Computers in Human Behavior xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

SENS: Network analytics to combine social and cognitive perspectives of collaborative learning

Dragan Gašević[a,b,*], Srećko Joksimović[c], Brendan R. Eagan[d], David Williamson Shaffer[d,e]

[a] Faculty of Education, Monash University, Clayton, VIC, Australia
[b] School of Informatics, The University of Edinburgh, Edinburgh, Scotland, United Kingdom
[c] Teaching Innovation Unit and School of Education, University of South Australia, Adelaide, SA, Australia
[d] School of Education, University of Wisconsin-Madison, Madison, WI, USA
[e] Aalborg University, Copenhagen, Denmark

ARTICLE INFO

Keywords:
Social network analysis
Epistemic network analysis
Collaborative problem solving
Learning analytics

ABSTRACT

In this paper, we propose a novel approach to the analysis of collaborative learning. The approach posits that different dimensions of collaborative learning emerging from social ties and content analysis of discourse can be modeled as networks. As such, the combination of social network analysis (SNA) and epistemic network analysis (ENA) analysis can detect information about a learner's enactment of what the literature on collaborative learning has described as a role: an ensemble of cognitive and social dimensions that is marked by interacting with the appropriate people about appropriate content. The proposed approach is named social epistemic network signature (SENS) and is defined as a combination of these two complementary network analytic techniques. The proposed SENS approach is examined on data produced in collaborative activities performed in a massive open online course (MOOC) delivered via a major MOOC platform. The results of a study conducted on a data set collected in a MOOC suggest SNA and ENA produce complementary results which can i) explain collaboration processes that shaped the creation of social ties and that were associated with different network roles; ii) describe differences between low and high performing groups of learners; and iii) show how combined properties derived from SNA and ENA predict academic performance.

1. Introduction

Collaborative learning has attracted much attention due to the growing importance of working in teams and networks to solve complex problems of contemporary life, work, and society (National Research Council (US), 2011). There are many accounts that demonstrate benefits of effective collaboration and social communication on productivity in the workplace, engineering and software development, mission control in aviation, and interdisciplinary research among scientists. These trends also highlighted a need for the integration of activities that promote collaborative learning in formal curricula. The development of technology has also increased opportunities for collaborative learning and attracted attention of researchers from the fields such as computer supported collaborative learning, online learning, psychology, sociology, and computer science. Benefits of technology-mediated or computer-supported collaborative learning can include examples such as increased sense of community (Dawson, 2008), creative potential (Dawson, Tan, & McWilliam, 2011), critical thinking (Garrison, Anderson, & Archer, 2001), academic performance (Joksimović et al., 2016), and integration into learning communities (Haythornthwaite, 2002).

Successful collaboration requires a complex mix of interrelated processes. According several sources (Griffin, McGaw, & Care, 2012; OECD, 2013), two key dimensions of collaboration can be recognized – social and cognitive. The cognitive domain is typically related to the existing literature on problem solving and self-regulated learning (Griffin et al., 2012; Winne & Hadwin, 1998; Zimmerman & Schunk, 2011) and includes task regulation and knowledge building. The social domain is focused on the processes necessary for productive collaboration (OECD, 2013). For example, Hesse, Care, Buder, Sassenberg, and Griffin (2015) posit that social domain include participation, perspective taking, and social regulation.

We posit that the two main dimensions of collaborative learning, as commonly theorized in the literature and used in formal assessment, can be modeled as networks. The existing literature presents many approaches to the analysis of collaborative learning (see Section 2 for

* Corresponding author. Monash University, Faculty of Education; 19 Ancora Imparo Way, Clayton, VIC 3800, Australia.
E-mail addresses: dragan.gasevic@monash.edu (D. Gašević), srecko.joksimovic@unisa.edu.au (S. Joksimović), beagan@wisc.edu (B.R. Eagan), dws@education.wisc.edu (D.W. Shaffer).

https://doi.org/10.1016/j.chb.2018.07.003
Received 15 February 2017; Received in revised form 28 May 2018; Accepted 2 July 2018

an overview). These approaches do not sufficiently capture the complexity of interaction between interrelated cognitive and social dimensions that emerge from social ties and collaborative discourse. To overcome limitations of existing approaches, this paper suggests a novel network analytic approach.[1] The social dimension of collaboration can be framed as a network structure that is formed through the interaction among actors of collaboration. That is, social processes are expressed through actor to actor interactions (Scott, 2012). Similarly, the cognitive dimension, typically extracted with content analysis of collaborative discourse, can be framed as a networked phenomenon, since knowledge and demonstration of competency do not function in isolation of one another (Shaffer et al., 2009). For instance, the understanding of a concept or performing a skill is dependent on other forms of understanding and the use of other skills and processes. The prerequisite processes, skills, and knowledge can be used as a form of network relationships.

To comprehensively capture these ensembles, we suggest a combined use of two well-established analytic techniques in the field of learning analytics (Dawson, Gašević, Siemens, & Joksimović, 2014; Gašević, Dawson, & Siemens, 2015) – ENA and SNA – to define SENS of collaborative learning. Although both techniques have been used for analysis of collaborative learning, their combined use has been limited and has not been proposed on the level of network analytic methods. This paper demonstrates an approach how the two methods can complement each other in the analysis of collaborative learning. The approach is illustrated on a case study conducted in the context of a MOOC. The main contributions of the combined use of SNA and ENA illustrated in the study reported are the prediction of

- the structure of social network ties with collaborative discourse,
- the students' role in group communications with collaborative discourse,
- collaborative discourse based on identification of high and low-achieving communities of learners, and
- academic performance.

The proposed SENS approach contributes to both fields of learning analytics and cognitive computing. On the one hand, learning analytics is often referred to as a bricolage field of research and practice (Gašević, Kovanović, & Joksimović, 2017) that borrows theories, techniques, and approaches from a wide range of related fields such as learning sciences, machine learning, and human computer interaction (Dawson et al., 2014). In this context, the paper provides a potential blue print how established analytic techniques can systematically be combined to address questions significance for learning analytics. Specifically, the paper offers a theory-informed, practical, and holistic approach to analytics of collaborative learning. On the other hand, cognitive computing aims to provide software and hardware systems that simulate human brain and advance human decision-making (Modha et al., 2011). This paper provides a concrete example of a holistic analytic approach that models collaborative learning by guiding integration of analytic methods with learning theory.

2. Background

2.1. Collaborative learning

There is a wide range of conceptualizations of collaborative learning (Roschelle & Teasley, 1995). In computer support-collaborative learning (CSCL), Stahl (2004) suggests that the process of knowledge construction and the development of shared understanding involves i) articulating and sharing ideas in statements and ii) reading and discussing ideas shared by others. An individual contributes to collaborative knowledge building by interacting with others to share their personal understanding. In a collaborative dialogue, actors need to be responsive to each other to create a cohesive conversation. By building on or challenging previous contributions of others, actors clarify and negotiate ideas to build collaborative knowledge as a group and internalize that knowledge as individuals. In online and distance education, collaborative learning is mostly conceptualized through the lenses of social constructivism with the emphasis on meaningful interaction (Garrison, 2011; Woo & Reeves, 2007).

The SENS method proposed in this paper is primarily related to the notion of emergent roles in the CSCL literature (Strijbos & Weinberger, 2010; Strijbos & de Laat, 2010). According to Strijbos and de Laat (2010), emergent roles are “roles that emerge spontaneously or are negotiated spontaneously by group members without interference by the teacher or researcher” (p. 496). According to Strijbos & de Laat, emergent roles are eventually determined by contributions made by group members and by the ways how group members participated in interaction with their peers. Strijbos & de Laat also distinguish between three levels of roles: micro – i) role is related to a specific task focused on a collaborative process or product; ii) meso – a role involves a pattern of several tasks focused on process, product and their combinations; and iii) macro – a role is determined by a stance composed of an individual's participation strategy. In this paper, we are primarily concerned with macro-level roles. Examples of emergent roles are reported by different authors mostly based on the participation patterns in online discussions such as communicative learners, silent learners, non-participants, superficial listeners and intermittent talkers, and concentrated listeners and integrative talkers (Hammond, 1999; Wise, Speer, Marbouti, & Hsiao, 2013).

SENS can also be used for the analysis of scripted roles (Strijbos & de Laat, 2010), although that is less relevant to the study reported in this paper. The literature in both CSCL and online learning also emphasizes the importance of the developmental nature of collaborative learning. The emerging script theory of guidance (Fischer, Kollar, Stegmann, & Wecker, 2013) suggests that high level collaboration can be promoted through scripting (as a scaffolding mechanism commonly used in CSCL), role assignment, and task structure. The communities of inquiry model (Garrison & Arbaugh, 2007) also highlights the role of teaching presence as critical to establish a supportive group climate that in order to reach high levels of knowledge construction through collaboration.

For analysis of collaborative learning, there is a need to consider critical aspects of group activities, including those articulated in different theoretical models of collaborative learning (Fischer et al., 2013; Garrison & Arbaugh, 2007; Stahl, 2004), their associated processes (Arastoopour, Shaffer, Swiecki, Ruis, & Chesler, 2016; Garrison et al., 2001; Gunawardena, Lowe, & Anderson, 1997; Sobocinski, Malmberg, & Järvelä, 2017; Tan, Caleon, Jonathan, & Koh, 2014; Weinberger, Stegmann, Fischer, & Mandl, 2007), general indicators of participation (Kovanović, Gašević, Joksimović, Hatala, & Adesope, 2015; Wise et al., 2013), and patterns of social interaction (Marcos-García, Martínez-Monés, & Dimitriadis, 2015; Martinez, Dimitriades, Rubia, Gomez, & de la Fuente, 2003). The literature on CSCL generally recognizes several key dimensions of the analysis of collaborative learning, including, capturing the influence of context, representing timing, calculating when timing matters, capturing community knowledge building, and designing to support and capture collaboration (Teasley, 2011). Commonly used analysis methods originate from social sciences, statistics, data mining, sequential analysis, text mining, information visualization, and SNA (Puntambekar, Erkens, & Hmelo-Silve, 2011). Of specific importance for the method propose in this paper are content analysis (De Wever, Schellens, Valcke, & Van Keer, 2006; Strijbos, Martens, Prins, & Jochems, 2006) and analysis of social structures that shape

[1] It is however important to note that the approach proposed suggests that cognitive and other relevant dimensions (e.g., affective and motivational) of collaborative learning are typically extracted using content analysis of collaborative discourse. Therefore, to suggest that SENS can also be used to account for those other dimensions (not just cognitive), we refer to content of collaborative discourse (or simply collaborative discourse) throughout the paper.

collaboration (Martinez et al., 2003; de Laat, Lally, Lipponen, & Simons, 2007).

2.2. Content analysis

Content analysis (De Wever et al., 2006; Strijbos et al., 2006) is a commonly used technique for analysis of transcripts of discussions generated in collaborative learning. Discourse is a rich source of information about collaboration and is the most comprehensive source of data, in addition to self-reports, about cognitive, metacognitive, motivational and affective dimensions of student engagement (Azevedo, 2015). Content analysis uses a coding scheme that can be either commonly used or recently proposed for the measurement of new features of collaboration. Examples of commonly used coding schemes for collaborative learning are those for argumentation (Weinberger et al., 2007), cognitive presence (Garrison et al., 2001), and knowledge construction (Gunawardena et al., 1997). Examples of recently proposed schemes for emerging topics in collaborative learning are those used for the measurement of engineering design thinking (Arastoopour et al., 2016), collective creativity (Tan et al., 2014), and self-regulated learning in group activities (Sobocinski et al., 2017). With the recent progress in text mining and natural language processing, approaches to automating the concept analysis process of collaborative learning have been proposed. While some authors propose to automate the coding process based on existing coding schemes such as Kovanović et al. (2016) for cognitive presence in online discussions, others use unsupervised text analysis methods for the detection of emerging themes in discourse as Yang, Wen, Kumar, Xing, and Rose (2014) did for the analysis of discussions in MOOCs.

The results of the coding process, as part of content analysis, are typically aggregated on either individual or group levels to understand to what extent certain process are being activated, measures according to a specific coding scheme. The results of coding process are also used in other (statistical) analysis as independent or dependent variables. For example, Schellens, Keer, Wever, and Valcke (2007) use the coding results, based on the knowledge construction scheme (Gunawardena et al., 1997), as dependent variables to determine the extent to which role assignment in group activities affected knowledge construction levels. Joksimović, Gašević, Kovanović, Riecke, and Hatala (2015) use codes of social presence as predictors of academic achievement. The literature on collaborative learning recommends using analytical methods that account for the nested nature of data, which can emerge from content analysis (Cress, 2008).

Methods commonly used for analysis of the results produced by content analysis cannot offer sufficient insights into the ways how processes represented by codes are interlinked, how links between processes can qualitatively be interpreted, and whether there are significant differences in the process links among different groups of collaborators. Rather than thinking of collaboration as a collection of processes, the CSCL literature suggests that being an effective collaborator means performing well in a role (Dillenbourg, Järvelä, & Fischer, 2009; Fischer et al., 2013). A role is an ensemble of different dimensions that assume interacting with the *right* people at the *right* times and in the *right* ways. Content analysis is also insufficient to analyze what kind of roles actors in collaborative learning played and social structures that emerge from collaboration. Therefore, this paper proposes a networked analysis approach to address the above limitations.

2.3. Epistemic network analysis

ENA is a network analysis technique that analyses logfile data and other records of individual and collaborative learning (Nash & Shaffer, 2012; Rupp, Gushta, Mislevy, & Shaffer, 2010; Rupp et al., 2010; Shaffer & Gee, 2012; Shaffer et al., 2009). ENA is an operationalization of the learning science theory of epistemic frames (Shaffer, 2004, 2006, 2008), which looks at expertise in complex domains not as a set of isolated processes, skills, and knowledge, but as a network of connections among knowledge, skills, values, and decision-making processes. ENA makes use of categories of action, communication, cognition, and other relevant features of group interaction that can be characterized with appropriate coding schemes used in content analysis as some of those mentioned in Section 2.2 (Shaffer, Collier, & Ruis, 2016). Specifically, an epistemic network is constructed by making use of the codes assigned to difference elements of collaborative discourse where nodes of the network are the codes themselves. Connections among nodes are established based on occurrence of the codes within a relevant unit of analysis (e.g., a discussion message or several messages).

In epistemic networks, the weights of the connections among nodes (i.e., the association structure between key elements of the domain) are a central point of interest. For example, ENA takes advantage of these features by using computational and statistical techniques to compare the salient properties of networks, including networks generated by different teams or by teams at different points in time, teams in different spatial locations, or teams engaged in different activities. Critically, though, these salient properties are not just modeled in terms of the general structure of the networks such as change in density. ENA also extracts properties relevant to the actual content of the network and traces of individual and collaborative processes activated during collaboration actives. The exiting literature showed that ENA can be used to a wide range of analyses about collaborative learning such as identification of critical patterns of interaction in expert and novice teams, successful and unsuccessful teams and individuals, and assessment of engineering design thinking in collaborative learning (Arastoopour et al., 2016; Chesler et al., 2015; Nash & Shaffer, 2012; Orrill, Shaffer, & Burke, 2013; Rupp, Gushta, et al., 2010; Rupp, Sweet, et al., 2010; Shaffer, 2013; Shaffer et al., 2009; Shaffer & Gee, 2012).

While ENA offers powerful mechanisms to analyze collaboration discourse and links among relevant features of collaborative learning, it does not have strong methods that can be used for analysis of roles actors occupy in collaborative activities, social structures formed in collaborative learning, sub-communities created during collaboration, and principles on which ties between actors of collaboration were formed (Morris, Handcock, & Hunter, 2008). These are precisely the types of insights that can be obtained by the use of SNA.

2.4. Social network analysis

The use of SNA has played a prominent role in the learning sciences and learning analytics for analysis of collaborative learning (Dawson et al., 2014). According to Haythornthwaite (1996), SNA provides insights into the following dimensions: cohesion – how a network is interlinked; centralization – the extent to which the network is depended on a small number of actors; structural equivalence – whether there are actors that have similar roles in the network; prominence – popularity of an actor in the network; range – the extent to which an actor is connected to others in the network; and brokerage – the extent to which an actor connects different parts of the network that are otherwise disconnected. Some authors show that some of these SNA dimensions are positively associated with outcomes such as academic performance, creative potential, and sense of community (Dawson, 2008; Dawson et al., 2011; Dowell et al., 2015; Gašević, Zouaq, & Janzen, 2013).

The dimensions provided by SNA have also been adopted by CSCL researchers to propose approaches to characterizations of roles played by different actors in collaborative learning activities (Capuano, Mangione, Mazzoni, Miranda, & Orciuoli, 2014; Marcos-García et al., 2015). For example, Marcos-García et al. (2015) suggest that the measures of SNA should be organized into “a qualitative combination of the relative ranges of values (high, medium, low and null)” (p. 338) of the SNA dimensions. They go on and offer an example of a student-animator (show initiative, set the pulse of collaboration, and encourage co-learners) who needs to have medium to high level of range,

promotes the development of their network with a high value of density, medium to high value for prominence, and low to medium for brokerage.

The emergence of novel statistical SNA methods (Morris et al., 2008) offer additional insights about emerging clusters of collaborators and social processes that shaped network formation. SNA can allow for the identification of communities of actors with mutually higher ties with each other than the rest of the network (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). Latent community structures are evident in many social networks emerging from interactions in online settings (Wang, Wang, Yu, & Zhang, 2015). With statistical SNA (Goodreau, Kitts, & Morris, 2009; Morris et al., 2008), it is possible get insights in a wide range of processes about the network formation such as a) the propensity of reciprocal ties which can be reflective of the equitable involvement in knowledge construction, b) the propensity of strong ties which is relevant for the opportunities afforded (e.g., high performance) by occupying certain roles in the network, and or c) the propensity to work with other actors who share similar characteristics (see Section 4.2 for more details).

Although SNA can reveal the role individual actors played in different collaborative tasks, the factors that define their positions are less clear from the use of SNA only. One critical reason for this lack of clarity is the fact that network dynamics are influenced by the type of information shared, content communicated, and social and cognitive processes activated among network actors while collaborating. Therefore, in this paper, we argue that social network models devoid of content are doomed to fail because group interactions are never “content neutral” (Shaffer, 2014). This is of importance for the measurement of relevant processes, as introduced in Section 2.2, whose assessment is rather limited by looking at social networks only. That is, we need to understand, simultaneously, the social network of the group, epistemic networks that guide the action of the individuals in the group, and the social network by which that action is accomplished. Thus, we argue that a critical step in creating a technique to monitor and support group performance is the development of network analysis techniques for assessing collaborative learning.

3. Social-epistemic network signature

A fundamental claim in this paper is that it is essential to consider both the semantic and conceptual content of what gets said during social interactions and the patterns of communication — of who talks to whom in a social network by bearing in mind the types of network ties. We argue that it is difficult to evaluate the quality of collaborative learning by examining either who is talking to whom without knowing what they are talking about, or by modeling what is being said without tracing the interactive contributions of the individuals involved. Although there have been previous proposals to combine content analysis and SNA for the study of collaborative learning (Martinez et al., 2003; de Laat et al., 2007), there has been limited previous work that combined the perspectives on the level of network analytic techniques. Therefore, we propose a combined use of SNA and ENA – named as social and epistemic network signature (SENS) – to measure collaborative learning.

SENS uses content of collaborative discourse and social ties to measure collaborative learning by combining the SNA and ENA techniques. In this paper, we demonstrate how SNA and ENA can be used jointly for the measurement of collaborative learning by addressing four research questions.

3.1. Research questions

Much of the existing literature suggests that learners who share similar characteristics are likely to collaborate with each other. For example, the use of statistical SNA methods such as exponential random graph models (ERGMs) have demonstrated the significant homophilic effects on network formation of a number of factors such gender, academic achievement, and nationality on the creation of network ties (Joksimović et al., 2016; Kellogg, Booth, & Oliver, 2014). Insights obtained through the analysis of homophilic ties allow for understanding the factors that shape the choices of collaborators. Although highly important for understanding of collaborative learning, there is a dearth of research that looks at the extent to which content of collaborative discourse is predictive of the factors that shape the structure of social network ties.

> RQ 1. Does an ENA analysis of the content of collaborative discourse predict the structure of social network ties? (That is, does what students talk about influence who they talk with?)

The study of roles individuals play in collaborative learning are well recognized in the current literature (Dillenbourg et al., 2009; Fischer et al., 2013). Much of the existing literature looks at the analysis of roles played by individuals with the use of either social network (Haythornthwaite, 1996; Marcos-García et al., 2015; Martinez et al., 2003) or content analysis (De Wever et al., 2006; Strijbos et al., 2006; Wise, Saghafian, & Padmanabhan, 2012). Although the benefits of roles played in social networks are well studied to predict academic performance (Gašević et al., 2013; Joksimović et al., 2016), it is much less understood whether and if so, to what extent the content students talk about is predictive of the students' roles in social networks.

> RQ 2. Does an ENA analysis of individual students' discourse predict students' centrality in the social network? (That is, is the content of students' talk related to their role in group communications?)

For example, the importance of homophilic ties based on academic performance (i.e., high achieving students connect with each other and likewise low achieving students with each other) is often reported (Joksimović et al., 2015; Kellogg et al., 2014). The presence of homophilic ties can be ultimately lead to the formation of communities/cliques composed mostly by either low or high performing students. However, it is much less understand to what extent network grouping of students is predictive of the content the students talk about.

> RQ 3. Does an SNA analysis of high- and low-performing sub-communities of students predict differences in the content of students' discourse? (That is, do groups of closely-linked students talk about different things depending on how well they do in class?)

Contributions of both roles played by students in social networks and processes identified by content analysis of collaborative discourse are reported in the literature to be associated with academic performance (Gašević et al., 2013; Joksimović et al., 2015, 2016; Schellens et al., 2007). However, it is much less understood whether and if so, to what extent a combination of factors identified by SNA and content analysis is more predictive than either SNA and ENA alone.

> RQ 4. Does a combined SNA and ENA model predict student outcomes better than an SNA or ENA model alone? (That is, is SENS an effective approach to modeling student performance?)

3.2. Methodological approach

Fig. 1 shows the methodological steps proposed for the combined use of SNA and ENA to form SENS. Initially, collaboration traces are used from a collaborative learning environment to extract collaboration discourse and social ties. Before ENA is applied, a content analysis is performed on the collaboration discourse to identify processes and themes pertinent to the objectives of a study. A social network is created based on the social ties before SNA is performed to identify roles actors played, network processes, and communities. SNA and ENA are

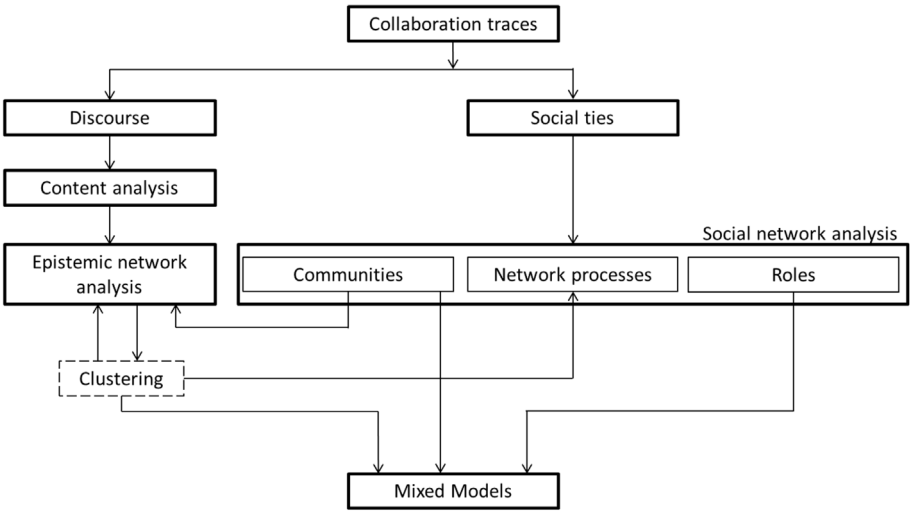

Fig. 1. The methodological steps followed for the combined use of SNA and ENA to form SENS.

combined to form SENS based on the outcomes of either of the analysis methods. The outputs of both SNA and ENA can also be additionally analyzed with methods that aim to identify subgroups of learners (e.g., cluster analysis) before further SNA, ENA, or SENS analysis is performed.

Specifically, research question 1 can be addressed by using the output of ENA to produce information about specific epistemic network of each network actor which statistical SNA can use to analyze social processes that influenced network formation. For example, this can be used to test whether actors who had similar epistemic networks had a statistically significant propensity to connect with each other.

Research question 2 can be addressed by using outputs produced by both SNA and ENA to i) test the extent to which SENS can explain or predict performance of network actors, ii) compare whether performance of learners with different roles and epistemic networks also statistically differs, and iii) check whether there is a significant association between epistemic networks and network positions. To enable the combined use of SNA and ENA in such cases, an external analysis method should be used such as regression analysis, mixed models, or statistical tests for comparisons of groups.

Research question 3 can be addressed by using a modularity analysis algorithm of SNA to identify relevant communities emerging from collaboration activities. Such communities can then be compared with ENA to check whether there are statistically significant differences in epistemic networks. In some cases, communities can be combined to test the extent to which certain properties of communities are associated with their epistemic networks. For example, epistemic networks of communities of high performing students and low performing learners can be compared.

Research question 4 can be addressed by using statistical analyzed commonly used for prediction of academic performance such as multiple linear regression or mixed models. The independent variables can be measures produced by both SNA and ENA, while the dependent variable can be student grades or other measures of academic performance.

4. Methods

The study was conducted with the sole purpose to provide a proof of concept for SENS that showcases the complementary use of the two analytical techniques – SNA and ENA – for the measurement of collaborative learning through a possible instantiation of the methodology outlined in Fig. 1. As such, the study was not conducted with the aim to have a generalization power and test specific theory-informed hypotheses. For analysis, we first performed SNA and ENA and then, the results of the two analyses are combined in SENS to address the four research questions.

4.1. Data

In this study, we analyzed forum discussions from the third iteration of a “Critical Thinking in Global Challenges” MOOC, delivered on the Coursera platform. The course aimed at developing and enhancing students' ability to think critically and develop justified arguments in the context of the global challenges facing today's society. The content of the course was delivered over five weeks, consisting of video lectures, quizzes and exercises. An additional final exam was provided for those students who wanted to obtain a certificate of accomplishment.

Initially, more than 61,000 students enrolled in the course. However, 29,466 students engaged with at least one course activity within the course, whereas 2660 students obtained a certificate of accomplishment. Additionally, 1989 students posted at least once to the course discussion forum. In total, we analyzed 6158 ($M = 3.10$, $SD = 7.07$) student generated posts across 1677 ($M = 3.67$, $SD = 9.54$) discussion threads. Moreover, we also extracted demographic data that were used to support SNA and ENA. Demographic data include students' age, gender, whether English is their first language or not, area of employment (e.g., Accountancy, banking and finance; Energy and utilities; and IT and information services), academic level (e.g., College, Undergraduate university, and Postgraduate university), intent (e.g., to get a certificate; learn new things; and improve my career options), as well as final course grade and achievement level (i.e., none – did not obtain certificate; normal – obtained certificate). Out of 1989 students who posted to a discussion forum, 1151 submitted a course survey. In order to obtain a more complete dataset, we searched for student data across surveys delivered in other courses provided by the same institution. This resulted in additional 63 entries, producing 1214 complete data records. The list of observed demographic variables was compiled from the standard Coursera questionnaire prior the analysis during which students' names were removed from the dataset and their unique identifiers were anonymized.

4.2. Social network analysis

4.2.1. Extraction of social ties

To investigate social processes, roles actors played in collaborative learning activities, and communities formed, we extracted a directed weighted graph to represent interactions occurring within a discussion forum. Specifically, we relied on the most commonly applied approach that considers each message as being directed to the previous one

(Dowell et al., 2015; de Laat et al., 2007). For example, if within a given thread, author A2 replied to a message posted by author A1, we would include a directed edge A2- > A1. Further, the Coursera platform also allows for posting comments to created posts; therefore if A3 posted a comment on A2's post, we would include another edge (i.e., A3- > A2). In case that A1, for example replied again to A2's post, we would increase the weight of the A1- > A2 edge by one.

Given the large number of small disconnected components (i.e., individual students or pairs of students) in the final graph, we decided to perform all the SNA on the largest connected component of the original network (Schluter, 2014; Whitelock, Field, Pulman, Richardson, & Van Labeke, 2014). Such a comprised graph included 79% of all the nodes (i.e., students – N_{nodes} = 1564) and 99% of edges (N_{edges} = 3303) from the original network.

4.2.2. Communities

We identified *communities* from the social graph by using the Louvain method (Blondel et al., 2008) that groups nodes in a network according to the strength of relationships between them (Steinhaeuser & Chawla, 2008). The community detection method determines *clusters* of nodes (i.e., course participants) that are "more densely connected to each other than with the rest of the network" (Deritei et al., 2014, p. 2). The extracted communities are then compared based on academic performance and the roles students played in the networks.

The community detection analysis resulted in 19 distinctive groups of students, where the two largest clusters comprised 8.70 (i.e., 136 participants) and 8.31 (i.e., 130 participants) percent of the graph, respectively. The five smallest groups included between 41 and 59 course participants, accounting for from 2.62 to 3.96 percent of the analyzed network. Despite an attempt to characterize communities with respect to students' age, gender, academic level, intent to complete the course, or experience with online learning, the most striking difference between emerging social groups was evident in the average final course grade. Fig. 2 shows that communities 3, 11, and 17 were composed of highest performing students on average, whereas communities 6, 8, and 12 included lowest performing students on average.

4.2.3. Roles

To examine roles actors played in the network and within communities, we relied on most commonly used SNA measures that capture various aspects of network structural centrality – weighted degree, closeness, and betweenness centrality (Wasserman, 1994). Weighted degree centrality accounts for the weight of edges a node has in the network. Closeness centrality explains the potential for control over communication in a network, measuring a distance of a given node to all other nodes in the network. Specifically, closeness centrality measures the potential of a node to connect easily with other nodes in the network. Finally, betweenness centrality is also related to the potential of control over communication; however, betweenness instead shows expected benefits for the nodes that bridge two or more distinct parts of the network (Wasserman, 1994). These measures are also proposed for role analysis in the CSCL literature (Marcos-García et al., 2015). The descriptive statistics of the centrality measures for the participants included in the entire graph used in the analysis (i.e., the largest connected component) are presented in Table 1.

The descriptive statistics of the network centrality measures for the high and low performing communities are given in Table 2.

4.2.4. Network processes

We applied methods of statistical network analysis to reveal social processes that framed the formation of the social network extracted from online discussions. In so doing, we relied on ERGMs that allowed for examining network formation mechanisms at the dyadic and triadic level (Morris et al., 2008). At the dyadic level, we aimed to investigate the effects of **selective mixing**, **reciprocity**, **popularity**, and **expansiveness**, as some of the most commonly used network statistics (Goodreau et al., 2009; Joksimović et al., 2016; Kellogg et al., 2014; Stepanyan, Borau, & Ullrich, 2010).

Selective mixing represents an effect of a students' propensity to interact with peers who share similar (demographic) characteristics (Goodreau et al., 2009; Morris et al., 2008). Here we examined two selective mixing dynamic types – *uniform homophily* (i.e., mixing across attribute categories), and *differential homophily* (i.e., propensity across individual categories within the observed attribute) (Goodreau et al.,

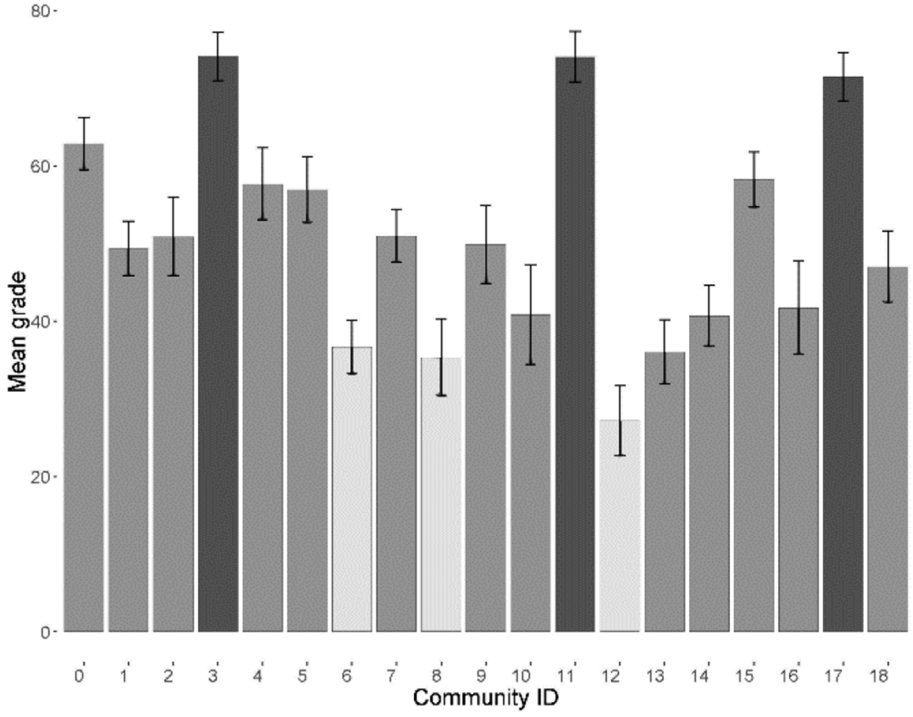

Fig. 2. Mean final course grade per network community extracted from the analyzed network.

Table 1
Summary statistics of network properties.

Graph/Cluster	Weighted degree		Betweenness centrality		Closeness centrality	
	Mean (SD)	Median (25–75%)	Mean (SD)	Median (25–75%)	Mean (SD)	Median (25–75%)
LCC graph	5.95 (15.08)	2.00 (1.00–5.00)	2098.00 (10,162.05)	0.00 (0.00–347.70)	3.86 (2.49)	4.42 (1.00–5.71)

Table 2
Summary statistics of the network centrality measures for the high (C.3, C.11, and C.17) and low (C.6, C.8, and C.12) achieving communities.

Community	Weighted degree		Betweenness centrality		Closeness centrality	
	Mean (SD)	Median (25–75%)	Mean (SD)	Median (25–75%)	Mean (SD)	Median (25–75%)
C.3	4.88 (5.33)	2.00 (1.00–4.00)	1227.89 (2908.36)	0.00 (0.00–700.00)	3.45 (2.59)	4.32 (1.00–5.71)
C.11	3.29 (9.01)	1.00 (1.00–2.00)	670.48 (3606.12)	0.00 (0.00–0.00)	4.23 (2.82)	7.04 (1.58–7.05)
C.17	3.08 (5.09)	2.00 (1.00–3.00)	129.02 (566.21)	0.00 (0.00–0.00)	1.79 (1.99)	1.00 (1.00–2.25)
C.6	5.38 (10.83)	2.00 (1.00–5.00)	1674.16 (6843.79)	0.00 (0.00–123.98)	3.65 (2.33)	4.56 (1.00–5.53)
C.8	4.59 (9.42)	2.00 (1.00–4.00)	1458.08 (6231.68)	0.00 (0.00–4.00)	4.24 (2.15)	5.01 (3.80–6.05)
C.12	3.42 (17.07)	2.00 (1.00–2.00)	1146.89 (9822.64)	0.00 (0.00–0.00)	6.13 (1.13)	6.44 (6.44–6.44)

2009). Specifically, we used uniform homophily to assess the propensity that students connect within the same gender, academic level, intent to complete the course, and based on the fact whether English is their first language or not (Goodreau et al., 2009; Kellogg et al., 2014; Lusher, Koskinen, & Robins, 2012). Differential homophily, on the other hand, was used to examine potential differences in students' propensity to interact with their peers across different epistemic groups – i.e., clusters of students defined based on their discourse. We also tested the main effect of the achievement level (i.e., obtained certificate or not) on the propensity to form ties (i.e., nodefactor).

Reciprocity reflects students' tendency to form reciprocal ties and cluster together (Morris et al., 2008). In networks that represent friendship formation in schools, for example, where students identify their peers as friends, such networks are considered cross-validated and potentially stronger than nominations characterized as one-way (Carolan, 2013). In the context of interactions within online discussion forums, such property should allow for revealing whether students tend to continue interaction with peers who replied to their posts.

Popularity and *expansiveness*, on the other hand, are network statistics that represent formations based on the students' tendency to "attract" a significant number of replies to their posts (i.e., to measure popularity) or whether there are students who are very active in replying to their peers' posts (i.e., to measure expansiveness). These network processes are also associated with the influence of roles played by learners as introduced in the CSCL literature and outlined in Section 2.4.

At the triadic level, we examined effects of **cyclical ties**, **triadic closure** and **Simmelian ties** formation. Our model included *cyclical ties* statistics in order to control for the potentially hierarchical nature of the network. We also modeled Simmelian ties (Krackhardt, 1999) to assess whether the network is characterized by formation of cliques of students that tend to form super-strong ties – i.e., to interact more often within a clique than with other peers in the network. Testing whether there is a significant propensity of Simmelian ties is important in order to assure whether benefits steepening from particular network roles can be hypothesized (Joksimović et al., 2016). As well-established in the SNA literature, benefits of occupying certain network positions occur in the presence of weak ties (Granovetter, 1982).

4.3. Content analysis

To prepare our data for ENA, a content analysis needed to be performed. In this study, we used an automated approach to content analysis and applied the Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). LDA is a probabilistic topic modeling technique, commonly applied in social sciences and humanities (Weingart & Meeks, 2012), that allows for extraction of prominent themes from a collection of text documents. Specifically, examining a co-occurrence of words in the documents analyzed, LDA identifies groups of words that are commonly used together and could potentially represent semantically different theme (Blei et al., 2003).

Prior to conducting the analysis we preprocessed the data by removing stopwords (i.e., commonly used words such as 'a', 'the, or 'be'), numbers, punctuation, and removing short words (i.e., less than three characters long). LDA also requires a number of topics to be specified in advance. There are various methods applied to identify an optimal number of topics and the selection of an appropriate approach depends on the research goals and the size of a dataset (Blei et al., 2003). Regardless the fact that we analyzed relatively large number of user generated posts (i.e., 6168), an average length of those posts was around 125 words (SD = 166.61), resulting in a manageable corpus that allowed us to evaluate a large number of LDA models in order to reveal an optimal number of topics.

The above procedure with the use of LDA identified 12 topics, including *Content* topics (i.e., topics about substance of the lessons in the course) such as *c.Energy*, *c.Population*, *c.Diseases*, *c.Course_related_interests*, *c.Critical_thinking*, and *c.Communities_social_groups* and *Process* topics (i.e., topics about general procedures of the course) such as *p.Course_expectations* and *p.Course_related_interests*.

4.4. Epistemic network analysis

Individual students were the unit of analysis to produce an epistemic network of each participant. We analyzed posts and comments within each thread using a moving window of three posts. That is, we looked for semantic connections between each post or comment and the previous three contributions in its thread, where semantic connection was defined as the co-presence of any pair of topics t_m and t_n from topics $t_{k=1\text{-}12}$ identified in the content analysis with LDA (see Section 4.3). We chose a moving window of three posts based on a qualitative assessment of the window size necessary to capture the largest percentage of meaningful connections made by students; more information on this approach are provided by Shaffer (2017). Based on the semantic connections in each window, we created NxN adjacency matrices for each post or comment that described the semantic connections the post or topic contained. N in our study was 12 as the LDA analysis – see Section 4.3 – identified 12 topics. That is, N corresponds to the number of topics or codes used in content analysis. The adjacency matrices were

then accumulated for each participant in the dataset into a cumulative adjacency matrix. The cumulative adjacency matrix for each participant was represented as a point in a high-dimensional space by taking each cell in the matrix as a dimension in the space. We used singular value decomposition (svd) to project the points into a lower dimensional space of orthogonal dimensions that maximized variance accounted for in the data. We used the first six dimensions ($svd_{i=1.6}$) as descriptors for study participants in the ENA space based on the amount of variance explained by the addition of subsequent dimensions. According to Shaffer et al. (2016), it is up to the researcher to assess which two dimensions to use in presenting an ENA space. It is also up to the researcher to decide if and how the dimensions can be interpreted as a high-fidelity model, or "fair sample" of the phenomenon of interest.

To produce two-dimensional EN graphs for each participant, we computed the location p_i for the epistemic network of each participant using svd_1 and svd_2. We positioned the nodes $N_{j=1\text{-}12}$ of the EN—which correspond to the topics $t_{k=1\text{-}12}$—and for each participant calculated the position of centroid c_i of his or her resulting network graph. We then optimized the positions of $N_{j=1\text{-}12}$ so as to minimize Σ_i (p_i–c_i). The position of the nodes in the svd_1 x svd_2 space were then used to interpret the significance of the spatial location of participants' networks—that is, the positions of the topics were used to explain the dimensions of the ENA space. The details about ENA are provided by Shaffer et al. (2016).

The first dimension of the ENA space (svd_1 or *Process or* X axis the for networks in Fig. 4) explained 46% of the variance in the ENA space: Participants with high *Process* scores were focused more on course processes and procedures. They made more connections between *process related topics* (i.e., *p.Course_expectations and p.Course_related_interests*) than to other content related topics (i.e., *c.Energy*, *c.Population*, *c.Climate_change*, *c.Humaity_challenges*, and *c.Critical_thinking*). Those with low *Process* scores made more connections between purely content related topics. The second dimension of the ENA space (svd_2 or *Content* or y-axis in the four networks in Fig. 4) explained 18% of the variance in the ENA space. Participants with high *Content* scores made more connections with *c.Energy*, *c.Population*, *c.Diseases*, and *c.Course_related_interests* to other topics. Participants with low *Content* scores made more connections between *p.Course_expectations*, *c.Critical_thinking*, and *c.Communities_social_groups.* Centroids of the individual epistemic networks of each student in the sample are plotted in Fig. 4.

4.5. Cluster analysis

A cluster analysis was performed on the svd values produced by ENA to prepare the outcomes of ENA to a) identify groups of students based on their epistemic networks, and b) check whether learners had a tendency to link with other learners who belonged to the same epistemic group as asked in research question 1, c) compare possible differences in roles played in the networks among students who belong to different epistemic network groups as asked in research question 2. Specifically, we used k-means cluster analysis (Jain, 2010) to identify relevant clusters of students based on the properties of their discourse characterized by the svd dimensions as calculated by ENA. The selection of the final number of clusters was based on two commonly used methods for extraction of the optimal number of clusters: comparing sum of squared errors (Peeples, 2011) and model-based methods for clustering (Fraley, Raftery, & others, 2007). The examination of scree plots and the solution obtained based on the probability models resulted in four clusters (Cluster 1 N = 220, Cluster 2 N = 305, Cluster 3 N = 272, and Cluster 4 N = 206) among participants who contributed at least 5 messages during the entire course; plus, a fifth cluster of participants (Cluster 5 N = 561) who posted fewer than 5 messages.

The cluster assignments of students – after performing K-means clustering on the svd values – are shown with different colors in Fig. 3. In total, cluster analysis identified five clusters. The epistemic networks for the four clusters are shown in Fig. 4. The epistemic network of the learners in ENA cluster 1 focused most on two process related topics (*p.Course_related_interests* and *p.Course_expectations*) and their links with one content topic (*p.Critical_thinking*), which was also a broad and generic topic considering the data were from a course on critical thinking. Learners in ENA cluster 2 focused on a process topic (*p.Course_related_interests*), two content topics (*c.Energy* and *c.Population*), and a generic topic (*c.Human_challenges*) considering that the topic was on critical thinking and global challenges. However, there was no dominant link(s) between any pair of topics in the epistemic network of the learners in ENA cluster 2. Learners from ENA cluster 3 mostly focused on content related topics and the most dominant ones included *c.Energy*, *c.Population*, and *c.Humanity_challenges*. They however did not have any dominant links between some of the topics. Finally, learners in ENA cluster 4 primarily focused on content topics *c.Climate_change* and its links with process topics *p.Course_related_interests* and *p.Course_expectation*.

Fig. 3 shows only four dots (i.e., centroids of the epistemic networks) for the participants in Cluster 5. This happened as we merged all those participants who posted less than 5 messages into a single "meta student" in the ENA. One of the dots represents all those who posted less than 5 messages and those other few show students whose contribution was low enough to be grouped within the Cluster 5.

4.6. SENS

To address research question 1, we used the ERG models. Akaike information criterion (AICc)[2] (Akaike, 2011) and Bayesian information criterion (BIC) (Neath & Cavanaugh, 2012) were used to compare the fitness of the two ERG models: i) one based on the SNA properties and demographics/academic attributes; and ii) another one that additionally included ENA cluster assignment as parameters of selective mixing. The estimates of SNA and ENA properties in the ERG models were used to compare their effects on the formation of social ties.

To address research question 2, we used tabularly and graphically compared the ENA cluster assignments on the three SNA network centrality measures. Non-parametric tests to determine whether the ENA clusters were significantly different based on their SNA centrality measures (weighted degree, closeness, and betweenness) were used. The non-parametric tests were used as the data were not normally distributed.

To address research question 3, we compared epistemic networks of the communities identified as described in Section 4.2.2. Specifically, we analyzed three communities that included highest performing learners and three communities that included lowest performing students. The epistemic networks of these two group communities are then compared by using ENA methods – including the interpretation of the differences in the epistemic networks of the two groups of communities, statistical differences of the centroids of the two groups of communities on both svd^1 and svd^2 dimensions with t-tests, and the subtraction of the epistemic networks of the two groups of communities.

Finally, to address research question 4, a multiple linear regression analysis was performed to determine whether SNA and ENA combined provided a better model of student performance than SNA and ENA alone. Because introducing additional parameters always increases the variance of a model (R^2), we compared models AICc[3]. All regression models included a predictor for whether the participants completed the pre-course survey. Previous MOOC studies have showed that this is a significant predictor of student performance and course completion (Joksimović, Poquet, et al., 2018).

[2] AICc was used with a correction for finite sample sizes which accounts for variance explained by a model controlling for the number of variables in the model.

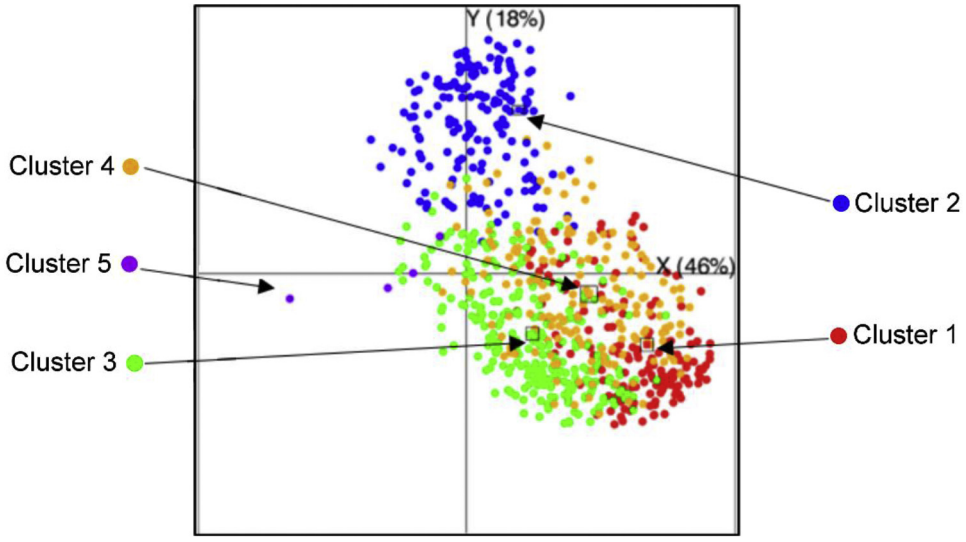

Fig. 3. The dots in the diagram represent the centroids of each student which were identified with ENA. Colors of the centroids represent cluster assignments. The rectangles pointed by the arrow heads from the five clusters represent the confidence intervals along X (i.e., svd_1) and Y (i.e., svd_2) axes of each of the five clusters. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

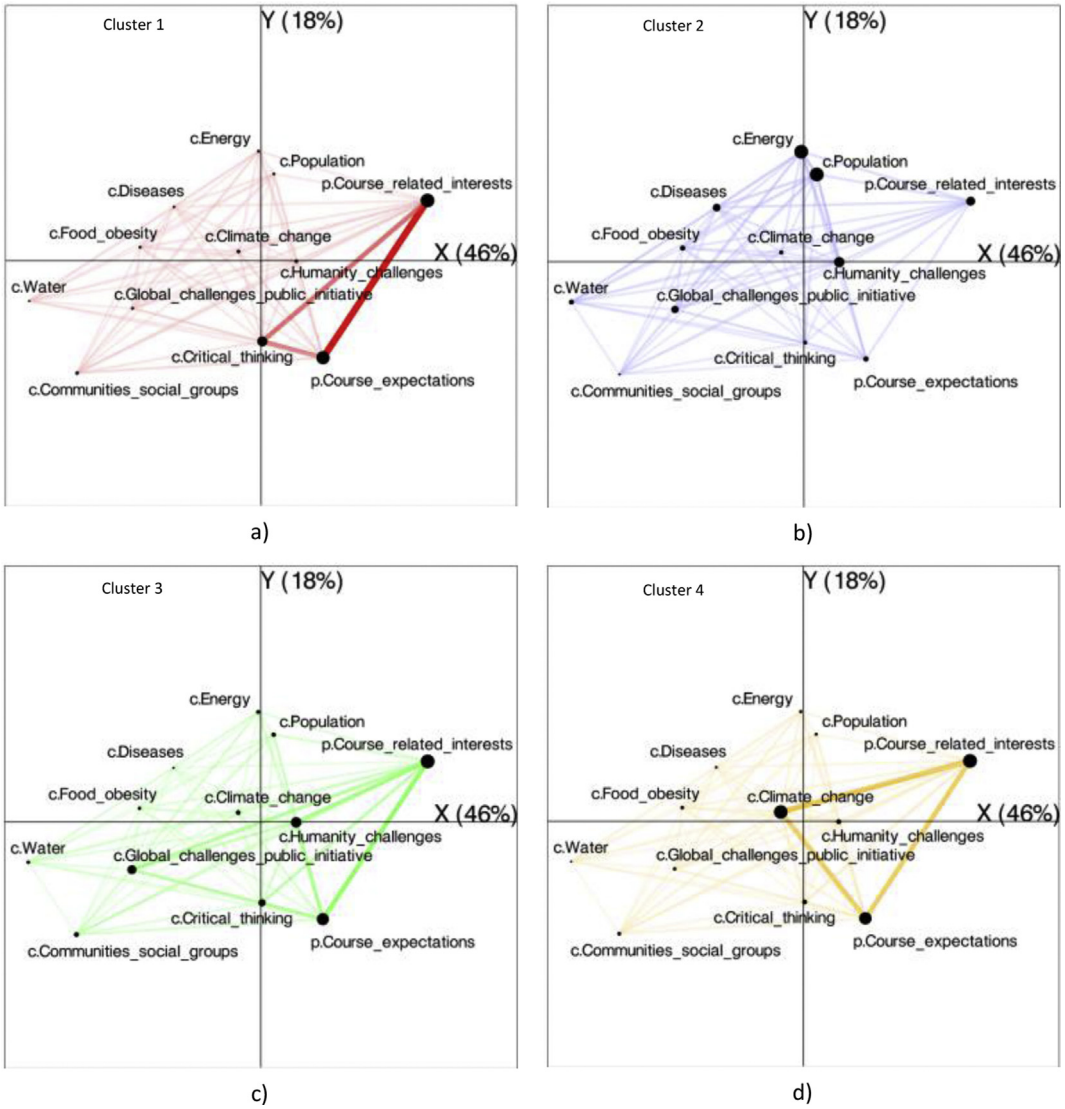

Fig. 4. Average ENA graphs of four ENA clusters. The x-axis is based on svd_1 and increase in the values represents greater focus on process related topics. The y-axis is based on svd_2 and is primarily focused on Content related topics.

Table 3
Analysis of the estimates for the two ERG models – without ENA properties and the model that included ENA properties.

	SNA only[1]		With ENA properties[2]	
	Estimate	SE	Estimate	SE
Baseline (Edges)	−7.050***	0.039	−7.051***	0.041
Selective mixing				
Achievement level (normal)	0.143***	0.022	0.080***	0.023
Gender	0.136***	0.040	0.138***	0.042
English first language	0.148***	0.038	0.127**	0.040
Academic level	0.146***	0.040	0.130**	0.043
Intent	0.065	0.040	0.035	0.042
Work area	0.207**	0.076	0.192*	0.080
ENA Cluster 1	–	–	0.942***	0.064
ENA Cluster 2	–	–	−0.051	0.094
ENA Cluster 3	–	–	1.003***	0.047
ENA Cluster 4	–	–	1.221***	0.049
ENA Cluster 5	–	–	−1.448***	0.122
Structural mechanisms				
Reciprocity	5.151***	0.069	4.836***	0.074
Cyclical ties	1.187***	0.034	1.058***	0.037
Expansiveness	−0.676***	0.079	−0.310***	0.080

Note: *p < .05. **p < .01. ***p < .001. [1]AICc = 45,654, BIC = 45,845; [2]AICc = 46,829, BIC = 46,956.

5. Results

5.1. Research question 1

Table 3 presents two best fitting ERG models (with and without ENA properties) selected by the AICc criterion. The SNA only analysis showed that reciprocity, cyclical ties, and expansiveness played significant roles in the network formation. The positive effect of reciprocity was highest on establishing social ties. The ERG modeling also showed that there were no Simmelian ties observed. Thus, significant associations of network centrality with student performance were expected (Joksimović et al., 2016), which was relevant for the analysis related to research question 4.

Further ERG models were fitted with ENA properties to address research question 1. The optimal ERG model with ENA properties had a better fit to data than the ERG model with SNA properties only – based on the AIC and BIC values (see Table 5). The analysis of selective mixing (i.e., homophilic relationships) in the ERG model with the ENA properties showed (see Table 5) that the assignment to ENA clusters, except for ENA cluster 2, had a significant effect. Students in ENA clusters 1, 3, and 4 were more likely to communicate to other students who were in ENA clusters 1, 3, and 4, respectively. Students in ENA cluster 5 were less likely to communicate to other students who were in the same ENA cluster 5. Finally, the values of the estimates for the ENA-based properties were much higher than those for SNA structural mechanisms and other attributes tested for selective mixing. The analysis also found much higher effects of the ENA properties for selective mixing than those found based on other demographic and academic attributes.

Table 5
Pairwise comparison of the ENA clusters on the measures of social network centrality.

ENA cluster pair	W. Degree z (df)	Closeness z (df)	Betweenness z (df)	Percent grade z (df)
1–2	7.60 (4)*	2.03 (4)	6.95 (4)*	−2.13 (4)
1–3	−2.62 (4)*	−1.92 (4)	−4.04 (4)*	−0.85 (4)
1–4	−4.11 (4)*	0.31 (4)	−4.35 (4)*	−2.85 (4) ***
2–3	−10.91 (4)*	−4.23 (4)*	−11.77 (4)*	1.35 (4)
2–4	−11.88 (4)*	−1.65 (4)	−11.49 (4)*	0.97 (4)
3–4	−1.75 (4)*	−2.29 (4)	−0.59 (4)	−2.16 (4)

Statistical significance codes:
*0.001,
**0.01,
***0.05.

5.2. Research question 2

The descriptive statistics of the five ENA clusters for the three centrality measures are shown in Table 4.

Nonparametric Kruskal-Wallis rank sum test tests (Table 5) found a significant association between ENA cluster assignment and all three network centrality measures – weighted degree ($\chi^2(4) = 649.50$, $p < .001$), closeness ($\chi^2(4) = 18.01$, $p < .001$), and betweenness ($\chi^2(4) = 471.62$, $p < .001$). For example, students in ENA cluster 4 had significantly higher betweenness and degree than their counterparts in ENA cluster 1. This significant association also translated into higher grades as could be hypothesized by the link between network centrality and performance in the absence of Simmelian ties (Joksimović et al., 2016; Krackhardt, 1999). The higher grades of ENA cluster 4 could also be expected due to its lower focus on course process and more on course content (i.e., lower Process scores than that of ENA cluster 1 as shown by Yang et al. (2014).

The results of the statistical tests for the centrality measures are further corroborated by the plots in Fig. 5.

5.3. Research question 3

To address research question 3 and understand the extent to which SNA and ENA can be used jointly to understand communities emerging in collaborative learning, we performed ENA for the three high archiving (Fig. 6a) and the three low achieving (Fig. 6b) communities. As shown in Fig. 6c, epistemic networks of the high performing communities were statistically significantly different from those of the low performing communities ($t(26) = -2.052$, $p = .045$ and Cohen's $d = -0.55$). They differed on the Process dimension of their epistemic networks (i.e., x-axis in the networks in Fig. 6) with the low performing communities being inclined more towards the process related dimension and high performing communities making more connections on the content end of the x-axis. There was no statistically significant difference on the Content (y-axis) dimension of the epistemic networks from Fig. 6 ($t(36) = 1.776$, $p = .081$, Cohen's $d = 0.45$), albeit a moderate effect size. The analysis also indicated that edges in the networks drove most of the differences and the subtracted epistemic networks of the

Table 4
Summary statistics of the network centrality measures for the ENA clusters.

Cluster	Weighted degree		Betweenness centrality		Closeness centrality	
	Mean (SD)	Median (25–75%)	Mean (SD)	Median (25–75%)	Mean (SD)	Median (25–75%)
Cluster 1	7.68 (18.27)	3.00 (2.00–6.25)	4048.98 (16,295.88)	0.00 (0.00–1126.69)	4.04 (2.13)	4.50 (3.70–5.33)
Cluster 2	2.79 (3.22)	2.00 (1.00–3.00)	360.67 (1327.42)	0.00 (0.00–0.00)	3.27 (2.74)	4.09 (1.00–5.69)
Cluster 3	10.61 (20.30)	5.00 (2.00–10.00)	4601.29 (12,349.60)	666.98 (0.00–3325.07)	4.63 (1.68)	4.49 (3.95–5.40)
Cluster 4	13.99 (25.27)	5.00 (3.00–13.00)	4839.19 (16,212.42)	613.96 (0.00–3037.9)	4.09 (1.84)	4.20 (3.60–5.06)
Cluster 5	1.77 (2.69)	1.00 (1.00–2.00)	57.87 (526.46)	0.00 (0.00–0.00)	3.66 (2.88)	4.51 (1.00–6.44)

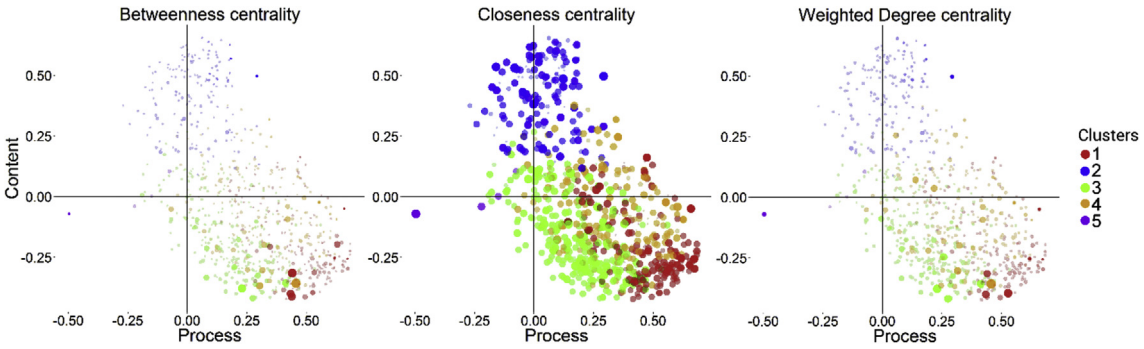

Fig. 5. Comparison of ENA clusters on the values of SNA centrality measures. Centrality measures are indicated by the size and color intensity of the points, while the clusters are marked with different colors. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

high and low performing communities highlighted the edges that drove the position of the means for each group (Fig. 6d). While the learners in both communities made connections to *p.Course_related_interests*, *p.Course_expectations*, *c.Humanity_Challenges*, and *c.Critical_thinking*, learners from the high performing communities also made more connections among more topics in total than the learners in the low performing groups.

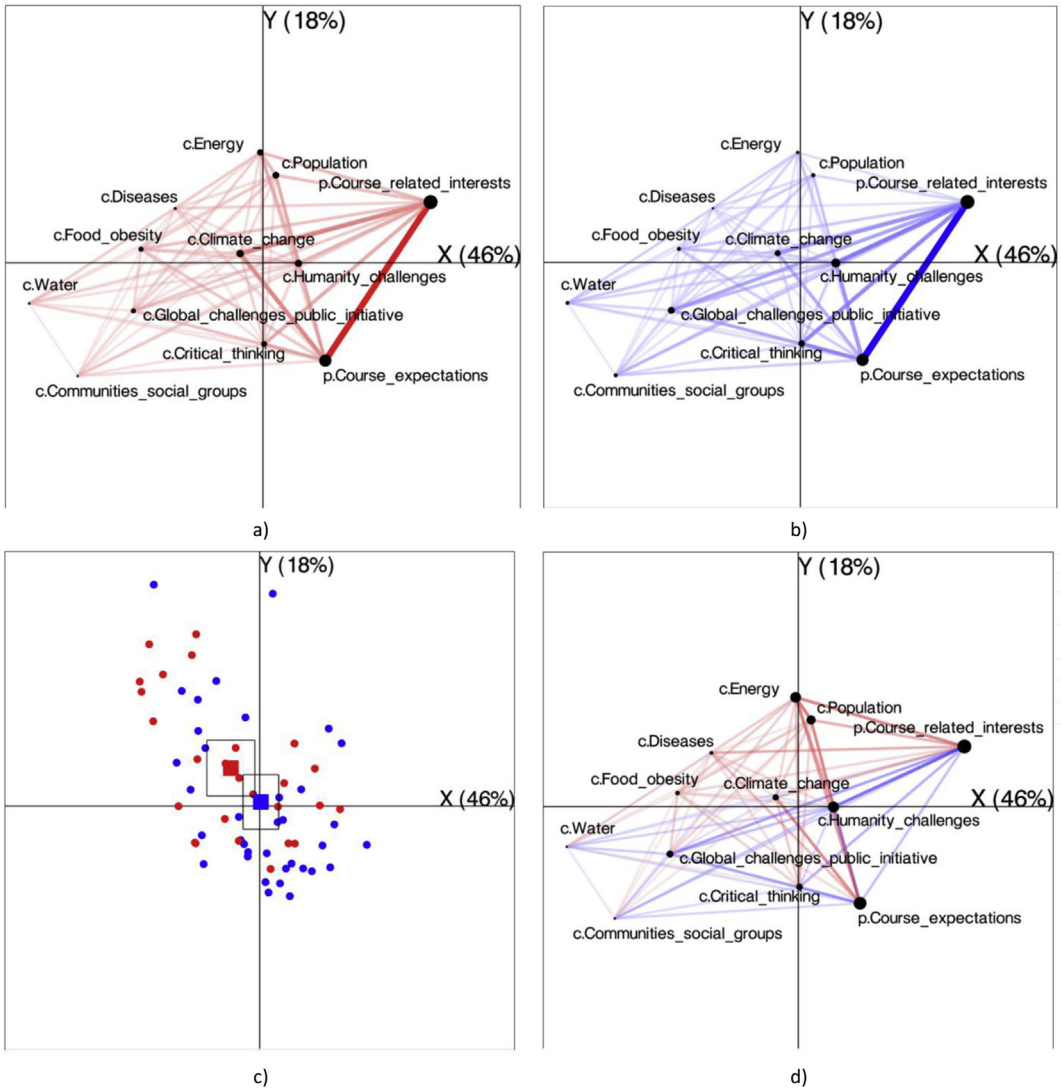

Fig. 6. a) Epistemic network of the three high performing communities; b) epistemic network of the three low performing communities; c) centroids of the epistemic networks of the learners included into high performing (red dots) and low performing (blue dots) communities along with centroids (filled red and blue squares) and confidence intervals (rectangles with black lines); and d) subtracted networks of high performing versus low performing communities. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 6
Multiple linear regression models with the course percent mark as the outcome and predictors as listed in the variable column.

Variable	Model 1 Est. (SE)	Model 2 Est. (SE)	Model 3 Est. (SE)	Model 4 Est. (SE)	Model 5 Est. (SE)
Survey (yes)	24.94 (1.92) ***	26.20 (1.92) ***	27.48 (1.94) ***	25.00 (1.92)***	26.18 (1.92)***
ENA Cluster 1	12.42 (2.90)***		-	13.06 (2.91)***	-
ENA Cluster 2	20.80 (2.60)***		-	20.38 (2.58)***	-
ENA Cluster 3	12.23 (2.72)***		-	14.23 (2.79)***	-
ENA Cluster 4	19.04 (2.98)***		-	20.94 (3.08)***	-
SVD1	-	12.97 (2.30)***	-	-	13.10 (2.34)***
SVD2	-	19.39 (3.39)***	-	-	18.58 (3.69)***
W. Degree	-		−20.08 (23.47)	−52.88 (23.64)*	−26.56 (23.35)
Betweenness	-		41.82 (26.18)	55.53 (25.69)*	50.08 (25.74)
Closeness	-		−10.06 (2.66) ***	−10.52 (2.64)***	−0.08 (2.64) ***
Adjusted R^2	0.16***	0.15***	0.12***	0.17***	0.152 ***
AICc	15,684.82	15,701.25	15,746.44	15,670.49	15,691.76

Significance codes: *** 0.001, ** 0.01, * 0.05.

5.4. Research question 4

A multiple linear regression models identified significant predictors of participants' course grades based on the variables derived from SNA and ENA. Specifically, to examine the contribution of SNA and ENA metrics to understanding of the factors that could predict learning in MOOCs, we specified five regression models (Table 6).

Model 1 included only the ENA clusters as predictors of the final course grade, along with the binary variable that indicated whether students submitted the course survey (Table 6). The model suggests that all of the ENA clusters performed better than the reference group composed of the people who had low participation (ENA cluster 5). This suggests that there might be multiple configurations of discourse patterns that students could use to be successful. Simplifying the interpretation of the model, the results suggest that students who were more active in the course performed better than those students who contributed less to the course discussions. Moreover, model 1 also suggests that students who were focused more on the course content, rather than on the course process, may have benefitted more from the interactions with their peers.

Model 2 further confirmed the importance of participation in discussion forum for better learning outcome (i.e., final course grade). Here, we included Process (svd_1) and Content (svd_2) scores from the ENA model of student discourse. The model revealed that those students whose discourse was characterized by either high level of Content or high level of Process or both had higher outcome scores.

Fitting the third model, we aimed at assessing the importance of social centrality, only, for predicting the final course grade. Model 3 (Table 3) showed a significant effect of closeness centrality only. The participants with lower closer centrality had higher outcomes. Because closeness centrality is a reversed measure, this means that those participants who were better connected with others performed better.

Model 4 used both ENA cluster assignments and SNA centrality measures of the participants and yielded the best fit, having the lowest AICc score and highest R^2 value. This suggests it was the most efficient model and implies that the SNA and ENA predictors are explaining different portions in the participants' variance in performance scores. Interestingly, model 4 showed significant and negative association between the weighted degree centrality and final course outcome.

Model 5 also used both Process (svd_1) and Content (svd_2) scores from the ENA model of student discourse and the SNA centrality measures. The model confirmed the findings of Models 2 and 3 and showed that the SNA and ENA predictors are explaining different portions in the participants' variance in performance scores.

6. Discussion

The results reported corroborated several complementary contributions of the network analytic approach that combines social and content perspectives of collaborative learning. The results addressing each of the four research questions showcased four different ways how the combined use of SNA and ENA advances potential findings that could not be obtained if either ENA or SNA was applied alone.

The results of ERG models related to research question 1 unveiled that processes (i.e., selective mixing) that drove the network formation were better explained when ENA properties are added than when only SNA properties and demographic and academic attributes are used. Moreover, effects of the ENA results on selective mixing were much higher than those by other demographic and academic attributes, which are commonly reported in the literature (Joksimović et al., 2016; Kellogg et al., 2014). This finding suggests that some students mostly chose to interact with peers who shared similar interests and perspectives. This was the case for the students who were in ENA clusters 1, 3, and 4. Epistemic networks of the students in these clusters showed strong focus on a few highly interlinked content and process topics. Shared interest was also found to have had negative effects on establishing social times based on selective mixing for ENA Cluster 5. These were the students who were least active in the discussions and were likely inclined to communicate to other, more active peers than those who had a low level of discussion activity. Finally, effect of the assignment to ENA cluster 2 was not significant on selective mixing. This is likely since students in ENA cluster 2 did not emphasize links between any of the 12 topics; instead they linked similarly all the concepts as shown in their epistemic analysis.

Our analyses also revealed a strong tendency to form reciprocal ties, significant positive effect of cyclical ties, and negative effect of expansiveness. As noted by Lusher et al. (2012), forming mutual relationships is one of the defining characteristics of networks emerging from online interactions, such as within MOOC discussion forums. Opposite to Krivitsky and Handcock (2014), for example, both our models showed a significant and positive tendency on cyclical ties formation. This further suggests a strong anti-hierarchical tendency in interaction between students (i.e., tie formation) within the observed networks (Krivitsky & Handcock, 2014; Krivitsky, 2012). Basically, this finding suggests students' tendency to freely share knowledge in small groups without expectation to be reciprocated (Zappa & Lomi, 2016). The negative and significant effect of expansiveness suggests that for any response to their peers, students' propensity to build new social times decreased. This finding indicates that students tended to form ties with a limited number of peers with whom they would have in-depth discussions with several rounds of responses. The negative effect of expansiveness is also observed and similarly interpreted by Joksimović et al. (2018) who studied the formation of social networks in two other MOOCs.

The findings of research question 1 indicate that for students to establish strong network positions and benefit from online discussions the most, they need to i) identify and link strongly a few key content and process-related topics in their online posts, ii) respond to the posts

of their peers, iii) identify and engage into deep conversations with a small group of peers; and iv) freely share knowledge within small groups without expectation to obtain something in return. These findings are consistent with the theoretical framing of collaborative learning as suggested by Stahl (2004) in terms of both social structures emerging and content discussed in collaborative knowledge building. In other words, the combination of SNA and ENA can detect information about a participant's enactment of what the CSCL literature has described as a role: an ensemble of cognitive and social domains that is marked by interacting with the appropriate people about appropriate content (Dillenbourg et al., 2009).

The results of the analysis related to research question 2 showed what kind of discussion makes students more central in a social network, and thus explaining a role they played in social interactions. Consistent with the results of the ERG models, the findings suggest that students (i.e., those in ENA clusters 1, 3, and 4) who focused strongly on a few highly interlinked content-and process related topics played the most central roles as brokers (i.e., strong closeness and betweenness) in the social network. Moreover, the use of ENA in combination with SNA allowed for detecting specific topics and links among topics that were associated with the strongest brokerage positions. For example, students in ENA cluster 4, who had the significantly highest brokerage role of all, focused on content topics *c.Climate_change* and its links with process topics *p.Course_related_interests* and *p.Course_expectation*. In practice, the identification of such topics may help both teachers and students to detect the key topics for discussion and regulate their future activities to increase benefits in the participation in a social network and collaborative learning.

The findings of research question 3 produced with the combined use of SNA and ENA insights into the possible reasons for the differences in performance among communities identified in the social network. The study showed that the students in low performing communities tended to focus mostly on course related topics and links with other issues in the course. The students in the high performing communities however tended to focus more on content-related topics and link them with course expectations. The focus on links with the course expectations likely indicated a stronger metacognitive monitoring and alignment of own learning with respect to the course standards. This higher monitoring likely in turn resulted in higher grades for the high performing communities in comparison to the low performing communities. This conclusion is consistent with the literature on self-regulated learning which also reported a positive association between engagement into metacognitive monitoring and performance (Greene & Azevedo, 2009).

The results of research question 3 suggest that learners need to pay attention to course expectations to meet assessment standards, which sometimes promote those topics that are not fully associated with students' personal interests. This finding is in connection with the conclusions drawn by Senko, Hulleman, and Harackiewicz (2011) in their analysis of benefits of performance vs mastery goal orientations. Senko et al. posit that students need to have a certain dose of performance goals (e.g., to get a good grade) in addition to being interested in mastering topics studied. Student goal orientations eventually are part of internal conditions that students use to monitory their learning and make decisions related to their study (Winne & Hadwin, 1998; Winne, 2013). The importance of goal-orientation of learners has already been reported in the CSCL literature in connection to the participation – reading and positing – in the online discussions (Wise & Chiu, 2014). Therefore, insights obtained by answering research question 3 highlighted the potential of SENS to identify latent constructs such as metacognitive monitoring and goal-orientation to inform teaching and provide custom-tailored feedback for different groups of students. In practice, if low performing communities are identified as it was the case in this study, instructors can be alerted about the direction in which they need to moderate discussions, stir the focus towards the course expectations, and help learners align their personal interest towards improved course outcomes.

Finally, the results of the analysis related research question 4, showed that combining structural and content features of collaborative learning – that is, both who is interacting with whom and what they are interacting about – provides a richer model of student learning in a collaborative setting. This is corroborated by regression model 4 which explained the highest proportion of variability in student grades and had the best fit to data based on AICc. Students who played brokerage roles (i.e., those with high betweenness scores), had close access to every other node in the network (i.e., those with low closeness scores), and tended to communicate with a small number of peers performed better overall in the course. As such, the association of network centrality measures with performance is consistent with the finding of the ERG models as done in research question 1 that found an insignificant presence of Simmelian ties (Joksimović et al., 2016).

A possible reason for the negative association of weighted degree could be found in the results related to research question 3, which warranted the use of SENS. Table 2 that points to the fact that students, grouped within low performing communities (C6, C8, and C12), had higher degree centrality, on average, than their counterparts in the high performing communities (i.e., C3, C11, and C17). The previous discussion related to research question 3 emphasized the importance of alignment of student discussions with course expectations to increase course performance as revealed by ENA. Moreover, the findings of RQ1 suggested the importance of discussing in small groups that were focused on a mix of process and content related topics to gain network benefits as a form of social capital (Joksimović, Dowell, et al., 2017). This stresses the importance of not just the role in the network in general, but rather the topics the learners discussed and types and numbers of people they interacted with. This is again consistent with the definition of role in the CSCL literature as previously referenced (Dillenbourg et al., 2009).

7. Limitations and future work

There are clearly some important limitations to this preliminary exploration of a single data set. First, the results above show that in this MOOC data, students did not have strong Simmelian ties—and therefore did not form tight-knit cliques. Since prior research has shown that such social ties have an impact on learning outcomes and processes (Joksimović et al., 2016), combining SNA and ENA analyses might produce different results in other datasets. Second, and equally important, the outcome variable used in this analysis was not directly related to collaborative learning: evaluations of student performance did not account for the collaborative components of their work. Thus, we can expect that combining SNA and ENA analyses in datasets with collaborative learning-based outcome measure might similarly produce different results.

Future studies can introduce several methodological and theoretical improvements. The use of SENS in the future can focus on the extraction of learning processes from discourse based on established constructs and theories in collaborative learning such as social presence and cognitive presence. Although this may imply that the content analysis will be more laborious to perform and/or require tools for automated content analysis (Kovanović et al., 2016), the granularity of insights, construct validity, and theoretical soundness of such studies can benefit profoundly. Applications of SNA can also take much more systematic approaches to role analysis as proposed in the CSCL literature (Marcos-García et al., 2015; Martinez et al., 2003; Strijbos & Weinberger, 2010; Strijbos & de Laat, 2010; de Laat et al., 2007). Of particular value would be approaches for role analysis that would be used in combination with ENA to triangulate the differences in epistemic networks among network actors who occupied different roles.

Despite these limitations, this research provides a proof of concept that discourse and social dimensions of collaborative learning can be modeled as networked in nature. Social interaction is expressed in network structures through which information is shared and knowledge

constructed; dimensions extracted from discourse – such as cognitive, metacognitive, social and affective – do not happen in isolation but rather co-occur with each other. This preliminary investigation of one particular data set suggests that by combining two analysis techniques well-known in educational research—SNA and ENA—we may be able to characterize collaborative learning as an interaction of social factors and factors extracted from discourse: a social and epistemic network signature (SENS) of collaborative learning.

The SENS approach, as outlined in Section 3.2, in its core is based on several data analysis methods that originate from different disciplines including data mining, statistics, and network science. Although the users of SENS need to have a wide range of skills to be able to use the method, none of the individual methods included in SENS is unknown to researchers and practitioners in CSCL, the learning sciences, and learning analytics. As shown in this paper, SNA and ENA are commonly used in the related literature (Chesler et al., 2015; Kovanović et al., 2016; Marcos-García et al., 2015; Martinez et al., 2003; Shaffer, 2006; de Laat et al., 2007). Likewise, cluster analysis and statistical methods included into SENS are commonly used in the related literature (Cress, 2008; Kovanović et al., 2015; Wise et al., 2013). In fact, most of these methods are used together; e.g., Joksimović et al. (2016) used statistical and descriptive SNA together with mixed-models and regression analysis to assess the effect of student centrality in social networks on performance. Similarly, Kovanović et al. (2015) and Wise et al. (2013) used cluster analysis with regression analysis to study effects of students' participation behaviors on social knowledge construction.

Researchers and practitioners who will use SENS will likely need to learn only an additional method and connect it with their exiting analysis toolkits. The main contribution of SENS is the link between SNA and ENA and methodological progress of the analysis steps needed to address questions that can benefit from a combined analysis of social ties and processes emerging from content analysis of collaborative discourse. Moreover, the methods used in SENS are all available in R and do not require the use of any proprietary tool. The SENS methodological approach should not however be limited just to the specific statistical, data and SNA methods used in the study described in the paper. Users of SENS are also encouraged to use other more sophisticated methods for the analysis types suggested in the SENS methodological approach if their use is better suited for questions they aim to answer (Jovanović, Gašević, Dawson, Pardo, & Mirriahi, 2017; Kovanović et al., 2015; Snijders, van de Bunt, & Steglich, 2010).

The work reported here suggests that further research can and should be done to broaden insights into the potential offered by the SENS approach. This approach would need to be validated across different environments of collaborative learning, pedagogical principles, and educational levels. The outcome of such a research program would be practical models for the use of SENS for teaching and assessment of collaborative learning skills across different learning environments and at different scales of enrollment, as well as guidelines to inform policy makers who define curriculum and assessment guidelines and vendors who develop technology for supporting collaborative learning.

Acknowledgements

This work was funded in part by the European Erasmus + program (562080-EPP-1-2015- BE-EPPKA3-PI-FORWARD and 586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP), Social Sciences and Humanities Research Council of Canada, The University of Edinburgh, Monash University, National Science Foundation (DRL-0918409, DRL-0946372, DRL-1247262, DRL-1418288, DUE-0919347, DUE-1225885, EEC-1232656, EEC-1340402, REC-0347000), the MacArthur Foundation, the Spencer Foundation, the Wisconsin Alumni Research Foundation, the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.

References

Akaike, H. (2011). Akaike's information criterion (pp. 25–25) In M. Lovric (Ed.). *International encyclopedia of statistical science*Berlin Heidelberg: Springer. Retrieved from http://link.springer.com/referenceworkentry/10.1007/978-3-642-04898-2_110.

Arastoopour, G., Shaffer, D. W., Swiecki, Z., Ruis, A., & Chesler, N. C. (2016). Teaching and assessing engineering design thinking with virtual internships and epistemic network analysis. *International Journal of Engineering Education, 32*(2).

Azevedo, R. (2015). Defining and measuring engagement and learning in Science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist, 50*(1), 84–94. https://doi.org/10.1080/00461520.2015.1004069.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Capuano, N., Mangione, G. R., Mazzoni, E., Miranda, S., & Orciuoli, F. (2014). Wiring role taking in collaborative learning environments. SNA and semantic web can improve CSCL script? *International Journal of Emerging Technologies in Learning (iJET), 9*(7), 30–38.

Carolan, B. V. (2013). *Social network analysis and Education: Theory, methods & applications* (1 edition). Los Angeles: SAGE Publications, Inc.

Chesler, N. C., Ruis, A. R., Collier, W., Swiecki, Z., Arastoopour, G., & Williamson Shaffer, D. (2015). A novel paradigm for engineering Education: Virtual internships with individualized mentoring and assessment of engineering thinking. *Journal of Biomechanical Engineering, 137*(2), 024701–024701 https://doi.org/10.1115/1.4029235.

Cress, U. (2008). The need for considering multilevel analysis in CSCL research—an appeal for the use of more advanced statistical methods. *International Journal of Computer-Supported Collaborative Learning, 3*(1), 69–84. https://doi.org/10.1007/s11412-007-9032-2.

Dawson, S. (2008). A study of the relationship between student social networks and sense of community. *Educational Technology & Society, 11*(3), 224–238.

Dawson, S., Gašević, D., Siemens, G., & Joksimović, S. (2014). Current state and future trends: A citation network analysis of the learning analytics field. *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 231–240). New York, NY, USA: ACM. https://doi.org/10.1145/2567574.2567585.

Dawson, S., Tan, J. P. L., & McWilliam, E. (2011). Measuring creative potential: Using social network analysis to monitor a learners' creative capacity. *Australasian Journal of Educational Technology, 27*(6), 924–942.

De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education, 46*(1), 6–28. https://doi.org/10.1016/j.compedu.2005.04.005.

Deritei, D., Lázár, Z. I., Papp, I., Járai-Szabó, F., Sumi, R., Varga, L., et al. (2014). Community detection by graph Voronoi diagrams. *New Journal of Physics, 16*(6), 063007. https://doi.org/10.1088/1367-2630/16/6/063007.

Dillenbourg, P., Järvelä, S., & Fischer, F. (2009). The evolution of research on computer-supported collaborative learning. In D. N. Balacheff, D. S. Ludvigsen, D. T. de Jong, D. A. Lazonder, & D. S. Barnes (Eds.). *Technology-enhanced learning* (pp. 3–19). Netherlands: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4020-9827-7_1.

Dowell, N. M., Skrypnyk, O., Joksimović, S., Graesser, A. C., Dawson, S., Gaševic, D., ... Kovanovic, V. (2015). Modeling learners' social centrality and performance through language and discourse. *Proceedings of the 8th international educational data mining society* (pp. 250–257). Madrid, Spain: IEDMS.

Fischer, F., Kollar, I., Stegmann, K., & Wecker, C. (2013). Toward a script theory of guidance in computer-supported collaborative learning. *Educational Psychologist, 48*(1), 56–66. https://doi.org/10.1080/00461520.2012.748005.

Fraley, C., Raftery, A. E., & others (2007). Model-based methods of classification: Using the mclust software in chemometrics. *Journal of Statistical Software, 18*(6), 1–13.

Garrison, D. R. (2011). *E-learning in the 21st century a framework for research and practice.* New York: Routledge.

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education, 15*(1), 7–23. https://doi.org/10.1080/08923640109527071.

Garrison, D. R., & Arbaugh, J. B. (2007). Researching the community of inquiry framework: Review, issues, and future directions. *The Internet and Higher Education, 10*(3), 157–172. https://doi.org/10.1016/j.iheduc.2007.04.001.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x.

Gašević, D., Kovanović, V., & Joksimović, S. (2017). Piecing the learning analytics puzzle: A consolidated model of a field of research and practice. *Learning: Research and Practice, 3*(2), 63–78. https://doi.org/10.1080/23735082.2017.1286142.

Gašević, D., Zouaq, A., & Janzen, R. (2013). "Choose your classmates, your GPA is at stake!" the association of cross-class social ties and academic performance. *American Behavioral Scientist, 57*(10), 1460–1479. https://doi.org/10.1177/0002764213479362.

Goodreau, S. M., Kitts, J. A., & Morris, M. (2009). Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks*. *Demography, 46*(1), 103–125. https://doi.org/10.1353/dem.0.0045.

Granovetter, M. (1982). The strength of weak ties: A network theory revisited. In P. V. Marsden, & N. Lin (Eds.). *Social structure and network analysis* (pp. 105–130). Beverly

Hill, CA: Sage.
Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. *Contemporary Educational Psychology, 34*(1), 18–29. https://doi.org/10.1016/j.cedpsych.2008.05.006.
Griffin, P., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st century skills*. Dordrecht: Springer Netherlands. Retrieved from http://link.springer.com/10.1007/978-94-007-2324-5.
Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of A Global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research, 17*(4), 397–431. https://doi.org/10.2190/7MQV-X9UJ-C7Q3-NRAG.
Hammond, M. (1999). Issues associated with participation in on line forums—the case of the communicative learner. *Education and Information Technologies, 4*(4), 353–367. https://doi.org/10.1023/A:1009661512881.
Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Library & Information Science Research, 18*(4), 323–342. https://doi.org/10.1016/S0740-8188(96)90003-1.
Haythornthwaite, C. (2002). Building social networks via computer Networks: Creating and sustaining distributed learning communities. *Building virtual communities*Cambridge University Press. Retrieved from https://doi.org/10.1017/CBO9780511606373.011.
Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin, & E. Care (Eds.). *Assessment and teaching of 21st century skills* (pp. 37–56). Netherlands: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-94-017-9395-7_2.
Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.
Joksimović, Srecko, Dowell, Nia, Skrypnyk, Oleksandra, Kovanovic, Vitomir, Gasevic, Dragan, Dawson, Shane, & Graesser, Arthur C. (2017). Exploring the Accumulation of Social Capital in cMOOC Through Language and Discourse. *The Internet and Higher Education, 36*, 5464. http://dx.doi.org/10.1016/j.iheduc.2017.09.004.
Joksimović, S., Gašević, D., Kovanović, V., Riecke, B. E., & Hatala, M. (2015). Social presence in online discussions as a process predictor of academic performance. *Journal of Computer Assisted Learning, 31*(6), 638–654. https://doi.org/10.1111/jcal.12107.
Joksimović, S., Jovanović, J., Kovanović, V., Gašević, D., Milikić, N., Zouaq, A., & van Staalduinen, J. P. (2018a). Comprehensive analysis of discussion forum participation: From speech acts to discussion dynamics and course outcomes. *IEEE Transactions on Learning Technologies* (under review).
Joksimović, S., Manataki, A., Gašević, D., Dawson, S., Kovanović, V., de Kereki, et al. (2016). Translating network position into Performance: Importance of centrality in different network configurations. *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 314–323). New York, NY, USA: ACM. https://doi.org/10.1145/2883851.2883928.
Joksimović, Srecko, Poquet, Oleksandra, Kovanovic, Vitomir, Dowell, Nia, Mills, Caitlin, Gasevic, Dragan, Dawson, Shane, Graesser, Arthur C., & Brooks, Christopher (2018b). How do we measure learning at scale? A systematic review of the literature. *Review of Educational Research, 88*(1), 43–86. http://dx.doi.org/10.3102/0034654317740335.
Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education, 33*, 74–85.
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), 263–289. https://doi.org/10.19173/irrodl.v15i5.1852.
Kovanović, V., Gašević, D., Joksimović, S., Hatala, M., & Adesope, O. (2015). Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *The Internet and Higher Education, 27*, 74–89. https://doi.org/10.1016/j.iheduc.2015.06.002.
Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., et al. (2016). Towards automated content analysis of discussion transcripts: A cognitive presence case. *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 15–24). New York, NY, USA: ACM. https://doi.org/10.1145/2883851.2883950.
Krackhardt, D. (1999). *The ties that torture: Simmelian tie analysis in organizations*.
Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. *Electronic Journal of Statistics, 6*, 1100–1128. https://doi.org/10.1214/12-EJS696.
Krivitsky, P. N., & Handcock, M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B, 76*(1), 29–46. https://doi.org/10.1111/rssb.12014.
de Laat, M., Lally, V., Lipponen, L., & Simons, R.-J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for social network analysis. *International Journal of Computer-Supported Collaborative Learning, 2*(1), 87–103. https://doi.org/10.1007/s11412-007-9006-4.
Lusher, D., Koskinen, J., & Robins, G. (2012). *Exponential random graph models for social Networks: Theory, methods, and applications*. Cambridge University Press.
Marcos-García, J.-A., Martínez-Monés, A., & Dimitriadis, Y. (2015). Despro: A method based on roles to provide collaboration analysis support adapted to the participants in CSCL situations. *Computers & Education, 82*, 335–353. https://doi.org/10.1016/j.compedu.2014.10.027.
Martinez, A., Dimitriades, Y., Rubia, B., Gomez, E., & de la Fuente, P. (2003). Combining qualitative evaluation and social network analysis for the study of classroom social interactions. *Computers and Education, 41*(4), 353–368.
Modha, D. S., Ananthanarayanan, R., Esser, S. K., Ndirango, A., Sherbondy, A. J., & Singh, R. (2011). Cognitive computing. *Communications of the ACM, 54*(8), 62–71.
Morris, M., Handcock, M. S., & Hunter, D. R. (2008). Specification of exponential-family random graph models: Terms and computational aspects. *Journal of Statistical Software, 24*(4), 1548.
Nash, P., & Shaffer, D. W. (2012). Epistemic trajectories: Mentoring in a game design practicum. *Instructional Science, 41*(4), 745–771. https://doi.org/10.1007/s11251-012-9255-0.
National Research Council (US) (2011). *Assessing 21st century Skills: Summary of a workshop*. Washington (DC): National Academies Press (US). Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK84218/.
Neath, A. A., & Cavanaugh, J. E. (2012). The bayesian information criterion: Background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics, 4*(2), 199–203. https://doi.org/10.1002/wics.199.
OECD (2013). *PISA 2015 collaborative problem solving framework*. Paris, France: OECD. Retrieved from http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf.
Orrill, C., Shaffer, D. W., & Burke, J. (2013). Exploring coherence in teacher knowledge using epistemic network analysis. *American educational research association annual meeting. San francisco, CA*. Retrieved from http://www.academia.edu/7247524/Exploring_coherence_in_teacher_knowledge_using_epistemic_network_analysis.
Peeples, M. A. (2011). *K-means analysis R-Script*. Retrieved October 17, 2016, from http://www.mattpeeples.net/kmeans.html.
Puntambekar, S., Erkens, G., & Hmelo-Silve, C. (2011). *Analyzing interactions in CSCL - methods, approaches and issues*, *Vol. 12*New York: Springer. Retrieved from http://www.springer.com/gp/book/9781441977090.
Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. *Computer supported collaborative learning* (pp. 69–97). Berlin, Heidelberg: Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-85098-1_5.
Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic Games: Measurement principles for complex learning environments. *The Journal of Technology, Learning, and Assessment, 8*(4), Retrieved from https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1623.
Rupp, A. A., Sweet, S. J., & Choi, Y. (2010). Modeling learning trajectories with epistemic network analysis: A simulation-based investigation of a novel analytic method for epistemic games. *Proceedings of the 3rd international conference on educational data mining* (pp. 319–320). (Pittsburgh, PA, USA).
Schellens, T., Keer, H. V., Wever, B. D., & Valcke, M. (2007). Scripting by assigning roles: Does it improve knowledge construction in asynchronous discussion groups? *International Journal of Computer-Supported Collaborative Learning, 2*(2–3), 225–246. https://doi.org/10.1007/s11412-007-9016-2.
Schluter, N. (2014). Centrality measures for non-contextual graph-based unsupervised single document keyword extraction. *Traitement Automatique des Langues Naturelles, TALN 2014, Marseille, France, 1-4 Juillet 2014, articles courts* (pp. 455–460). . Retrieved from http://aclweb.org/anthology/F/F14/F14-2012.pdf.
Scott, J. (2012). *Social network analysis*. SAGE.
Senko, C., Hulleman, C. S., & Harackiewicz, J. M. (2011). Achievement goal theory at the Crossroads: Old controversies, current challenges, and new directions. *Educational Psychologist, 46*(1), 26–47. https://doi.org/10.1080/00461520.2011.538646.
Shaffer, D. W. (2004). Epistemic frames and islands of Expertise: Learning from infusion experiences. *Proceedings of the 6th international conference on learning sciences* (pp. 473–480). Santa Monica, California: International Society of the Learning Sciences. Retrieved from http://dl.acm.org/citation.cfm?id=1149126.1149184.
Shaffer, D. W. (2006). Epistemic frames for epistemic games. *Computers & Education, 46*(3), 223–234. https://doi.org/10.1016/j.compedu.2005.11.003.
Shaffer, D. W. (2008). *How computer games help children learn* (2006 edition). New York: Palgrave Macmillan.
Shaffer, D. W. (2013). *Data mining the Common Core Standards: What Epistemic Network Analysis reveals about complex thinking skills in mathematics and English language arts*. Bill and Melinda Gates Foundation.
Shaffer, D. W. (2014). *User guide for epistemic network analysis web*(Games and Professional Simulations No. Technical Report 2014-1). Madison, WI: University of Wisconsinversion 3.3.
Shaffer, D. W. (2017). *Quantitative ethnography*. Madison, WI: Cathcart Press.
Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network Analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics, 3*(3), 9–45. https://doi.org/10.18608/jla.2016.33.3.
Shaffer, D. W., & Gee, J. P. (2012). The right kind of gate: Computer games and the future of assessment. In M. C. Mayrath, J. Clarke-Midura, G. Schraw, & D. H. Robinson (Eds.). *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 211–228). Charlotte, NC: Information Age Publications.
Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., ... Mislevy, R. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media, 1*(2), 33–53. https://doi.org/10.1162/ijlm.2009.0013.
Snijders, T. A. B., van de Bunt, G. G., & Steglich, C. E. G. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks, 32*(1), 44–60. https://doi.org/10.1016/j.socnet.2009.02.004.
Sobocinski, M., Malmberg, J., & Järvelä, S. (2017). Exploring temporal sequences of regulatory phases and associated interactions in low- and high-challenge collaborative learning sessions. *Metacognition and Learning*, 1–20. https://doi.org/10.1007/s11409-016-9167-5.
Stahl, G. (2004). Building collaborative knowing. In J.-W. Strijbos, P. A. Kirschner, & R. L. Martens (Eds.). *What we know about CSCL* (pp. 53–85). Netherlands: Springer. Retrieved from http://link.springer.com/chapter/10.1007/1-4020-7921-4_3.
Steinhaeuser, K., & Chawla, N. V. (2008). Community detection in a large real-world

social network. *Social computing, behavioral modeling, and prediction* (pp. 168–175). Boston, MA: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-0-387-77672-9_19.

Stepanyan, K., Borau, K., & Ullrich, C. (2010). A social network analysis perspective on student interaction within the twitter microblogging environment. *2010 10th IEEE international conference on advanced learning technologies* (pp. 70–72). . https://doi.org/10.1109/ICALT.2010.27.

Strijbos, J.-W., & de Laat, M. F. (2010). Developing the role concept for computer-supported collaborative learning: An explorative synthesis. *Computers in Human Behavior, 26*(4), 495–505. https://doi.org/10.1016/j.chb.2009.08.014.

Strijbos, J.-W., Martens, R. L., Prins, F. J., & Jochems, W. M. G. (2006). Content analysis: What are they talking about? *Computers & Education, 46*(1), 29–48. https://doi.org/10.1016/j.compedu.2005.04.002.

Strijbos, J.-W., & Weinberger, A. (2010). Emerging and scripted roles in computer-supported collaborative learning. *Computers in Human Behavior, 26*(4), 491–494. https://doi.org/10.1016/j.chb.2009.08.006.

Tan, J. P.-L., Caleon, I. S., Jonathan, C. R., & Koh, E. (2014). A dialogic framework for assessing collective creativity in computer-supported collaborative problem-solving tasks. *Research and Practice in Technology Enhanced Learning, 9*(3), 411–437.

Teasley, S. D. (2011). Thinking about methods to capture effective collaborations. *Analyzing interactions in CSCL* (pp. 131–142). Boston, MA: Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-1-4419-7710-6_6.

Wang, M., Wang, C., Yu, J. X., & Zhang, J. (2015). Community detection in social Networks: An in-depth benchmarking study with a procedure-oriented framework. In: *Proc. VLDB Endow. 8*(10), 998–1009In: . https://doi.org/10.14778/2794367.2794370.

Wasserman, S. (1994). *Social network Analysis: Methods and applications.* Cambridge University Press.

Weinberger, A., Stegmann, K., Fischer, F., & Mandl, H. (2007). Scripting argumentative knowledge construction in computer-supported learning environments. In F. Fischer, I. Kollar, H. Mandl, & J. M. Haake (Eds.). *Scripting computer-supported collaborative learning* (pp. 191–211). US: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-0-387-36949-5_12.

Weingart, S. B., & Meeks, E. (2012). The digital humanities contribution to topic modeling. *Journal of Digital Humanities, 2*(1), Retrieved from http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/.

Whitelock, D., Field, D., Pulman, S., Richardson, J. T. E., & Van Labeke, N. (2014). Designing and testing visual representations of draft essays for higher education students. *2nd international workshop on discourse-centric learning analytics, 4th conference on learning analytics and knowledge (LAK2014). Indianapolis, Indiana, USA*. Retrieved from https://dcla14.files.wordpress.com/2014/03/dcla14_whitelock_etal.pdf.

Winne, P. H. (2013). Learning strategies, study skills, and self-regulated learning in postsecondary education. In M. B. Paulsen (Ed.). *Higher Education: Handbook of theory and research* (pp. 377–403). Netherlands: Springer.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.). *Metacognition in educational theory and practice* (pp. 277–304). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Wise, A. F., & Chiu, M. M. (2014). The impact of rotating summarizing roles in online discussions: Effects on learners' listening behaviors during and subsequent to role assignment. *Computers in Human Behavior, 38*, 261–271. https://doi.org/10.1016/j.chb.2014.05.033.

Wise, A. F., Saghafian, M., & Padmanabhan, P. (2012). Towards more precise design guidance: Specifying and testing the functions of assigned student roles in online discussions. *Educational Technology Research & Development, 60*(1), 55–82. https://doi.org/10.1007/s11423-011-9212-7.

Wise, A. F., Speer, J., Marbouti, F., & Hsiao, Y.-T. (2013). Broadening the notion of participation in online discussions: Examining patterns in learners' online listening behaviors. *Instructional Science, 41*(2), 323–343. https://doi.org/10.1007/s11251-012-9230-9.

Woo, Y., & Reeves, T. C. (2007). Meaningful interaction in web-based learning: A social constructivist interpretation. *The Internet and Higher Education, 10*(1), 15–25. https://doi.org/10.1016/j.iheduc.2006.10.005.

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

Zappa, P., & Lomi, A. (2016). Knowledge sharing in organizations: A multilevel network analysis. *Multilevel network analysis for the social sciences* (pp. 333–353). Cham: Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-24520-1_14.

Zimmerman, B. J., & Schunk, D. H. (2011). *Handbook of self-regulation of learning and performance.* New York: Routledge.