Using Process Mining (PM) and Epistemic Network Analysis (ENA) for comparing processes of collaborative problem regulation

Nadine Melzner[1][0000-0002-8801-1016], Martin Greisel[2][0000-0002-9586-5714], Markus Dresel[3][0000-0002-2131-3749] and Ingo Kollar[4][0000-0001-9257-5028]

[1] Universität Augsburg, Universitätsstraße 10, 86159 Augsburg, Germany
nadine.melzner@phil.uni-augsburg.de

Abstract. Learning Sciences research often concerns the analysis of data from individual or collaborative learning processes. For the analysis of such data, various methods have been proposed, including Process Mining (PM) and Epistemic Network Analysis (ENA). Both methods have advantages and disadvantages when analyzing learning processes. We argue that a concerted use of both techniques may provide valuable information that would be obscured when using only one of these methods. We demonstrate this by applying PM and ENA on data from a study that investigated how students regulate collaborative learning when faced with either motivational or comprehension-related problems. While PM showed that collaborative learners are more incoherent (i.e. more heterogeneous in their chosen activities) when regulating motivational problems than comprehension-related problems at the beginning, ENA revealed that in later stages of their learning process, they focus on fewer activities when being confronted with motivational than with comprehension-related problems. Thus, a combination of the two approaches seems to be warranted.

Keywords: Epistemic Network Analysis, Process Mining, Self-Regulation, Collaborative Learning, Co-Regulation, Shared Regulation.

1 Problem Statement

Learning Sciences research is often concerned with the analysis of how learning processes emerge over time [1]. Traditionally, research typically used a coding-and-counting approach to analyze such processes (e.g., summing up frequencies by which learners employ certain strategies).

However, the problem with this routine is that it does not account for the dynamics of the learning process, i.e. for the fact that learners' engagement in different learning processes may change over time. Researchers have thus called for methods that consider learning processes in their temporal sequence [2]. One approach to do this is to use process mining (PM). PM uses mathematical algorithms to inductively discover

sequences of processes in event traces by visualizing them in process models. Based on Petri nets, process models are illustrations of systematically connected codes and transitions between codes and serve to uncover hidden information on the processes of interest.

For instance, using PM, [3] found that successful self-regulators initially prepare their learning before deeply processing information, whereas less successful learners did not show the described shift towards in-depth information processing. Yet, PM also has limitations such as partially producing "spaghetti-like" models that run the danger of becoming too complex for visual comparison. Additionally, PM does not provide statistical tests that check for differences between processes of different groups on a global level [4]. Furthermore, PM includes the individual activities of all subjects as an influence on the same process model with equal weighting. For example, if we want to investigate how groups regulate motivational as opposed to comprehension-related problems, it might be that one single person may be accountable for most loops on a single code in one situation (e.g., the person repeatedly applies an elaboration strategy), whereas in the other situation, such loops might be more evenly distributed across persons. If these loops are not weighted, this may lead to a distorted picture of the regulatory differences between situations resp. between different groups.

An approach that may help overcome these challenges is Epistemic Network Analysis (ENA) – a network analysis method based on a dimensional reduction procedure for tagging, extracting, and plotting meaningful compounds of activities by considering regulation processes as a network of coherent activities [5]. Some of its advantages are that ENA provides global statistical tests to compare models from different conditions (e.g., regulation processes in groups that experience motivational vs. regulation processes in groups that experience comprehension-related problems), that it provides information on the relatedness of codes within a specific window size that can consider more than just two successive codes (as it is the case in PM), and routines such as rotating networks in space for visually highlighting group differences, or normalizing vectors to check whether differences between two models are caused by single individuals within a group, but rather by a concerted (i.e., more or less evenly distributed) effort of the group.

ENA has lately been used to analyze data from a wide variety of different contexts ([6]; [7]). Despite its advantages, though, ENA still faces challenges: Since the networks drawn by ENA are based on so called adjacency matrices that include sums of counted code-code connections, it ignores start and end points and self-loop information. Additionally, it simply highlights connections between certain activities, rather than the direction of transitions. Thus, when visually and statistically comparing two models with ENA, these characteristics are not considered. Given the mutual strengths and weaknesses of the two approaches, we argue that a concerted use of PM and ENA might help to better understand the temporal structure of learning processes. We test this assumption by applying both methods to the analysis of data from a study on how learners cope with different kinds of collaborative regulation problems.

2 Method

2.1 Participants and Design

N=82 students (61 female, M_{Age}=21.79, SD_{Age}=4.86) who were on average in their 2nd semester (M_{Stud}=2.12, SD_{Stud}=0.57) of studies participated in this study. They received a booklet with four vignettes (in randomized order) that described a self-organized study group preparing for an exam that faced different kinds of regulation problems. One of the four vignettes described the group as experiencing no regulation problems at all, another one described the group as experiencing solely motivational problems, a third one said the group would experience solely comprehension-related problems, and a fourth one describing the group as experiencing both motivational and comprehension-related problems. That way, we established a 2x2-factorial within-subjects design with the independent factors "motivational problems" (with vs. without) and "comprehension-related problems" (with vs. without).

For example, in the condition "motivational problems", the vignette read: "*Imagine you are part of a study group with three fellow students. You meet regularly and are a well-rehearsed team. Currently, you prepare with your group for an exam that is in three weeks. Concerning the content to be learnt for the exam, all group members have high knowledge and low learning motivation*". In the vignette "comprehension-related problems", for example, "high knowledge" and "low learning motivation" were turned into "*low knowledge*" and "*high learning motivation*".

Due to lack of space, in this paper we focus our analysis on the conditions "with motivational problems/without comprehension-related problems" and "with comprehension-related problems/without motivational problems".

2.2 Variables

After each vignette, students received open-ended questions that asked them to indicate (a) what *types of strategies* they would apply if they were a member of the group, and (b) at what *social level* they would apply each of those strategies.

To measure the *types of strategies*, after each vignette, participants had to write down the exact sequence of actions they would perform to ensure high quality of learning in each situation (1. At first…, 2. After that…, 3. After that…, After that…, and so on). Open answers were coded by means of a coding scheme based on strategy classification schemes of [8] and [9]. This coding scheme differentiated between 1. elaboration strategies, 2. surface-oriented strategies, 3. metacognitive strategies, 4. resource-oriented motivational strategies, 5. resource oriented-non motivational strategies, 6. other strategies, and 7. no strategies (see Table 1). Two independent coders rated ten percent of the data and reached a sufficient level of interrater reliability (Cohen's Kappa=0.73).

To measure the *social level* at which participants would apply those strategies, we provided three tick boxes after each strategy that asked them to indicate whether they would apply the respective strategy to (a) regulate their own learning ("self-level"), (b) to regulate some other group member's learning ("co-level"), or whether the person would negotiate about that strategy with all group member ("shared level"; [10].

Table 1. Coding scheme for regulation activities along with examples.

Strategy type code	Example (in brackets the social level at which the answer was mentioned)
Elaborative	"[After that] I try to understand my part" (Self), "[After that], other members ask their questions" (Co), "[After that] joint elaboration of a summary" (Shared)
Surface oriented	"[After that] I skim through the material independently" (Self), "[After that], everyone learns their notes by heart" (Co), "[After that] everyone repeats the content independently" (Shared)
Metacognitive	"[After that] I also check if I am more motivated" (Self), "[After that] I ask who needs help with topics which the others perceived to be difficult" (Co), "[After that], we'll see if we've completed all that we had planned to learn." (Shared)
Motivational	"[First], I formulate a bond between knowledge and my life" (Self), "[After that], I try to bring humor into the learning situation" (Co), "[After that], the contents are asked together in plenary and made playful" (Shared)
Non Motivational	"[After that] I start to prepare independently: I structure my learning materials" (Self), "[After that] I ask the group what thoughts about it they had" (Co), "[After that] we make fixed dates so that we are "forced" to come" (Shared)
Other	"[First] I make an appointment" (Self), (no example provided for Co), "[After that], everyone goes home" (Shared)
No	"[After that] I write the exam with my already collected knowledge" (Self), (no examples provided for Co and Shared)

2.3 Data Preparation

Strategy type codes were paired with social level codes to generate meaningful codes (e.g., "Motivational Shared" indicates a motivational strategy that a participant reported to apply at the shared level) for each condition. Thus, from each of the seven strategy codes mentioned above, 3 "strategy type"—"social level" pair codes (= 18 codes in total) were generated.

Since sample size was insufficient to perform dimensionality reduction through ENA with all 18 codes, the aforementioned code pairs with their absolute and relative frequency were listed in descending order so that 7 codes, each accounting for at least five per cent of all pairs, could be selected for data analysis (see below). By choosing this threshold (= selection criterion), we arrived at almost complete models.

For example, of the codes that met this condition in the “motivational problems” condition, the *Elaboration Shared* code had the highest relative frequency of 0.21, while the *Elaboration Self* code reached the lowest relative frequency of 0.07. In the "comprehension-related problems" condition, also the *Elaboration Shared* code was most frequent (with a relative frequency of 0.25), but different to the other problem condition, also the *Metacognitive Self* code met the selection criterion with a relative frequency of exactly 0.05. It is noteworthy that in both conditions, five times the same of these seven codes fulfilled the inclusion criterion (the two exceptions: *Motivational Shared* in the motivational problem condition (relative frequency=0.16 in this condition) and *Metacognitive Self* in the comprehension-related problem condition) (since PM is based on event logs, artificial timestamps with identical time intervals between all consecutive codes were added to two event log files we created before conducting PM).

Process Mining. For plotting regulation sequences with PM, we used the R package “bupaR” (version 0.4.2; [11]). The PM algorithm generated one precedence matrix per condition by using the absolute frequencies of antecedent and consequent codes (activities) and flow of each person within a condition (= “absolute_case”). As the data was stored in the data.frame format it had to be transformed into an eventlog object before the process map could be computed based on this object.

Epistemic Network Analysis. For plotting the regulation sequences with ENA, we used the ENA 0.1.0 online tool ([12]) and included the following codes: *Elaboration Self*, *Elaboration Co*, *Elaboration Shared, Motivational Shared*, *Non motivational Shared*, *Metacognitive Self*, and *Metacognitive Shared*. We defined the units of analysis as the lines associated with a single value per condition (i.e., motivational problems, comprehension-related problems) associated with each participant’s case ID (= subset). Resultantly, one unit consisted, as an example, of the lines associated with the “motivational problems” condition and the participant with Case ID 42. The ENA algorithm counted the frequencies of each of two “strategy type”—“social level” pairs (= binary summation) based on a moving stanza window size of three (each of three lines plus the two previous ones) within a given conversation [5]. That way, each person received one value within the 28 dimensional vector space (for seven codes the space is calculated 7+6+5+4+3+2+1) represented by the matrices per condition.

To represent these values in the lower-dimensional vector space (= dimensional reduction), only the first seven dimensions were used as descriptors svd_i=1.7. Equally, the respective node positions were calculated based on the summed adjacency matrices within each condition, N_j=1-7, while the centroid values of the network graphs were calculated based on the weighted connections of the nodes. A final optimization routine served to minimize the difference between the plotted points and the corresponding network centroids (Σ_i (p_i-c_i)), while an additional means rotation minimized the network’s distance towards the x-axis in order to make possible group differences visible.

The projection of all subsequent dimensions, on the other hand, was done using a singular value decomposition, which produces orthogonal dimensions that maximize the variance explained by each of these dimensions.

3 Results

Process Mining. Process models (see Figure 1 and 2) show that students with motivational problems tend to start off with one of two kinds of strategies, both at the shared level: *Motivational* and *Metacognitive Shared.* In the comprehension-related problems condition, students clearly prefer starting off with *Metacognitive Shared* regulation, while Motivational Shared regulation does not play a large role in that condition at all. In both conditions, *elaborative shared regulation* seems to particularly be chosen later in the process.

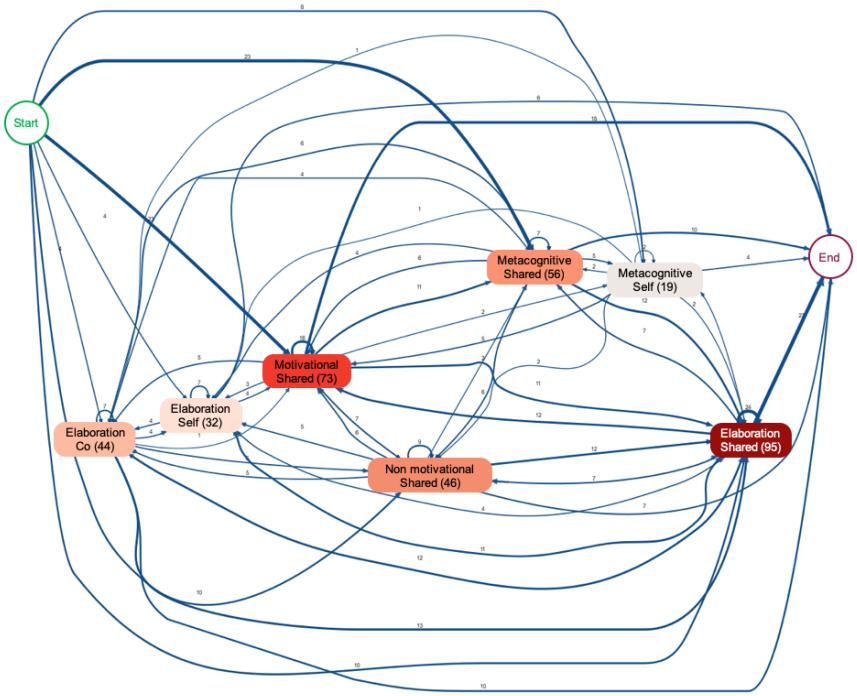

Fig. 1. Process Model for regulating motivational problems with absolute frequencies of all codes (boxes), as well as all observed directional code-code connections (arrows). Darker box colors indicate higher absolute code frequencies which means that the corresponding activities were observed more frequently, indicating that several persons have progressed from the corresponding first to the corresponding second activity (or, in the case of self-loops, that one person has performed the same activity several times in succession).

When experiencing motivational problems, students appear to switch more often between *Motivational* and *Metacognitive Shared* regulation, between *Non Motivational*

and *Elaborative Shared* regulation, and between *Elaborative* and *Motivational Shared* regulation, as compared to situations with comprehension-related problems.

When experiencing comprehension-related problems, in turn, students seem to switch more often between *Metacognitive Shared* and *Elaborative Self*-regulation, between *Metacognitive* and *Elaborative Shared* regulation, and between *Elaborative Co-* and *Shared* regulation.

The fact that weaker and stronger connections are more distinct in this process model might give rise to the interpretation that students tend to regulate comprehension-related problems in a more coherent way than they tend to regulate motivational problems. Further, it is noticeable that the temporal arrangement of codes in both conditions appears to be the same, and that only the "Motivational Shared" code comes in earlier in case of motivational problems.

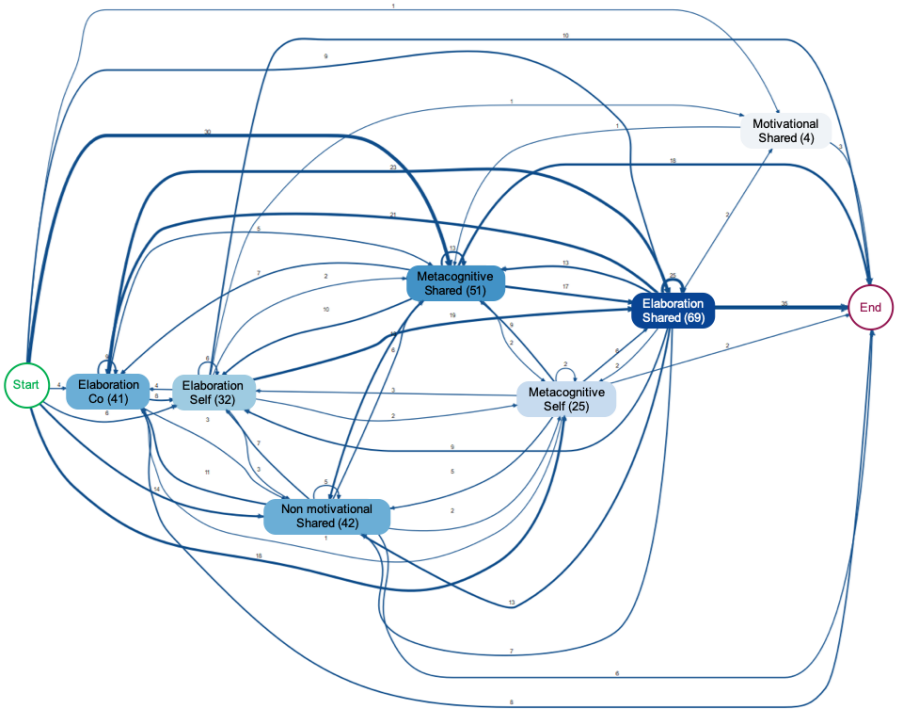

Fig. 2. Process Model for regulating comprehension-related problems with absolute frequencies of all codes (boxes), as well as all observed directional code-code connections (arrows). Again, darker box colors indicate higher absolute code frequencies which means that the corresponding activities were observed more frequently, indicating that several persons have progressed from the corresponding first to the corresponding second activity (or, in the case of self-loops, that one person has performed the same activity several times in succession).

Epistemic Network Analysis. The first ENA model (see Figure 3a) shows that the first component mr1 represented by the x-axis explained 10.20% of the variance in the ENA parameter space, while the second component svd2 represented by the y-axis accounted for 15.10% of the variance.

a) b)

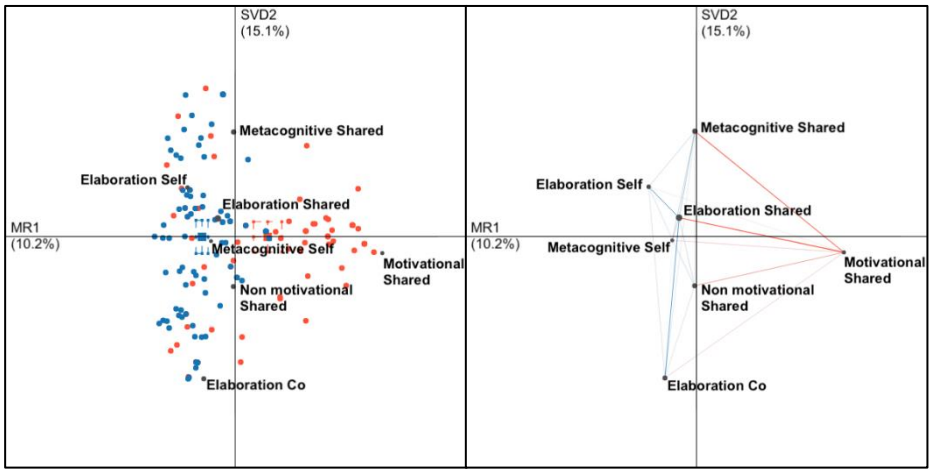

Fig. 3. (a) Networks of students in the conditions "motivational problems" (red) and "comprehension-related problems" (blue) with mean values (squares) and confidence intervals (boxes around squares). The x-axis is based on the descriptor mr1: values on this axis increase as participants demonstrate a higher emphasis on motivational regulation. The y-axis is based on svd2 and primarily focuses on (meta-)cognitive regulation. (b) Subtracted (= contrasted) networks for the conditions „motivational problems" (red) and „comprehension-related problems" (blue) which were generated by subtracting both networks' nodes and connection weights from each other. They serve to represent the differences between the two network graphs and illustrate what makes regulation of motivational problems in collaborative learning different to the regulation of comprehension-related problems.

Visualization of subtracted networks (see Figure 3b) shows that with the relatively stronger connections between metacognitive, elaboration, and non-motivational shared strategies with motivational shared strategies, the center of mass of the motivational problems condition network shifts to the right, while the relatively stronger links between elaboration and metacognitive strategies at all levels place the center of mass of the comprehension-related problems condition network to the left quadrants.

In addition, the subtracted network retains the differences found by PM: When encountering motivational problems, students show higher scores along the x-axis (mr1 can be seen as representative for motivational shared regulation) than when encounter-

ing comprehension-related problems. It also reveals higher relative co-occurrences between *Metacognitive Self*-regulation and *Elaborative Shared* regulation in the condition with comprehension-related problems, but does not retain the higher frequencies between *Non-Motivational* and *Elaboration Shared* in the motivational problem condition any more that was shown by PM.

Moreover, students in both conditions scored similarly on the y-axis (svd2 is representative for meta-cognitive activities). This was statistically confirmed by a paired t-test along the y-axis that failed to reject the null hypothesis as no statistical differences were found between the centroids in the condition with motivational (M=0.00, SD=0.60) and with comprehension-related problems (M=0.00, SD=0.66, t(81)=0.00, p=1.00). Nonetheless, a paired t-test along the x-axis revealed that the centroid of the motivational problem condition (M=0.29, SD=0.56) was significantly different from the centroid of the comprehension-related problems condition (M=-0.29, SD=0.24, t(81)=8.76, p=.00). On a more general level, these results illustrate that there are differences in how students in groups regulate motivational problems and in how they regulate comprehension-related problems. On a more specific level, they illustrate the shift of the regulation focus to motivational group activities in situations with motivational problems and to (meta-)cognitive activities at different social levels in situations with comprehension-related problems.

Additional analyses to converge findings from PM and ENA. Apart from the fact that ENA, unlike PM, cannot consider start and end points as codes, ENA also lacks to consider self-loop frequencies that may differ between conditions. To make sure taking the loops into account would not have resulted in completely different results of the t-tests, we proved that at least the self-loops of codes that were not plotted close to the x-axis by ENA did not significantly differ between conditions. Thus, we used exact Fisher's tests for count data to compare the cell frequencies of self-loops of all codes between both conditions that were already revealed by PM for significant differences (as the cell frequencies for the Motivational Shared code in the comprehension-related problems condition had zero counts, we have corrected for all cell frequencies based on a proposed procedure by [13]).

Since results showed no significant differences of self-loop frequencies between conditions except for the Motivational Shared code which was higher in the motivational than in the comprehension-related problem condition (this code was already plotted close to mr1 in the ENA), M_{Mot}=32, M_{Comp}=1, p=.00, OR = 36.33 (95% CI: 6.02, 1472.99), we take this as an indication that the group differences we found regarding mr1 would have maximally been even larger if the global test had also taken into account the self-loops on the motivational shared code beside the higher frequencies of this code in the motivational compared to the comprehension-related problems condition.

4 Discussion

This paper intended to demonstrate a procedure for comprehensively testing differences between regulation processes by aid of PM and ENA. At the same time, it intended to depict ways to bypass the drawbacks of each technique.

When performing PM and ENA individually, we encountered some of the problems of the two methods that are already discussed in literature. For example, when preparing the data for analysis, it turned out that our sample size was appropriate for PM, but too small for ENA. Therefore, less frequent codes had to be excluded from ENA (we also excluded these codes from PM as to better demonstrate the extent results of both methods converge). Also, PM created rather confusing models which were barely visually comparable due to the representation of all observed paths. The visual comparison in PM is also generally impeded by the fact that process models include all person-specific regulation paths with same weight irrespective to the person specific activity rate (for what ENA offers a solution). In addition, with PM, global differences between the models could not be verified by a statistical test. Interestingly enough, the arrangement of the codes in both models showed differences only in terms of the "Motivational Shared" code, which was positioned earlier in the process when motivational problems were present. Thus, PM showed that motivational problems are primarily regulated motivationally and metacognitively in the beginning, whereas for comprehension-related problems, the initial focus is on metacognitive regulation.

These findings – which ENA failed to reveal – might indicate that students more or less automatically activate different motivational strategies to solve motivational problems, whereas they seem to be more analytical (and coherent) when faced with comprehension-related problems (see [14]). However, that elaborative shared regulation was rather chosen at the end of the process (in both process models), which seems to be in line with Boekaerts' [8] three-layered model, claiming that goals and resources need to be regulated before learning and which is adheres to the findings of [3] which are described above.

However, ENA allowed for a global statistical verification of the differences between the compared processes that could not be gained by PM. Additionally, while PM would have required further reductions of codes or code-connections to clearer visualize regulation processes, ENA was able to clearly visualize group differences through differently weighted code-code-connections. The clear visualization of group differences was further due to the subtraction of networks and to the provided rotation of networks, which cannot be done in PM. Even though the shift towards joint motivational group efforts for motivational and towards activities closer to the learning process for comprehension-related problems were already observed in PM, ENA revealed that students with comprehension-related problems frequently control their own regulation when acquiring knowledge: because students constantly switched to motivational shared regulation in the regulation of motivational problems, but did not that much switch to one specific cognitive activity in the comprehension-related problems condition, ENA's visual comparison of conditions revealed that students regulate comprehension-related problems comparatively incoherently.

This latter observation would not have been apparent from PM. We specifically revealed this through the ENA's normalization of adjacency vectors. This routine, which is not implemented for PM, ensured that the number of each participant`s activities did not affect the structure and thickness of networks. In return, however, it removed the effect of the person-specific regulatory length, which PM considers, but does not clearly visualize. Furthermore, as ENA disregards start and end points of processes, as well as the order of selected paths and self-loop information (all information provided by PM), exact Fisher's tests for count data based on the self-loop information provided by PM complied with the findings of the global comparison.

Overall, when critically appraising the use of each of the two methods and what they add to our understanding of the differences between motivational vs. comprehension-related problem regulation in groups, we argue that the concerted use of PM and ENA has high potential in comparing regulatory processes and is superior to using only either technique. As an outlook for further research with process data remains to say that researchers are already working on the development and refinement of so-called directed ENA; [15]).

References

1. Csanadi, A., Eagan, B., Kollar, I., Shaffer, D.W., Fischer, F.: When coding-and-counting is not enough: using epistemic network analysis (ENA) to analyze verbal data in CSCL research. International Journal of Computer-Supported Collaborative Learning 13(4), 419–438 (2018). doi: 10.1007/s11412-018-9292-z
2. Hadwin, A.F., Järvelä, S., Miller, M.: Self-Regulated, Co-Regulated, and Socially Shared Regulation of Learning. In: Zimmerman, B., Schunk, D. (eds.) Handbook of Self-Regulation of Learning and Performance, pp. 65–84. Routledge, New York (2011).
3. Bannert, M., Reimann, P., Sonnenberg, C.: Process mining techniques for analysing patterns and strategies in students' self-regulated learning. Metacognition and Learning 9(2), 161–185 (2014). doi: 10.1007/s11409-013-9107-6
4. Bolt, A.J., van der Aalst, W.M.P., de Leoni, M.: Finding process variants in event logs (short paper). In: Panetto, H., Debruyne, C., Gaaloul, W., Papazoglou, M., Paschke, A., Agostino Ardagna, C., Meersman, R. (eds.) On the Move to Meaningful Internet Systems. OTM 2017 Conferences: Confederated International Conferences: CoopIS, C&TC, and ODBASE 2017, Rhodes, Greece, October 23-27, 2017, Proceedings, Part I. LNCS, vol. 10573, pp. 45–52. Springer, Dordrecht (2017). doi: 10.1007/978-3-319-69462-7_4
5. Shaffer, D.W.: Quantitative Ethnography. Cathcart Press, Madison (2017).
6. Ruis, A.R., Rosser, A.A., Quandt-Walle, C., Nathwani, J.N., Shaffer, D.W., Pugh, C.M.: The hands and head of a surgeon: Modeling operative competency with multimodal epistemic network analysis. The American Journal of Surgery 216(5), 835–840 (2018). doi: 10.1016/j.amjsurg.2017.11.027
7. Zhang, S., Liu, Q., Cai, Z.: Exploring primary school teachers' technological pedagogical content knowledge (TPACK) in online collaborative discourse: An epistemic network analysis. British Journal of Educational Technology (2019). doi: 10.1111/bjet.12751
8. Boekaerts, M.: Self-regulated learning: where we are today. International Journal of Educational Research 31(6), 445– 457 (1999). doi: 10.1016/S0883-0355(99)00014-2

9. Friedrich, H.F., Mandl, H.: Lernstrategien: Zur Strukturierung des Forschungsfeldes. In: Mandl, H., Friedrich, H.F. (eds.) Handbuch Lernstrategien, pp. 1–23. Hogrefe, Göttingen (2006).
10. Hadwin, A., Oshige, M.: Self-regulation, coregulation, and socially shared regulation: exploring perspectives of social in self-regulated learning theory. Teachers Coll. Rec. 113(2), 240-264 (2011)
11. Janssenswillen, G.: bupaR: Business Process Analysis in R. R package version 0.4.2 (2019).
12. Marquart, C.L., Hinojosa, C., Swiecki, Z., Shaffer, D.W.: Epistemic Network Analysis version 0.1.0 (2018).
13. Dureh, N., Choonpradub, C., Tongkumchum, P.: An alternative method for logistics regression on contingency tables with zero cell counts. Songklanakarin J. Sci. Technol. 38(2), 171–176 (2016). doi: 10.14456/sjst-psu.2016.23
14. Melzner, N., Greisel, M., Dresel, M., Kollar, I.: Effective Regulation in Collaborative Learning: An Attempt to Determine the Fit of Regulation Challenges and Strategies (long paper). In: Lund, K., Niccolai, G., Lavoué, E., Hmelo-Silver, C., Gweon, G., Baker, M. (eds.) A Wide Lens: Combining Embodied, Enactive, Extended, and Embedded Learning in Collaborative Settings: Proceedings of the 13th International Conference on Computer Supported Collaborative Learning, CSCL, vol. 1, pp. 312–319. International Society of the Learning Sciences, Lyon (2019).
15. Marquart, C.L., Swiecki, Z., Collier, W., Eagan, B., Woodward, R., Shaffer, D.W.: rENA: Epistemic Network Analysis. R package version 0.1.6.1 (2019).