions and relationship in MOOC discussion forums.

One notable characteristic of activities in MOOC forums is that the discussion posts are made on highly diversified topics. Unlike more traditional formal online learning discussion forums which are usually designed and used for targeted discussions about the course content, MOOC discussion forums are generally open and thus host posts on topics ranging from clarification of course content to logistical questions about assignments, and from sharing deep personal connections with the learning material to lightweight social connections (Stump et al. 2013). The diversity of forum activities offers one possible explanation for the lack of clear connection to learning outcomes found thusfar, as different mode of interaction may serve different purposes. For example, discussions directly about the course content can play a different role in the learning process than those of a social nature (Wise et al. 2004). Distinctions may be more nuanced as well; for example past work from the higher education literature indicates that academically-related social interactions are more impactful for retention than purely social ones (Kuh 2002). This highlights network partitioning as a critical but under-addressed research area in applying SNA to MOOC data.

Another outstanding characteristic of MOOC discussion forum is the diversity of participation configurations. Unlike in formal online learning environments where learners generally participate in consistent ways with a regular group of people following requirements set by the course, MOOC learners can initiate, join or abandon discussions at any time and for any reason. These different conditions lead to a voluminous number of threads of vastly different sizes with shifting configurations of participants. This in turn has implications for the meaning of and what should be taken as an indication of "interaction" and makes defining the nature of social ties in these environments even more challenging than in conventional online learning environments. MOOC studies that use SNA methods have adopted several tie definitions to conceptualize and operationalize social ties (Gruzd and Haythornthwaite 2008; Zhu et al. 2016); however there has not yet been a clear articulation of the typology of these different definitions, the rationales for them, and the implications for the results and interpretations of the analysis.

2.1 SNA and MOOC Discussion Forums: Whole Network Approaches

SNA has been used to investigate learning interaction in MOOC discussion forums from many perspectives. For instance, Kellogg et al. (2014) studied the structure of peer support networks formed in two MOOCs designed for educators (on digital learning and mathematics learning) and examined factors that might account for such structures. The discussions in each course were studied as a single network of interaction with a directed edgelist constructed based on the exact reply structure. The study found some cross-course consistency in general network measures and participation patterns: both networks had clear core-periphery structure and low edge weight; learner's participation fell into four patterns, including mutual interactions, extensive but nonmutual interactions, thread initiation, and unresponded interaction attempts; reciprocity was found in both networks. However, the associations between network connections and demographic factors were largely inconsistent across courses: the tendency that learners connect more with those who have the same activity pattern, teach at the same schooling level (elementary vs high school), or live in the same state or country were only found in the course on digital learning.

Joksimović et al. (2016) also investigated the factors that influence social connections. The study was conducted on two instances of a programming MOOC, offered in English and Spanish respectively. A directed social network was extracted from the whole forum, based on the exact reply structure. Like Kellogg et al. (2014), this work found some consistent cross-offering results, such as reciprocity and performance-based homophily; they also failed to find cross-offering consistency in the association between social connection and learner's similarity in geographical location. In the same study, Joksimović et al. (2016) examined the association between social centrality degree (the number of direct connections a node has), closeness (average of the shortest path lengths from a node to all other nodes in the network), betweenness (the number of times a node is part of the shortest path between any two other nodes in the network) and academic performance (operationalized as completion and distinction). Weighted degree was found to be significantly associated with learning outcome across offerings; the effect of betweenness and closeness was only found in the Spanish offering.

Jiang et al. (2014) also examined the association between social centrality and academic performance (operationalized as no certificate, completion, or distinction). They conducted the study on an algebra MOOC and a finance MOOC. Undirected social networks were extracted from the whole discussion forums based on copresence in the same thread or subthread (how these distinctions were made was not explained). The results found from the two courses were inconsistent: degree and betweenness were positively correlated with learning performance in the algebra course while no significant correlation was found between any centrality index and learning performance in the finance course. Contradicting to the findings of Joksimović et al. (2016), Jiang et al. (2014) found learners tend to talk to those who are in different performance groups than themselves.

In summary, these studies' findings about social networks and learning are largely inconsistent or contradictory. One possible explanation is MOOC discussion forums are used for highly diversified purposes, such as understanding learning materials, clarifying course policy, and developing social connections (Cui and Wise 2015; Stump et al. 2013). Consequently, analyzing the discussion forum as one social network may compile interactions with distinct natures together, confounding relationships and concealing important patterns. Another possible explanation relates to the different decisions made around how to define ties in the network.

2.2 Partitioning Social Networks

2.2.1 Exogenous Partitioning

To address the problems with studying MOOC discussion forum as a whole network, some researchers have examined the interactions and networks in a more refined manner by paying specific attentions to defining network boundaries. One straightforward approach to doing so uses the presence of sub-forum designations for categorization. For example, Gillani and Eynon (2014) created an exogenously defined boundary by partitioning the social network in a business MOOC based on the presence of seven sub-forums, including Readings, Lectures, Cases, Final Project, Course Material Feedback, Technical Feedback, and Study Groups. Gillani and Eynon (2014) also examined sub-forum networks week by week. The networks were extracted based on copresence within a thread. It was found although all sub-forum networks followed a similar trend in structural shift across the different weeks (from centralized to dispersed), they differed in several ways: sub-forums were participated in largely by distinct groups of learners and only a small percentage of learners participated across sub-forum: participants in different sub-forums showed different levels of persistence over time, with the Cases sub-forum for discussing learning material showing the highest level of persistence.

In another study, Gillani et al. (2014) explored how social network structures influenced information diffusion in two successive offerings of a business MOOC. Social networks were extracted based on copresence in thread and partitioned based on eight sub-forums, including Readings, Lectures, Cases, Final Project, Questions for Professor, Course Material Feedback, Technical Feedback, and Study Groups. In addition, networks formed based only on frequent ties were compared with the unfiltered networks. Interesting cross-offering results were found. First, in both offerings, the proportion of one-off ties (edges with a weight of one) was different across sub-forums: the Feedback sub-forum used mostly for technical support and gratitude expressions had the highest proportion of one-off ties while the Cases sub-forum for discussing learning material had the fewest. Second, sub-forums demonstrated different densities of interaction: the Cases and Final Project sub-forums were more cohesive than the Study Groups sub-forum. Taken together, these results indicate that patterns of interaction vary across different kinds of forum activities; notably there seemed to be key differences between sub-forums whose purpose directly related to the learning of course content and those which did not. Similarly, Hecking et al. (2016) used sub-forums to explore social networks in a finance MOOC. They mapped networks in sub-forums dedicated to course content related issues (lectures, exercises and quizzes) to investigate user roles from both social and semantic perspectives.

These studies examined networks in MOOC forums in a more nuanced way by partitioning the whole forum into the component sub-forums. However, segmenting networks in a structural way based on sub-forums is a non-optimal approach for two reasons. First, each MOOC sets different course-specific sub-forums, therefore the generalizability of findings based on such divisions are limited. Second, prior studies have shown that cross posting is common in MOOCs (Rossi and Gnawali 2014). Thus, there is no guarantee that learners make posts in the appropriate designated sub-forum. Consequently, social networks built based on sub-forums may not always accurately reflect the nature of relationships formed in forum interactions.

2.2.2 Endogenous Partitioning

In contrast to using sub-forums as pre-defined boundaries, Dowell et al. (2015) created an endogenous boundary by conducting separate network analyses for all learners and only active learners that made more than 4 posts. The social network was constructed based on the direct reply structure. Dowell et al. (2015) examined five discourse dimensions of learners' forum contributions: narrativity, deep cohesion, referential cohesion, syntactic simplicity, and word concreteness. They then conducted mixed-effects modeling to identify the association between discourse features and social centrality (indexed by degree, closeness, and betweenness), as well as the association between discourse features and the final course grade. Discourse features were found useful for predicting learners' performance and social centrality, and the model worked better for active users than for all users. It was also found learners with higher social centrality had different discourse features than those with higher performance, such as lower referential cohesion, less abstract words and simple syntactic structure. The distinction between socially central learners and high performers and the association of the social learners with features indicting more superficial discourse dramatically indicates the need to consider forum activities directly related to the learning of course content and those that focused on social purposes separately.

In another study, Poquet and Dawson (2016) also focused on the type of activities learners participated in while examining the development of social networks and analyzed separately all users and regular users who participated in at least three weeks in a MOOC on solar energy. An undirected social network was extracted based on copresence in the thread. Adding an additional layer of endogenous boundaries to their network, they classified the posts contributed by regular users into five topic categories: “cognitive task” (comments about quizzes and assignment), “social task” (learner emotions about tasks), “cognitive non-task” (learners engage with subject outside of assignments), “social non-task” (purely social aspects), and an additional category for administrative and technical issues. They then analyzed what qualitative attributes were associated with regular participants’ network formation. First, regular users were found to have different social networks than all learners when examined at both the individual and group levels. Second, social non-task and cognitive task were the dominant type of posts in the discussion forum. Poquet and Dawson (2016) also found the topic of conversations were not significant for network formation modeling. The authors suggested this might have to do with the fact that the importance of topics may vary in different stages in the course, but interactions in different course stages were not modeled separately.

Findings from the above two studies as well as Gillani and Eynon (2014) and Gillani et al. (2014) disagree on the relationship between how students interact in discussions and whether or not the interactions are related to the course content, highlighting the need for further investigation. Content-related and non-content interactions refer to different genres of topics that play different roles in the learning process. Content-related interactions are directly related to the learning of the course material while non-content interactions may serve a highly diversified array of purposes (for example logistic or social) that can contribute to, but are generally more distal in their impact on learning. As shown in Dowell et al. (2015) these different ways of engaging may attract distinct collections of learners. Even when learners participate in both kinds of interactions, they may play different roles and show different participation patterns in the two contexts. Furthermore, from a theoretical perspective there is an expectation that participation in content-related interactions and social networks would be more predictive of learning outcomes than participation in non-content interactions. Thus for both conceptual and empirical reasons, it makes sense to analyze the social networks formed based on these activities separately.

2.3 Defining Social Ties

As noted previously SNA studies of MOOCs have used varying definitions to construct social ties. For example, the studies cited in this paper thusfar have used copresence and (directed and undirected) direct reply to construct ties. Tie definition is critical for SNA studies because different ways of establishing ties carry different assumptions about the nature of the interaction in social networks that have implications for network outcomes and their interpretation. This issue, however, has not been well addressed in the literature. The majority of studies simply establish their tie definition without giving any explanation or rationale.

Two types of social networks have been commonly constructed using the reply structure of threaded discussions in discussion forums that adopt multi-level post structure. The first type is the network of speaking interaction. In this type of network, a tie is defined as "speaking to" someone. This type of network formation mechanism is based strictly on the reply relationship in the forums. One approach that adopts this mechanism is direct reply relationship, in which a tie is constructed only if there is a direct reply relationship between two nodes in the same thread, either between the thread starter and a reply post addressed to it, or between a reply post and its reply, which is commonly seen in discussion forums that support multiple levels of posts. Direct Reply only maps the speaking interaction between users, without making any assumption about others who may have been informed by the post but not replied to it. This is perhaps the most straightforward approach to tie formation in online discussion and is used by many studies (Joksimović et al. 2016; Kellogg et al. 2014). However, an important problem with this scheme is there is no guarantee that when making a post, forum users always choose the appropriate location and level. In addition, some discussion forums may only support limited levels of posts. For instance, discussion forums on the Coursera platform only support three levels of posts. When a response is made to a level 3 post, it is displayed as another level 3 post. How accurately the reply structure produced in such forums reflects the actual relations among learners is questionable. As Direct Reply tie definition extracts social ties strictly according to the logged reply structure, it is most vulnerable to such problems.

To work around such concerns, Zhu et al. (2016) proposed Star tie definition for network extraction when investigating the relationships between course engagement, performance and social connectivity. Star also defines a tie as "speaking to" someone; but different than Direct Reply which distinguishes multiple levels of posts and maps ties as direct contact between nodes on different levels, Star considers all posts in the same thread being tied only to the thread starter. The rational is even if a reply post was not addressed directly to the starter, it was made in the context of the thread and should address the topic set by it, thus the tie is considered a traceable contact within the scope of a thread. Star structure highlights the importance of thread starter; however, as it does not distinguish between different levels of replies, it overlooks connections formed between learners within the same thread.

A hybrid scheme that combines Direct Reply and Star was introduced by Gruzd and Haythornthwaite (2008) in their investigation of social structure in online discussion forums using natural language processing and SNA methods. This scheme also makes use of traceable reply relationships to map ties. Ties are constructed both between posts on different levels and between each post and the thread starter. This scheme produces a more comprehensive network than the first two schemes alone, but still only considers the act of speaking in a threaded discussion. As a learner could access multiple posts made by people interested in the same topic, social relationships could thus form among them through "listening to" each other. Therefore, the methods that strictly follow the speaking contacts leave out the interactions and relationships between learners who never speak to each other directly but have spoken on the same topic in the same thread.

A different approach that addresses this issue is to create a copresence network that embodies coparticipation relationships among nodes. In a copresence network, a tie is defined as "being present" in the same part of a discussion. Two nodes can have a tie as long as they are in the same thread or subthread, without the need to have directly replied to each other. Thus ties are formed both hierarchically between child-parent nodes and horizontally between two nodes on the same level in the reply structure. This type of network formation mechanism represents the notion of online discussions as collective conversations rather than single streams of individual replies. Within the genre of copresence networks, the generally used scheme is total copresence where any two nodes in the same thread are considered having a tie (Gruzd and Haythornthwaite 2008). However, when this scheme is used to map interaction, the size of thread can be problematic, especially when many distinct people are involved as is the case in MOOCs. It might be reasonable to assume that a participant in a thread with a small number of replies or a large number of replies made by a small number of people has ties with all others in the same thread through reading their posts, but this assumption becomes less reasonable when the number of replies and people involved is very large. To address this problem, we propose another scheme which sets a reasonable number of posts in the same thread or subthread that a participant would read and use this as a measure of limited copresence. It would also be possible to assign deceased weight to ties in large threads. The five tie definitions are summarized in Figure 1.

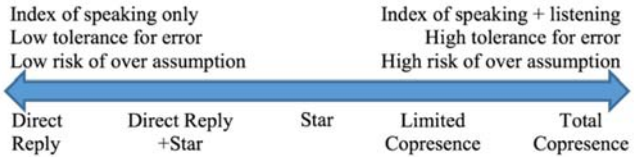


Fig. 1 Tie definitions on a continuum

3 STUDY FRAMING

In this study, we examine the effects of partitioning a MOOC forum social network according to whether or not the discussion interactions are related to the course content. This can bring many benefits. First, examining the two networks separately can provide understanding of network properties and formation mechanisms for different kinds of interactions and allows for more accurate attribution of factors that influence such processes. Second, it can reveal the characteristics of participants in the two networks and help to identify learner subpopulations so as to allow for a more specialized understanding of different learning needs. Third, it may allow for more nuanced prediction of learning outcomes. Finally, as content-related MOOC forum activities share many characteristics with conventional online learning forums, this approach allows for more aligned comparisons with previous online discussion SNA research findings. In addition, we use five tie formation definitions to investigate their impact on the consequent networks extracted. To allow for parallel comparisons across tie definitions, this phase of the work examines only undirected networks.

3.1 Research Questions

RQ1: What effects do different tie formation definitions have on the resulting network characteristics and interpretation?

RQ2: In what ways do social networks extracted from content-related and non-content activities show distinct characteristics from each other and the overall network?

RQ3: What differences in the threads and messages in content-related and non-content discussions may account for the observed differences in network structure?

These three research questions were investigated in a two phase approach. The first phase addresses both RQ1 and RQ2 as whole network comparisons are made across content-related, non-content and unpartitioned networks and across each of the five tie definitions. The second phase provides additional answers to RQ2 and addresses RQ3 with a deeper exploration of differences between key individuals and communities across content-related and non-content networks.

4 PHASE ONE: NETWORK CONSTRUCTION AND COMPARISON

4.1 Methods

4.1.1 Data Source

This study used data from StatMed'14, a completed MOOC offered in 2014 on Stanford open-source platform Lagunita. The course is an introductory course on probability and statistics with a special focus on statistics in medical studies. The course provided a discussion forum for interaction in nine topic areas, including General, Video, Homework, Course Material Feedback, External Resources, Tech Support, Introductions, Study Group, and Platform Feedback. Learners were invited to post questions and comments for response by peers, the TA and the instructor. Forum information provided in the dataset included the following: thread id; post id; user id; post position in thread (thread starting post, reply post, or reply to reply post); parent post; post text; post creation date and time; and number of votes post received. Thread titles were not included. The discussion forum was participated by 568 unique users. They made 817 thread starting posts, 1277 reply posts, and 1035 reply-to-reply posts in the forums. Of the 817 threads, 117 received no reply. Five reply posts that contained non-English language or only punctuation were removed, leaving a total corpus of 817 threads with 2307 replies made by 567 users.

4.1.2 Thread Classification

The 817 threads were classified as either being content-related or non-content using a unigram and bigram based-model built on manually-coded starting posts from a prior offering of the course (Wise et al. 2016). In previous work, the model demonstrated good reliability on StatMed'14 data for both thread starting and reply posts (accuracy > .81) (Wise et al. 2017; Wise et al. 2016). This model was used in conjunction with dynamic interrelated post and thread categorization DIPTiC (Cui et al. in press) to categorize threads by comparing the model classification of thread starting post and distribution of replies. This additional step increases the estimation of classification accuracy to .88 (Cui et al. in press). Using this comprehensive characterization method, a total of 468 threads containing 1446 replies were labeled as content-related and a total of 349 threads containing 861 replies were labeled as non-content.

4.1.3 Network Participants

The nodelist was extracted from discussion forum data using user id of posts. It was found that of the 567 forum users extracted from the cleaned data, 178 participated only in content-related threads, 232 participated only in non-content threads, and 157 participated in both kinds of threads. Thus the number of nodes for the unpartitioned, content-related, and non-content networks were 567, 335, and 389 respectively.

4.1.4 Tie Extraction

Edgelists were extracted using five different tie definitions for the unpartitioned, content-related, and non-content networks.

Direct Reply: The author of each post is connected with the author of its parent post; this represents the actual reply structure (see Figure 2a). Using this definition, a total of 2307 ties were extracted from the unpartitioned network; after removing 286 self-loops, 2021 ties representing 1086 unique edges remained, including 1249 content-related ties representing 625 unique edges and 772 non-content ties representing 551 unique edges.

Star: The author of each reply and reply-to-reply post is connected with the author of the thread starting post (see Figure 2b). Using this definition, a total of 2307 ties were extracted from the unpartitioned network; after removing 502 self-loops, 1805 ties representing 1116 unique edges remained, including 1092 content-related ties representing 625 unique edges and 713 non-content ties representing 558 unique edges.

Direct Reply + Star: Ties defined in both Direct Reply and Star were included but the same tie was never counted more than one time (see Figure 2c). Using this definition, a total of 3339 ties were extracted from the unpartitioned network; after removing 683 self-loops, 2656 ties representing 1292 unique edges remained, including 1697 content-related ties representing 747 unique edges and 959 non-content ties representing 643 unique edges.

Total Copresence: All authors in the same thread are connected with each other (see Figure 2d). Using a VBA script written for this definition, a total of 15299 ties were extracted from the unpartitioned network; after removing 1992 self-loops, 13307 ties representing 5578 unique edges remained, including 7018 content-related ties representing 1133 unique edges and 6289 non-content ties representing 4641 unique edges.

Limited Copresence: All users in small threads (<5 replies) are connected to each other; in larger threads users are connected to all other users in their sub-thread and the thread starter only (see Figure 2e). Of all 489 subthreads in larger threads, only 69 (14%) have more than four posts. Using a VBA script written for this definition, a total of 5313 edges were extracted from the unpartitioned network; after removing 1066 self-loops, 4247 ties representing 1456 unique edges remained, including 2879 content-related ties representing 848 unique edges and 1368 non-content ties representing 724 unique edges.

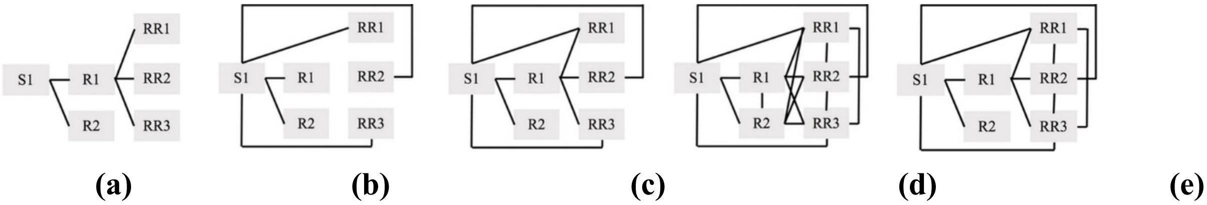


Fig. 2 Ties based on five definitions **a** Direct Reply; **b** Star; **c** Direct Rply+Star; **d** Total Copresence; **e** Limited Copresence

S = thread starting post R = reply post RR = reply to reply post

Solid lines represent ties extracted using this definition.

4.1.5 Network Construction and Network Properties

The edgelists and nodelists for each network were imported into Gephi 0.9.1. for mac to construct undirected weighted networks, compute network measures, and visualize the networks employing the Force Atlas layout algorithm. For each of the fifteen networks (unpartitioned, content-related, non-content for each of the five tie definitions), the number of edges, average node degree, average edge weight, and graph density were computed. Community detection was performed through modularity maximization using the Louvain method. Randomization was used to improve decomposition with an acceptable tradeoff in computation time; ten runs were conducted for each of the fifteen networks. For twelve of the fifteen networks the resulting major modules were consistent. For the Limited Copresence unpartitioned and non-content networks, and the Total Copresence content-related network, the runs produced two different module structures, one with two of the key nodes (u1 and u417) in the same module and one with them in separate ones. Each structure was produced multiple times, necessitating a principled choice about which to interpret. The decision was made to use the structure with u1 and u417 in separate modules in order to allow for better examination of potential differences between them.

4.2 Results and Discussion

The properties of the fifteen network graphs are reported in Table1 and the graphs are shown in Figure 3.

4.2.1 Comparing Tie Definitions

For both partitioned and unpartitioned networks, there are clear trends across tie definitions in the number of edges with the more liberal tie definitions producing higher values (see Table 1). This shows the expected relationship between how ties are defined and the number of connections presumed to have taken place. Interestingly, while this trend is replicated in average node degree (more liberal tie definitions indicated a greater breadth of others with whom users were considered to have interacted), the average edge weights (indexing the strength of connection) based on the different definitions did not follow the same pattern: Total Copresence generated the *highest* average edge weight of all tie definitions for the content-related network, but it generated one of the *lowest* average edge weights of any definition for the non-content network. The network graphs extracted using the Total Copresence definition are dramatically different than those extracted using other definitions. Specifically, all of the unpartitioned and non-content networks extracted based on Direct Reply, Star, Direct Reply+Star, and Limited Copresence have three major modules, each dominated by a single node of high centrality (u1, u2, and u417, see Figure 3 a1-d1, a3-d3). In contrast, the unpartitioned and non-content networks extracted using Total Copresence definition have only two major modules, one containing both u1 and u417, and the other a "balloon" of many similar-degree interconnected nodes (see Figure 3 e1 and e3). Examination of the post text contributed by u1 and u417 revealed them both to be members of the instructional team. Examination of the post-text data that contributed to the "balloon" showed that this was due to a single socializing thread started by learner u2 at the beginning of the course which received a total of 92 replies. When this superthread was removed from the non-content network and the properties were recalculated, the average node degree was 4.40 (SD = 12.70), still the highest among all definitions; the average edge weight was 1.85 (SD = 1.96), slightly lower than that of Limited Copresence. This suggests Total Copresence is problematic for calculating network properties because one or two superthreads can inflate average node degree and deflate average edge weight dramatically.

Looking globally, differences between Direct Reply and Star definitions were smaller than expected, both in terms of network properties (see Table 1) and graph appearance (see Figure 3). Direct Reply+Star and Limited Copresence, although based on different conceptual tie definitions, appear to produce networks with similar properties (see Table 1). These findings suggest that social network analysis is more robust to tie definition than was expected, except for Total Copresence.

Total Copresence is useful for identifying inflated social status due to initiating superthread. It can be used in combination with other tie definition to bring another perspective to understanding activities in the network.

In addition, the contrast between the networks constructed based on Direct Reply and Total Copresence, the two most commonly used tie definitions in the literature, revealed the influence of tie definitions on the resulting networks, highlighting the importance of choosing tie definition appropriate for the research purpose.

Table 1 Network measures of five overall networks

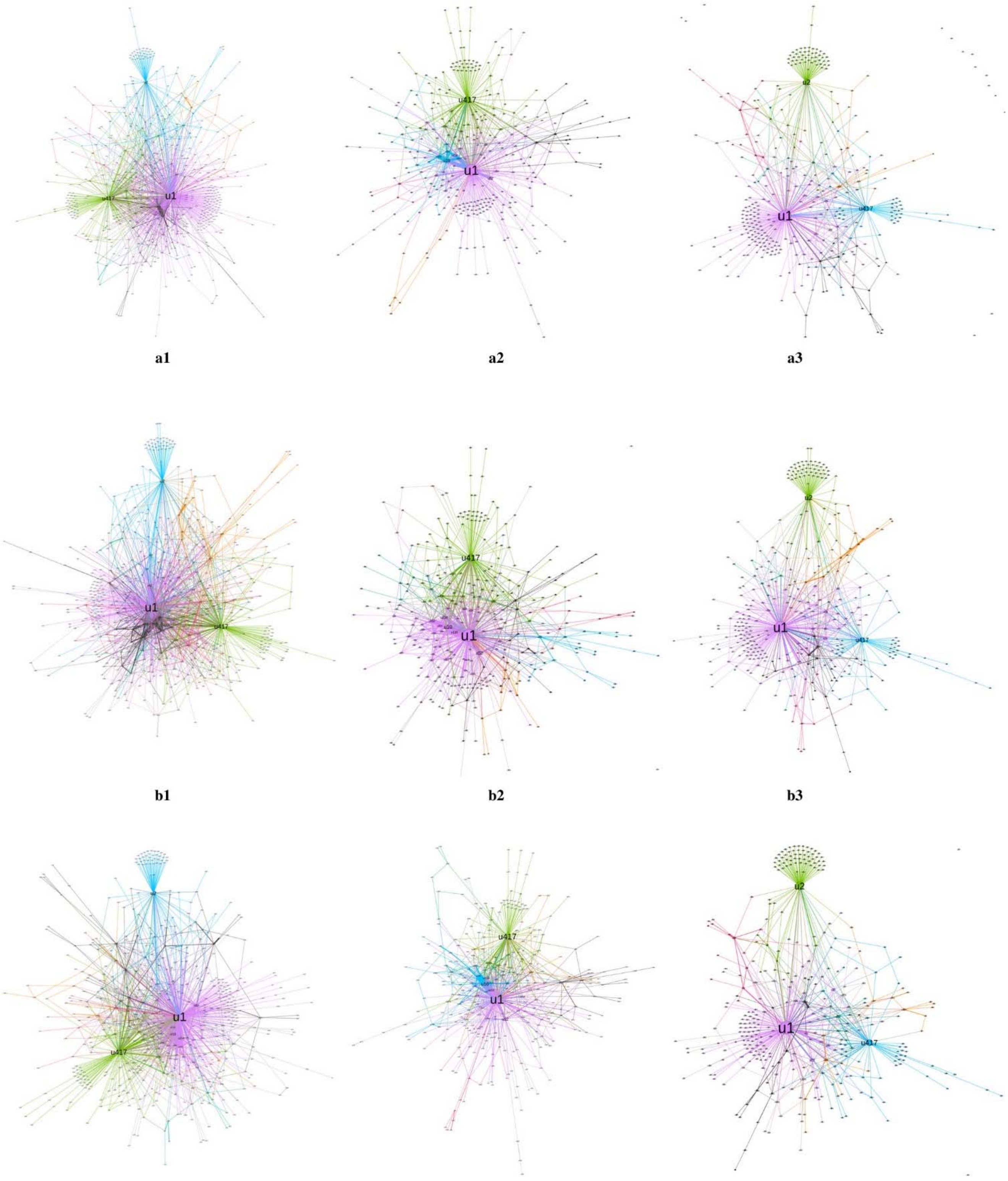
	Unpartitioned (N=567)				Content-related (N=335)				Non-content (N=389)			
	# of edges	Avg node degree (SD)	Avg edge weight (SD)	Graph density	# of edges	Avg node degree (SD)	Avg edge weight (SD)	Graph density	# of edges	Avg node degree (SD)	Avg edge weight (SD)	Graph density
DR	1086	3.83 (14.54)	1.86 (3.38)	0.007	625	3.73 (11.33)	2.00 (3.62)	0.011	551	2.83 (10.47)	1.40 (1.16)	0.007
S	1116	3.94 (12.92)	1.62 (2.50)	0.007	625	3.73 (9.59)	1.75 (2.89)	0.011	558	2.87 (9.50)	1.28 (0.79)	0.007
DR+S	1292	4.56 (15.48)	2.06 (4.36)	0.008	747	4.46 (12.04)	2.27 (4.90)	0.013	643	3.31 (11.20)	1.49 (1.41)	0.009
LC	1456	5.14 (16.69)	2.92 (9.49)	0.009	848	5.06 (13.20)	3.40 (11.10)	0.015	724	3.72 (12.09)	1.89 (2.75)	0.010
TC	5578	19.68 (36.63)	2.39 (14.98)	0.035	1133	6.76 (15.64)	6.19 (31.73)	0.020	4641	23.86 (38.06)	1.36 (1.53)	0.061

DR = Direct Reply; S = Star; DR+S = Direct Reply+Star; LC = Limited Copresence; TC = Total Copresence

4.2.2 Comparing Content-Related and Non-Content Networks

Content-related and non-content networks were found to have several distinct characteristics. First, both average node degree and average edge weight were higher in the content-related network than in the non-content network for all tie definitions except Total Copresence (see Table 1). In Total Copresence, the non-content network had dramatically higher average node degree and lower average edge weight than the content-related network. However, this was largely due to the fact that the superthread with 88 participants in the non-content network inflated the average node degree and deflated the average edge weight. When the superthread in the non-content network was removed, both the recalculated average node degree (4.40, SD = 12.70) and average edge weight (1.85, SD = 1.96) were lower than those of the content-related network. These cross-definition trends indicate that learners interact with more people and have more repeated interactions with the same people in content-related discussions than in non-content discussions.

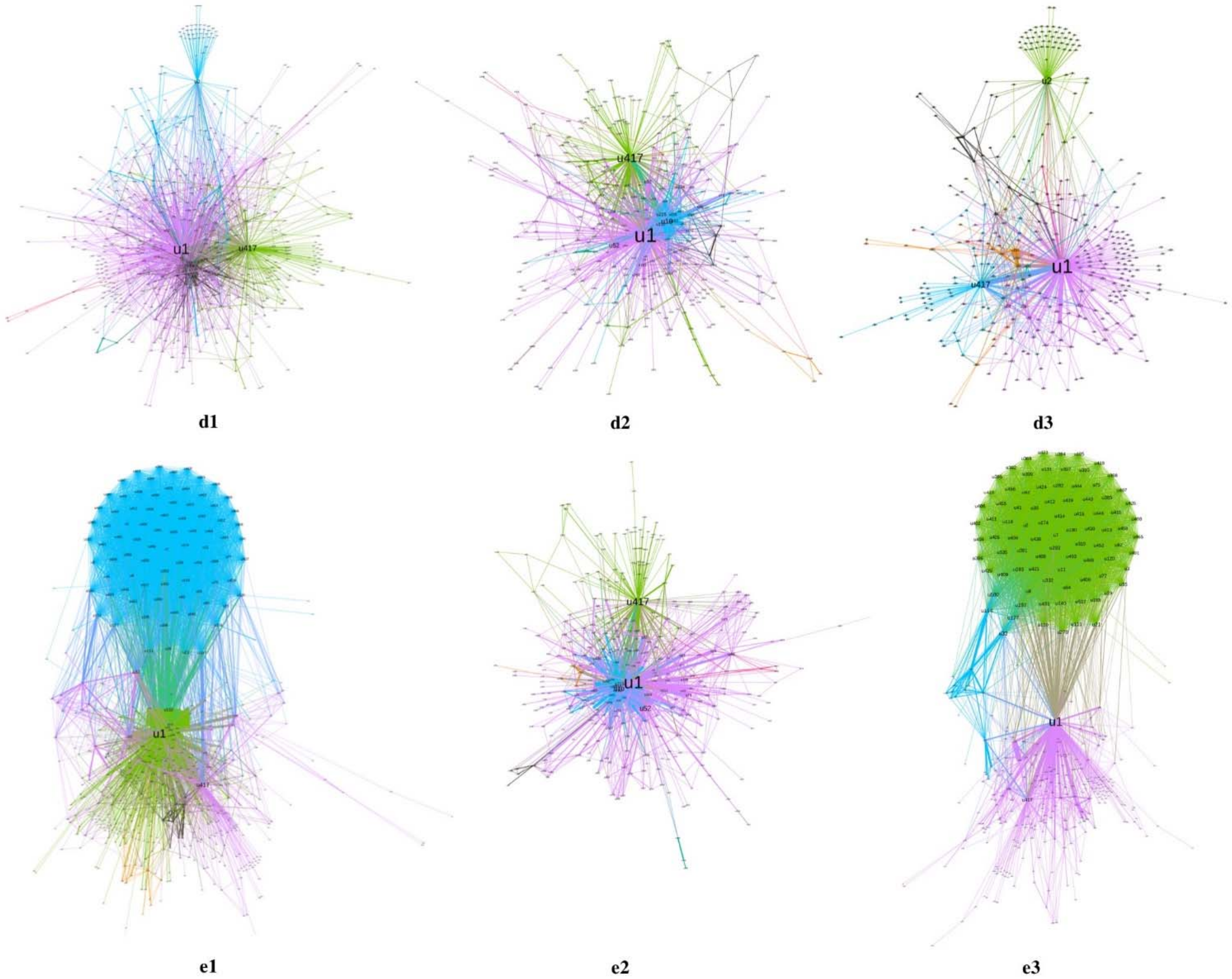
Second, the two kinds of networks showed different community patterns (see Figure 3). All of the content-related networks contained two major modules, each dominated by one of the instructors (u1, u417) with substantial connections between the two modules. In contrast, all of the non-content networks (except Total Copresence) showed three major modules with fewer links between them. Even when the superthread was removed from the non-content network, this pattern still held, with another learner-only module taking the place of the balloon.



a1 a2 a3

b1 b2 b3

c1 c2 c3



Only the primary components are shown in this figure. Node size represents degree. Color indicates module.

Fig. 3 Social networks constructed using **a** Direct Reply; **b** Star; **c** Direct Reply + Star; **d** Limited Copresence; **e** Total Copresence for **1** unpartitioned; **2** content-related; **3** non-content discussions

5 PHASE TWO: In-Depth Comparison of Individuals and Communities in Content-Related and Non-Content Networks

5.1 Methods

The second phase of analysis compares the characteristics of individuals and communities in the content-related and non-content networks of the MOOC discussions. The Limited Copresence tie definition was used to construct the networks as conceptually it encapsulates both speaking and listening activities, and empirically is less sensitive to posting error than direct-reply approaches but avoids the disproportionate influence of large threads in the Total Copresence approach.

Learner – Community Level: Top Learner Module Extraction, Calculations, and Thread Examination

All learner-only modules that contained more than 5% of the total participants in the content-related network or non-content network were extracted for further investigation. From the content-related network one learner-only module (Content-Related Learner Module 1, hereafter as CL1) was extracted; from the non-content network two learner-only modules (Non-Content Learner Modules 1 and 2, hereafter as NL1 and NL2) were extracted. For each module the number of nodes, number of edges, average node degree, average edge weigh, and graph density were calculated. All threads that contributed to the connections in each learner module were extracted and examined. Prior to analysis, threads were manually checked to verify if they had been properly categorized as content / non-content using the DIPTiC method. Only one of the 30 threads labelled as content-related and none of the threads labelled as non-content were determined to be miscategorized. Exclusion of the miscategorized thread did not substantively change interpretation; thus with the eventual aim of automated analyses where such miscategorization might not be detected, results for the full data set are reported. Threads were analyzed intact to maintain the context of interaction for interpretation; this required the inclusion of some posts in these threads made by learners who did not belong to the module (23%, 31%, and 21% of posts in CL1, NL1, and NL2 respectively). For the threads contributing to each module, the following were calculated: number of contributing threads, average thread length (number of posts per thread), average number of subthreads per thread, average subthread length (number of posts per subthread). The contents of the posts in each thread were also examined qualitatively to probe the characteristics, similarities and differences between interactions contributing to the top learner modules for the content-related and non-content networks. Threads were first color coded by participant id, then read and summarized according to the expressed purpose of the thread initiation, the overall dynamics of replies that resulted, any other purposes that emerged, and if these purpose(s) were eventually fulfilled. In addition, predominant interaction techniques (e.g., providing factual information, sharing personal understanding, asking leading questions) and social presence indicators (e.g., greetings, addressing people by name) (Wise, Chang, Duffy, & Del Valle, 2004) were noted for each thread. These summaries were then examined for patterns across threads contributing to each module. As this was an exploratory analysis, one researcher conducted the bulk of the analysis with consultation with a second researcher.

Learner – Individual Level: Top Individual Learner Identification Calculations, and Post Examination

Degree was calculated for all learners in the unpartitioned, content-related, and non-content networks. Learners that ranked in the top 10 for any of the three networks were selected for further investigation, including examination of which network(s) they participated in (content-related, non-content, both) and whether they also ranked in the top 10 list for any other network. For the learners in the top 10 for both the content-related and non-content networks, activities in the two networks were further examined and compared. All posts made by each learner in both networks were extracted and the following were calculated for each learner: the number of threads they participated in, average number of posts they made in each thread, and average number of words in the post. The contents of the posts were also examined qualitatively to probe the characteristics, similarities, and differences between their activities in different networks. Posts were read and summarized according to the topics discussed, interaction techniques, and social presence indicators.

Instructor – Community and Individual Levels: Instructor Module Extraction, Calculations and Post Examination

All modules containing a member of the instructional team were extracted for further investigation. This included two modules in the content-related network (Content-Related Instructor Module 1 and 2 containing u1 and u417 respectively, hereafter as CI1 and CI2) and two in the non-content network (Non-Content Instructor Module 1 and 2 containing u1 and u417 respectively, hereafter as NI1 and NI2). For each module the number of nodes, number of edges, graph density, average node degree, average edge weigh, and the instructor's degree were calculated. To investigate the instructors' influence on learning activities directly related to the course content, the post texts contributed by the instructors in the content-related threads were examined qualitatively to probe the characteristics, similarities, and differences between the instructors' intervention activities. Posts were read and summarized according to the instructor's intervention techniques (e.g. providing straight forward answers, providing hints, asking leading questions) and their use of social presence cues.

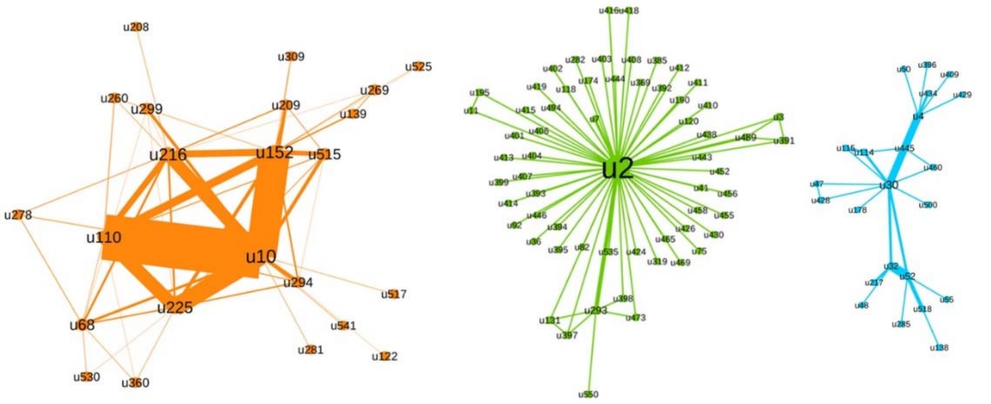
5.2 Results and Discussion

5.2.1 Learners - Community Level

Several differences were found at community level between content-related and non-content learner modules. First, while module CL1 had the same number of members (nodes) as module NL2, it had twice the number of edges, leading to a higher average node degree and density (see Table 2) indicating that participants interact more broadly with different peers in the content-related network. While it is not proper to make direct comparisons of density with NL1 due to its larger size, it can be seen that the average node degree was more similar to that of NL2 than CL1. The distribution of these differences in connectivity can be seen in the network graphs (see Figure 4). In CL1 there was a web of connections with the majority of learners connected to at least two others. NL1 had a wheel structure with the majority of learners connected only to the hub; NL2 displayed an elongated branching chain where again the majority of learners were connected to only one other (though in this case there was greater diversity in which other learner they were connected to). In addition, looking at the intensity of connection, the average edge weight in CL1 was dramatically higher than that in NL1 and NL2 (see Table 2), indicating that learners had substantially more repeated interactions with the same peers in the content-related module. Again, the distribution of edge weights can be seen in the network graphs (see Figure 4). In CL1 a central clique {u10, u110, u152, u216, u225} accounted for much of the increase in average edge weight, but there were also multiply weighted ties connecting these learners with others. Such high-weighted edges were absent in NL1 and NL2 indicating the majority of participants in the module did not have repeated interactions with the same peers.

Table 2 Thread structures of learner modules (ties based on Limited Copresence)

	CL1	**NL1**	**NL2**
# of nodes	23	62	23
# of edges	57	71	28
Graph density	0.225	0.038	0.111
Node degree: ave (SD)	4.96 (4.08)	2.29 (7.44)	2.44 (2.39)
Edge weight: ave (SD)	14.67 (32.61)	1.06 (0.29)	1.89 (1.63)



a. CL1 b. NL1 c. NL2

Fig. 4 Comparison of **a** CL1; **b** NL1; **c** NL2 (ties based on Limited Copresence)

Looking at the specific threads contributing to the connections in these modules (see Table 3), CL1 had over twice the number of contributing threads as either of the non-content learner module, indicating that interactions were more distributed across the discussion forums. Differences were also found in thread size and structure. Threads contributing to CL1 had more than twice the number of posts on average with relatively longer sub-threads compared to NL1 and NL2.

Table 3 Structure of threads contributing to learner-only modules (ties based on Limited Copresence)

	CL1	NL1*	NL2
# of contributing threads	30	2	11
Avg # of posts in thread (SD)	13.4 (14.85)	6, 93	6 (5.75)
Avg # of subthreads per thread (SD)	2.73 (2.05)	5, 82	3 (3.41)
Avg # of post per subthread (SD)	4.80 (2.76)	1.2, 1.13	2.57 (1.51)

* This module had only two contributing threads of very different characters, thus results are reported for each thread separately

Qualitative examination of thread contents revealed such structural differences were associated with several differences in learners' information seeking and giving behaviors in content-related and non-content contexts. First, the differences in overall thread length may be explained in the kinds of questions asked in content-related versus non-content threads. Non-content starting posts often asked for straightforward factual information which could be easily provided without extended conversation. For example:

> ***U48****: I included the % sign by mistake in the answer box… Is there any possibility to counteract this problem? Please...*

Also:

> ***U32****: I remember in the introduction to the course, there have been a note that you can't amend your answer after submission (this is applied to homework only and not quiz). I hope if there is a way someone will help you.*

In contrast, content-related threads commonly involved asking for help with problem-solving or understanding abstract / complicated concepts which required multiple rounds of back-and-forth comments to resolve. In these threads, it was common for participants to use diverse interaction techniques such as paraphrasing or clarifying the question, giving explanation, examples, and comparisons, asking follow-up questions and using leading questions to coach their peers. For example:

> ***U209****: Can anybody help me with quest 10 of unit 4. Do we have to consider the mean = proportion = 112/200 = 0.56? Then assuming X=112, mean= 0.56, SD = 0.035, the Z score is coming something abnormal. where is the fault?*
>
> ***U167****: Good morning, the question states that you should use the normal approximation to the binomial. I would go back and look at slides 139-141 for this unit and double check the equation that you're using for the mean. The mean is not a proportion, it is =n* p. The problem wordings gives you both of the values for those variables you just need to plug them in!"*
>
> …
>
> ***U209****: Thanks but I'm still confused. Don't we have to use the statistics of proportion here? 112/200 =0.56 and if I'm using the formula mean= n*p, and X = 112, then the z score is coming to zero. Does that make any sense?*
>
> ***U502****: p of flip a coin is 0.5, X=112, mean(u)=n*p=0.5*200. You can calculate SD using sigma^2=np(1-p) and z=(x-u)/sd. and use Standard Normal Distribution Table.*
>
> ***U10****: Hint: Don't forget to take the square root in the denominator when solving for the SD (a mistake I made!).*

Second, the difference in sub-thread length may be also related to the different ways that threads were initiated in the content-related and non-content discussions contributing to the modules. Almost half of all starting posts in the content-related threads were expansive in nature, either making a general call for discussion or asking together about several homework questions. The responses to such starting posts consequently often broke into separate sub-threads focused on specific homework questions or topics. For example:

> ***U10****: Has anyone begun on this homework? I'd love to start some threads like we had going in homework 7. Thanks!*

The other half of content-related starting posts raised specific questions that could be addressed in single subthread, but in many cases the participants initiated new subthreads to expand the discussion to include follow-up or other relevant questions. In contrast, non-content discussions did not tend to break into multiple topics and when sub-threads were used, the structure did not seem to indicate anything substantive about the conversation.

Differences were also seen in how learners participated in the threads contributing to the different modules. Participants in the threads contributing to CL1 made an average of 2.41 (SD = 1.62) posts per thread, while those in the threads contributing to NL1 and NL2 contributed 1.13 (SD = 0.07) and 1.42 (SD = 0.53) posts per thread respectively. These findings indicate participants in CL1 revisited the threads more often than those in NL1 and NL2 did. This pattern can be seen in the qualitative examination of the post texts which showed that in many of the content-related threads examined, the thread initiator revisited their threads to interact with peers who responded to them. In addition, after their initial questions had been answered, learners often posted in the same subthread or other subthreads in the same thread to provide help for other peers, which often led to multiple rounds of talks; while as for non-content conversations, once the answer/solution was provided, the interaction usually stops. For example:

> ***U110**: Hello [u10], For Question 8: Referring to Table 2, the unadjusted beta coefficient for exclusive/near exclusive breast feeding for visual reception is 4.4 (0.7, 8.2). What model was used to generate this beta coefficient? My logic: The unadjusted beta (4.4) is pertaining to the visual reception right? In the equa]tion I should quantify the breast feeding score in terms of the intercept and the visual reception score? For 9-11 I saw [the instructor's] videos of how to analyse the data for confounders and then answered them. Do you think I am on the right track? For 3 I actually made a pseudo data like the scatter plot and then checked the assumptions that were being violated/followed for linear regression. Do you think I am using the right logic? Best Wishes.*
>
> ***U10**: [u110], Q8: Are you asking if Y = visual reception and then what is the equation to estimate Y in regards to the intercept and the categories of infant feeding? For the equation problems, when there were categorical predictors, I went back to the slides to see how [the instructor] handled that. e.g., slide 124, 128. Q9-11 Only one of those questions seems to be referring to confounding, the other two were about interpretation, I would just say to review an example of her interpreting a continuous vs categorical predictor. Q3 Probably module 3 is the best one to use for this question, I know you have been discussing this one with Josh also. When I was considering the question, I looked closely at the graph and all the dots over the years. I think you can answer the question by looking at the graph. Good luck and feel free to ask me more questions :).*
>
> ***U110:** Thank you [u10]. :) One generic question if we have categorical data it can represented as a series of binary combinations. In a study we can only identify categorical variables as binary if they have been dealt as binary right? Otherwise they can be identified as categorical. Best Wishes.*
>
> ***U10**: Hi [u110], We create dummy codes when there are more than two categories, but I believe that a binary variable is a categorical variable with only 2 categories, but a variable with more than 2 categories is never binary - but can be dummy coded in the binary variables... OK - how confusing was that? Actually - I think I just reworded what you said!*
>
> ***U110**: So [u10], does Linear Regression always handle categorical predictors as binary (2 predictors simply as binary and more than two dummy coded binary)?*
>
> ***U10**: [u110], I would not say that LR handles anything, we have to create the new variables if they have more than 2 categories and then enter them in the model correctly, and with the referent group of our choice. IF I am following your question.... I think it is also important, here - for the homework, how we interpret the effect of a continuous variable on Y and the effect of a binary variable on Y.*
>
> ***U110**: [u10] Thank s for the advice. I am using your advice and Module 5 to make a decision.*

The above excerpt also shows there were clearer signs of sprouting community in CL1 than NL1 and NL2. In many cases in the examined content-related threads, participants called on peers to have group discussions and expressed longing for study partners:

> ***U299**: Hi everybody, I am enjoying this course so much, I can't wait to apply my newly acquired stat skills to my work. However, I ran into an unexpected complication and would like to summon your help (instructors and participants) to sort it out…*

> *Participants in the content-related threads also addressed their questions to the community:*
>
> ***U209**: Can anybody help me with quest 10 of unit 4?*
>
> *They also social presence indicators when interacting with each other in CL1, which was less common in NL1 and NL2. For example:*
>
> ***U360**: [U68], I didn't understand it either, so I looked for the original article. It refers to scenes where the characters are drinking alcohol, so the viewers are the ones being exposed to alcohol.*

In particular, a conversation among the learners connected by high-weighted edges in CL1 (u10, u110, u152, u216, u225) exemplifies that besides learning, they developed strong interpersonal bonds through participating in learning-related discussions:

> ***U10**: I am done! I missed one we didn't even discuss, because I just read the response wrong or the selection moved on me by accident, but anyways, OVER! And so is the exam... sigh... this was the toughest MOOC I have taken - grading wise.*
>
> ***U225**: Congrats [u10]! Yes, it has been hard, but fun, and we learned an awful lot, right?*
>
> ***U110**: Great! Everyone it was a pleasure to work with you. Thank you….*
>
> ***U10**: YES [u225]! And [u110] - the test was scary - I thought of my discussion board friends often!!*
>
> ***U216**: Thanks, thanks so much to [u10], [u152], [u110], [u225] and everybody who helped us to understand this beautiful course! And in my case also for writing many posts, I see I have improved my English skills and my statistics vocabulary!!!*
>
> ***U225**: [u10], [u216], [u152], [u110], [u515] and everyone, your discussions helped me so much. I was always a few days behind you in homework - glad I was able to catch up in the last weeks and participate a little bit….*

Moreover, the description of the self-introduction thread in NL1 revealed several interesting findings about socializing activities. This thread was initiated by u2 for learners to make self-introduction and was joined by 87 other participants, many of whom expressed interest in interacting with others. For example:

> ***U32:** Hi everyone, I’m a doctoral student in laboratory medicine, from Egypt. My specialty is Clinical Chemistry. As you may guess, it is part of my study to have a good knowledge about medical statistics. I'm looking forward to complete this course, hopefully with distinction. Hope to be friend with all of you guys. Best wishes for all.*

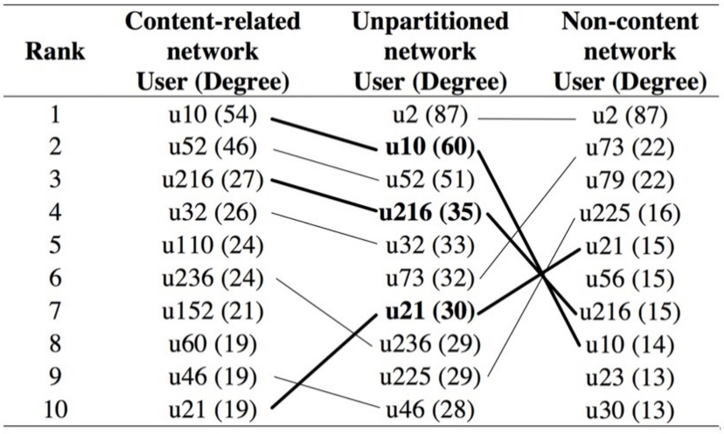
However, most participants just made a monologue self-introduction and never participated in this thread again; only 5 participants posted more than once in this thread. This indicates despite of the participants’ intention to have interactions with others, little interaction actually took place in this purely socializing-oriented thread.

5.2.2 Learners - Individual Level

Cross-network differences

First, except for u2 and u30 who only participated in the non-content network, other top learners all participated to some extent in both networks. However, for the 15 remaining distinct learners on the lists, only u10, u216, and u21 appeared on the high degree lists for both the content-related and non-content networks while the rest of the top ranked learners had high degree in one network but not both (see Table 4). This indicates top players in the two networks were largely different people and that even participating in both content-related and non-content threads, learners who were highly connected in one network were not necessarily highly connected in the other, and vice versa.

Table 4 Top 10 learners ranked by degree centrality in content-related, unpartitioned, and non-content networks (ties based on Limited Copresence)



Rank	Content-related network User (Degree)	Unpartitioned network User (Degree)	Non-content network User (Degree)
1	u10 (54)	u2 (87)	u2 (87)
2	u52 (46)	**u10 (60)**	u73 (22)
3	u216 (27)	u52 (51)	u79 (22)
4	u32 (26)	**u216 (35)**	u225 (16)
5	u110 (24)	u32 (33)	u21 (15)
6	u236 (24)	u73 (32)	u56 (15)
7	u152 (21)	**u21 (30)**	u216 (15)
8	u60 (19)	u236 (29)	u10 (14)
9	u46 (19)	u225 (29)	u23 (13)
10	u21 (19)	u46 (28)	u30 (13)

Second, u10, u216, and u21 who were on the top learner lists for both the content-related and non-content networks demonstrated differences in their participation across the two networks. All three learners had higher degree in the content-related network than in the non-content one (see Table 4), and interacted with more people through the former network. Differences were also found in how they participated in the discussions (see Table 5): two of the three learners (u10 and u216) participated in dramatically more content-related threads than non-content ones, and both made more posts per thread in the content-related threads than in the non-content ones. Together this shows both a greater breadth of participation (across multiple threads) and depth of engagement (revisiting the same thread more than once). In addition, in most cases, the posts (both starters and replies) made by the three learners in the content-related threads contained more words than those they made in the non-content threads, which is a potential indicator that their participation in the content-related network involved greater depth and complexity.

Table 5 Forum contributions made by u10, u216, and u21 in the content-related and non-content networks (ties based on Limited Copresence)

	U10		**U216**		**U21**	
	Content-related	Non-content	Content-related	Non-content	Content-related	Non-content
# of threads participated in	49	16	25	12	9	8
# of post per thread	2.84	1.44	2.48	2	1.11	1.13
Avg # of words per starter (SD)	134.65 (85.23)	54.4 (56.70)	68.2 (37.85)	70.8 (79.81)	38.75 (7.46)	10.5 (0.5)
Avg # of words per reply (SD)	50.82 (48.41)	37.77 (43.15)	34.1 (29.55)	26.89 (26.51)	37.5 (26.11)	20 (16.37)

Qualitative investigation of the three learners' post texts revealed characteristics that help explain these differences. In the non-content threads, the three learners' interactions with others mostly involved exchanges of factual information, for example:

***U10**: Hello, I am wondering if there is - or could be - a handout to accompany the math review video. Thank you.*

In contrast, in content-related threads, their interactions involved not only exchanges of factual information, but also more complicated activities, such as sharing solution steps and thinking processes, explaining concepts and addressing confusions, as well as discussing abstract and complicated problems. For example:

> ***U21:** I think it is similar to mean and standard deviation. If within 3 standard deviation[s] from [the] mean, we have 99% [of the] data. So here is median, and 3 quantile[s] (1 for 3/4 quantile and 1/4 quantile plus Q3 + 1.5 * IQR and Q1 - 1.5 * IQR) from median. But why choosing 1.5, I also want to know.*

It is worth noting that this finding is in conformity to the community level findings in Section 5.2.1, indicating that the differences between the content-related and non-content networks were due not only to participation by largely different people, but also the same people behaving differently in the two networks.

Inter-learner comparison

The qualitative investigation also revealed inter-learner differences in their interaction behaviors and techniques. Of the three learners, u10 expressed the strongest enthusiasm for interacting with others and utilized the most sophisticated interaction techniques. He/she initiated and took part in multiple group discussion threads; shared information and thoughts related to the learning of course content to invite responses; addressed his/her posts to not only the instructors but also the peers; and used social presence indicators frequently when communicating with others.

> ***U10**: Hi ~ I got the same answer for numbers 3 and 4 and wonder if anyone else did, and also if anyone wants to compare steps taken to solve the problem. This is how I approached it…*

Also:

> ***U10**: Dear [instructors] and Classmates, have you seen the popular press (mags, TV, on line) announcements about running as little as 5 minutes a day leading to 45% less all-cause mortality and 30% less cardiovascular (CVD) death, compared to nonrunners? I found the study in the Journal of the American College of Cardiology and read i[t] carefully. I did this because the headlines don't sound right to me. Also, I wanted to practice some of the skills we have been acquiring in this class. Definitely, this is an observational study so it is incorrect to say that running any amount of time is 'causing' the reductions in adverse outcomes, but the study authors themselves are using the 'as little as five minutes a day' line. This bothers me a lot…. I really want to understand this study so that I can share the results in a responsible way - also, I love running. TY! [Abstract]: http://content.onlinejacc.org/article.aspx?articleID=89600*

Moreover, this learner demonstrated instructor-like behaviors when interacting with the peers, such as providing hints and asking leading questions. For example:

> ***U10**: Hi [Peer learner's name], you make the determination based on the CI related to the beta. Do you remember the range of beta that is possible? Does the CI cross the null boundary?*

U216 also demonstrated strong interest in interacting with peers and used good interaction skills. He/she provided help and contributed ideas in threads initiated by others. A big proportion of his/her posts involves socializing and expressing gratitude. Like u10, u216 also used social presence indicators in some cases when communicating with others. However, a notable characteristic of u216 is his/her strong interest in interacting with the instructors. The majority of the threads this learner initiated were addressed explicitly to the instructors for help, hint, or information. For example:

> ***U216**: Dear [Instructor], good evening. Thanks much for your last help (extra-material always is good). I have been working on HWK5 Q11 - Q12 many hours but I think I am missing something in the slides because I am not sure about my answer with proportions. I was reading about [Learner's name] and you posted but I think I need more hints! Thanks in advance.*

Compared to u10 and u216, u21 made a much smaller number of posts; his/her contributions to the content-related and the non-content threads were of similar proportion. It is noteworthy that unlike u10 and u216, u21 rarely posted multiple times in the same thread. In addition, U21 never addressed any post explicitly to any specific person or audience, and never used social presence indicators. However, such behavior should not be interpreted as lacking interest in connecting with others. Although all of the content-related threads he/she started are help-seeking questions, u21 did provide help to others in both content-related and non-content threads initiated by other learners.

> ***U21**: I do not know the answer of your question, and I hope someone could help us for your problem. But I think here the Prof. just illustrated how we could handle category variables when we use linear regression (e.g., dummy coding).*

Interestingly, among the three learners, u21 is the only one that took part in the self-introduction thread initiated by u2 at the beginning of the course.

5.2.3 Community and Individual Level Analysis for Instructors

Cross-network comparisons

The biggest difference was found in the dominance of instructor modules between the two networks. In the content-related network, CI1 and CI2 together contained 77.32% of all nodes; in contrast, in the non-content network, NI1 and NI2 together contained 55.27% of all nodes (see Table 6). This means compared with the non-content network, the content-related network was more dominated by interactions that happened around the instructors. Second, average node degree and average edge weight for all nodes in CI1 and CI2 are both greater than those in the corresponding modules in the non-content network, indicating that community members in the content-related modules interacted with more learners and had more repeated interactions with the same learners.

Table 6 Network measures of instructor modules (ties based on Limited Copresence)

	CI1	NI1	CI2	NI2
# of nodes (% in network)	184 (54.93%)	168 (43.19%)	75 (22.39%)	47 (12.08%)
# of edges (% in network)	400 (47.17%)	315 (43.51%)	105 (12.38%)	55 (7.6%)
Graph density	0.024	0.022	0.038	0.051
Avg node degree (SD)	4.35 (11.06)	3.75 (11.18)	2.8 (7.56)	2.34 (6.03)
Avg edge weight (SD)	2.23 (3.21)	2.11 (2.48)	1.83 (1.72)	1.20 (0.44)
Instructor Degree	145	144	67	43

Inter-instructor comparisons in the content-related network

First, CI1 has 2.5 times of nodes and 4 times of edges as CI2 (see Table 6), indicating that the community formed around u1 contained more interactions than that formed around u417. In addition, the degree of u1 was more than twice that of u417 (see Table 6), indicating that u1 interacted directly with many more learners than u417.

Second, the module graphs showed that in both modules, a large proportion of learners were only connected to the instructor and not any learner in the module (see Figure 5), indicating that a large number of learners only had interactions with the instructor. Comparatively, there were more interconnections among learners in CI1 than in CI2. This is echoed by the finding that the average node degree in CI1 was one and a half times of that in CI2. These characteristics indicate participants in CI1 interacted with more other learners than those in CI2.

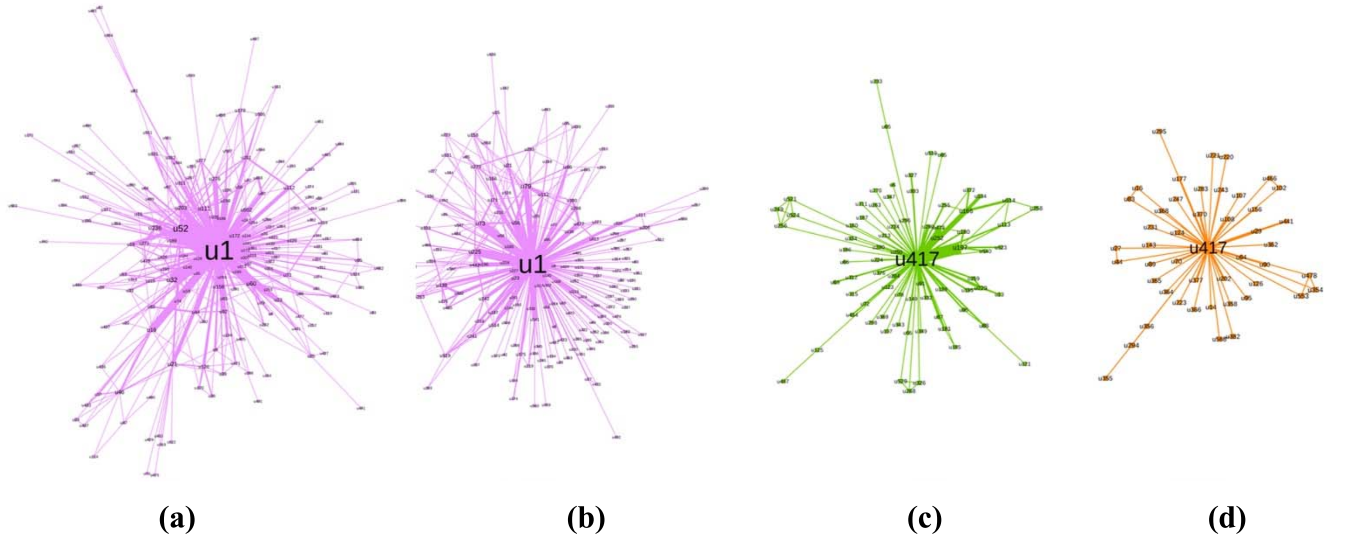


Fig. 5 Comparison of **a** CI1; **b** NI1; **c** CI2; **d** NI2 (ties based on Limited Copresence)

Third, the average edge weight in CI1 is somewhat higher than that in CI2 (see Table 6), indicating that participants in the former module had more repeated interactions with the same other learners than those in the latter module. It is noteworthy that the average edge weight of the content-related learner-only module CL1 is much higher (6-8 times) as that of the two content-related instructor modules (see Table 2 and Table 6), indicating that learners in CL1 had dramatically more repeated interactions with the same learners than those in the instructor modules did.

Moreover, the posts made by the two instructors revealed two distinct characteristics of instructors' forum activities. First, the two instructors had distinct posting patterns. Instructor u1 made 353 posts in 240 content-related threads, including 216 reply posts to thread starting posts and 137 reply posts to subthread starting posts or other posts in subthreads. This diverse posting pattern showed that the instructor not only revisited the threads that he/she has participated in, but also commented on other learners' replies to learner-initiated threads. In contrast, instructor u417 made 121 replies to 119 content-related thread starting posts and never made any reply to subthreads. This indicates that u417 only addressed the subset of all leaners that initiated threads and not learners who participated only as responders.

Second, the instructors appeared to use distinct intervention techniques and elements of social presence in posting. In many cases, instructor u1 used diverse techniques to encourage and help learners to work out the answer or solution themselves. For example, he/she often gave hints instead of answering a question outright:

> ***U1****: Looks like you are making great progress! ... You are correct about dealing with categorical data, and the observations are certainly correlated! Think about it again using the hint and let me know if you have any other questions?"*

Instructor u1 also used leading questions to help learners work through problems and figure out solutions themselves:

> ***U1****: That is correct - Nice! So how would you use this to solve the question?*

In addition, u1 used a variety of social presence indicators such as greetings and addressing learner's by name in his/her messages:

U1*: Hello [learner' s name], do you think you could clarify your example data set a little more?*

In contrast, instructor u417 tended to provide straightforward answers or instructions to learners' questions and used social presence cues infrequently. For example:

> ***U417****: A bell shape is not necessary. You could have a 'bimodal' distribution (with two distinct peaks in the distribution) where the two groups do not follow a bell shape.*

6 GENERAL DISCUSSION

6.1 Differences in content-related and non-content network activities

In this study, we found the content-related and non-content social networks showed distinct characteristics; participants' interactions in the two network also showed different characteristics. First, participants in the content-related network interacted with more people and developed stronger ties with the same people than those in the non-content network did. A similar finding was reported by Gillani et al. (2014) in the networks formed in two successive offerings of a business MOOC using the Total Copresence tie definition. In that study the proportion of one-off ties was lowest in the sub-forum for discussing course materials and highest in the one that contained mostly conversations related to technical support and gratitude expression. Second, learners appeared to use more social presence cues such as greetings and addressing each other by name in the content-related discussions. This is interesting and unexpected given that many of the non-content threads were considered "social" in nature. We hypothesize that repeated interactions resulted in stronger social connections between learners, which encouraged the use of social presence cues and subsequent additional participation, resulting in more opportunities to interact with additional other people. This process may have helped the participants to develop a sense of community, which could lead to higher persistence. This may help to explain previous findings about the association between different kinds of social connections and retention in learning environments. For instance, Kuh (2002) found that social connections developed in academic activities help college student retention rate while other types of connections are not helpful; Gillani et al. (2014) found the content-related sub-forum showed the highest rate of participant persistence among all sub-forums. Moreover, leaner's participation in content-related and non-content discussions has also been found to associate differently with learning outcome. For instance, Romero et al. (2013) found that learner's outdegree centrality calculated based on content-related forum contributions and the features of this type of contributions are more useful for predicting learning outcome in a computer science course. Wang et al. (2015) found that the quantity of students' on-task discourse is a significant predictor of their learning gains in an introductory psychology MOOC. Although no causal relations can be made based on our findings, the association between retention and content-related interactions is a worthwhile direction for future study.

Differences in participant's behavior in the content-related and non-content networks may be explained in two ways. First, the differences may have to do with who choose to participation in which network. We found the two networks were participated by largely distinct people. A similar observation was made by Gillani and Eynon (2014) who found that in a business MOOC where different sub-forums were designated for content-related and non-content discussions, cross-sub-forum participation was rare. It is possible that the differences we observed in the two networks were to some degree due to intrinsic differences in the participants. In addition, when zooming in on learners who gained high degree centrality in the content-related and non-content networks, we found that they were largely distinct people. This may to some extent help to explain the findings from Dowell et al. (2015) that higher learning performance and higher social centrality are associated with different linguistic features, as it is highly possible that the two groups were formed by different people. Second, the nature of discussions in the two contexts may have affected participant's behavior. A big proportion of the content-related discussions involved more complicated communication topics, such as problem solving or understanding abstract / complicated concepts, which often took several rounds of talks and needed the contributions from multiple learners. We hypothesize that in this process the learners not only interacted with the same people repeatedly, but were exposed to opportunities to interact with additional other peers. In contrast, the non-content discussions usually involved exchange of factual information, which often could be accomplished straightforwardly with input from a smaller number of learners. In fact, the cross-network difference observed in the activities of the three learners who gained high degree centrality in both network exemplified the influence of context. Consequently, learners' different behaviors in the two contexts led to the difference in the social networks: in the content-related network, learners developed stronger ties with a greater number of people.

Implications

First, the finding that content-related discussions resulted in deeper and wider learner interactions can inform MOOC course design. For instance, the courses that aim to promote learner connections can include learning activities that can promote content-related interactions among learners, such as open-ended discussion and problems that can be addressed from multiple perspectives.

Second, our findings reiterate the value in content-based partitioning in MOOC research. For one thing, as the two kinds of networks are participated by largely different people, content-based partitioning helps researchers to identify more accurately the individuals or subgroups that they are interested in. For instance, for studies that want to identify community TAs (Papadopoulos et al. 2014) and influential learners for disseminating the content of instructor's intervention (Jiang et al. 2015), it is useful to distinguish the learners who gained high social status in the content-related network from those who are only prominent in the non-content network. Similarly, the marginalized learners in the two networks may have different needs for support and intervention (Brugha and Restoule 2016). For another, as content-related and non-content activities were found to associate differently with other variables in learning, such as retention and learning outcome (Kuh 2002; Romero et al. 2013; Wang et al. 2015), content-based partitioning enables researchers to select the kind of variables that are most useful for their research purpose.

6.2 Tie definitions

In this study we improved the understanding of five tie definitions and their influence on the resultant networks and interpretation. First, we found that network formation was more robust to tie definition than expected. Except for Total Copresence, the networks constructed using the other four definitions did not differ dramatically in structure or properties. Interestingly, although Direct Reply+Star and Limited Copresence are based on different assumptions about speaking and listening activities, the resultant networks for this dataset were quite similar. This may in part be accounted for by the two definitions' adjacency on the definition continuum. It is also worth investigating to what extent this is dependent on the technical limitations of the MOOC to three levels of threading as well as the actual patterns of post distribution in this particular data set. This might usefully be done with synthetically generated data designed to have various different properties. In addition, it would also be interesting to examine how different limits for the maximum number of shared posts impacts the structure of the resulting Limited Copresence network. Second, Direct Reply and Total Copresence, the two most used tie definitions in SNA studies on MOOC discussion forums (Gillani and Eynon 2014; Gillani et al. 2014; Jiang et al. 2014; Joksimović et al. 2016; Kellogg et al. 2014) were found to produce dramatically different network in terms of structure and properties. This may help to partially explain the discrepancies we found in MOOC literature regarding results using SNA method. Third, it was found Total Copresence is particularly sensitive to superthreads, reflected by the inflated node degree and deflated edge weight. This characteristic of Total Copresence may qualify the findings from studies using this tie definition. For instance, Jiang et al. (2014) built Total Copresence networks for the discussion forums in an algebra MOOC and a finance MOOC. It was found that learner's degree and betweenness were positively correlated with learning performance only in the algebra course but not the other. Based on the current finding, if one of the two datasets contained superthreads, learners' degree would have been inflated and thus could have potentially distorted the results. With that said, as the balloon-shaped structure formed due to superthread is easy to identify in network graph, Total Copresence is useful for detecting superthreads in the dataset and thus can be a useful additional lens for examining social networks.

Implications: aligning tie definition with the nature of interaction and relationship

Our study revealed the importance of choosing tie definitions appropriate for the nature of network relationship to be examined. Many kinds of relationships can be studied using social network analysis, such as similarity, social relations, interactions and flows (Borgatti et al. 2009). The network relationships in MOOC discussion forums has been studied from different perspectives, such as interest (Poquet et al. 2016), reciprocity (Kellogg et al. 2014), and information transmission (Jiang et al. 2015). The nature of the relationship to be examined should not only determine the guiding theories for the study, but also the appropriate tie definition. For instance, when studying reciprocity in posting behavior, Direct Reply can be a legitimate option as social connection is considered as one participant "speaking" to another (Kellogg et al. 2014); however, for connections between forum participants formed through both "speaking" and "listening", Limited Copresence is more appropriate as it takes in to account both kinds of activities. Moreover, when choosing a tie definition, it is also important to explicitly acknowledge its strength and limitation. For instance, Direct Reply traces visible interactions among participants and has a low risk for constructing false connections, however researchers have noted that the limited number of posting levels in MOOC forums (such as EdX and Coursera) may impact the resulted structural information extracted from threads (Chandrasekaran et al. 2015; Rossi and Gnawali 2014). In addition, this definition is also more vulnerable to data error caused by participant's behavior such as liberal use of the posting structure. In contrast, Total Copresence constructs learners' connections based on their presence in the same threaded discussion and thus can capture "invisible" connections, thus it is suitable for constructing interest networks. However, the risk of overestimating connections should be accounted for due to the fact as the discussion evolves, it may change direction or involve new topics that not all participants in the thread are interested in (Stump et al. 2013). At a higher level, our study on the effects of tie definition is a showcase that highlights the important of many micro decisions researchers make. Similar examples of the micro decisions include our choice of methods for investigating modules in the networks, such as using Louvain method for calculating modularity and examining modules as non-overlapping clusters. Such operational decisions contain presumptions about the interactions and relationships in the learning environment as well as the data structure. It is important that researchers reflect on and acknowledge the effects of such decisions on the findings and interpreting of the findings so as to inform future studies.

6.3 Instructor's intervention

Instructor's intervention in discussion forum is perceived as an important factor in online learning by instructors and learners (Dennen et al. 2007; Hew 2015). However the empirical effect of instructor intervention on learning remains controversial: some studies found it positively associated with retention, quantity of student participation, and learning outcome (Baker 2010; Dennen 2005) while others reported null or negative associations (Mazzolini and Maddison 2007; Tomkin and Charlevoix 2014).

In this study, we adopted social analytic perspective and found two important findings about instructor's activities. First, interactions in the examined discussion forum were mostly instructor-centered: the two instructor modules in the content-related network (CI1 and CI2) contained 77% of the total participants and 60% of the connections in the network, indicating that most of the interactions happened around the instructors. Second, the instructors directly interacted with many learners, but their intervention did not lead to active learner-learner interaction. Both CI1 and CI2 networks showed hub-and-spoke structure with the instructor being the only central node and little learner-learner interactions shown in the network. In addition, the connections formed in both modules were weak. Similar hub-and-spoke structure was reported in a case study on the discussion forum of a small cohort course on computer science in Brooks et al. (2014). In that study, the instructor posted topics about weekly readings and required the students to post their understandings. The resultant social network for the discussion forum showed that interactions were mostly centered around the instructor and little student-student interaction took place.

The phenomenon that instructor intervention did not always result in vigorous student activities has been reported in literature on discussion forums in both MOOC and non-MOOC environments. Mazzolini and Maddison (2007) studied online discussion forums in postgraduate astronomy courses for lifelong learners. They found that greater number of instructor postings and instructor initiated threads did not lead to more posts from students and were associated with shorter conversations (indexed by the number of posts per thread). Tomkin and Charlevoix (2014) conducted an a/b test on the discussion forum of a MOOC on sustainability in which the learners were assigned to two groups: the intervention groups received instructor intervention including forum feedbacks and weekly summary of highlight forum activities and contributors emailed to participants. To complete the course, the students could choose to be assessed in one of three ways: final project, weekly quizzes, or forum participation. It was found that instructor intervention did not cause more students to participation in the forum, view the videos, or take the quizzes; it did result in higher completion rate in the forum-based assessment intervention subgroup, but not the other two subgroups. These two studies show the time and effort spent on intervention do not necessarily result in more learner participation.

However, it would be too assertive to claim that instructor should abstain from intervening. In our study, we found that the two instructors both intervened but differently, and their intervention behaviors are associated with difference in network structure and learner behavior. The first intervention difference is instructor u417 only responded to thread starting posts while u1 responded to both initial posts and replying posts. The second difference was found in the techniques used by the instructors: u417 mostly provided straightforward answers to learner's questions and seldom used social presence cues; u1 used diversified intervention techniques, such as providing hints and asking leading questions, and frequently used social presence cues. As reported in Section 5.2.3, the u1 module contained more learner-learner interactions than the u417 module; learner interacted with more peers and had more repeated interactions with the same peers in the u1 module than those in the u417 module. It is likely that u1's intervention style encouraged and resulted in more opportunities for the learners to prolong their engagement in the discussions and interact with other peers, as well as to develop a stronger sense of community; while in contrast, by providing straightforward answers, u417 largely eliminated the need for learners to linger on (which actually can be considered as efficient intervention in certain learning environments). This may help to explain the findings from Mazzolini and Maddison (2007). Although the Astronomy program in the study advocates constructivist approaches and most of the instructors indicated that they intervened by combining answers with follow-up questions, manual coding revealed that nearly only 12% of the instructors' intervention posts were combination of answer and follow-up question while 68% were just answers. Given this, it is possible that instructors' intervention in certain cases might have removed the necessity for learners to have more conversations and subsequent participation. Similarly, the instructor intervention reported in Tomkin & Charlevoix (2014) mainly involved positive feedbacks to high quality postings, which does not necessarily create the need for more subsequent content-related conversations among learners; the weekly email that summarized highlighted discussions and participants might have motivated some learners to participate, but at the same time could have demotivated some learners to go to the forum as they could get the essence of the discussion from the more authoritative (summarized by the instructional team) and less time-consuming source.

Implications: informing and supporting instructor intervention

These findings and observations lead us to ask how research can better inform and support instructor intervention in real-time, an important goal of learning analytics (Ifenthaler et al. 2014). A promising approach is to provide comprehensible learning analytic results to instructors to help them improve understanding of their intervention behavior and the effects. Literature shows that instructors do not always make accurate judgement about their teaching behavior. For instance, in Mazzolini and Maddison (2007) most instructors reported performing constructivist intervention but only 12% of the interventions fell into this category. The benefit of providing learning analytic results to instructors was showcased in the case study in Brooks et al. (2014): when informed by the social analytic results that the discussion forum activities were high instructor centric, the instructor made pedagogical adjustment and successfully promoted learner-learner interaction. Similarly, by showing instructors the association between intervention styles and learner-learner interactions found in our study, we can help the instructors to make sense of their intervention behaviors and inform their pedagogical decision. With that said, it is also important to adopt appropriate evaluation framework for instructor intervention in different learning contexts. Literature shows that instructor intervention has been examined from many perspectives, such as retention (Hernández-García et al. 2015), quantity and quality of student activities (Kellogg et al. 2014), as well as student satisfaction and learning outcome (Dennen 2005; Tomkin and Charlevoix 2014). As a primary role of educational discussion forum is to facilitate learning, we argue the fundamental criteria for instructor intervention should address how effectively the intervention supports learners to achieve the expected learning outcome, while factors such as the quantity of instructor's or learner's forum contributions are not necessarily informative measures of intervention quality. In addition, as MOOCs in different disciplines can vary substantially on assumptions about learning and teaching (Ross et al. 2014), it is important that studies align evaluation of instructor intervention with the type of learning and learning activities valued in the specific subject and discipline.

6.4 Finding community in the crowd

MOOC discussion forum's potential in realizing effective collaborative learning and peer support is controversial. While some MOOC studies claim discussion forum can be leveraged to foster social networks and facilitate peer-supported learning (Kellogg et al. 2014), others posit that MOOC forum participants are dispersed crowds rather than communities of learners, evidenced by the findings that in the modularized and short-lived discussion groups, learners do not move from peripheral participation to playing important roles such as promoting collaboration among the peers (Gillani and Eynon 2014).

In this study, we found that learner-learner interaction is overweighed by instructor-centered interaction. However, it is noteworthy that the learner-only module we found in the content-related network (CL1) is of rudimental community characteristics. First, members in this module had common interest in participating in learning-related collaboration. As revealed by the qualitative examination of their post texts, not only that their communications were mostly content-related, the participants also felt they benefited from the collaborative discussions. Second, the social connections in this module were more cohesive than other modules. For one thing, CL1 members had stronger ties with a larger number of peers. Moreover, they used social presence cues frequently, indicating a well-developed sense of community. Third, they interacted with more peers and formed stronger inter-learner connections than learners in other modules did. Moreover, a small group of learners played core role and made the connections in CL1 more robust than other modules. The social network of CL1 was a web structure that can be divided into two parts: one is the core formed by several central learners who were connected via strong ties with other central members and had interactions with a large number of peers; the other part is the periphery made up of other learners who had weaker ties with a smaller number of peers. A big difference between this core-peripheral web structure and the hub-and-spoke modules (e.g. CI1 and CI2) is that in CL1 multiple strongly connected learners form the central part in the collaborative interactions in the network. Similar findings were reported in Kellogg et al. (2014). When examining two MOOCs on digital learning and mathematics learning, Kellogg et al. (2014) removed postings from the instructional members to simulate learner-only social networks. They found in both MOOCs there were several individuals with a disproportionate number of ties compared to their peers. In addition, using CORR algorithm, they divided the network into a densely connected subgroup of core actors and a peripheral group of sparsely connected learners. They found that in the two courses respectively 13% and 21% of the learners belong to the core group. This group of learners played an important role of making the module more robust and sustainable in that even if one or two central members stop participating other members of the module can still keep their social connections with each other (Gillani et al. 2014). Future study can further investigate how such distributed core structures develop by analyzing the overtime evolvement of the social network.

In the scope of this study, we were not able to investigate in-depth the central learners' influence on learning outcome and community building. However learners with similar characteristics identified in the literature may shed light for future study. For instance, the central learners we identified have similarities with the superposters identified by Huang et al. (2014). They examined the learners whose total number of forum contributions ranked among top 5% in 44 MOOC courses and found that these superposters have better performance, respond more quickly to discussions, and gain more upvotes than regular learners. Moreover, in contrast to the findings of Mazzolini and Maddison (2007) and Tomkin and Charlevoix (2014) that frequent instructor intervention was associated with less contributions from the learners, Huang et al. (2014) found superposters' prolific behaviors did not suppress their peers' contributions, but were associated with greater quantity and better quality (indexed by the number of upvotes discussions received) of activities in the forums. The central learners we identified in CL1 were all among the top 5% of learners in term of post quantity. This provides an additional perspective to study these learners' influence on peer learners and their role in community development.

Implication: facilitating learner community

Participants in MOOCs are often seen as big and dispersed crowds (Gillani et al. 2014), but the small group of strongly connected learners identified in our study indicates small communities can form in MOOC discussion forums. Moreover, the finding that learner-learner interaction associates differently with instructor intervention and central learners' activities indicates the potential in enhancing peer support in MOOC discussion forum through facilitating learning community. Different to connecting people who ask questions with the answers (Agrawal et al. 2015; Yang et al. 2014), cultivating a learner community provides a route to ongoing peer support and social connections. Various strategies can be adopted to cultivate learner community. For instance, the courses can include learning activities that involves group collaboration and sharing of personal experience (Shackelford and Maxwell 2012); the instructor can model and encourage the use of discussion facilitation techniques, such as asking leading questions and using social presence cues (Rovai 2000); learners can be asked to volunteer as peer facilitators for the units that they are interested in (Zydney 2014); the instructor can monitor the dynamics of learner interaction and phase out his/her facilitation once the learner community is robust enough to sustain (Brooks et al. 2014).

7 LIMITATIONS AND FUTURE STUDY

First, our analysis was conducted on networks constructed with little refinement of node and edges. SNA literatures have suggested the usefulness of network refining techniques such as filtrating nodes and edges and assigning differentiated weight to edges (Dowell et al. 2015; Gillani et al. 2014). Future research could experiment filtering out edges with extremely low weight to screen out the one-off connections between forum users in Total Copresence to see if this reveals different characteristics of interactions. In addition, future research could use differentiated edge weights to reflect the special status of some nodes. For instance the edge between a thread starter and a reply post could be weighted as 1 while that between a reply post and it reply as 0.5. In addition, edge weight could also be used to differentiate the strength of connections in threads of different sizes by treating the total strength of connection in a thread as one and computing edge weight among nodes through dividing total strength by the number of nodes in it.

Second, although this study used SNA method in combination with content analysis, network analysis was focused mainly on the structural properties revealed from visualizations. Future studies could make use of network property indices in combination with other analysis methods to investigate the association between content-related network properties and learning. For instance, ERGM could be used to investigate the association of node-level properties such as degree centrality, closeness centrality, and betweenness centrality with learning performance. Topic modelling could be used to investigate the interest of different modules in content-related network.

Third, the scope of this study is limited to examining end-of-course networks in a single MOOC. Future study could make dynamic network analysis to investigate the longitudinal formation of content-related network and its association with pedagogical influences overtime. Moreover, data from other courses and other domains could be used to test generalizability of the findings.

8 CONCLUSIONS

Online discussion forums are commonly provided in MOOCs as a central medium for learning support and interaction. To bring them to the full potential, it is important to understand the interactions that happen in discussion forums and their relationship to learning. SNA is a useful method for investigating interactions in online learning environment (Cho et al. 2007). When applied to MOOCs, it is important that SNA methods take into account the complexity and distinct characteristics of MOOC discussion activities.

This study examines the influence of content-based network partitioning and tie definition on social network structures and interpretation for MOOC discussion forums. Results showed the social networks partitioned based on content-relatedness of the activities have distinct characteristics. Network properties were less sensitive to differences in tie definition than expected with the exception of Total Copresence, which showed distinct characteristics useful for detecting inflated social status due to "superthread" initiation. The partitioned networks are distinct in terms of participating members, the complexity of activities, participant's behavioral patterns, and their interaction techniques. This work makes both theoretical and practical contributions to large-scale online discussions. Theoretically, the study enhances understanding of the tie definitions' influence on social network construction and the importance of content-based partitioning of social networks. The findings about the characteristics of learner's and instructor's activities and their effects on learning interactions can inform both course design and intervention decisions.

REFERENCES

Agrawal, A., Venkatraman, J., Leonard, S., & Paepcke, A. (2015). YouEDU: Addressing confusion in MOOC discussion forums by recommending instructional video clips. In *Proceedings of the 8th International Conference on Education Data Mining* (pp. 297-304). ACM, New York, NY, USA.

Baker, C. (2010). The impact of instructor immediacy and presence for online student affective learning, cognition, and motivation. *The Journal of Educators Online, 7*(1). doi:10.9743/JEO.2010.1.2.

Borgatti, S., Mehra, A., Brass, D., & Labianca, G. (2009). Network analysis in the social sciences. *Science, 323*(5916), 892-895.

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom research into edX's first MOOC. *Research & Practice in Assessment,* 8, 13-25.

Brugha, M., & Restoule, J. P. (2016). Examining the learning networks of a MOOC. In ElAtia, S., Ipperciel, D., & Zaïane, O. R (Eds.), *Data Mining and Learning Analytics: Applications in Educational Research* (pp. 121-138). Wiley.

Chandrasekaran, M. K., Kan, M. Y., Tan, B. C., & Ragupathi, K. (2015). Learning instructor intervention from MOOC forums: Early results and issues. In *Proceedings of the 8th International Conference on Education Data Mining* (pp. 218-225). ACM, New York, NY, USA.

Cho, H., Gay, G., Davidson, B., & Ingraffea, A. (2007). Social networks, communication styles, and learning performance in a CSCL community. *Computers and Education, 49*(2), 309-329.

Cui, Y., & Wise, A. F. (2015). Identifying content-related threads in MOOC discussion forums. In *Proceedings of the 2nd ACM Conference on Learning @ scale* (pp. 299-303). ACM, New York, NY, USA. doi:10.1145/2724660.2728679.

Cui, Y., Wise, A. F., & Jin, W.Q. (in press). Humans and machines together: Improving characterization of large scale online discussions through dynamic interrelated post and thread categorization (DIPTiC).

Dawson, S. (2010). 'Seeing' the learning community: An exploration of the development of a resource for monitoring online student networking. *British Journal of Educational Technology*, *41*(5), 736-752.

Dennen, V. P. (2005). From message posting to learning dialogues: Factors affecting learner participation in asynchronous discussion. *Distance Education*, *26*(1), 127-148.

Dennen, V. P., Aubteen Darabi, A., & Smith, L. J. (2007). Instructor–learner interaction in online courses: The relative perceived importance of particular instructor actions on performance and satisfaction. *Distance education*, *28*(1), 65-79.

Dowell, N., Skrypnyk, O., Joksimović, S., Graesser, A. C., Dawson, S., Gašević, D., Vries, P. d., Hennis, T., & Kovanović, V. (2015). Modeling learners' social centrality and performance through language and discourse. In *Proceedings of the 8th International Conference on Educational Data Mining* (pp. 250-257). ACM, New York, NY, USA.

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

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

Gruzd, A.A., & Haythornthwaite, C. (2008). Automated discovery and analysis of social networks from threaded discussions. In *Proceedings of the International Network of Social Network Analysts 2008, St. Pete Beach* (St. Pete Beach, USA.2008). Retrieved September 28, 2016, from http://hdl.handle.net/10150/105081.

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 and Knowledge* (pp. 198-207) ACM New York, NY, USA. doi:10.1145/2883851.2883924.

Hernández-García, Á., González-González, I., Jiménez-Zarco, A. I., & Chaparro-Peláez, J. (2015). Applying social learning analytics to message boards in online distance learning: A case study. *Computers in Human Behavior*, *47*, 68-80.

Hew, K. F. (2015). Student perceptions of peer versus instructor facilitation of asynchronous online discussions: Further findings from three cases. *Instructional Science*, *43*(1), 19-38.

Hew, K.F., & Cheung, W.S. (2014). Students' and instructors' use of massive open online courses (MOOCs): motivations and challenges. *Educational Research Review*, *12*, 45-58. doi:10.1016/j.edurev.2014.05.001.

Huang, J., Dasgupta, A., Ghosh, A., Manning, J., & Sanders, M. (2014). Superposter behavior in MOOC forums. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 117-126). ACM.

Ifenthaler, D., Adcock, A. B., Erlandson, B. E., Gosper, M., Greiff, S., & Pirnay-Dummer, P. (2014). Challenges for education in a connected world: Digital learning, data rich environments, and computer-based assessment - Introduction to the Inaugural Special Issue of Technology, Knowledge and Learning. *Technology, knowledge and learning*, *19*(1-2), 121.

Jacobsen, D. Y. (2017). Dropping out or dropping in? A Connectivist approach to understanding participants' strategies in an e-learning MOOC pilot. *Technology, Knowledge and Learning*, doi:10.1007/s10758-017-9298-z.

Jiang, S., Fitzhugh, S. M., and Warschauer, M. (2014). Social positioning and performance in MOOCs. In *Proceedings of Graph-Based Educational Data Mining Workshop at the 7th International Conference on Educational Data Mining* (pp. 55-58). CEUR-WS.

Jiang, Z., Zhang, Y., Liu, C., & Li, X. (2015). Influence analysis by heterogeneous network in MOOC forums: What can we discover?. In *Proceedings of the 8th International Conference on Education Data Mining* (pp. 242-249). ACM, New York, NY, USA.

Joksimović, S., Manataki, A., Gašević, D., Dawson, S., Kovanović, V., & De Kereki, I. F. (2016). Translating network position into performance: Importance of centrality in different network configurations. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge* (pp. 314-323). ACM New York, NY, USA. doi:10.1145/2883851.2883928.

Kellogg, S., Booth, S., & Oliver, K. (2014). A social network perspective on peer supported learning in MOOCs for educators. *The International Review of Research in Open and Distributed Learning*, 15, 5. doi:10.19173/irrodl.v15i5.1852.

Khalil, H., & Ebner, M. (2013). “How satisfied are you with your MOOC?” - a research study on interaction in huge online courses. In *Proceedings of EdMedia 2013* (pp. 830-839). AACE.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (pp. 170-179). ACM New York, NY, USA. doi: 0.1145/2460296.2460330.

Kuh, G. (2002). From promise to progress: How colleges and universities are using student engagement results to improve collegiate quality. *National Survey of Student Engagement Annual Report*. Bloomington, IN: Indiana University.

Mazzolini, M., & Maddison, S. (2007). When to jump in: The role of the instructor in online discussion forums. *Computers & Education*, *49*(2), 193-213.

McGuire, R. (2013). Building a sense of community in MOOCs. *Campus Technology*, *26*(12), 31-33.

Oshima, J., Oshima, R., & Matsuzawa, Y. (2012). Knowledge building discourse explorer: A social network analysis application for knowledge building discourse. *Educational Technology Research and Development*, *60*(5), 903-921.

Papadopoulos, K., Sritanyaratana, L., & Klemmer, S. R. (2014). Community TAs scale high-touch learning, provide student-staff brokering, and build esprit de corps. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 163-164). ACM.

Poquet, L., & Dawson, S. (2016). Untangling MOOC learner networks. In *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (pp. 208-212). ACM New York, NY, USA. doi:10.1145/2883851.2883919.

Rabbany, R., Takaffoli, M., & Zaïane, O. R. (2011). Analyzing participation of students in online courses using social network analysis techniques. In Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero, C., & Stamper, J. (Eds.), *Proceedings of the 4th International Conference on Educational Data Mining* (pp. 21-30). EDM.

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458-472.

Rosé, C. P., & 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.

Ross, J., Sinclair, C., Knox, J., & Macleod, H. (2014). Teacher experiences and academic identity: The missing components of MOOC pedagogy. *Journal of Online Learning and Teaching*, *10*(1), 57-69.

Rossi, L.A., & Gnawali, O. (2014). Language independent analysis and classification of discussion threads in Coursera MOOC forums. In *Proceedings of 2014 IEEE 15th International Conference on Information Reuse and Integration* (pp. 654-661). IEEE. doi:10.1109/IRI.2014.7051952.

Rovai, A. P. (2000). Building and sustaining community in asynchronous learning networks. *The Internet and higher education*, *3*(4), 285-297.

Santos, J.L., Klerkx, J., Duval, E., Gago, D., & Rodríguez, L. (2014). Success, activity and drop-outs in MOOCs: An exploratory study on the UNED COMA courses. In *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (pp. 98-102). ACM New York, NY, USA. doi: 10.1145/2567574.2567627.

Shackelford, J. L., & Maxwell, M. (2012). Sense of community in graduate online education: Contribution of learner to learner interaction. *The International Review of Research in Open and Distributed Learning*, *13*(4), 228-249.

Stump, G. S., DeBoer, J., Whittinghill, J., & Breslow, L. (2013). Development of a framework to classify MOOC discussion forum posts: Methodology and challenges. In *Proceedings of NIPS 2013 Workshop on Data Driven Education* (pp. 1-20). NIPS Foundation.

Tomkin, J. H., & Charlevoix, D. (2014). Do professors matter?: Using an a/b test to evaluate the impact of instructor involvement on MOOC student outcomes. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 71-78). ACM.

Trentin, G. (2000). The quality-interactivity relationship in distance education. *Educational Technology, 40*(1), 17-27.

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains. In *Proceedings of the 8th International Conference on Educational Data Mining*. International Educational Data Mining Society.

Wise, A. F., Chang, J., Duffy, T. M., & del Valle, R. (2004). The effects of teacher social presence on student satisfaction, engagement, and learning. *Journal of Educational Computing Research, 31*(3), 247-271.

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 and Knowledge* (pp. 188-197) ACM New York, NY, USA. doi: 10.1145/2883851.2883916.

Wise, A. F., Cui, Y., Jin, W.Q., & Vytasek, J. (2017). Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling. *The Internet and Higher Education*, 32, 11-28.

Wise, A., Chang, J., Duffy, T., & Del Valle, R. (2004). The effects of teacher social presence on student satisfaction, engagement, and learning. *Journal of Educational Computing Research, 31*(3), 247-271. doi:10.2190/V0LB-1M37-RNR8-Y2U1

Yang, D., Piergallini, M., Howley, I., and Rose, C. 2014. Forum thread recommendation for massive open online courses. In *Proceedings of the 7th International Conference on Educational Data Mining* (pp. , 257-260). ACM, New York, NY, USA.

Yusof, N., & Rahman, A. A. (2009). Students' interactions in online asynchronous discussion forum: A social network analysis. In *Proceedings of 2009 International Conference on Education Technology and Computer* (pp. 25-29). IEEE.

Zhu, M., Bergner, Y., Zhang, Y., Baker, R., Wang, Y., & Paquette, L. (2016). Longitudinal engagement, performance, and social connectivity: a MOOC case study using exponential random graph models. In *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (pp. 223-230). ACM, New York, NY, USA. doi: 10.1145/288385.

Zydney, J. (2014). Strategies for creating a community of inquiry through online asynchronous discussions. *Journal of online learning and teaching*, *10*(1), 153-165