The Influence of Different Measurement Approaches on Student Affect Transitions Using Ordered Networks

Nidhi Nasiar[1](✉), Andres Felipe Zambrano[1], Jaclyn Ocumpaugh[1], Alex Goslen[2], Jonathan Rowe[2], Jessica Vandenberg[2], Jordan Esiason[2], and Stephen Hutt[3]

[1] Graduate School of Education, University of Pennsylvania, Philadelphia, PA, USA
nasiar@upenn.edu
[2] North Carolina State University, Raleigh, NC, USA
[3] University of Denver, Denver, CO, USA

Abstract. Affect detection is integral to creating affect-sensitive learning systems, but the impact of measurement methods needs further research. This paper uses ordered network analysis (ONA) to compare the affect dynamics of two suites of affect detectors trained on complementary data (i.e., labels from an in-the-moment self-reporting (SR) tool vs labels from field observations) in game-based learning. We then use ONA difference models to assess how divergence in learning and motivational measures impact affect dynamics.

Keywords: Affect Dynamics · Epistemic Network Analysis · Transition Analysis

1 Introduction

Epistemic emotions emerge during cognitive activities and play a vital role in learning [1]. Typically, learning analytics research has focused on 5 emotions: boredom, confusion, frustration, engaged concentration, and delight [1, 3, 4]. Prior work explores the cumulative effect of each emotion [2], its distribution across the learning experience, and the effect of different transitions [1]. However, affect dynamics results have been difficult to interpret even when data is reanalyzed to correct for sampling biases and align methodological choices [4].

One methodological concern that has not been fully studied concerns the ground truth labels that train affect detectors. Exceptions exist [5], but types of labels are rarely compared. Other understudied concerns relate to duration effects and the effects of student-level characteristics (e.g., motivational factors, learning indicators). For example, highly anxious students undergo longer frustration bouts [6], which may mean that common practices of removing self-transitions [7] obscure results.

This study addresses these gaps with a study of Crystal Island, a game-based learning system for middle school microbiology. First, we compare the dynamics of affect detectors [9] trained with ground-truth labels collected using BROMP observations [8]

Y. J. Kim and Z. Swiecki (Eds.): icqe 2024, CCIS 2279, pp. 195–203, 2024.
https://doi.org/10.1007/978-3-031-76332-8_16

to detectors trained on data from a self-report tool employed during those observations. Specifically, we use an ordered network analysis (ONA), a type of epistemic network analysis (ENA) that accounts for directionality, to compare the networks from each set of detectors. We then use ONA difference models to explore how learning gains and situational interest, change affect transitions in these models. We do this to address two research questions (RQ): **RQ1**: Which affective sequences emerge as methodological consequences of ground truth labels (i.e., SR vs. BROMP-based detectors)? **RQ2:** Which sequences emerge due to participant sampling (i.e., differences in students with respect to (a) learning gains and (b) situational interest)?

2 Related Work

Detecting student affect helps us to understand its impact on learning, engagement, and motivation [1, 10]. The development of affective detectors often relies on supervised machine learning (ML) [3], which requires ground truth labels from either the subject (self-report) or a third party (observer), each with strengths and weaknesses. Self-reports can be hampered by self-presentation effects and metacognitive difficulties but have been used to create effective affect detectors in education. Observation-based measurements—like Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP; [8])—have informed the development of affect detectors in many learning systems [10] but are limited to what coders can discern.

In addition to concerns about labeling strategies, there have also been concerns about what different affective states (or combinations thereof) might mean for learning. [1] hypothesized two main affective pathways, which have been explored using L statistic and ENA [15]. These studies show connections between those pathways and learning, but the hypothesized pathways occur infrequently. Other studies [11] find that the rates of occurrence and recurrence of affect are important. Finally, research suggests that affect dynamics differ for different groups of students. Pathways are different among those with high and low learning [1] and differences are also associated with trait-anxiety [6]. It seems reasonable that we should expand these efforts to other motivational constructs linked to students' emotions, like situational interest, and also learning gains.

3 Methods

This study investigates learning in Crystal Island (CI), an open-world, game-based learning system aligned to 8th-grade microbiology state standards. Students act as investigative scientists to identify a mysterious disease spread on the remote island, interacting with NPCs, reading educational materials, and testing in-game items, while tracking their hypotheses and results. This study uses data from 124 middle schoolers who played CI during their regular science class in the southeastern US. Students are well-balanced for male (N = 53) and female (N = 66), with 5 others who prefer to describe themselves. They include a large share of students from historically marginalized communities (46% Black, 16% Hispanic, 5% Asian, 5% Multiracial, 1% Native American). Students answered a series of pre-surveys (i.e., a demographic questionnaire, knowledge pre-test, and situational interest scale). After 2 days of gameplay, they answered post-surveys (i.e., knowledge post-test and several other motivational measures). While interacting with the game, affect was simultaneously labeled using 2 distinct methods: BROMP observations [8] and a self-report.

3.1 Measures Used in the Study

Affect Detectors using Self-report and BROMP-based Data. This study uses previously published [9], cross-validated, interaction-based detectors of epistemic affective states developed from labels generated from (a) in-game self-reports (SR) and (b) BROMP-based observations. Interaction-based detectors model what students do within a system when they are experiencing an emotion, allowing us to infer these emotions in real-time (see reviews in [12, 13], including a review of BROMP [8]). Students self-reported 1 of 6 affective states: boredom, focused, confusion, happiness, frustration, and nervousness [9]. BROMP classroom observations [8] included boredom (BOR), engaged concentration (ENG), confusion (CON), delight (DEL), and frustration (FRU), but not nervousness, which is hard to observe directly. Corresponding SR labels were aligned with standard BROMP labels adopting a child-friendly language (i.e., focus for engaged concentration; happiness for delight).

BROMP and SR detectors were developed independently from one another. Each set of affective data was used to train a distinct set of ML-based affect detectors, cross-validated to ensure generalization to new populations. As in prior work using BROMP [10], both suites of detectors are trained using 20-s clips of students' log files. Then, each suite is applied and analyzed separately, assigning 2 labels to each clip—from the SR-based detector and BROMP-based detector with the highest probability. Previous work suggests that SR- and BROMP-based detectors capture complementary information about students' affective states [9]. Hence, the differences produced by each are likely to be meaningful.

Learning Gains (LG) and Situational Interest (SI). Students completed identical pre and post-tests of domain knowledge. Normalized learning gains, which scale from 1- to 1, were calculated. Of 124 students, 46 answered both and we excluded students within a standard deviation of the mean (avg = −.10, SD = .35). By ensuring that the groups are distinct, this approach contrasts students with the lowest (LG $<$ −.45, N = 9) and highest (LG $>$ .25, N = 6) LGs, avoiding spurious results that can occur when splitting at the mean. SI surveys were adapted for middle school science from [14], featuring 9 items rated on a 7-point scale. Of 124 students, 122 completed this scale, and we excluded students within a standard deviation of the mean (avg = 3.78, SD = 1.05), comparing those with high (SI $>$ 4.83, n = 20) and low SI (SI $<$ 2.73, n = 16).

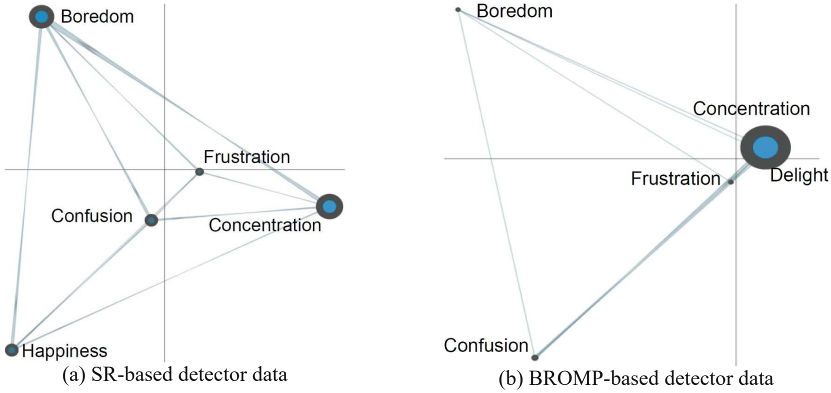
3.2 Ordered Network Analysis (ONA)

We study affective sequences using Ordered Network Analysis (ONA; [16]), building on work using ENA to explore affect transitions [4] and in-game behaviors. We chose ONA for its ability to capture directionality and self-transitions. In *standard uses of ONA*, codes are visualized as nodes in a network, and edges have line weights (LW) of 0 to 1, representing connection strength. Self-transitions are shown with increased node size and LWs. In *ONA difference models,* LWs from one model are subtracted from another (referred to as LW_{diff}). This study selects *ONA parameters* to best understand affective transitions, including self-transitions. Each line of data corresponds to 20-s of gameplay, with 2 labels are applied to each line, one from each suite of detectors. Interrater reliability is unnecessary here as validated detectors provided codes. The *conversation* variable is the full gameplay session (avg = 41.6 min, SD = 15.6) and a *moving stanza window* of size one, meaning only connections between the current and previous line (i.e., affective state) are considered, as standard when analyzing affective transitions [15].

4 Results

4.1 Comparing SR-Based and BROMP-Based Affect Detectors Using ONA

We first use ONA to assess differences in the affective transitions identified by the 2 sets of detectors across all students. Figure 1 shows SR-based models (1a, left) and BROMP-based models (1b, right). Line weights (Table 1) determine node size for self-transitions (e.g., BOR → BOR).



(a) SR-based detector data

(b) BROMP-based detector data

Fig. 1. Ordered network of affect transitions for SR-based and BROMP-based detectors.

For both sets of detectors, self-transitions (where the affective state repeats across two 20-s clips) are most common. For SR models, boredom (LW = .51) is followed by concentration (LW = .43), confusion (LW = .21), happiness (LW = .21), and frustration (LW = .12). For BROMP models, the distribution is more skewed towards concentration (LW = .95) than the other four states (LW = .01 to .07).

Within the SR models, common pathways *between different* affective states are much rarer. Pathways between boredom and concentration (BOR → ENG: LW = .04; ENG → BOR: LW = .03) occur slightly more often than pathways involving concentration, confusion, and happiness (LW = .02 for ENG → CONF, CONF → ENG, HAP → ENG, and ENG → HAP). All other transitions have lower LWs. Similarly, in BROMP models, self-transitions (LW = .01 to .95) effectively mute other possible transitions and only shifts between engaged concentration and confusion occur above .01 (CONF → ENG: LW = .03; ENG → CONF: LW = .02).

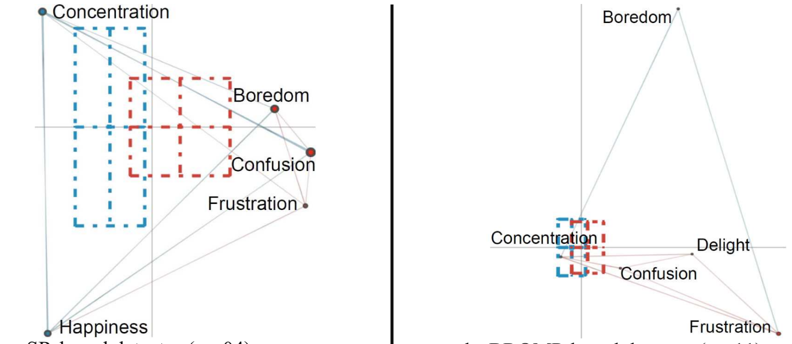
Table 1. Line Weights for ONA Models shown in Fig. 1. LW > = .10 are in bold.

Self-Transitions	SR	BROMP	Other Transitions	SR	BROMP
BOR → BOR	**0.51**	0.05	BOR → ENG	0.04	<.01
ENG → ENG	**0.43**	**0.95**	ENG → BOR	0.03	<.01
DEL/HAP → DEL/HAP	**0.22**	0.01	DEL/HAP → ENG	0.02	<.01
CONF → CONF	**0.21**	0.07	CONF → ENG	0.02	0.03
FRU → FRU	**0.12**	0.05	ENG → DEL/HAP	0.02	<.01
			ENG → CONF	0.02	0.02

4.2 ONA Difference Models for Learning Gains (LG)

Figure 2 uses ONA difference models to compare students with high and low LG, with models for SR (2a) and BROMP (2b) data. Here, LWs indicate the difference between the two student groups. Transitions that are most common among those with high LG are shown in blue, with low LG in red. Network variability (confidence interval) is shown as a rectangle surrounding a point that represents the average network position (mean). Mann-Whitney U tests show statistically significant differences for LG in SR models but not in BROMP models (U = 9.0, p = .04, r = .67 vs U = 13.0, p = .11, r = .52).

In SR models, sustained states show the largest differences. Low LG learners are more likely to sustain boredom (BOR → BOR: LW = .35 for high vs LW = .45 for low, producing $\text{LW}_{\text{diff}} = .25$) or confusion (CONF → CONF: LW = .09 vs LW = .32, producing $\text{LW}_{\text{diff}} = .23$). Six other transitions have LW ≥ .02 (ENG → CONF, CONF → ENG and ENG → HAP) or LW ≥ .01 (BOR → HAP, CONF → HAP, and HAP → BOR) (Table 2).



a. SR-based detector (p=.04)

b. BROMP-based detector (p=.11)

Fig. 2. Difference models for students with high (blue) and low LG (red) for each set of data.

In BROMP models, the largest differences are also for sustained states. High LG students spent less time in sustained frustration or boredom (FRU → FRU: LW = .01 vs LW = .08, $\text{LW}_{\text{diff}} = .06$; BOR → BOR: LW = .06 vs LW = .09, $\text{LW}_{\text{diff}} = .03$), and less

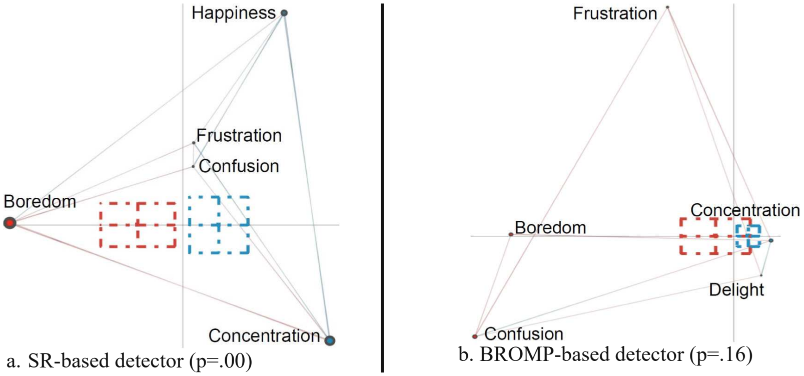
Table 2. LWs for ONA difference models in Fig. 2. LW and $LW_{diff} > = .10$ are in bold.

	SR			BROMP			SR				BROMP		
Self-Trans	HiLG	LoLG	|Diff|	HiLG	LoLG	|Diff|	Other Trans	HiLG	LoLG	|Diff|	HiLG	LoLG	|Diff|
BOR	**0.25**	**0.45**	**0.25**	0.06	0.09	0.03	BOR → ENG	0.02	0.02	<.01	<.01	<.01	-
ENG	**0.66**	**0.49**	**0.17**	**0.98**	**0.97**	0.02	ENG → BOR	0.01	0.01	<.01	0.01	<.01	<.01
DEL/HAP	**0.38**	**0.24**	**0.14**	<.01	0.02	0.02	DEL/HAP → ENG	0.03	0.01	0.02	<.01	<.01	-
CONF	0.09	**0.32**	**0.23**	0.04	0.05	0.02	CONF → ENG	0.03	0.01	0.02	0.02	0.02	<.01
FRU	0.03	**0.13**	**0.10**	0.01	0.08	0.06	ENG → DEL/HAP	0.03	0.01	0.02	<.01	0.01	<.01
							ENG → CONF	0.03	0.01	0.02	0.02	0.03	0.01
							CONF → FRU	<.01	<.01	<.01	-	0.01	0.01

time in sustained happiness (HAP → HAP: $LW < .01$ vs $LW = .02$, yielding $LW_{diff} = .02$). Differences in 2 other transitions are small but at .01 (ENG → CONF: $LW = .02$ vs $LW = .03$, $LW_{diff} = .01$; CONF → FRU: $LW = 0$ vs. $LW = .01$, $LW_{diff} = .01$).

4.3 ONA Difference Models for Situational Interest (SI)

ONA is also used to examine how affect transitions differ based on SI, and detector type influences these results. High and low SI groups are statistically different in SR models but not in BROMP models (Mann Whitney $\alpha = .05$, $U = 68.0$, $p = .00$, $r = .56$ vs $U = 115.5$, $p = .16$, $r = .28$). In Fig. 3, blue lines show the transitions most common among high SI learners, while red show those for low SI learners (Table 3).



a. SR-based detector (p=.00)

b. BROMP-based detector (p=.16)

Fig. 3. Difference models for high SI (blue) vs. low SI (red) for SR and BROMP data.

Within SR models, the largest differences are again related to self-transitions. High SI learners are less likely to sustain boredom (BOR → BOR: $LW = .37$ vs $LW = .72$, $LW_{diff} = .35$) and more likely to sustain concentration and happiness (ENG → ENG: $LW = .59$ vs $LW = .32$, $LW_{diff} = .27$; HAP → HAP: $LW = .25$ vs $LW = .1$, $LW_{diff} = .15$) and even frustration and confusion (FRU → FRU: $LW_{diff} = .04$; CONF → CONF $LW_{diff} = .02$). Differences in transitions between distinct affective states are also found

Table 3. Line weights for ONA difference models in Fig. 3. LW and LW_{diff} >= .10 are in bold.

	SR			BROMP				SR			BROMP		
Self Trans	HiSI	LoSI	|Diff|	HiSI	LoSI	|Diff|	Other Trans	HiSI	LoSI	|Diff|	HiSI	LoSI	|Diff|
BOR	**0.37**	**0.72**	**0.35**	-	0.07	0.07	BOR → ENG	0.03	0.05	0.02	-	<.01	<.01
ENG	**0.59**	**0.32**	**0.27**	**0.98**	**0.9**	0.08	ENG → BOR	0.02	0.04	0.02	-	0.01	0.01
DEL/HAP	**0.25**	**0.10**	**0.15**	0.01	<.01	<.01	DEL/HAP → ENG	0.03	0.01	0.01	<.01	<.01	<.01
CONF	**0.20**	**0.18**	0.02	0.06	**0.16**	**0.10**	CONF → ENG	0.02	0.03	0.01	0.02	0.02	<.01
FRU	**0.13**	0.09	0.04	0.06	0.09	0.03	ENG → DEL/HAP	0.03	0.01	0.02	0.01	<.01	<.01
							ENG → CONF	0.02	0.02	<.01	0.02	0.02	<.01

in the SR models. High SI learners are more likely to transfer from engaged concentration to happiness (e.g., ENG → HAP: LW = .03 vs. LW = .01) and less likely to transfer back and forth between boredom and engagement (BOR → ENG: LW = .03 vs. LW = .05; ENG → BOR: LW = .02 vs LW = .04). Likewise, high SI learners are less likely to show two transitions involving confusion (BOR → CONF: LW = .01 vs. LW = .02; CONF → ENG: LW = .02 vs. LW = .03).

In BROMP models, the largest differences are also in self-transitions, but sustained confusion (i.e., not the most common code, ENG) shows the largest difference. High SI learners are less likely to sustain confusion (CONF → CONF: LW = .06 vs LW = .16, LW_{diff} = .1) and boredom, which was notably absent among high SI learners (BOR → BOR: LW_{diff} = .07). Instead, they are more likely to sustain concentration (ENG → ENG: LW = .98 vs LW = .9, LW_{diff} = .08). Smaller differences are driven by absent transitions among high SI learners, including the only transition between 2 different states with $\text{LW}_{\text{diff}} > .01$ (CONC → BOR), and 5 transitions with $\text{LW}_{\text{diff}} < .01$ (BOR → ENG, BOR → CONF, CONF → BOR, CONF → DEL, and FRUS → DEL).

5 Discussion and Conclusions

Since affect modeling is key to supporting learning and motivation, it is important to examine the effects of methodological differences. This study uses 2 sets of cross-validated, interaction-based affect detectors, trained with self-reports and BROMP observations, to examine differences in affective dynamics using ONA. Further, ONA difference models are used to study how affect dynamics are shaped by two student-level characteristics, learning gains and situational interest.

Results from RQ1 show that self-transitions are more common regardless of detector type, but the SR-trained data are far more distributed (LW = .51 to .12, BOR and FRU) than the BROMP-trained data, where sustained concentration is markedly more common (LW = .95) than other self-transitions. That is, even though self-transitions are important in both sets of detectors, there are still major persistence-rate differences that could significantly impact intervention design.

Results from RQ2 show the relevance of student-level characteristics in affect dynamics research. Specifically, we compare students at the extreme ends of normalized learning gains (LG), and situational interest (SI). For LG, SR-trained data produced large differences. All 5 self-transitions showed $\text{LW}_{\text{diff}} > .1$, with boredom and confusion more likely for low LG students. In BROMP-trained data, self-transitions were also more common than transitions between 2 distinct states. Sustained frustration has the largest difference, with low LG learners experiencing the most frustration. For SI, similar patterns emerge when comparing the 2 sets of detectors. Differences are larger (a) in SR models than BROMP models and (b) for self-transitions, chiefly boredom and concentration. In SR models, low SI learners undergo sustained boredom at twice the rate of others. BROMP models show similar patterns: sustained boredom is entirely absent in high SI learners. High SI learners also exhibit higher rates of concentration, though this effect is strongest in SR models.

More generally, we emphasize the need to study differences related to learning and motivation. Combined with results from the initial ONA analyses, the difference models show the relative importance of self-transitions, often excluded in previous work. We hope that these results motivate similar future investigations.

Acknowledgments. This study was funded by NSF Grants IIS-2016943 and IIS-2016993.

Disclosure of Interests.. The authors have no competing interests to declare for this article.

References

1. D'Mello, S., Graesser, A.: Dynamics of affective states during complex learning. Learn. Instr. **22**(145–157), 9 (2012)
2. Cloude, E.B., Wortha, F., Dever, D.A., Azevedo, R.: How do emotions change during learning with an intelligent tutoring system? metacognitive monitoring and performance with MetaTutor. In: CogSci (2020)
3. Calvo, R.A., D'Mello, S.K.: Affect detection: an interdisciplinary review of models, methods, and their applications. Affective Comput., IEEE Trans. **1**(1), 18–37 (2010)
4. Karumbaiah, S., Baker, R.S., Ocumpaugh, J., Andres, A.: A re-analysis and synthesis of data on affect dynamics in learning. IEEE Trans. on Affective Computing (2021)
5. Kai, S., et al.: A comparison of video-based and interaction-based affect detectors in physics playground. In: International Conference on Educational Data Mining, pp. 77–84 (2015)
6. Andres, J.M.A.L., Hutt, S., Ocumpaugh, J., Baker, R.S., Nasiar, N., Porter, C.: How anxiety affects affect: a quantitative ethnographic investigation using affect detectors and data-targeted interviews. In: International Conference on Quantitative Ethnography (2021)
7. Karumbaiah, S., Andres, J.M.A.L., Botelho, A.F., Baker, R.S., Ocumpaugh, J.: The implications of a subtle difference in the calculation of affect dynamics. In: 26th International Conference for Computers in Education (2018)
8. Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T.: Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical & Training Manual. Working Paper (2015)
9. Zambrano, A., et al.: Says Who? How different ground truth measures of emotion impact student affective modeling. In: Proceedings of the 17th International Conference on Educational Data Mining (2024)
10. Jiang, Y., et al.: Expert feature-engineering vs. deep neural networks: Which is better for sensor-free affect detection? Proceedings of the 19th Conference of AIED, pp. 198–211 (2018)

11. Baker, R.S., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C.: Better to be frustrated than bored: the incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. Int. J. Hum. Comput. Stud. **68**(4), 223–241 (2010)
12. Baker, R.S., Ocumpaugh, J., Andres, J.M.A.L.: BROMP quantitative field observations: a review. In: Learning Science: Theory, Research, and Practice, pp. 127–156 (2020)
13. Baker, R.S.J.d., Ocumpaugh, J.: Interaction-based affect detection in educational software. In: Calvo, R.A., D'Mello, S.K., Gratch, J., Kappas, A. (eds.) The Oxford Handbook of Affective Computing. Oxford University Press, Oxford, UK (2014)
14. Linnenbrink-Garcia, L., et al.: Measuring situational interest in academic domains. Educ. Psychol. Measur. **70**(4), 647–671 (2010)
15. Karumbaiah, S., Baker, R.S.: Studying affect dynamics using epistemic networks. In: Ruis, A.R., Lee, S.B. (eds.) ICQE 2021. CCIS, vol. 1312, pp. 362–374. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67788-6_25
16. Tan, Y., Ruis, A., Marquart, C., Cai, Z., Knowles, M., Shaffer, D.: Ordered Network Analysis. In: Damşa, C., Barany, A. (eds.) Advances in Quantitative Ethnography. ICQE 2022. Communications in Computer & Information Science, vol. 1785 (2023)