MORF ENA – A Tool for Making MOOC Discussion Forum Data More Accessible for Epistemic Network Analysis

Zhanlan Wei, Xiner Liu, Amanda Barany, Kirk Vanacore, Ryan S. Baker, Stefan Slater,
University of Pennsylvania, zhanlanw@upenn.edu, xiner@upenn.edu, amanda.barany@gmail.com,
kirk.vanacore@gmail.com, ryanshaunbaker@gmail.com, slater.research@gmail.com
Jade Pratt, Pepperdine University, jade.pratt@pepperdine.edu
Mamta Shah, Elsevier, University of Pennsylvania, m.shah@elsevier.com
Michael Mogessie, Carnegie Mellon University, michaelmogessie@cmu.edu

Abstract: MOOC discussion forum data can help researchers answer questions about learners' engagement, thought processes, and knowledge development. However, this data is often inaccessible due to technical and privacy constraints. We introduce MORF-ENA, an automated tool that enables researchers to conduct Quantitative Ethnographic research using Epistemic Network Analysis on MOOC data without direct access to learner data. Developed through a participatory design approach, MORF-ENA enables researchers to visualize and explore learners' behaviors within discussions. We present the tool's capabilities with an example, examining how learners' engagement with MOOCs changed before, during, and after the COVID-19 pandemic. The results of this design process point toward future directions for the tool's development and use in learning sciences as well as for improving access to learning data.

Introduction

Increased access to and use of Massively Open Online Courses (MOOCs) have created new opportunities to better understand and support online learning. MOOC datasets diversity, granularity, and size, allows researchers to identify patterns in learner engagement (Topali et al., 2024), predict academic success (Gardner & Brooks, 2018a), and develop interventions aimed at impacting student learning processes (Cobos & Ruiz-Garcia, 2021). However, data access and scalability issues have limited MOOC data use. Sensitive user data must be stored in secure databases requiring advanced technical skills to access. Although, some data types, such as clickstream logs, can be more easily de-identified and shared (Crossley et al., 2016), other forms of data, such as discussion forum posts, have limited availability due to the challenges of full de-identification (Zeide & Nissenbaum, 2018). Through complex data enclaves, this rich data has slowly become available to a small number of researchers with high degrees of technical skill, but overall access remains limited (Hutt et al., 2022).

Second, MOOC datasets are massive, containing interactions from tens or even hundreds of thousands of learners across multiple course dimensions over weeks. The complexity and richness of MOOC data require sophisticated algorithms and scalable analyses (e.g., Chen & Poquet, 2022). Many machine learning methods – such as predictive analytics – leverage data to accurately predict student learning (e.g. Gardner & Brooks, 2018a), but by themselves often do not explain learning processes (Conati et al., 2018). This has led to calls both for more explainable artificial intelligence methods (Khosravi et al., 2022) and for the use of mixed methods approaches that combine large-scale algorithms with more interpretive methods. For example, Quantitative Ethnography (QE) analyzes large datasets focused on making meaning of student interactions and learning. QE systematically quantifies complex qualitative data, enabling visualization and comparison of patterns (Shaffer, 2017). While QE techniques such as Epistemic Network Analysis (ENA; Marquart et al., 2021) can be used to analyze large-scale MOOC datasets, existing tools require direct dataset access. To bridge this gap, we introduce and demonstrate the application of a designed tool – MORF-ENA – that securely connects the MOOC Replication Framework (MORF; Gardner et al., 2018b) – a large-scale data enclave – to QE tools that can automate the coding and visualization of student discourse practices. The MORF-ENA tool allows researchers to explore, create, and compare epistemic networks of patterns in MOOC discussion forums in ways that preserve the security of raw datasets while automating elements of data processing and analysis to support greater access.

In this work, we describe the design of MORF-ENA, from codebook development to user testing, and outline how developers generate ENA models for external researchers. Then, we demonstrate the application of MORF-ENA to address the ways in which learners' online engagement and interaction patterns on MOOCs vary across three distinct periods: before (2019), during (2020), and after the COVID-19 pandemic (2022). This work demonstrates how secure access to large-scale datasets can yield insights into MOOC learning. We conclude with participant researchers' reflections on the process, and a discussion of the advantages and constraints of the MORF-ENA approach.

Literature review

Quantitative Ethnography (QE) is a methodological approach that combines quantitative and qualitative techniques to analyze and model complex patterns within data-rich environments (Shaffer, 2017). At its core, QE seeks to address how meaning is constructed in complex datasets by examining connections between identified themes. Given these affordances, QE research has seen increasing adoption in the learning sciences (Kaliisa et al., 2021). However, QE research, such as ENA, also required rich, contextualized data on student activities, behaviors, and processes within the learning system, thus posing challenges for data capture, processing, and storage. The learning sciences community has taken significant strides to make high-quality data available for research by a broader community of scholars (e.g., Prihar et al. 2022). However, some forms of data are difficult to fully anonymize (Hutt et al., 2022), and there remains a lack of supportive technologies for research using educational data that ensure student privacy and agency.

To address privacy concerns, richer educational data is often de-identified using PII word libraries (Bosch et al., 2020) or large language models (Singhal et al., 2024), though both methods can miss PII. The MORF-ENA tool offers a potential solution by identifying cross-cutting themes in MOOC discussion boards and automating coding so that researchers may explore patterns of discourse while the underlying data remains secure.

Research context

The MOOC Replication Framework (MORF) is an open-source system designed to facilitate reproducible research using MOOCs data (Gardner et al., 2018b). It not only provides researchers with access to large-scale data from MOOC platforms and other intelligent tutor systems but also supports comprehensive analyses and the replication of previous results across multiple datasets. To design and test the MORF-ENA tool, we used a fully deidentified data sample of 3503 forum posts from 2882 students, collected between 2012 and 2015 from nine courses on MOOCs. These courses covered topics such as accounting, calculus, design, gamification, business trends, poetry, mythology, probability, and vaccines. Student posts were distributed relatively evenly across courses. All potential PII was manually redacted by three human coders prior to further analysis. GPT was used as a further deidentification check, catching a small number of cases that the human coders had missed (Singhal et al., 2024). While these deidentification processes are not feasible for all data in the broader MORF enclave (which has data from millions of learners), this training set reflects the diversity of MOOC topics on MORF and allowed participant researchers to view data examples when supporting qualitative coding and tool development.

Method

We began by identifying cross-cutting themes in the MOOC discussion training set that could be applied to the wider MORF database. We developed a codebook of qualitative themes that appear in discussion boards across all courses in the data sample using the procedure outlined by Weston and colleagues (2001), which includes code conceptualization, generation, continuing review, and refinement. Two human coders divided discussion posts and independently reviewed the sample. They identified inductive themes that meet two criteria: (1) recurring and broadly applicable across the sample of nine courses, and (2) concrete enough to be identified through a set of regular expressions, which include strings of words or characters that could be used to identify the themes in the dataset automatically. Researchers then compared, discussed, and consolidated codes into a codebook of 9 constructs: (1) *Apologies*, (2) expressions of *Gratitude*, (3) personal *Introductions*, (4) discussions of *Course Logistics*, (5) *External Resource Sharing* (e.g., links, readings), (6) discussions of course *Evaluations* (e.g., tests, essays), (7) discussions of *Lecture Videos*, and general (8) *Positive Expressions* and (9) *Negative expressions*.

We then used two approaches to automate the application of codes to the discussion board data. First, Codey, an automated tool for semantic text analysis, (Rietz & Maedche, 2021) was used to construct an initial set of semantic categories used for coding. This was achieved using regular expressions (regex; e.g. "sorry" for *Apologies*) and string searches within student discussion board posts. Regex allows for pattern-based text matching and the identification of relevant text across large datasets in cases where specific use of verb tense, pluralization, or other characteristics make exact string matching difficult. While Codey excelled at systematically applying predefined rules, it was less flexible when it came to handling complex patterns. Therefore, additional regular expressions were generated using ChatGPT-4 through interactively designed prompts based on a subset of de-identified forum posts and construct definitions. These expressions achieved interrater reliability comparable to that of human coders after validation.

Each code was associated with multiple regex patterns, then refined iteratively by human coders to account for variations and contextual usage across the data. Each forum post was coded based on regex presence, with a binary assignment of 1 (present) or 0 (absent). Human coders first manually coded a new sample of 200 lines for each code to establish ground truth. They then met to resolve any discrepancies to finalize coding.

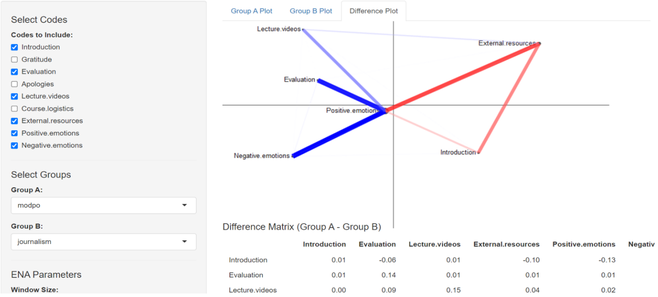
Acceptable human inter-rater agreement (Cohen's κ>0.70) was achieved for all codes before social moderation. The regular expressions were then applied to automatically code the same data and refined until human-automated inter-rater agreement thresholds were reached (Cohen's κ>0.70). The MORF-ENA tool will then automatically apply these expressions to new discussion data drawn from the MORF database. It is worth noting that our approach – using GPT to generate regular expressions based on de-identified data and then applying those regular expressions locally – avoids data privacy concerns that could arise from uploading raw student data to GPT.

Tool design

MORF-ENA is a shiny app built using R (source available at https://cran.qe-libs.org/codey/) that allows researchers to request specific courses from the MORF repository. MORF-ENA imports the data into a secure environment for subsequent analysis while maintaining strict data confidentiality. MORF-ENA then automatically extracts forum post titles and their content into a single variable for subsequent coding. Automated classifiers are then applied to the data to code for the selected constructs (1 if present, 0 if absent).

As shown in Figure 1, MORF-ENA draws on relevant meta-data (e.g., course types, dates, users, etc.) and the automated codes to generate epistemic network visualizations based on the user's specifications. The tool provides options for single-course analysis or difference models (comparing two courses) of student connection-making between the selected constructs. The tool also provides plot descriptions, related statistics (e.g., code frequencies, significance tests), tables of normalized code connection weights, and additional summaries based on user-selected variables, such as weekly post counts. Throughout the entire process, users do not have direct access to the raw data and instead generate plots directly through the tool based on selected parameters. This approach addresses the current Web ENA tool's limitation of requiring users to upload their own datasets for subsequent analysis. Figure 1 shows the interface of the MORF-ENA tool (version 1). To gather feedback on current and possible future tool functions, we further engaged in participatory design research with QE scholars.

Figure 1
MORF-ENA tool with a difference network for two courses, line weights based on user selection.



Select Codes

Codes to Include:

- ☑ Introduction
- ☐ Gratitude
- ☑ Evaluation
- ☐ Apologies
- ☑ Lecture.videos
- ☐ Course.logistics
- ☑ External.resources
- ☑ Positive.emotions
- ☑ Negative.emotions

Select Groups

Group A: modpo

Group B: journalism

ENA Parameters

Window Size:

Difference Matrix (Group A - Group B)

	Introduction	Evaluation	Lecture.videos	External.resources	Positive.emotions	Negativ
Introduction	0.01	-0.06	0.01	-0.10	-0.13	
Evaluation	0.01	0.14	0.01	0.01	0.01	
Lecture.videos	0.00	0.09	0.15	0.04	0.02	

Participatory design

Participatory research centers on collaborating with communities to co-create knowledge and address social inequalities (Bergold & Thomas, 2012; Vaughn & Jacquez, 2020), while participatory design applies these principles to iteratively co-develop tools through shared goals and feedback (Spinuzzi, 2005). This study integrates both research approaches, in which all participants are researchers (including the MORF-ENA developers), but we intentionally included the inputs of learning science research stakeholders who use the tool and datasets to answer research questions that may be relevant to their current and future work. First, we connected with QE collaborators who participated in the annual Quantitative Ethnography Data Challenge, where scholars meet online for rapid, week-long collaboration around a research topic and dataset of their choosing. These scholars used the MORF-ENA tool in their analyses and later published their work at the 2024 International Conference of Quantitative Ethnography. After the data challenge, a tool developer (author 3) held a virtual session with research stakeholders (authors 6 and 7) to gather insights on the tool's strengths and weaknesses, design improvements, and challenges in automating QE research. Stakeholder feedback was then implemented in the next iteration of tool development and directly contributed to refinement to better align with stakeholder needs.

Participatory design research results

The research stakeholders who participated in the QE Data Challenge were conference attendees looking for collaborators who had datasets that they could use to explore their topics of interest. MORF-ENA designers met with the team and introduced the features of the tool (e.g., existing codes and definitions) and underlying dataset (e.g., course types, years, meta-data). From there, the team brainstormed research questions and explored how

MORF-ENA might help provide answers. Research stakeholders noted that the database contained multiple iterations of the same course over time, which prompted discussions on how students' online interaction patterns may have shifted due to COVID-19. After considering several MOOCs available on MORF-ENA, the research stakeholders selected a widely used Design course that was popular over a long period on Coursera. They then reviewed the list of variables made available through MORF-ENA to build epistemic networks illustrating discussion posts differences across time periods. Key variables such as course name, course date, and anonymous user IDs for each post were used, while the less relevant variables, such as class size, were excluded. The research stakeholders opted to include all provided automated codes in the model to explore discourse variations over time.

The resultant dataset consisted of 4254 forum posts from the Design course collected between 2019 and 2022. Results included difference models between course data from pre-pandemic (2019), during the pandemic (2020), and post-pandemic (2022). Findings revealed changes in 2020, including a surge in posts, shorter posts, and increased peer evaluations and external resources. In 2022, post-pandemic interaction patterns shifted toward course logistics, reflecting evolving learner support needs in online education. While research stakeholders felt that patterns revealed in the networks were contextually meaningful, they shared that their ability to fully explore, understand, or communicate their findings was limited without example discussion posts to help them close the interpretive loop. In other words, the de-identified dataset provided broad context but lacked specificity for their research questions. Therefore, they requested that the MORF-ENA design team share a small set of discussion examples from each year. Following this request, the MORF-ENA team manually extracted, manually de-identified, and shared these data, which provided needed contextual information while protecting privacy.

The researcher stakeholders identified several key strengths of the MORF-ENA tool, particularly its ability to provide easy access to large-scale MOOC data while streamlining analysis. They appreciated the tool's accessibility and flexibility for users with varying levels of experience, emphasizing how it allows users to explore different research questions by grouping and visualizing relationships between multiple constructs within the network. They saw this functionality as serving as an intuitive entry point, especially for researchers less familiar with network analysis methods, making the exploration of MOOC learning patterns more approachable. One stakeholder noted how the tool makes it *"easier to examine the complex dynamics of student engagement"*. They also highlighted the advanced features that are easy to access for users who already have in-depth experience with ENA: "*the ability to adjust line weight and tabulating line weights and occurrence frequency is very convenient*."

Stakeholders also provided feedback for improving the tool. First, they proposed features to aid researchers in writing and presenting findings derived from MORF-ENA analyses, such as a methodological template to help researchers describe the ENA model creation, and a download button for frequency tables and pre-titled ENA model figures. These improvements would streamline the integration of MORF-ENA outputs into academic manuscripts. Another recommendation was to provide resources for new users unfamiliar with ENA, specifically a step-by-step tutorial for generating models and interpreting results, similar to the Web ENA tool. They also provided an alternate design idea, creating a community-shared playlist of worked ENA design examples with remixable options for diverse research needs. Lastly, they emphasized the importance of bridging the gap between quantitative network models and real-world learning behaviors, proposing an interface for requesting de-identified text cases, to provide the necessary context for interpreting specific discourse patterns.

Discussion and future work

Our stakeholder discussions led to two enhancements that will be developed in further work. First, designers will offer a small deidentified dataset to give users a broader context of the data and coding categories. Second, designers will build a feature in the MORF-ENA tool that will allow users to request specific qualitative examples directly through the tool's interface. Upon request, the MORF team will extract, redact (automated followed by human review to ensure privacy), and share a small sample of learners' forum posts. Future versions of the tool will also include new plot types and visualizations that address user needs and contexts and offer more support in the selection of data subsamples such as course week or course iteration.

Our work shows how MORF-ENA enables researchers to study sensitive, large-scale learner data while preserving privacy, offering pre-developed codes and visualizations to analyze online engagement and learning patterns. However, the current MORF-ENA tool has several limitations. While it generates ENA models of MOOC learning patterns, access to raw data and qualitative examples remains limited, which may result in a loss of contextual nuance needed for qualitative analysis. Similarly, the use of regular expressions ensures consistency in coding, but may fail to capture all aspects of the desired constructs. Recent work using LLMs for qualitative coding (Liu et al., 2024) may address this concern, as the quality of open-source and locally runnable LLMs rapidly increases. Finally, the labor and time-consuming code validation process may also delay analysis and reduce flexibility in adapting to evolving research needs. Despite these limitations, tools like MORF-ENA enable use of data by a broader range of scholars, contributing to broader uptake of ENA methods for more types of data.

References

Bergold, J., & Thomas, S. (2012). Participatory research methods: A methodological approach in motion. *Historical Social Research/Historische Sozialforschung*, 191-222.

Bosch, N., Crues, R. W., Shaik, N., & Paquette, L. (2020). "Hello, [REDACTED]": Protecting student privacy in analyses of online discussion forums. *13th Intl. Conf. on Educational Data Mining*, 39-49.

Chen, B., & Poquet, O. (2022). Networks in learning analytics: Where theory, methodology, and practice intersect. *J. of Learning Analytics, 9*(1), 1-12.

Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). AI in Education needs interpretable machine learning: Lessons from open learner modeling. *3rd Workshop on Human Interpretability in Machine Learning.*

Cobos, R., & Ruiz-Garcia, J. C. (2021). Improving learner engagement in MOOCs using a learning intervention system: A research study in engineering education. *Computer Applications in Engineering Education, 29*(4), 733-749.

Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. *Intl. Conf. on Learning Analytics & Knowledge,* 6-14.

Gardner, J., & Brooks, C. (2018a). Student success prediction in MOOCs. *User modeling and user-adapted interaction*, 28, 127-203.

Gardner, J., Brooks, C., Andres, J. M., & Baker, R. S. (2018b). MORF: A framework for predictive modeling and replication at scale with privacy-restricted MOOC data. *2018 IEEE Intl. Conf. on Big Data,* 3235-3244.

Hutt, S., Baker, R. S., Ashenafi, M. M., Andres-Bray, J. M., & Brooks, C. (2022). Controlled outputs, full data: A privacy-protecting infrastructure for MOOC data. *British J. Educational Technology, 53*(4), 756-775.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3, 100074.

Kaliisa, R., Misiejuk, K., Irgens, G. A., & Misfeldt, M. (2021). Scoping the emerging field of quantitative ethnography: Opportunities, challenges and future directions. *Intl. Conf. on Quantitative Ethnography,* 3-17.

Liu, X., Zhang, J., Barany, A., Pankiewicz, M., Baker, R. S. (2024) Assessing the potential and limits of large language models in qualitative coding. *Intl. Conf. on Quantitative Ethnography*, 89-103.

Marquart, C. L., Hinojosa, C., Swiecki, Z., Eagan, B., & Shaffer, D. W. (2021). Epistemic network analysis (Version 1.7.0). http://app.epistemicnetwork.org.

Marshall, R., Pardo, A., Smith, D., & Watson, T. (2022). Implementing next generation privacy and ethics research in education technology. *British J. of Educational Technology, 53*(4), 737-755.

Prihar, E., Syed, M., Ostrow, K., Shaw, S., Sales, A., & Heffernan, N. (2022). Exploring common trends in online educational experiments. *15th Intl. Educational Data Mining Conf*, 27-38.

Rietz, T., & Maedche, A. (2021). Cody: An AI-based system to semi-automate coding for qualitative research. *2021 CHI Conf. on Human Factors in Computing Systems*, 1-14.

Spinuzzi, C. (2005). The methodology of participatory design. *Technical communication, 52*(2), 163-174.

Shaffer, D. W. (2017). *Quantitative ethnography*. Madison, WI: Cathcart Press.

Singhal, S., Zambrano, A. F., Pankiewicz, M., Liu, X., Porter, C., & Baker, R. S. (2024). De-identifying student personally identifying information with GPT-4. *17th Intl. Conf. on Educational Data Mining*, 559-565.

Topali, P., Asensio-Pérez, J. I., Ortega-Arranz, A., Martínez-Monés, A., Villagrá-Sobrino, S. L., & Dimitriadis, Y. (2024). Unveiling the role of learning design on feedback in MOOCs. *18th Intl. Conf. of the Learning Sciences-ICLS 2024*, 1151-1154.

Vaughn, L. M., & Jacquez, F. (2020). Participatory research methods–choice points in the research process. *J. of Participatory Research Methods, 1*(1).

Zeide, E., & Nissenbaum, H. (2018). Learner privacy in MOOCs and virtual education. *Theory and Research in Education, 16*(3), 280-307.

Acknowledgments

This research was supported by National Science Foundation Grants # DRL1931419 and # DUE-2153481.