Mining Unstructured Text Online Discussion Data to Understand Group Collaboration: Mixed and Multi-Methods Field Study

California State University Channel Islands Seminar Series

Dr. Evren Eryilmaz

11/09/2021

About Me

Name

Evren Eryilmaz

Work Experience

- Assistant Professor of Management Information Systems at CSUS
- Faculty Coordinator for the Center for Small Business at CSUS
- Adjunct Professor at CSUCI

Achievements

Probationary Faculty Grant 2021 at CSUS

Education

Ph.D. in Information Systems & Technology from Claremont Graduate University

Hobbies

- Baking
- Gardening
- Learning French

Overview

- Motivation and problem identification
- Objectives and special issues/constraints
- A short literature review on
    - ❖Big data research perspectives
    - ❖Learning analytics
    - ❖Community of inquiry framework
- Research questions
- Major Findings
- Comments & questions

Motivation and problem identification

- Criteria-based outcome assessment to transform post-COVID education, diversity, and student success faculty learning committee
- Academic information technology committee
- Establishing successful service-learning project teams is difficult in online settings
- Asynchronous online discussions (AODs) can support developing shared understandings and cultivating a sense of community

Objectives

- Combine the analytical efficiency and scalability of topic modeling, social network analysis, and cluster analysis with theory-driven qualitative content analysis to obtain a comprehensive picture of group collaboration in AODs

- Establish the boundaries of an intermediate cluster within a learning community

Special Issues/Constraints

- Assessment needs to center on educational theories
- Aspects of the final product can be integrated into canvas in the future

Literature Review: Two Big Data Research Perspectives

- Data-driven big data research: Provides answers to situated practical or tactical questions
- Theory-driven big data research: theoretical foundations developed can guide big data research through focus such as variable selection and search for patterns in data

Maass, W., Parsons, J., Purao, S., Storey, V. C., & Woo, C. (2018). Data-driven meets theory-driven research in the era of big data: opportunities and challenges for information systems research. *Journal of the Association for Information Systems*, *19*(12), 1.

Johnson, S. L., Gray, P., & Sarker, S. (2019). Revisiting IS research practice in the era of big data. *Information and Organization*, *29*(1), 41-56.

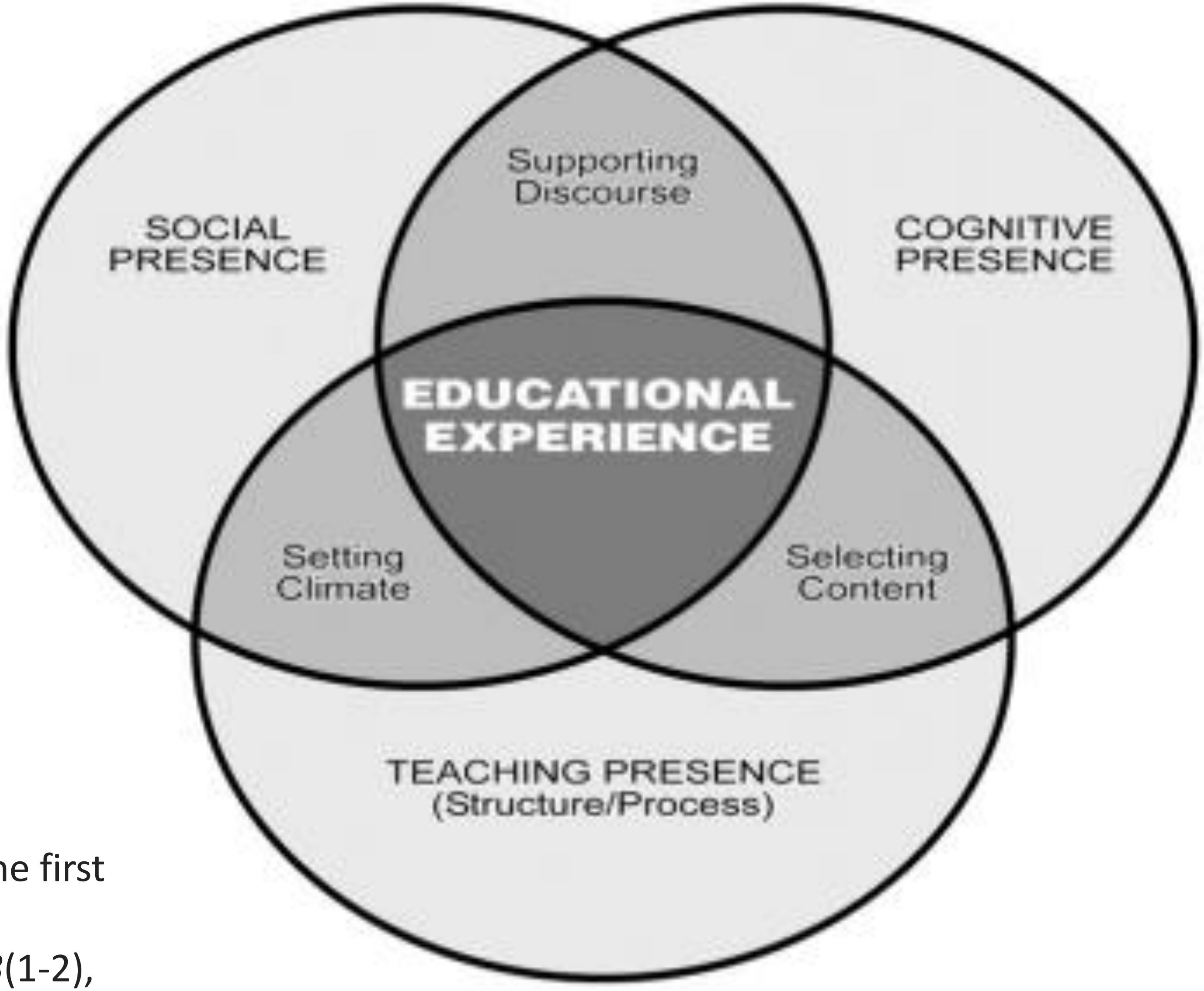
Literature Review: Learning Analytics

- Learning Analytics: *The measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs*
  - Process Focus: Aligns well with constructivism and experiential learning
  - Outcome Focus: Aligns well with behaviorist theory of learning (i.e., test scores)

Siemens, G.; and Long, P. Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, *46*, 5 (2011), 30.

Deeva, G., Willermark, S., Islind, A. S., & Oskarsdottir, M. (2021, January). Introduction to the Minitrack on Learning Analytics. In *Proceedings of the 54th Hawaii International Conference on System Sciences* (p. 1507).

Literature Review: Community of Inquiry



Garrison, D. R., Anderson, T., & Archer, W. (2010). The first decade of the community of inquiry framework: A retrospective. *The internet and higher education*, *13*(1-2), 5-9.

Research Questions

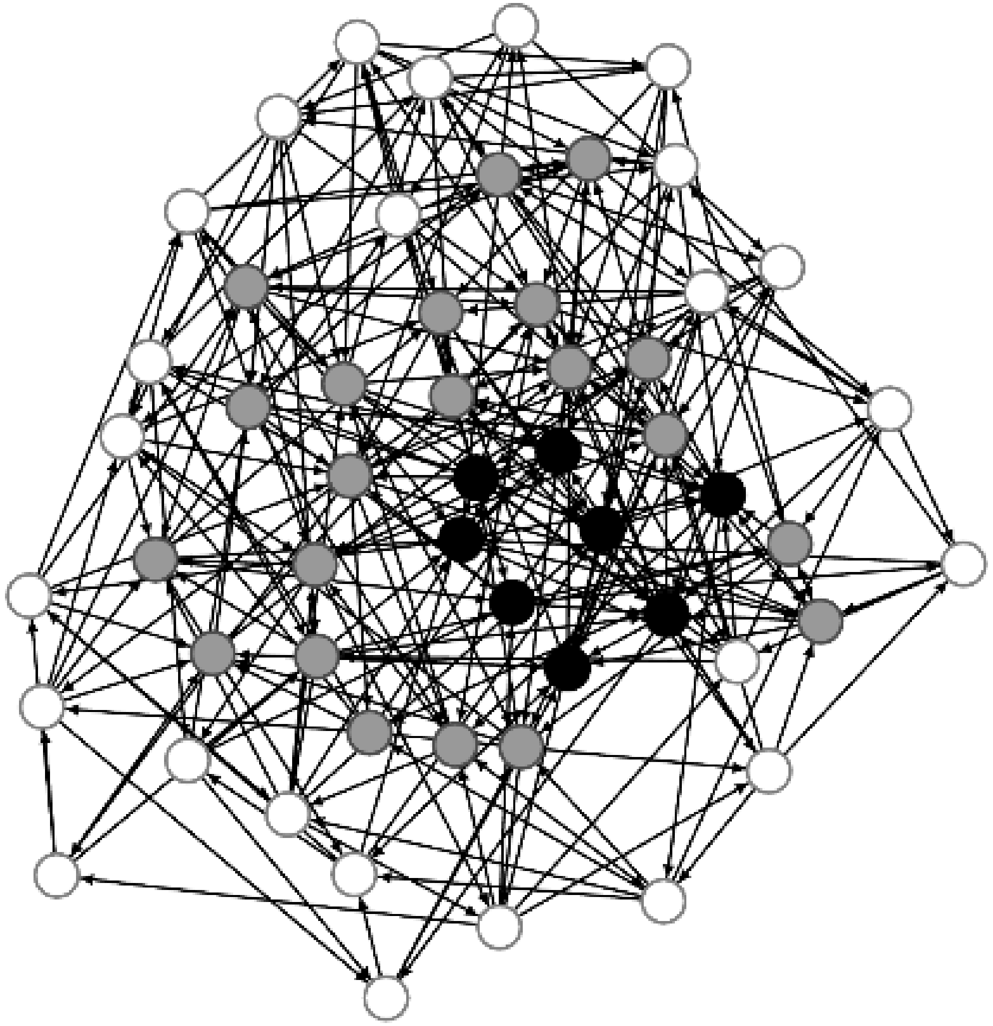
1. What is the social network structure of a COI facilitated by the Canvas AOD tool?
2. What are differences of topics among a COI's clusters via topic modeling?
3. How and to what extent topic modeling results relate to the COI model's cognitive presence message-coding schema among a COI's clusters?

Field Study

- 54 senior undergraduate management information systems students in a service-learning project based capstone course
- Male: 58% Female: 42%
- Average age: 21.87 (SD= 3.23)
- Total messages: 470 (M=8.70, SD= 0.96)
- Average number of words per message: 121.48 (SD=32.54)

Learning Community's Sociogram

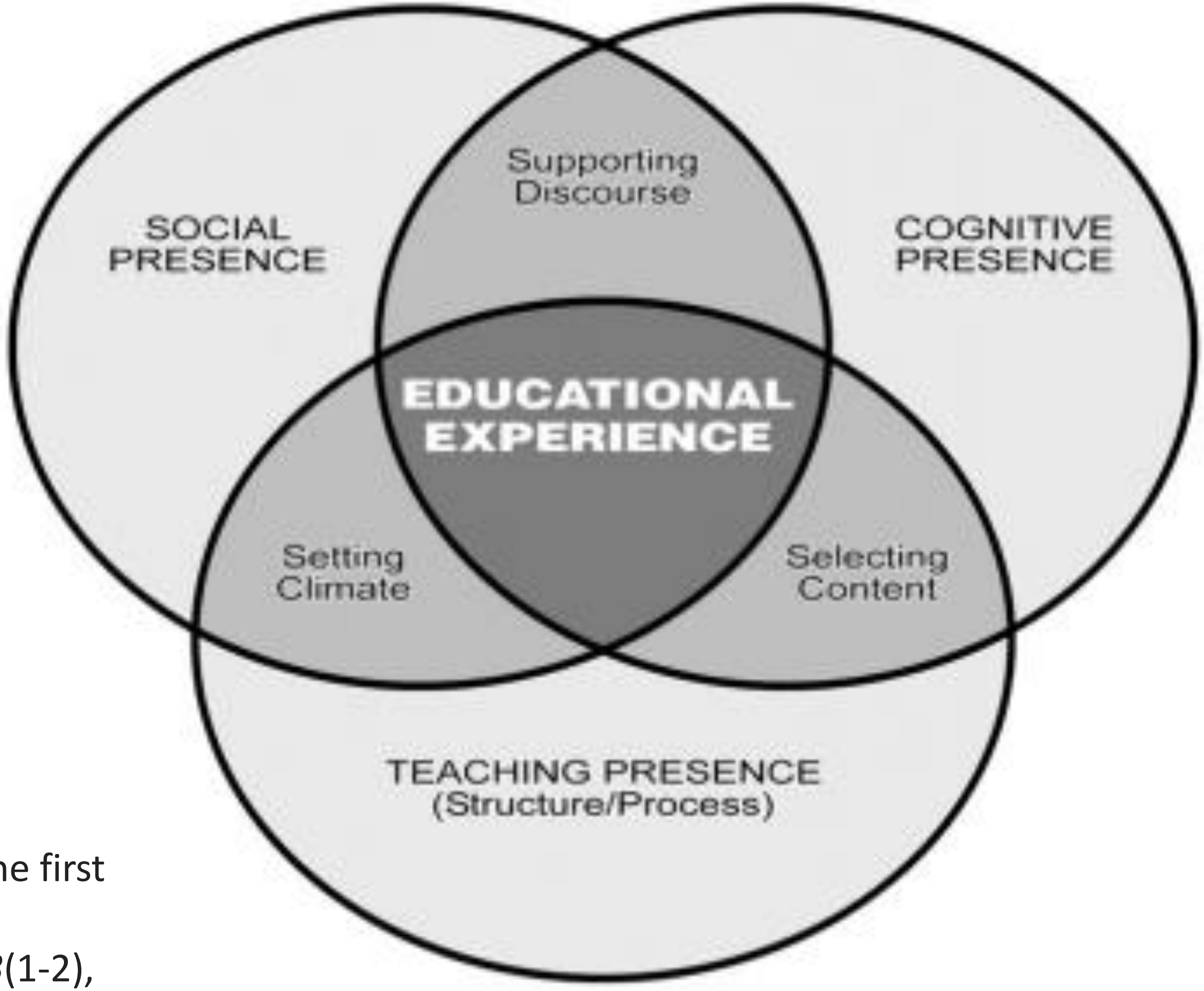
Learning Community (n =54)		
	M	**SD**
In-degree	5.26	2.14
Out-degree	5.26	0.80
Closeness	0.40	0.02
Betweenness	79.44	33.03



Cluster Analysis Results

Learning Community (n =54)		
Clusters	**Frequency**	**Proportion**
Peripheral Members	26	0.48
Intermediate Members	21	0.39
Central Members	7	0.13

Literature Review: Community of Inquiry



Garrison, D. R., Anderson, T., & Archer, W. (2010). The first decade of the community of inquiry framework: A retrospective. *The internet and higher education*, *13*(1-2), 5-9.

Topic Modeling Algorithm

Among different algorithms, I employed latent Dirichlet allocation (LAD) because

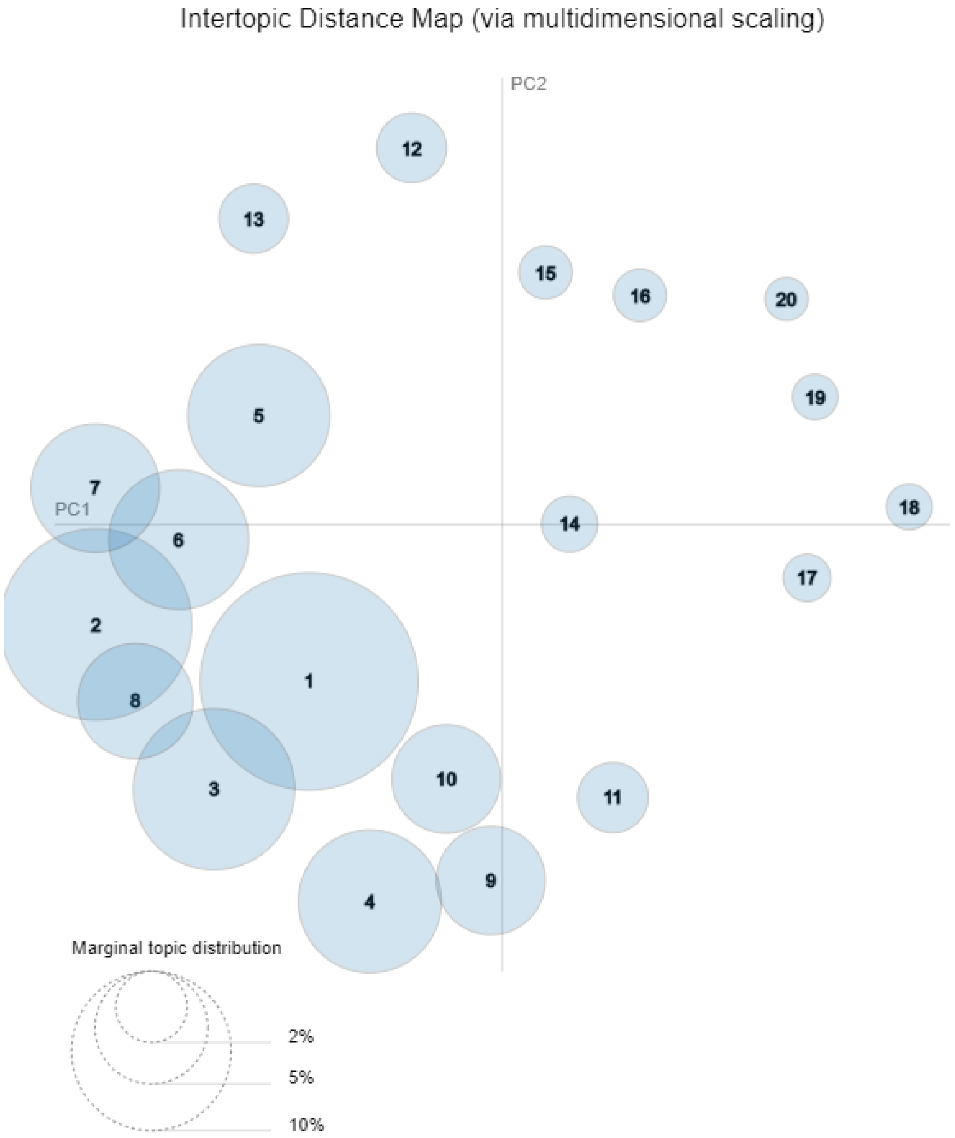
- There are many guides on how-to-aspects of LDA topic models
- LDA's outputs are easy to visualize

Palese, B., & Piccoli, G. (2020). Evaluating Topic Modeling Interpretability Using Topic Labeled Gold-standard Sets. *Communications of the Association for Information Systems*, *47*(1), 16.

https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

Topic Modeling Algorithm

- Perplexity: Captures a model’s uncertainty to predict unobserved documents
- Topic Coherence: Captures the degree of semantic similarity among a topic’s top words



https://fwd.delabapps.eu/topic_modelling.html

Topic Modeling Results: Peripheral Cluster

Label	Most Frequent Words	Distribution Of Topics
Access to Kaiser and AAA's insurance programs	kaiser, aaa, insurance, health, access	18%
Cost of personal health information	information, health, sell, personal, google	17%
Digital divide	Individuals, low-income, smartphone, patient, access	15%
HealthATM system usability	usability, people, healthatm, learn, easy	13%
Medical records confidentiality	privacy, security, information, healthcare, records	11%
Off-topic	smart, toilets, lives, weird, comment	10%
Persuasive design	encouragement, help, specific, people, behaviors	8%
PHR adoption in underserved communities	underserved, populations, system, health, phr	8%
Total Within the Peripheral Members Cluster		100%
Coherence Score		0.53
Perplexity Score		-5.87

Topic Modeling Results: Intermediate Cluster

Label	Most Frequent Words	Distribution Of Topics
Rapid application development	rapid, application, development, authors, approach	20%
Usability issues	application, constraint, users, phr, problems	16%
Waterfall development	phase, system, waterfall, authors, development	15%
Building an information system with Google's API	api, application, google, example, program	13%
Gamification systems	gamification, phr, keep, track, service	13%
HealthATM system usability and usefulness	people, useful, healthatm, find, easy	10%
Samsung health application	phone, app, Samsung, health, information	7%
Gaps in healthcare	issue, patient, health, care, gap	6%
Total Within the Peripheral Members Cluster		100%
Coherence Score		0.59
Perplexity Score		-6.14

Topic Modeling Results: Central Cluster

Label	Most Frequent Words	Distribution Of Topics
HIPAA Requirements	hipaa, laws, records, privacy, important	23%
Rapid application development	sdlc, rad, sounds, used, since	20%
Waterfall development	Waterfall, agree, method, determining, needs	18%
System Security	api, google, microsoft, security, used	18%
Design phase in system development lifecycle	sdlc, phase, design, model, system	11%
Patient activation measure score in healthatm software	pam, healthatm, score, patients, software	10%
Total Within the Peripheral Members Cluster		100%
Coherence Score		0.62
Perplexity Score		-5.63

Community of Inquiry Message Coding Schema: Peripheral Cluster

	Peripheral Members (n=26)		Intermedia Members (n=21)		Central Members (n=7)		ANOVA Test Results
Message Category	M	SD	M	SD	M	SD	
Connect ideas from course content/reading	0.43	0.13	0.33	0.07	0.30	0.07	$F(2,51) = 7.27$, $p = 0.002$, $\eta_p^2 = 0.53$

Message Category	Cluster Pairs	Tukey HSD Q Statistic	Tukey HSD Inference
Connect ideas from course content/reading	Peripheral vs Intermediate	4.47	** $p < 0.01$
	Peripheral vs Central	4.24	* $p < 0.05$
	Intermediate vs Central	1.12	insignificant

Community of Inquiry Message Coding Schema: Peripheral Cluster

	Peripheral Members (n=26)		Intermedia Members (n=21)		Central Members (n=7)		ANOVA Test Results
Message Category	M	SD	M	SD	M	SD	
Information exchange (i.e., a factual question, answer, or clarification)	0.20	0.18	0.09	0.08	0.10	0.05	*F*(2,51) = 4.30, *p* < 0.02, $\eta_p^2 = 0.41$

Message Category	Cluster Pairs	Tukey HSD Q Statistic	Tukey HSD Inference
Information exchange (i.e., a factual question, answer, or clarification)	Peripheral vs Intermediate	3.95	* *p* < 0.05
	Peripheral vs Central	2.42	insignificant
	Intermediate vs Central	0.29	insignificant

Community of Inquiry Message Coding Schema: Intermediate Cluster

	Peripheral Members (n=26)		**Intermedia Members (n=21)**		**Central Members (n=7)**		**ANOVA Test Results**
Message Category	M	SD	M	SD	M	SD	
Expressing puzzlement from instructional materials	0.05	0.07	0.14	0.08	0.08	0.08	$F(2,51) = 9.05$, $p < 0.001$, $\eta_p^2 = 0.59$

Message Category	**Cluster Pairs**	**Tukey HSD Q Statistic**	**Tukey HSD Inference**
Expressing puzzlement from instructional materials	Peripheral vs Intermediate	6.00	** $p < 0.01$
	Peripheral vs Central	1.36	insignificant
	Intermediate vs Central	2.70	insignificant

Community of Inquiry Message Coding Schema: Intermediate Cluster

	Peripheral Members (n=26)		Intermedia Members (n=21)		Central Members (n=7)		ANOVA Test Results
Message Category	M	SD	M	SD	M	SD	
Discussion of comprehension issues and alternate views	0.04	0.07	0.16	0.07	0.12	0.07	$F(2,51) = 18.47$, $p < 0.001$, $\eta_p^2 = 0.85$

Message Category	Cluster Pairs	Tukey HSD Q Statistic	Tukey HSD Inference
Discussion of comprehension issues and alternate views	Peripheral vs Intermediate	8.53	** $p < 0.01$
	Peripheral vs Central	3.66	* $p < 0.05$
	Intermediate vs Central	2.16	insignificant

Community of Inquiry Message Coding Schema: Central Cluster

	Peripheral Members (n=26)		Intermedia Members (n=21)		Central Members (n=7)		ANOVA Test Results
Message Category	M	SD	M	SD	M	SD	
Seeking to reach consensus\ understanding	0.05	0.07	0.03	0.05	0.12	0.01	$F(2,51) = 5.49$, $p = 0.007$, $\eta_p^2 = 0.42$

Message Category	Cluster Pairs	Tukey HSD Q Statistic	Tukey HSD Inference
Seeking to reach consensus\ understanding	Peripheral vs Intermediate	1.79	insignificant
	Peripheral vs Central	3.57	* $p < 0.05$
	Intermediate vs Central	4.69	** $p < 0.01$

Community of Inquiry Message Coding Schema: Central Cluster

	Peripheral Members (n=26)		Intermedia Members (n=21)		Central Members (n=7)		ANOVA Test Results
Message Category	M	SD	M	SD	M	SD	
Offer solution to comprehension issues	0.04	0.07	0.03	0.05	0.11	0.09	*F*(2,51) = 4.18, *p* = 0.02, η_p^2 = 0.43

Message Category	Cluster Pairs	Tukey HSD Q Statistic	Tukey HSD Inference
Offer solution to comprehension issues	Peripheral vs Intermediate	0.78	insignificant
	Peripheral vs Central	3.58	* $p < 0.05$
	Intermediate vs Central	4.02	* $p < 0.05$

Summary of Key Findings

- Peripheral Cluster (n=26)
  - Participants focused on the topics: Access to Kaiser and AAA’s insurance programs, cost of personal health information, and digital divide
  - Their messages connected these topics to their personal experiences and involved factual questions, answers, clarifications
- Intermediate Cluster (n=21)
  - Participants focused on the topics: Rapid application development, usability issues, and waterfall development
  - Their messages expressed puzzlements. They discussed comprehension issues/alternative viewpoints

Summary of Key Findings

- Central Cluster (n=7)
  - ❖Participants focused on the topics: HIPAA requirements, rapid application development, waterfall development, and system security
  - ❖ Their messages offered potential solutions to the comprehension issues and they tried to reach consensus on those solutions

Message Lexical Complexity

	Central Members' Messages (n=91)		**Intermediate Members' Messages (n=150)**		**Peripheral Members' Messages (n=61)**		**ANOVA Test Results**
	M	SD	M	SD	M	SD	
Message Lexical Complexity Score	5.21	1.24	5.36	1.26	5.48	1.33	F(2,299) = 0.80, p = 0.45, η_p^2 = 0.07

B. Thoms, E. Eryilmaz, N. Dubin, R. Hernandez, S. Colon-Cerezo, "Real-Time Visualization to Improve Quality in Computer Mediated Communication," Web Intelligence Journal, September, 2019.

Thank you
For Your Attention

Your Comments and Questions are welcomed.

Please address feedback to:

evren.eryilmaz@csus.edu