What is the Effect of a Dominant Code in an Epistemic Network Analysis?

Rafael Ferreira Mello[1](✉) and Dragan Gašević[2]

[1] Universidade Federal Rural de Pernambuco, Recife, Brazil
rafael.mello@ufrpe.br
[2] Monash University, Clayton, Australia
dragan.gasevic@monash.edu

Abstract. This paper investigates how different configuration of epistemic network analysis parameters influence the examination of student interactions in asynchronous discussions in online learning environments. Specifically, the paper investigates strategies for dealing by unintended consequences of a dominant node in epistemic network analysis (ENA). In particular, the paper reports on a study that explored the effects of two different strategies including (i) the use of different dimensions calculated with singular value decomposition (SVD), and (ii) exclusion of a dominant code. Our results showed that the use of different SVDs did not change the influence of a dominant code in the graph. On the other hand, the exclusion of the dominant code led to an entirely different configuration in ENA. The practical implications of the results are further discussed.

Keywords: Epistemic network analysis · Dominant code · Graph analysis

1 Introduction

The technological advancements experienced over the past decade led to the increased adoption of digital learning environments that support online and blended learning [21]. These environments provide several tools that enable interactions between instructors and students. Among these, asynchronous online discussions represent one of the most commonly adopted tools for supporting social interactions and social-constructivist pedagogies [1]. While online discussions are widely used to support student learning, there are many challenges associated with their effective use by teachers and students. As with any learning tool, the simple provision of the tool does not mean that students will know how to use it, and more importantly, how to use it effectively [8].

In this regard, the Community of Inquiry (CoI) model [14] represents a pedagogical model that seeks to understand how asynchronous online communication shapes student learning and cognitive development. The CoI model defines three dimensions (called presences) that together provide a holistic overview of online learning experience: (1) Cognitive presence captures the development of desirable learning outcomes such as critical thinking and knowledge (co-)construction, (2) Social presence outlines the essential role of humanizing relationships among online course participants, and

B. Eagan et al. (Eds.): ICQE 2019, CCIS 1112, pp. 66–76, 2019.
https://doi.org/10.1007/978-3-030-33232-7_6

(3) Teaching presence describes the instructors' role before (i.e., course design) and during (i.e., facilitation and direct instruction) the course.

Several studies have attempted to analyze online discussions under the CoI perspective using manual content analysis [6], natural language processing techniques [11], social network analysis [7], among others. Although they provide valuable results, the existing research still lacks a more qualitative understanding of the discussions. More recently, several works [4, 13] proposed the adoption of Epistemic Network Analysis (ENA) [18] to produce in-depth insights in communities of inquiry. For instance, Rolim et al. [13] demonstrated how ENA could be used to identify links between indicators of cognitive and social presence across different groups of learners. Rolim et al. [13] also showed how the development of the relationship between social and cognitive presence changes over time in an online discussion.

ENA offers a promising direction to evaluate social interactions, with increasing adoption numbers [2, 4, 10]. Nevertheless, it is necessary to define several configurations in order to apply ENA adequately. An ENA configuration includes parameters related to the domain of the application, which the user needs to configure depending on the application (e.g., unit of analysis and codes), and the method, which can be changed in order to improve the interpretability of the results (e.g., dimensions identified through singular value decomposition). However, there have been fewer known examples of how changes in the ENA parameters could affect results and the interpretation of the results.

This paper presents the results of an exploratory study about the impact of different parameter configurations, related to the method, on the analysis of student interactions analyzed in terms of the CoI model. Specifically, the study analyzed the approaches that can be used to mitigate the impact of the dominance of a node in an epistemic network. The approaches included the use of different dimensions obtained with singular value decomposition (SVD) and the removal of the dominant code.

2 Background

Epistemic Network Analysis (ENA) [18] is a graph-based analysis technique, built on the theory of epistemic frames [17] and used to investigate associations between concepts. It was created to analyze problems with a relatively small set of concepts characterized by highly dynamic and dense interactions. The first application was to evaluate the student progress on epistemic games [19], which was an environment that simulates novices training to be professionals, further called virtual internships. Within this environment, the students were requested to interact with peers and mentors to develop prototypes for fictitious companies producing content, which was automatically analyzed with ENA.

ENA investigates the relationships among different concepts (called codes) for each analysis unit (e.g. individual students). Two codes are considered related if they appear in the same chunk of text, called stanza, which in [19], is the message of each student within the virtual internships environment. One or more stanzas can be collapsed into a single conversation (stanza window) in order to be analyzed together. Unlike other network analysis techniques, ENA was primarily designed for problems with a relatively small set of concepts characterized by highly dynamic and dense interactions.

To visualize epistemic networks, a two-dimensional representation of the analytics space, called projection space, is derived through a dimensionality reduction algorithm called singular value decomposition (SVD). Moreover, the networks of code relationships for a particular analysis unit (or group of analysis units) are also visualized as undirected graphs. It can also be used to compare the differences between different groups of analysis units in a visualization called subtraction network.

ENA has been largely used in educational settings [3, 4, 7, 10]. For instance, it can be used to examine students' cognitive connections during problem-solving [10], or dynamics and interactions in students' group discourse [3]. In general, ENA is used to investigate the associations between codes of a coding scheme (e.g., the topics extracted from students interactions in virtual internships [19]) where the coding scheme is applied to analyze transcripts of online discussions [3, 7].

Recently, Ferreira et al. [4] and Rolim et al. [12] proposed the adoption of ENA to assess the relationship between cognitive and social presence with the course topics within communities of inquiry. These papers qualitatively investigated the difference between two different groups of students under the perspective of ENA graphs. With this analysis, it was possible to investigate not only the statistical differences between groups of students but also the main characteristics of the students that led to these differences. Moreover, Rolim et al. [13] used ENA to investigate the connections between cognitive and social presence and how those developed over time.

Besides the application of ENA to several domains, the literature reports works related to the configuration of the analysis. For instance, Siebert et al. [20] highlighted the importance of adopting moving windows to established links between codes in ENA. The study showed that for groups interactions (e.g., student teams), the moving windows captured more information than the use of a single utterance as a stanza. Furthermore, Ruis et al. [15] reported the results of an experimental study which revealed that more relevant connections were captured when a moving window with a length of seven was adopted.

However, as ENA is a relatively new analytic technique, several aspects could be further explored in order to improve the readability and effectiveness of its results. In this paper, we address the issue of having a dominant code in an ENA model.

3 Problem Statement and Research Questions

Although [20] and [15] studied the influence of the moving window in the outcome of ENA, other parameters should be taken into consideration. For instance, there is a lacuna in the literature that investigates the impact of the usage of different codes and dimensions calculated through SVD on ENA. Specifically, in this work, we studied different parameter configurations in ENA when there are one or more dominant codes.

In this study, we consider a dominant code when it possesses one or more of the following characteristics:

- It has many more instances, at least the double, than the other codes in dataset used in ENA;
- It is the most connected node in the network according to the results of ENA;
- It is the main factor that explains one axis of the final graph generated by ENA.

This paper reports on a study that reproduced the same analysis performed by Rolim et al. [13], but with different input parameters to avoid dominant codes. The first strategy to reach this goal was the evaluation of different dimensions identified with SVD. As a hypothesis, we intend to evaluate if a different combination of dimensions of SVD could eliminate the code dominance in terms of explanation of one axis. As such, our first research question was:

Research Question 1:
What is the effect of using different dimensions of SVD on the final network when dealing with the problem of a dominant code?

In addition to examining the different SVD combinations, we were interested in exploring whether the exclusion of a dominant code (with more instances and connections) could provide additional insights into the outcome of an ENA. Thus, our second research question was:

Research Question 2:
Does the exclusion of a dominant code impact the information provided by the ENA?

4 Method

4.1 Data

The data used consisted of six offerings of a graduate level research-intensive online course in software at a Canadian public university between 2008 and 2011. In those six offerings, a total of 81 students posted 1,747 messages. As part of the assessment, the students were required to select one research paper on a course topic, record a video presentation, and post a URL to a new course online discussion, in which the 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. These students composed the control group in this study, which consisted of 37 students who produced 845 messages. After the first two course offers, the scaffolding of discussion participation through role assignments and clear instructions were adopted. In total, 44 students (treatment group) were exposed to this instructional intervention and produced a total of 902 messages [6].

The dataset was coded according to the indicators of social presence and the phases of cognitive presence. Initially, the coders annotated 1 and 0 for the presence and absence of an indicator of social presence following the scheme defined by Rourke et al. [14]. The coders achieved a high level of agreement, with all of the indicators reaching a percentage of agreement of at least 84% [9]. It is important to highlight that each message could have more than one indicator. Thus, the final number of annotations was 3,770, instead of 1,747 (the total number of messages). Besides, some of the indicators (i.e., Continuing a thread, Complementing, and Vocatives) were excluded from the analysis due to a disproportionately large number of messages with such codes [9]. Table 1 details the distribution of messages per indicator.

Table 1. Distribution of social presence indicators [9]

Category	Indicator	Messages	Percentages
Affective	Expression of emotions	288	16.5%
	Use of humor	44	2.52%
	Self-disclosure	322	18.4%
Interactive	Continuing a thread	1,664	95.2%
	Quoting from others messages	65	3.72%
	Referring explicitly to other's messages	91	5.21%
	Asking questions	800	45.8%
	Complementing, expressing appreciation	1391	79.6%
	Expressing agreement	243	13.9%
Cohesive	Vocatives	1,433	82%
	Addresses or refers to the group using	144	8.24%
	inclusive pronouns		
	Phatics, salutations	1,281	73.3%

Cognitive presence was coded in one of the four phases of cognitive presence and the absence of any was coded as other. Initially, the messages were coded by two expert coders according to the phases of cognitive presence as suggested by Garrison et al. [5]. The coders achieved an excellent level of agreement for both presences reaching a (percentage of agreement = 98.1%, Cohens k = 0.974) with a total of only 32 disagreements which were resolved through discussion. Table 2 presents the number of messages coded into the cognitive presence's phases.

Table 2. Distribution of cognitive presence.

ID	Phase	Messages	Percentages
0	Other	140	8.01%
1	Triggering event	308	17.63%
2	Exploration	684	39.15%
3	Integration	508	29.08%
4	Resolution	107	6.13%
Total		1,747	100.00%

4.2 Epistemic Network Analysis

This paper addresses the problem of a dominant code using the work proposed by Rolim and colleagues [13]. Rolim et al. propose the adoption of ENA to analyze the relationship between social and cognitive presences in asynchronous online discussions.

The ENA initial configuration used binary codes representing the presence and the absence of the social presence indicators and the 5 categories (4 phases + others) of cognitive presence as ENA codes. The units of analysis and stanzas were individual students and students' discussion messages, respectively.

This configuration was also used to address the research questions studied in this work. Initially, we explored different configurations of SVD dimensions of for the mean network of all students together. We decided to investigate all the combinations of SVD1, SVD2 and SVD3 dimensions because they each explained a variance higher than 10%. Then, we evaluated a different configuration removing the most dominant codes. Finally, we analyzed the students in the treatment/control groups before and after the removal of the most dominant codes. The differences between the student groups on different configurations of SVDs were then compared by using a series of the Mann-Whitney [16] with $\alpha = 0.05$. The subtraction network was used to explain the qualitative differences between the student groups.

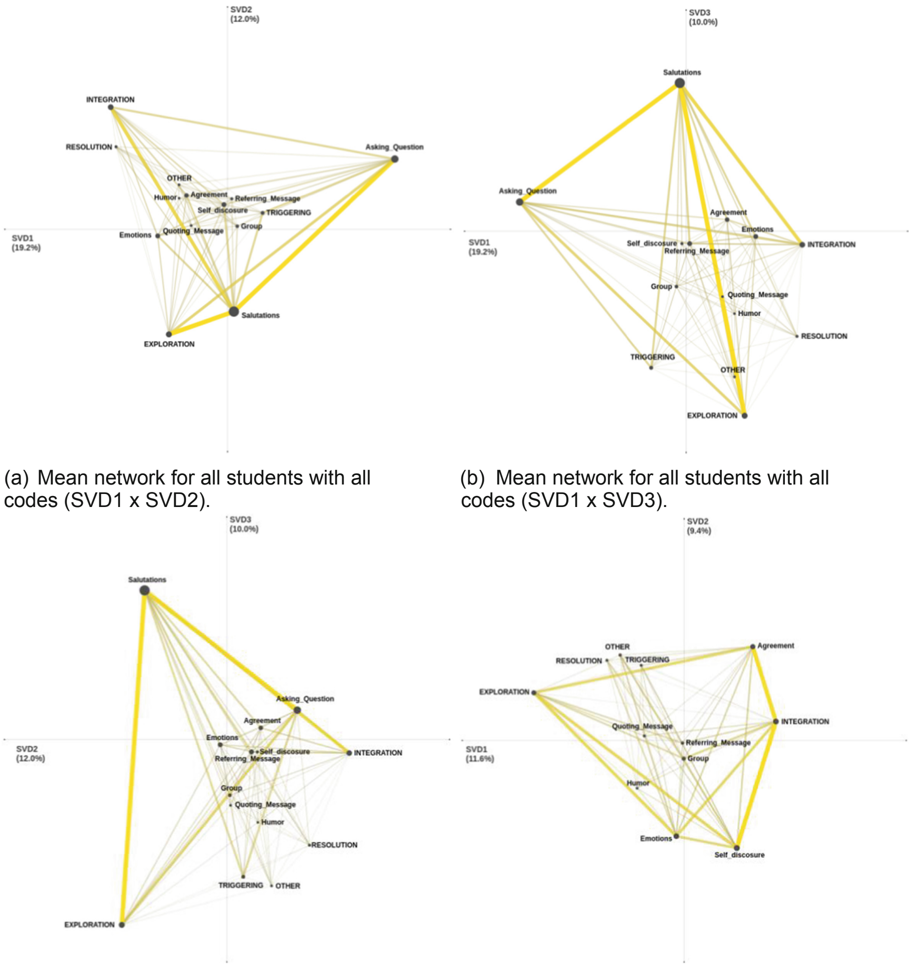
5 Results

As described before, for the purpose of this analysis, the dominant code has more instances and is highly connected in comparisons to other codes, or it is the main factor that explains one axis of the ENA.

Figure 1(a) presents the mean network produced by original ENA configuration for all students in our dataset using SVD1 and SVD2. It reveals the predominance of the codes salutation (higher number of instances and connections) and asking question (which explains the right side of SVD1). The rest of the indicators of social presence were plotted at the center of the graph. This configuration could bias the final analysis as these codes are dominating the analysis regarding the social presence.

Figures 1(b) and (c) show the networks for SVD1 x SVD3 and SVD2 x SVD3, respectively. In these alternative networks, all social presence indicators continue to be plotted in the middle of the graph while asking question and salutation have a predominant position. Finally, Fig. 1(d) presents the same network without the dominant codes. In contrast to the previous graphs, this one shows the indicators of social presence well divided into the network. On the other hand, cognitive presence changed from a well-distributed arrangement to a concentrated plot where four out of five categories were in the same quadrant.

We evaluated the differences between control and treatment groups in all configuration. For the initial three set-ups, different SVDs configuration, the statistical analysis reached the same result $U = 2165.50$; $p = 0.001$; $r = 0.37$ using Mann-Whitney test. On the other side, there was a slightly change in the results after the removal of the dominant code ($U = 2327.00$; $p = 0.001$; $r = 0.32$), but the difference



(a) Mean network for all students with all codes (SVD1 x SVD2).

(b) Mean network for all students with all codes (SVD1 x SVD3).

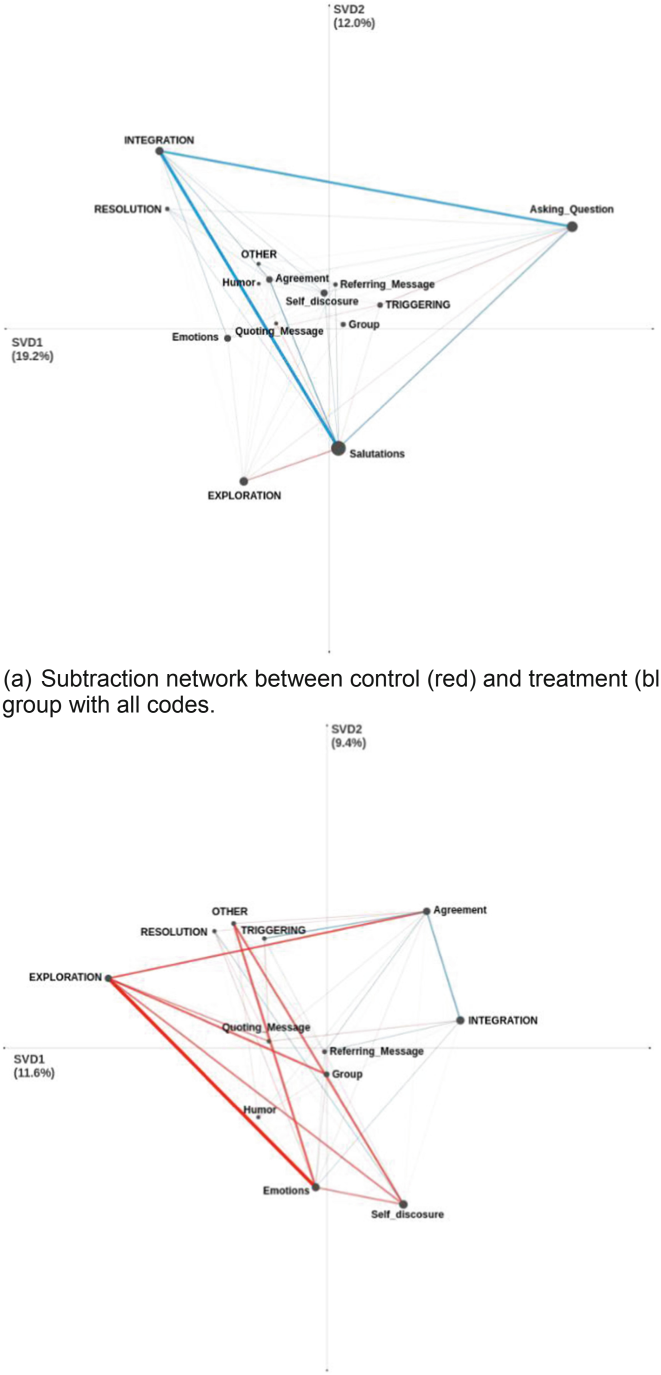
(c) Mean network for all students with all codes (SVD2 x SVD3).

(d) Mean network for all students without Salutation and Asking Question (SVD1 x SVD2).

Fig. 1. ENA network with different parameters configuration.

between the groups continued to be statistically significant. Figures 2(a) and (b) present the main difference, between the ENA original version (with SVD1 x SVD2) and the version without the dominant codes, in more details using subtraction networks between control (red) and treatment (blue) groups.

The analysis of the network graphs considering only the connections showed that in Fig. 2(a) the treatment group had stronger connections, while Fig. 2(b) presents that the control group had more dominant links. Moreover, these figures presented different behaviors related to the ENA codes, while the Fig. 2(a) has a distribution of cognitive presence over different quadrants, in Fig. 2(b) the social presence is more spread over the graph.



(a) Subtraction network between control (red) and treatment (blue) group with all codes.

(b) Subtraction network between control (red) and treatment (blue) group without Salutation and Asking Question.

Fig. 2. ENA subtraction network with different parameters configuration. (Color figure online)

6 Discussion

The initial set-up of the ENA (Fig. 1(a)) led to the interpretation of the graphs only based on the categories of cognitive presence, as they were well-distributed across the different quadrants of the final graph. Thus, the evaluation of different configurations of the ENA could increase the readability of the final graph.

The study showed that even by changing the SVDs, the codes asking question and salutation remained to be dominant. The rest of the indicators of social presence continue to be plotted at the center of the graph. This result indicated that the replacement of the SVD dimensions did not change significantly the interpretation of the original results which were drawn by plotting SVD1 and SVD2. In addition to not reducing the effect of the dominant code, this approach of using other SVD dimensions (SVD1 x SVD3 and SVD2 x SVD3) decreased the variance explained by each axis used in the diagrams. On the other hand, after excluding the dominant codes, the ENA graph improved the presentation of the distribution of the codes related to social presence, as they are plotted in different quadrants (Fig. 1(d)) of the graph instead of just the middle as presented in Fig. 1(a). This new graph allows a better analysis about what is happening to social interactions on the online discussion.

The comparison of the control and treatment groups showed the impact on the final ENA network after the of removing dominant codes. Although the differences between the groups continued to be significant, the subtraction network revealed a shift in the most dominant relationships among codes from the treatment group to the control group. This happened because the students in the treatment group increased the numbers of the integration and resolution messages with the asking question and salutation indicators. Thus, the removal of these two codes (asking question and salutation) decreased the impact of the cognitive presence phases on the results. Figures 1(d) and 2 (b) groups exploration, resolution, triggering and other categories in the top-left corner of the graph. This grouping reduced the impact of the codes of cognitive presence and emphasized the role of the codes social presence in the ENA graphs.

Therefore, the results reveal that the best approach to deal with the problem posed in this study is the removal of the dominant code. However, these findings need to be further validated on different datasets.

7 Final Remarks

This paper analyzed the behavior of student interaction in online discussions under the perspective of the community of inquiry framework using different configurations of ENA parameters. The results showed that the variation of SVD dimensions had low influence on the final results. The removal of dominant codes presented a high impact on the analysis of different groups performance. The key implication of our results is that the removal of the codes can lead to an improved understanding of one of the two key constructs in the CoI model (social presence) but reduced insights into the other one (cognitive presence).

Based on these results, we can suggest for studies that aim to look at the holistic associations of different constructs (in our study - social and cognitive presences), the approach that removes dominant codes is undesirable. However, when the insight into the construct that had dominant codes in epistemic networks (i.e., social presence) is the goal, the removal approach could lead to plausible results.

There are some limitations of the present study which should be acknowledged. First, the data is from a single course at a single institution, although from six-course offerings. This can negatively affect the degree of potential generalization of the analysis. Besides, the method required many methodological decisions, such as deciding on the coding used for cognitive and social presences and the parameters that we variate. It might be the case that different decision would lead to different findings. To address these limitations, we intend to perform the same analysis using data from another scenario and applying different parameters as input to the ENA.

As future work, we intend to: (i) reproduce the same analysis using a different context of the application, such as data from other courses, and analyze other phenomena different from the model of community of inquiry; (ii) explore what is the impact of change of other ENA parameters such as the length of the sliding window for stanza, and the order which the instances were presented in the input file; and (iii) apply statistical tools to determine whether two graphs, for instance graphs generated with different SVDs, are significantly different, following for instance the methodology used by Swiecki and colleagues [22].

References

1. Anderson, T., Dron, J.: Three generations of distance education pedagogy. Int. Rev. Res. Open Distrib. Learn. **12**(3), 80–97 (2011)
2. Arastoopour, G., Shaffer, D.W., Swiecki, Z., Ruis, A., Chesler, N.C.: Teaching and assessing engineering design thinking with virtual internships and epistemic network analysis. Int. J. Eng. Educ. **32**(2), 1492–1501 (2016)
3. Cai, Z., Eagan, B., Dowell, N., Pennebaker, J., Shaffer, D., Graesser, A.: Epistemic network analysis and topic modeling for chat data from collaborative learning environment. In: Proceedings of the 10th International Conference on Educational Data Mining (2017)
4. Ferreira, R., Kovanović, V., Gašević, D., Rolim, V.: Towards combined network and text analytics of student discourse in online discussions. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10947, pp. 111–126. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93843-1_9
5. 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
6. 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)
7. Gašević, D., Joksimović, S., Eagan, B.R., Shaffer, D.W.: SENS: network analytics to combine social and cognitive perspectives of collaborative learning. Comput. Hum. Behav. **92**, 562–577 (2019)
8. Gašević, D., Mirriahi, N., Dawson, S., Joksimović, S.: Effects of instructional conditions and experience on the adoption of a learning tool. Comput. Hum. Behav. **67**, 207–220 (2017)

9. Kovanovic, V., Joksimovic, S., Gasevic, D., Hatala, M.: What is the source of social capital? the association between social network position and social presence in communities of inquiry (2014)
10. 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)
11. Neto, V., et al.: Automated analysis of cognitive presence in online discussions written in Portuguese. In: Pammer-Schindler, V., Pérez-Sanagustín, M., Drachsler, H., Elferink, R., Scheffel, M. (eds.) EC-TEL 2018. LNCS, vol. 11082, pp. 245–261. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98572-5_19
12. Rolim, V., Ferreira, R., Kovanović, V., Gašević, D.: Analysing social presence in online discussions through network and text analytics. In: 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), vol. 2161, pp. 163–167. IEEE (2019)
13. Rolim, V., Ferreira, R., Lins, R.D., Gǎsević, D.: A network-based analytic approach to uncovering the relationship between social and cognitive presences in communities of inquiry. Internet High. Educ. **42**, 53–65 (2019)
14. Rourke, L., Anderson, T., Garrison, D.R., Archer, W.: Assessing social presence in asynchronous text-based computer conferencing (1999)
15. Ruis, A., Siebert-Evenstone, A., Pozen, R., Eagan, B.R., Shaffer, D.W.: A method for determining the extent of recent temporal context in analyses of complex, collaborative thinking. In: International Society of the Learning Sciences, Inc. [ISLS] (2018)
16. Ruxton, G.D.: The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. Behav. Ecol. **17**(4), 688–690 (2006)
17. Shaffer, D.W.: Epistemic frames for epistemic games. Comput. Educ. **46**(3), 223–234 (2006)
18. Shaffer, D.W., Collier, W., Ruis, A.: A tutorial on epistemic network analysis: analyzing the structure of connections in cognitive, social, and interaction data. J. Learn. Anal. **3**(3), 9–45 (2016)
19. Shaffer, D.W., et al.: Epistemic network analysis: a prototype for 21st-century assessment of learning. Int. J. Learn. Media **1**(2), 1–21 (2009)
20. Siebert-Evenstone, A.L., Irgens, G.A., Collier, W., Swiecki, Z., Ruis, A.R., Shaffer, D.W.: In search of conversational grain size: modeling semantic structure using moving stanza windows. J. Learn. Anal. **4**(3), 123–139 (2017)
21. Siemens, G., Gašević, D., Dawson, S.: Preparing for the digital university: A review of the history and current state of distance, blended, and online learning (2015)
22. Swiecki, Z., Ruis, R., Shaffer, D.: Does order matter? investigating sequential and cotemporal models of collaboration. In: 13th International Conference on Computer-Supported Collaborative Learning (2019)