Towards Combined Network and Text Analytics of Student Discourse in Online Discussions

Rafael Ferreira[1,2(✉)], Vitomir Kovanović[3], Dragan Gašević[2,4], and Vitor Rolim[1]

[1] Federal Rural University of Pernambuco, Rua Dom Manuel de Medeiros, Recife, Brazil
{rafael.mello,vitor.rolim}@ufrpe.br
[2] University of Edinburgh, Old College, South Bridge, Edinburgh EH8 9YL, UK
{rafael.ferreira,dragan.gasevic}@ac.ed.uk
[3] University of South Australia, 160 Currie Street, Adelaide, SA 5000, Australia
Vitomir.Kovanovic@unisa.edu.au
[4] Monash University, 19 Ancora Imparo Way, Clayton, VIC 3800, Australia
dragan.gasevic@monash.edu

Abstract. This paper presents a novel method for the evaluation of students' use of asynchronous discussions in online learning environments. In particular, the paper shows how students' cognitive development across different course topics can be examined using the combination of natural language processing and graph-based analysis techniques. Drawing on the theoretical foundation of the community of inquiry model, we show how topic modeling and epistemic network analysis can provide qualitatively new insight into students' development of critical and deep thinking skills. We also show how the same method can be used to investigate the effectiveness of instructional interventions and its effect on student learning. The results of this study and its practical implications are further discussed.

Keywords: Community of inquiry model
Epistemic network analysis · Content analysis
Instructional interventions · Online discussions

1 Introduction

Asynchronous online discussions represent one of the most commonly used tools for supporting social interactions within online and blended courses [3]. With the increased shift towards social learning models and constructivist pedagogies, there is a rising need to understand how asynchronous online discussions can be used to foster learning and knowledge (co-)construction by the group of learners. In this regard, the Community of Inquiry (CoI) model [13] represents one of the most widely used and researched pedagogical models which tries to

C. Penstein Rosé et al. (Eds.): AIED 2018, LNAI 10947, pp. 111–126, 2018.
https://doi.org/10.1007/978-3-319-93843-1_9

understand how asynchronous online communication impacts student learning and cognitive development. Using the CoI model, a large body of research examined the effect of different instructional approaches to student engagement and learning outcomes [15]. However, the majority of research evidence provides a high-level overview of student learning by focusing solely on the *learning process*, with insufficient understanding of how it relates to *learning content*. Moreover, there is limited evidence on how various instructional interventions affect the development of positive learning outcomes concerning different components of both learning processes and learning content.

In this paper, we show how the combination of network-based analysis and natural-language processing (NLP) techniques [24] can be used to provide more detailed insights into student learning in asynchronous online discussions. Using the data from six offers of a fully-online graduate-level course in software engineering, we show how a combination of Epistemic Network Analysis (ENA) [32] and Latent Dirichlet Allocation (LDA) [7], can provide rich insights into the effect of instructional scaffolding on student learning of particular course topics. Study results and practical implications are further discussed.

2 Background

2.1 The Community of Inquiry Model

While numerous models and approaches to understanding students' online learning have been proposed, the Community of Inquiry (CoI) model represents arguably the most widely-used and researched theoretical framework that outlines the important facets of students' online learning [15]. The CoI model defines three dimensions (called presences) that together provide a holistic overview of online learning experience: (1) *Cognitive presence*, which captures the development of desirable learning outcomes such as critical thinking, problem-solving, and knowledge (co-)construction [12,14,17], (2) *Social presence*, which models the social climate within learners' group (i.e., cohesion, affectivity, and open communication) [28], and (3) *Teaching presence*, which concerns the instructors' role before (i.e., course design) and during (i.e., facilitation and direct instruction) the course [4].

To assess the levels of CoI presences, researchers can use either a set of coding schemes for content analysis of discussion transcripts [13] or a 34-item survey instrument by Arbaugh et al. [5]. The CoI model and its instruments have been extensively used and validated in a large number of studies, both in traditional online [15], and MOOC contexts [21]. Finally, there has been some research focusing on the development of automated tools for identification of cognitive presence phases [18,22,35] which show the potential of data mining and natural language processing (NLP) techniques for assessing students' learning within communities of inquiry.

From the standpoint of the present study, the most important construct is the cognitive presence, which captures the development of critical and deep-thinking skills [13,14]. Rooted in the inquiry-based conceptions of learning of Dewey [9] and

Lipman [23], cognitive presence is operationalized using a four-phase inquiry process: (1) *Triggering event*, where a problem or dilemma is identified and conceptualized, (2) *Exploration*, in which students' explore and brainstorm potential solutions to the issue at hand, (3) *Integration*, during which students (co-)construct new knowledge by synthesizing the existing information, and (4) *Resolution*, in which students evaluate the newly-created knowledge through hypothesis testing or vicarious application to the problem/dilemma that triggered the learning cycle [14]. The four phases of cognitive presence are theorized as being differentiated across two orthogonal dimensions: (1) *perception-awareness* dimension, which captures the differences between early and late stages of cognitive presence, and (2) *deliberation-action* dimension, which differentiates between phases that primarily occur in the shared world of student discourse (triggering event and resolution) and the ones that happen in the private world of reflection (exploration and integration).

The recent studies of the CoI model recognised self-regulated learning (SRL) [36], metacognition [2] and computer-supported collaborative learning (CSCL) [33] as central to understanding students' online learning. For example, Gašević et al. [16] examined the use of role assignments – a key CSCL technique for fostering student deep and collaborative learning [11] – in combination with externally-facilitated regulation [6] as a scaffolding for student discussion participation. What Gašević et al. [16] results show is that give clear instructions of discussion participation, and role assignments increased students' cognitive presence, as evidenced by the higher percentage of their contributions showing the presence of later stages of inquiry cycle, namely integration and resolution phases.

2.2 Research Questions

It should be noted that the CoI model focuses on the *learning process*, (rather than learning outcomes) and how different course designs and interventions affect students' learning experience, such as the one by Gašević et al. [16]. However, there has been limited research that tried to examine how students' cognitive presence develops concerning different course topics, or how different instructional interventions affect students' cognitive development around the different course topics. We suspect that this is primarily due to the very time-consuming and labour-intensive work required to assess cognitive presence on such fine-grained level. However, with the rapid development of sophisticated NLP, text mining, and graph analysis techniques, there is a potential to use them to identify different course topics and levels of cognitive presence [22,35] (semi-) automatically and examine how they relate to each other. In this paper, we utilized LDA, a popular NLP method [32], and ENA, a graph analysis technique [7], to provide insights into students cognitive presence development with regards to different course topics. As such, our first research question is:

Research Question 1:
What is the relationship between students' cognitive presence and different course topics? Can we use the proposed NLP and graph-based method to uncover cognitive presence development with respect to different course topics?

In addition to examining the overall relationship between cognitive presence and the course topics, we are interested in exploring whether the proposed approach can provide additional insights into the effects of externally-scaffolded interventions such as the one presented in Gašević et al. [16] study. While Gašević et al. [16] provide general evidence of the effectiveness of the role-assignment scaffolding, it is essential to examine how the intervention affected the cognitive presence development in relation to different course subjects. As such, our second research question is:

Research Question 2:
What is the effect of instructional scaffolding intervention on students cognitive presence of particular course topics? Can we use the proposed model to assess the effectiveness of role assignment intervention on the development of students' cognitive presence?

By addressing proposed research questions, the paper contributes to the existing body of knowledge in the literature on the CoI model and provides a novel method for assessing students' communities of inquiry through the data-informed lenses.

3 Method

3.1 Data and Course Design

The analysis presented in this paper builds upon the same dataset used in the Gašević et al. [16] study. The primary benefit of using the same dataset is that it makes it simple to directly compare the "traditional" analysis and the proposed approach based on a graph- and text-mining. The data encompasses six offerings of a masters level research-intensive course in software engineering offered entirely online at a Canadian public university between 2008 and 2011. The course covers 14 different topics related to software requirements, design, implementation, evolution, and process which are grouped into six modules (Table 1).

The course consisted of four tutor marker assignments (TMA1–4) which accounted for 15% (presentation of a published peer-reviewed paper), 25% (literature review paper), 15% (project proposal), and 30% (project) of the final grade, respectively. Discussion participation in the course accounted for the remaining 15% of the grade [for further details on the course design, see 16]. As part of the TMA1, students were required to select one research paper on a software engineering topic, record a video presentation, and post a URL to a new course

Table 1. Course topics by weeks.

Module	#	Section	Label
Software engineering scope	1	Introduction	intro.scope
Software requirements	2	Introduction	reqs.intro
	3	Methods	reqs.meth
Software design	4	Introduction	desig.intro
	5	Structure, architecture and quality	desig.qual
	6	Methods and strategies	desig.meth
Software construction approaches	7	Introduction	const.intro
	8	Languages	const.lang
	9	Component-based and product-line engineering	const.meth
Software maintenance	10	Introduction	maint.intro
	11	Reverse engineering and knowledge management	maint.manag
Software engineering process	12	Introduction	proc.intro
	13	Stages	proc.stages
	14	Key issues	proc.issues

online discussion, where other students would engage in the debate around their presentation. During the first two offerings of the course, student participation was primarily driven by the extrinsic motivational factors (i.e., course grade), with limited scaffolding support. In our study, this group – referred to as *control group* – consisted of 37 students who produced 845 messages (Tables 2 and 3). After the first two course offers, the scaffolding of discussion participation through role assignments and clear instructions were implemented. In total, 44 students – referred to as *treatment group* – were exposed to this instructional intervention, and produced the total of 902 messages (Tables 2 and 3).

Table 2. Distribution of cognitive presence phases in control and treatment groups.

ID	Phase	Messages					
		Control		Treatment		*Total*	
0	Other	71	8.40%	69	7.56%	140	8.01%
1	Triggering event	196	23.19%	112	12.10%	308	17.63%
2	Exploration	390	46.15%	294	33.39%	684	39.15%
3	Integration	152	17.98%	356	39.26%	508	29.08%
4	Resolution	36	4.28%	71	7.69%	107	6.13%
	Total	845	100%	902	100%	1,747	100.00%

Table 3. Number of students and messages across the two study conditions.

Condition	Course offering	Students	Messages
Control	Winter 2008	15	212
	Fall 2008	22	633
	Total	37	845
Treatment	Summer 2009	10	243
	Fall 2009	7	63
	Winter 2010	14	359
	Winter 2011	13	237
	Total	44	902

The complete study dataset consists of 1,747 discussion messages produced by 81 students that were coded by two expert coders using the content analysis instrument for cognitive presence [14]. The coders achieved an excellent level of agreement (percent agreement = 98.1%, Cohen's $\kappa = 0.974$) with a total of only 32 disagreements which were resolved through discussion. The distribution of cognitive presence phases for both control and treatment groups are shown in Table 2, while the details about course offerings are shown in Table 3. While both conditions produced a similar number of messages, the difference in the distribution of four phases of cognitive presence is clearly visible, in particular relating to the higher number of integration messages and a lower number of triggering event and exploration messages for the treatment group.

3.2 Analysis Procedure

Natural Language Processing. As the first step of our analysis, we used NLP techniques to detect topics from student discussion messages. In particular, we used Latent Dirichlet Allocation (LDA) [7], a widely used NLP method for topic modelling, which uses a statistical approach to discover prominent themes in the document corpora. More precisely, through the analysis of word co-occurrences, LDA models *topics* as statistical distributions across all words in the corpora (i.e., model vocabulary) and *documents* (i.e., posts) as distributions across all identified topics. LDA has been extensively used in social sciences [26] and humanities [1], including education [cf. 19], where it has frequently been used for problems such as the detection of topics in students' discussions [37], course evaluations [27], and public media discussions [20].

Before applying LDA, we pre-processed the dataset to enable for a more accurate topic detection, as commonly done in NLP [24]. We (1) converted all words into lowercase, (2) removed *stop words*, which were the words with little or no value for topic detection (e.g., prepositions, articles), (3) removed all words shorter than three characters and numbers, and (4) adopted a stemming algorithm to remove the derivational affixes of words [10].

It is important to mention that LDA requires the number of topics specified in advanced. While there are several methods to identify the optimal number of topics [34], since we analysed the corpus of manageable size (1,747 messages) from a well-designed course (see Sect. 3.1), it allowed us to evaluate several models with the number of topics close to the number of topics defined in the course syllabus (Table 1). In the end, 15 topics were identified as optimal, which corresponded to 14 content-related topics from the syllabus and one additional topic related to course logistics (gen.com). The output of LDA algorithm was a $1,747 \times 15$ matrix, with each row containing the relevance of all 15 topics for a given discussion message.

3.3 Epistemic Network Analysis

After we identified prominent topics in the discussion corpus, we examined the relationship between students' cognitive presence and course topics through Epistemic Network Analysis (ENA) [32]. ENA is a graph-based analysis technique which can be used to examine rich relationships between a set of concepts. In educational settings, ENA is typically used to examine the relationships between different elements in a coded dataset, such as coded discourse transcripts. For example, ENA can be used to examine students' cognitive connections during problem-solving [25], or the dynamics and interactions in students' group discourse [31]. Building on the theory of epistemic frames [29], ENA can be used to model complex domains as networks of connections among relevant constructs [30]. Unlike other network analysis tools, ENA was primarily designed for problems with a relatively small set of concepts characterized by highly dynamic and dense interactions. It can also be used to compare the differences between different groups of analysis units – such as between the control and treatment groups in the present study.

Within ENA, the connections among the different concepts (i.e., codes) are derived for each *analysis unit* (e.g., study subject) based on the concept co-occurrences in data subsets called *stanzas* (e.g., sentence, paragraph, document). From code co-occurrences, ENA first creates a high-dimensional representation, called *analytic space*, of all analysis units. The analysis units are then projected in a lower-representational space, called *projection space*, which is derived from analytic space through single-value decomposition. It should be noted that besides binary codes that represent presence/absence of a particular code, ENA can also be used for codes that represent strength or probability of a given code. In this case, the analytic space is not constructed from code co-occurrence, but the weighted product of codes' values; the weighting can be done as (1) direct product (*direct product* method), (2) square root of the direct product (*square root* method), or (3) natural log of the direct product (*natural log* method). In the end, the output of ENA is a series of graph models which capture the relationships between different coding categories [31], in our case four cognitive presence phases and course topics.

In the present study, each discussion message was coded with four binary codes capturing presence/absence of cognitive presence phases and fifteen codes representing the relevance of all extracted topics (0.00–1.00). Both units of analysis and stanzas were students (i.e., all student messages) within control and treatment groups. The use of students as units of analysis enabled us to see for each student his connections between phases of cognitive presence and the different topics, as well as between different phases of cognitive presence. Since our codes for topics were ratio variables measuring the presence of each topic in every message, we examined all three types of weighting (direct product, square root, and natural log) and conducted the statistical comparison between control and treatment groups in all three cases. However, given the space limitations, in the next section, we report only the network models created with the direct product weighting method. The results obtained with other two weighting methods were comparable.

4 Results

Figure 1 shows the projection of individual students' networks with relationships between cognitive presence and course topics. The analytic space was created using the *direct product* method (see Sect. 3.3) while visualisation was done using svd_1 and svd_2, which accounted for 30 and 18% of variability between students' network models, respectively. The differences between the students in the treatment and control groups can be observed visually, with the key difference along

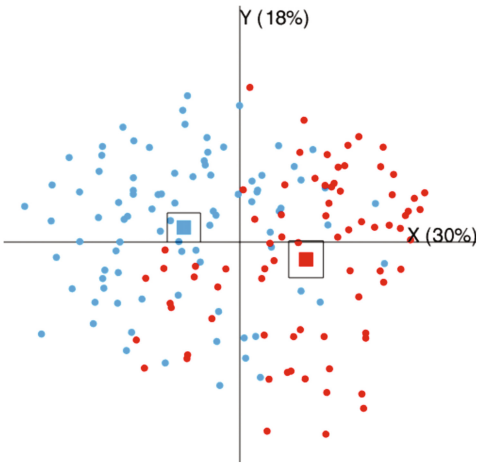
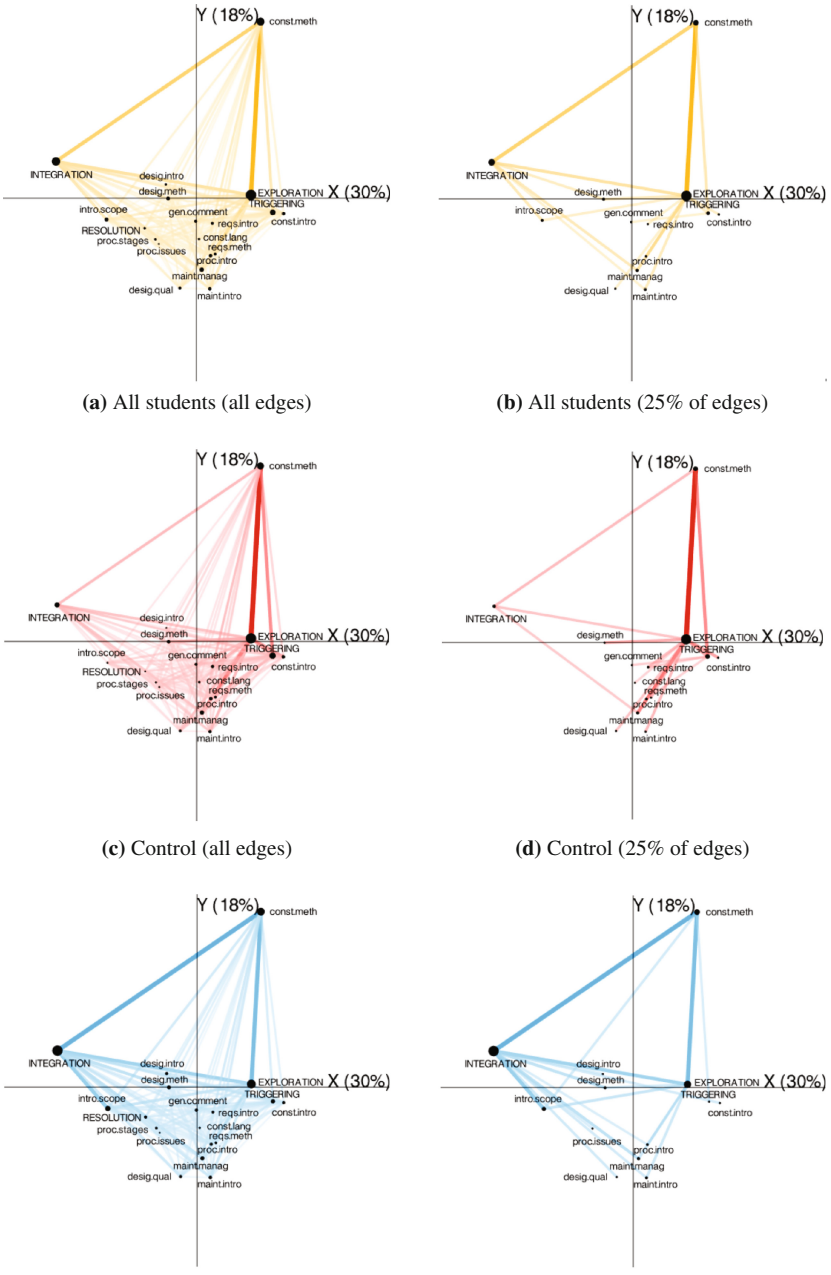


Fig. 1. ENA projection of students' networks between cognitive presence and course topics. The model is created using *direct product method* and shows first (svd_1) and second (svd_2) singular values, respectively axis X and Y. Networks of students from control (red) and treatment groups (blue) are shown as colored circles, while their group means are shown as colored squares (95% CI are outlined around the group means). (Color figure online)

the X-axis which represents the first singular value of the analytic space. The circles represent students in control (red) and treatment (blue) groups, while rectangles represent group-average networks (95% CI are also outlined). The group-average graphs for all students (Figs. 2a and b) indicate that svd_1 primarily distinguishes between students focusing more on early phases (triggering event and exploration) or later phases (integration and resolution) of cognitive presence. The exploration phase is especially related to the topic of software construction methodology, indicating the unique importance of this topic in the course and is primarily captured by the Y-axis.

The visual comparison of the graphs of the control and treatment groups (Figs. 2c and e) shows that the exploration phase was the principal focus of student discourse before the intervention in the course, whereas the discussion shifted towards the integration phase after the role-based and externally-facilitated regulation scaffolding was introduced. Not surprisingly, this is well aligned with the results reported by Gašević et al. [16] on the same dataset. The abbreviated network models for two groups (i.e., with the top 25% of the strongest edges, Figs. 2d and f) reveal the shift towards the integration phase more clearly, with far more topics being connected to the integration phase than before the intervention. Interestingly, in both abbreviated models, the resolution phase got removed, indicating smaller differences between the two groups with regards to the resolution phase. Finally, Fig. 3 which shows the subtracted graph (i.e., the difference between the two networks) and indicates the far more connections of the topics to the integration and resolution phases in the treatment group than in the control group; likewise, there is a higher number of links to the triggering event and exploration phases for the control group than in the treatment group.

In addition to visual examination of the differences between network models for control and treatment groups, we examined the statistical difference between network models for control and treatment groups using all three weighting methods for constructing analytic space (Table 4). Thus, independent-samples t-tests were conducted to compare the differences in students scores among the X and Y axes, using each of the three weighting methods. With respect to X-axis, there were statistically significant differences between the control and treatment groups using all three weighting methods ($p < .001$) with the average effect sizes of 1.56 Cohen's d which is considered a large effect size [8]. These findings suggest that there is on average 1.56 standard deviation a higher number of connections among cognitive presence and course topics in the treatment group than in the control group. With regards to Y-axis, the differences between the control and treatment groups were significant for the direct product and natural log methods $p < .05$ and with effect sizes of 0.43 and 0.35 Cohen's d, respectively. The observed differences can be considered small to medium effect sizes [8].



(a) All students (all edges)

(b) All students (25% of edges)

(c) Control (all edges)

(d) Control (25% of edges)

(e) Treatment (all edges)

(f) Treatment (25% of edges)

Fig. 2. Group-average ENA networks between cognitive presence and course topics. (Color figure online)

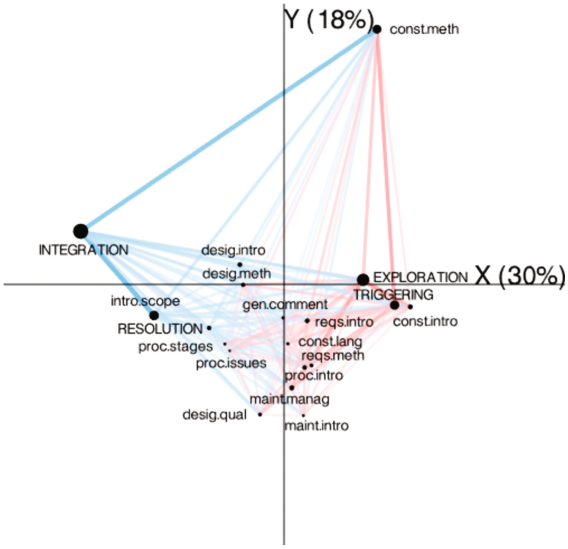


Fig. 3. Differences between control (red) and treatment (blue) group-average networks of cognitive presence and course topics. (Color figure online)

5 Discussion

5.1 Discussion of the Results in Relation to Research Questions

Based on the study results presented in the paper, we can see that the proposed approach provides important and novel insights into the development of cognitive presence with regards to the different course topics, answering the first research question. As expected, exploration and integration are the most prominent cognitive presence phases; this finding aligns well with their distribution (Table 2). However, the direct associations between triggering event and exploration phases and between exploration and integration phases provide additional qualitatively new insights into the students' cognitive presence development which are above the message-level which is provided by CoI content analysis instrument. While cognitive presence has been typically evaluated by looking at number of messages in each phase, the presented approach provides richer insights into the cognitive presence development, by showing the relationship between different cognitive presence phases that reveal the *quality* of the learning process and not its outcome (i.e., number of messages in each phase).

In terms of course topics, we see that software engineering construction methods were particularly well-represented in the data, while other topics received less attention. General course logistics also showed on the graph (Fig. 2a), primarily relating to triggering event phase. This is not surprising since they are typically procedural questions relating to course organization, expectations and activities.

Table 4. Comparison of networks models between control and treatment groups.

Weighting method	*X-Axis* (svd_1)			*Y-Axis* (svd_2)		
	Control N = 73	Treatment N = 90	*Cohen's d*	Control N = 73	Treatment N = 90	*Cohen's d*
Direct product	0.179**	−0.151**	1.580	−0.047*	0.039*	0.428
Square root	−0.153**	0.129**	1.546	−0.003	0.003	0.031
Natural log	0.158**	−0.133**	1.555	0.033*	−0.028*	0.348

Note: * indicates $p < .05$, ** indicates $p < .001$

Figure 2 shows that the X-axis primarily differentiates between early and later phases of cognitive presence. This is well aligned with the conceptualisation of cognitive presence [14] provided in Sect. 2. In this regard, the X-axis primarily differentiates across the deliberation-action dimension of the practical inquiry model that underlies the conceptualization of cognitive presence, while the Y-axis differentiates across the perception-awareness dimension. However, we can see that differences in this second dimension were much smaller than on the first one, which explains the higher portion of variability (30%).

As posted in the second research question, an essential aspect of our analysis focused on the effect of instructional scaffolding intervention on the students' cognitive presence development in different course topics. Our results showed the effect instructional intervention had, not just on an overall cognitive presence level, but also on particular course topics. The observed effect sizes for svd_1 (X-axis) differences were substantial, suggesting significant differences, in particular, relating to the frequencies of exploration and integration messages.

It should also be noted that with the move towards the integration phase, students also substantially changed the topics of their discussion. For example, process issues, software engineering scope, and introduction to design methodologies appear more prominently in the treatment group, whereas software construction process, software engineering requirements, and construction languages being more common before the intervention. Given that students had a choice of papers to present and discuss, the differences in topics are probably unsurprising. However, the navigation through the course and course expectations could be affected by the intervention. Indeed, messages related to the topic of course logistics were more prominent before the intervention, signalling that the instructional intervention was associated with the reduction of the students' need to seek help regarding course practicalities. Similar findings were also shown using subtracted network model (Fig. 3 albeit in a more condensed form).

5.2 Study Limitations

While this proposed approach shows a promise in addressing significant issues in the research around communities of inquiry, there are some important limitations of the present study which should be acknowledged. First, while we used the data from six-course offerings, the data is still from a single course at a single

institution which can negatively affect the broader usability of the presented approach and the generalizability of our study findings. Second, it is possible that the generalizability of the findings from the present study is somewhat limited, given the specifics of the adopted course design and instructional intervention. Finally, as most data mining analysis involves making many methodological decisions, such as deciding on various algorithm tuning parameters, it might be that the findings in our study would be different with a different set of tuning parameters.

5.3 Study Contributions and Conclusions

The primary contribution of the present study is a novel analytical method for assessment of cognitive presence concerning different course topics. Through NLP and graph-based analysis techniques, we showed how it is possible to provide more in-depth insights into students cognitive presence development with regards to different course topics. Moreover, by examining cognitive presence on the student level instead of the message level, we were able to gain rich understanding of students' cognitive presence development which goes beyond simple message counts. In practical terms, the presented method can be used to enable instructors better facilitation of students' course participation by analysing cognitive presence concerting different key course topics. Moreover, the tool can be used to assess the quality of students inquiry-based learning by examining students' connections between different cognitive presence phases.

Another substantial contribution of the present study is the examination of the effects of instructional scaffolding through externally-facilitated regulation and role assignment on students' cognitive presence of different course topics. While the general benefits of this instructional scaffolding in the same dataset were already explored by Gašević et al. [16], the present study shows how it impacted students' cognitive presence of different course topics. This and similar types of analyses have a strong potential to provide relevant research evidence on the benefits of different instructional interventions in social learning environments, where participation in asynchronous online discussions represents a principal learning activity.

Finally, there are some important ways in which this work can be extended and improved. In the future, we intend to investigate the trajectory shift from triggering phase to integration phase by exploring additional analytic features available in the epistemic network analysis. Using a similar approach as presented in this study, we also aim to examine social and teaching presences as well as their relationships. By doing this, we hope to provide a comprehensive analysis of student online learning experience which builds upon the potentials provided by the collected educational data.

References

1. Blei, D.M.: Topic modeling in digital humanities. J. Digit. Humanit. **2**(1) (2012). Special Issue
2. Akyol, Z., Garrison, D.R.: Assessing metacognition in an online community of inquiry. Internet High. Educ. **14**(3), 183–190 (2011). https://doi.org/10.1016/j.iheduc.2011.01.005
3. Anderson, T., Dron, J.: Three generations of distance education pedagogy. Int. Rev. Res. Open Distance Learn. **12**(3), 80–97 (2010). http://www.irrodl.org/index.php/irrodl/article/view/890/
4. Anderson, T., Rourke, L., Garrison, D.R., Archer, W.: Assessing teaching presence in a computer conferencing context. J. Asynchronous Learn. Netw. **5**, 1–17 (2001)
5. Arbaugh, J., Cleveland-Innes, M., Diaz, S.R., Garrison, D.R., Ice, P., Richardson, J.C., Swan, K.P.: Developing a community of inquiry instrument: testing a measure of the community of inquiry framework using a multi-institutional sample. Internet High. Educ. **11**(3–4), 133–136 (2008). https://doi.org/10.1016/j.iheduc.2008.06.003
6. Azevedo, R., Moos, D.C., Greene, J.A., Winters, F.I., Cromley, J.G.: Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? Educ. Technol. Res. Dev. **56**(1), 45–72 (2008). https://doi.org/10.1007/s11423-007-9067-0
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003). http://dl.acm.org/citation.cfm?id=944919.944937
8. Cohen, J.: The analysis of variance. In: Statistical Power Analysis for the Behavioral Sciences, pp. 273–406. L. Erlbaum Associates, Hillsdale (1988)
9. Dewey, J.: My pedagogical creed. Sch. J. **54**(3), 77–80 (1897)
10. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cambridge (2007)
11. Fischer, F., Kollar, I., Mandl, H., Haake, J.M.: Scripting Computer-Supported Collaborative Learning: Cognitive. Computational and Educational Perspectives. Springer Science & Business Media, New York (2007). https://doi.org/10.1007/978-0-387-36949-5
12. Garrison, D.R.: Cognitive presence for effective asynchronous online learning: the role of reflective inquiry, self-direction and metacognition. Elem. Qual. Online Educ. Pract. Dir. **4**(1), 47–58 (2003)
13. Garrison, D.R., Anderson, T., Archer, W.: Critical inquiry in a text-based environment: computer conferencing in higher education. Internet High. Educ. **2**(2–3), 87–105 (1999). https://doi.org/10.1016/S1096-7516(00)00016-6
14. Garrison, D.R., Anderson, T., Archer, W.: Critical thinking, cognitive presence, and computer conferencing in distance education. Am. J. Distance Educ. **15**(1), 7–23 (2001). https://doi.org/10.1080/08923640109527071
15. Garrison, D.R., Anderson, T., Archer, W.: The first decade of the community of inquiry framework: a retrospective. Internet High. Educ. **13**(1–2), 5–9 (2010). https://doi.org/10.1016/j.iheduc.2009.10.003
16. Gašević, D., Adesope, O., Joksimović, S., Kovanović, V.: Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. Internet High. Educ. **24**, 53–65 (2015). https://doi.org/10.1016/j.iheduc.2014.09.006
17. Heo, H., Lim, K.Y., Kim, Y.: Exploratory study on the patterns of online interaction and knowledge co-construction in project-based learning. Comput. Educ. **55**(3), 1383–1392 (2010). https://doi.org/10.1016/j.compedu.2010.06.012

18. Kovanović, V., Joksimović, S., Gašević, D., Hatala, M.: Automated cognitive presence detection in online discussion transcripts. In: Proceedings of the Workshops at the LAK 2014 Conference Co-located with 4th International Conference on Learning Analytics and Knowledge (LAK 2014), Indianapolis, IN (2014). http://ceur-ws.org/Vol-1137/
19. Kovanović, V., Joksimović, S., Gašević, D., Hatala, M., Siemens, G.: Content analytics: the definition, scope, and an overview of published research. In: Lang, C., Siemens, G., Wise, A., Gašević, D. (eds.) Handbook of Learning Analytics and Educational Data Mining, pp. 77–92. SoLAR, Edmonton (2017). https://doi.org/10.18608/hla17.007
20. Kovanović, V., Joksimović, S., Gašević, D., Siemens, G., Hatala, M.: What public media reveals about MOOCs: a systematic analysis of news reports. Br. J. Educ. Technol. **46**(3), 510–527 (2015). https://doi.org/10.1111/bjet.12277
21. Kovanović, V., Joksimović, S., Poquet, O., Hennis, T., Čukić, I., de Vries, P., Hatala, M., Dawson, S., Siemens, G., Gašević, D.: Exploring communities of inquiry in massive open online courses. Comput. Educ. **119**, 44–58 (2018). https://doi.org/10.1016/j.compedu.2017.11.010
22. Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., Siemens, G.: Towards automated content analysis of discussion transcripts: a cognitive presence case. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK 2016), pp. 15–24. ACM, New York (2016). https://doi.org/10.1145/2883851.2883950
23. Lipman, M.: Thinking in Education. Cambridge University Press, Cambridge (1991)
24. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing, vol. 999. MIT Press, Cambridge (1999)
25. Nash, P., Shaffer, D.W.: Mentor modeling: the internalization of modeled professional thinking in an epistemic game. J. Comput. Assist. Learn. **27**(2), 173–189 (2011). https://doi.org/10.1111/j.1365-2729.2010.00385.x
26. Ramage, D., Rosen, E., Chuang, J., Manning, C.D., McFarland, D.A.: Topic modeling for the social sciences. In: NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond, Whistler, Canada (2009)
27. Reich, J., Tingley, D., Leder-Luis, J., Roberts, M.E., Stewart, B.: Computer-assisted reading and discovery for student generated text in massive open online courses. J. Learn. Analytics **2**(1), 156–184 (2014). http://epress.lib.uts.edu.au/journals/index.php/JLA/article/view/4138
28. Rourke, L., Anderson, T., Garrison, D.R., Archer, W.: Assessing social presence in asynchronous text-based computer conferencing. J. Distance Educ. **14**(2), 50–71 (1999). http://www.ijede.ca/index.php/jde/article/view/153
29. Shaffer, D.W.: Epistemic frames for epistemic games. Comput. Educ. **46**(3), 223–234 (2006). https://doi.org/10.1016/j.compedu.2005.11.003
30. Shaffer, D.W.: Epistemic frames and islands of expertise: learning from infusion experiences. In: Proceedings of the 6th International Conference on Learning Sciences, ICLS 2004, pp. 473–480. International Society of the Learning Sciences, Santa Monica (2004). http://dl.acm.org/citation.cfm?id=1149126.1149184
31. Shaffer, D.W., Collier, W., Ruis, A.R.: A tutorial on epistemic network analysis: analyzing the structure of connections in cognitive, social, and interaction data. J. Learn. Analytics **3**(3), 9–45 (2016). https://doi.org/10.18608/jla.2016.33.3

32. Shaffer, D.W., Hatfield, D., Svarovsky, G.N., Nash, P., Nulty, A., Bagley, E., Frank, K., Rupp, A.A., Mislevy, R.: Epistemic network analysis: a prototype for 21st-century assessment of learning. Int. J. Learn. Media **1**(2), 33–53 (2009). https://doi.org/10.1162/ijlm.2009.0013
33. Stahl, G., Koschmann, T., Suthers, D., Sawyer, R.K.: Computer-supported collaborative learning: a historical perspective. In: The Cambridge Handbook of the Learning Sciences, pp. 409–426. Cambridge University Press, Cambridge, New York (2006)
34. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning, ICML 2006, pp. 977–984. ACM, New York (2006). https://doi.org/10.1145/1143844.1143967
35. Waters, Z., Kovanović, V., Kitto, K., Gašević, D.: Structure matters: adoption of structured classification approach in the context of cognitive presence classification. In: Zuccon, G., Geva, S., Joho, H., Scholer, F., Sun, A., Zhang, P. (eds.) AIRS 2015. LNCS, vol. 9460, pp. 227–238. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-28940-3_18
36. Winne, P.H., Hadwin, A.F.: Studying as self-regulated learning. In: Hacker, D.J., Dunlosky, J., Graesser, A.C. (eds.) Metacognition in educational theory and practice. The Educational Psychology Series, pp. 277–304. Lawrence Erlbaum Associates Publishers, Mahwah (1998)
37. Yang, D., Wen, M., Rose, C.: Towards identifying the resolvability of threads in MOOCs. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 21–31 (2014). http://www.aclweb.org/anthology/W/W14/W14-41.pdf#page=28