Discourse: MOOC Discussion Forum Analysis at Scale

Alexander Kindel[1], **Michael Yeomans**[2], **Justin Reich**[3], **Brandon Stewart**[1], **Dustin Tingley**[2]
[1]Princeton University, [2]Harvard University, [3]MIT
akindel@princeton.edu, myeomans@fas.harvard.edu, jreich@mit.edu,
bms4@princeton.edu, dtingley@gov.harvard.edu

ABSTRACT
We present *Discourse*, a tool for coding and annotating MOOC discussion forum data. Despite the centrality of discussion forums to learning in online courses, few tools are available for analyzing these discussions in a context-aware way. Discourse scaffolds the process of coding forum data by enabling multiple coders to work with large amounts of forum data. Our demonstration will enable attendees to experience, explore, and critique key features of the app.

Author Keywords
Discussions; content analysis; reply mapping; MOOCs

INTRODUCTION
Many kinds of social interactions are now mediated through technology — online forums, chatrooms, social media, and other collaborative platforms. These discussions produce rich datasets that can help researchers discover new answers to fundamental questions about the nature of communication. For example, online courses bring together large, diverse groups of students to discuss political or controversial topics, and record a wealth of forum data from their discussions [1,2].

Discussion data pose unique challenges for qualitative data analysis, because the meaning of individual posts are dependent on the posts that came before. However, the tools most commonly used for qualitative coding do not take this context into account. Some tools only show one document at a time (e.g. NVivo), which hides the necessary context of the focal post. Other tools display all of the data at once (e.g. Excel), which puts the burden of parsing context on the coder.

Here we present a tool that was built for efficient, context-sensitive qualitative coding of discussion data at scale. Our system allows many discussions to be simultaneously coded by many coders. Furthermore, the posts are automatically embedded in the relevant context of the discussion, so that coders can more easily understand the posters' intent.

L@S 2017, April 20-21, 2017, Cambridge, MA, USA
ACM 978-1-4503-4450-0/17/04
http://dx.doi.org/10.1145/3051457.3053967

DESIGN AND ADVANTAGES
Discourse is a specialized tool for the qualitative analysis of discussion data that offers two key advances over existing content analysis software. First, Discourse is aware of the structure of discussion data, rather than treating each post as an atomic document. Specifically, each post is presented along with its "parents" - the posts to which it replies, which are higher up in the thread structure and posted before the focal post. This allows each coder to efficiently understand the context in which the focal post was made, to more accurately label the content of the post (see Fig 1).

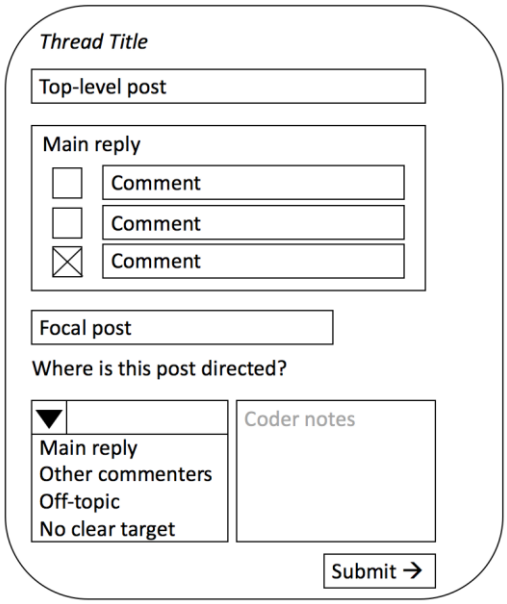

Figure 1. Coder-level view (schematic). Focal post is presented below the parent posts that preceded, along with a menu of coding choices and optional text box for coder notes.

Second, Discourse automates the management of coding. Coders select from a researcher-defined set of codes in a drop-down menu, which minimize the possibility of typos or mis-targeted entries. The app automatically provides coders only the data they need to code, and each code is logged in a relational database immediately upon submission. Coding tasks can be dynamically assigned to coders based on workflow and availability, increasing potential throughput for large coding tasks. The app also simplifies the task of merging and exporting data at the conclusion of coding. Rather than collecting and collating

multiple spreadsheets or retrieving data from proprietary analysis software, data can be selected and exported directly from the app's database.

USE CASE

Discourse is currently supporting data analysis on discussion forum data from two edX MOOCs: *Introduction to American Government* and *Saving Schools*. To date, coders have generated approximately 40,000 codes on 140 discussion threads from these two courses. Discourse is compatible with forum data from any edX course, facilitating cross-course analyses and comparisons.

So far, coders have been focused on a task we call "reply mapping." This task is built to circumvent a limitation in edX (and other) forum data: forum participation is recorded in threads with finite (i.e. 3-level) depth. These metadata do not necessarily reflect the true reply structure of discussions. Comments at level 3 can (and often do) talk to one another rather than each responding to the level 2 post (see Fig. 2). This ambiguity makes it difficult to use the forum metadata to identify which users are interacting with one another.

For reply mapping, coders are tasked with identifying the previous posts (if any) towards which the focal post is directed. Discourse streamlines this task by displaying only those posts to which a comment could plausibly have responded. Preliminary results based in part on these codes have been submitted for publication [2].

FUTURE DIRECTIONS

At the time of submission, we have planned a number of features for the tool designed to make it easier to manage and reconcile results from multiple coders working on the same corpus of forum data. Dashboards for automatically calculating coder progress, reliability measures and agreement statistics provide research managers with a synoptic overview of data analysis as it progresses. Where a single canonical coding is needed, "reconciliation" or "tiebreaking" tasks permit managers to assign a third coder to only those forum posts where a pair of coders disagreed. By reducing the manual labor associated with merging and comparing codes, these features will simplify the task of managing qualitative data analysis on forum data.

In the future, we also plan to extend the app to handle any kind of discussion data. The unique advantages of our analysis tool—sensitivity to document context and automation of routine research management tasks—are applicable to a broad range of structured discussion data, including group meetings, chat logs, or forum data with infinite depth (e.g. Reddit). We look forward to seeking feedback from conference attendees on these planned features, as well as suggestions for other desirable analytic tools for forum data.

DEMONSTRATION

To demonstrate the app, we plan to briefly show the app's basic features. Additionally, a number of guest accounts on Discourse will be made available throughout the conference so that attendees may test out the app on their own devices. We also plan to seek user feedback through the demonstration in order to improve the usability and functionality of the app.

ACKNOWLEDGMENTS

We gratefully acknowledge grant support from the Spencer Foundation's New Civics initiative and the Hewlett Foundation. We also thank the course teams from *Saving Schools* and *American Government*, and the Harvard VPAL-Research Group for research support. Finally, we appreciate the research assistance from our coders: Ben Schenck, Elise Lee, Jenny Sanford, Holly Howe, Jazmine Henderson & Nikayah Etienne.

REFERENCES

1. Justin Reich, Brandon Stewart, Kimia Mavon, and Dustin Tingley. 2016. "The Civic Mission of MOOCs: Measuring Engagement across Political Differences in Forums." *Proceedings of the Third Annual ACM Conference on Learning at Scale, L@S 2016.* Edinburgh, UK.
2. Michael Yeomans, Justin Reich, Brandon Stewart, Kimia Mavon, Alex Kindel, and Dustin Tingley. 2016. "The Civic Mission of MOOCs: Engagement across Political Differences in Online Forums." *Under review.*

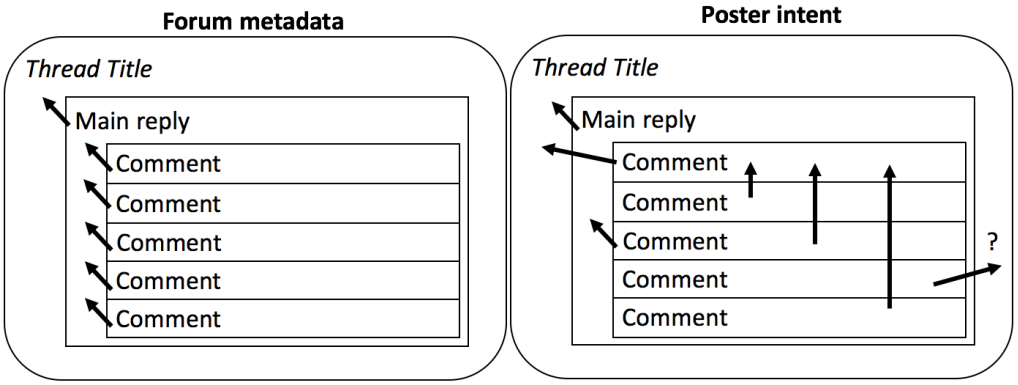

Figure 2. Recorded forum metadata (left) vs. true poster intent (right) structure of edX MOOC forum discussions. Arrows indicate comment target. Replies may skip earlier replies, respond directly to a top-level post, or be completely off-topic.