IRRODL
THE INTERNATIONAL
REVIEW OF RESEARCH IN
OPEN AND DISTANCE LEARNING

Towards an Integration of Text and Graph Clustering Methods as a Lens for Studying Social Interaction in MOOCs

Diyi Yang, Miaomiao Wen, Abhimanu Kumar, Eric P. Xing, and Carolyn Penstein Rosé
Carnegie Mellon University, United States

Abstract

In this paper, we describe a novel methodology, grounded in techniques from the field of machine learning, for modeling emerging social structure as it develops in threaded discussion forums, with an eye towards application in the threaded discussions of massive open online courses (MOOCs). This modeling approach integrates two simpler, well established prior techniques, namely one related to social network structure and another related to thematic structure of text. As an illustrative application of the integrated technique's use and utility, we use it as a lens for exploring student dropout behavior in three different MOOCs. In particular, we use the model to identify twenty emerging subcommunities within the threaded discussions of each of the three MOOCs. We then use a survival model to measure the impact of participation in identified subcommunities on attrition along the way for students who have participated in the course discussion forums of the three courses. In each of three MOOCs we find evidence that participation in two to four subcommunities out of the twenty is associated with significantly higher or lower dropout rates than average. A qualitative post-hoc analysis illustrates how the learned models can be used as a lens for understanding the values and focus of discussions within the subcommunities, and in the illustrative example to think about the association between those and detected higher or lower dropout rates than average in the three courses. Our qualitative analysis demonstrates that the patterns that emerge make sense: It associates evidence of stronger expressed motivation to actively participate in the course as well as evidence of stronger cognitive engagement with the material in subcommunities associated with lower attrition, and the opposite in subcommunities associated with higher attrition.

We conclude with a discussion of ways the modeling approach might be applied, along with caveats from limitations, and directions for future work.

Introduction

The contribution of this paper is an exploration into a new methodology that provides a view into the evolving social structure within threaded discussions, with an application to analysis of emergent social structure in massive open online courses (MOOCs). In the current generation of MOOCs, only a small percentage of students participate actively in the provided discussion forums (Yang et al., 2013; Rosé et al., 2014). However, social support exchanged through online discussions has been identified as a significant factor leading to decreased attrition in other types of online communities (e.g., Wang, Kraut, & Levine, 2012). Thus, a reasonable working hypothesis is that if we can understand better how the affordances for social interaction in MOOCs are functioning currently, we may be able to obtain insights into ways in which we can design more socially conducive MOOCs that will draw in a larger proportion of students, provide them with needed social support, and ultimately reduce attrition. In this paper we focus on the first step down this path, namely developing a methodology that can be used to gain a bird's eye view of the emerging social structure in threaded discussion. As such, this is a methods paper that describes a modeling approach, and illustrates its application with a problem that is of interest to the online and distance education community.

Current research on attrition in MOOCs (Koller et al., 2013; Jordan, 2013) has focused heavily on summative measures rather than on the question of how to create a more socially conducive environment. Some prior work has used clustering techniques applied to representations of clickstream data to identify student practices associated with levels of engagement or disengagement in the course (Kizilcec, Piech, & Schneider, 2013). Our work instead focuses on social interaction within the MOOC exclusively. In particular, the motivation is that understanding better the factors involved in the struggles students encounter and reflect to one another along the way can lead to design insights for the next generation of more socially supportive MOOCs (Yang et al., 2013; Rosé et al., 2014). As large longitudinal datasets from online behavior in MOOCs are becoming easier to obtain, a new wave of work modeling social emergence (Sawyer, 2005) has the potential to yield valuable insights, grounded in analysis of data from learning communities as they grow and change over time. Powerful statistical frameworks from recent work in probabilistic graphical models (Koller & Friedman, 2009) provide the foundation for a proposed new family of models of social emergence (Sawyer, 2005). This paper particularly focuses on integration of two well established prior techniques within this space, namely one related to social network structure (Airoldi et al., 2008) and another related to thematic structure of text (Blei et al., 2003).

From a technical perspective, we describe how the novel exploratory machine learning modeling approach, described in greater technical detail in our prior work (Kumar et al., 2014), is able to identify emerging social structure in threaded discussions. Our earlier account of the approach focused on the technical details of the modeling technology and an evaluation of its scalability in an online cancer support community and an online Q&A site for software engineers. This paper instead focuses on a methodology for using the approach in the context of research on MOOCs.

In the remainder of this paper, we begin by describing our methodology in qualitative terms meant to be accessible to researchers in online education and learning analytics. Next, we present a quantitative analysis that demonstrates that the detected subcommunity structure provided by the learned models predicts dropout along the way across three different Coursera MOOCs. Specifically, we describe how this modeling approach provides social variables associated with emerging subcommunities that students participate in within a MOOC’s threaded discussion forums. We evaluate the predictive validity of these social variables in a survival analysis. We then interpret the detected subcommunity structure in terms of the interests and focus of the discussions highlighted by the model’s representation. We conclude with a discussion of limitations and directions for future research, including proposed extensions for modeling emerging community structure in cMOOCs (Siemens, 2005; Smith & Eng, 2013).

Method

Data

In preparation for a partnership with an instructor team for a Coursera MOOC that was launched in fall of 2013, we were given permission by Coursera to extract the discussion data from and study a small number of courses. Altogether, the dataset used in this paper consists of three courses: one social science course, “Accountable Talk™: Conversation that works”, offered in October 2013, which has 1,146 active users (active users refer to those who post at least one post in a course forum) and 5,107 forum posts; one literature course, “Fantasy and Science Fiction: the human mind, our modern world”, offered in June 2013, which has 771 active users who have posted 6,520 posts in the course forum; and one programming course, “Learn to Program: The Fundamentals”, offered in August 2013, which has 3,590 active users and 24,963 forum posts. All three courses are officially seven weeks long. Each course has seven week specific subforums and a separate general subforum for more general discussion about the course. Our analysis is limited to behavior within the discussion forums. We will refer to the three data sets below as Accountable Talk, Fantasy, and Python respectively.

Modeling Emerging Subcommunities

Model overview.

The aim of our work is to identify the emerging social structure in MOOC threaded discussions, which can be thought of as being composed of bonds between students, which begin to form as students interact with one another in the discussion forums provided as part of many xMOOCs (e.g., MOOCs provided by Coursera, EdX, or Udacity). The structure of cMOOCs (Siemens, 2005; Smith & Eng, 2013) is more complex, and we address in the conclusion how the approach may be extended for such environments.

The unique developmental history of MOOCs creates challenges that can only be met by leveraging insights into the inner-workings of the social interaction taking place within those contexts. In particular, rather than evolving gradually as better understood forms of online communities, MOOCs spring up overnight and then expand in waves as new cohorts of students arrive from week to week to begin the course. Students may begin to form weak bonds with some other students when they join, however, massive attrition may create challenges as members who have begun to form bonds with fellow students soon find their virtual cohort dwindling.

Within these environments, students are free to pick and choose opportunities to interact with one another. As students move from subforum to subforum, they may take on a variety of stances as they interact with alternative subsets of students in discussions related to different interests, goals, and concerns. From the structure of the discussion forums, it is possible to construct a social network graph based on the post-reply-comment structure within threads. This network structure provides one view of a student's social participation within a MOOC, which may reflect something of the values and goals of that student. A complementary view is provided by the text uttered by the students within those discussions. In our modeling approach, we bring both of these sources of insight together into one jointly estimated integrated framework with the goal of modeling the ways in which the linguistic choices made by students within a discussion reflect the specific stances they take on depending upon who they are interacting with, and therefore which subcommunities are most salient for them at that time.

Just as Bakhtin argues that each conversation is composed of echoes of previous conversations (Bakhtin, 1981), we consider each thread within a discussion forum to be associated with a mixture of subcommunities whose interests and values are represented within that discussion. This mixture is represented by a statistical distribution. Whenever two or more users interact in a thread, they each do so assuming a particular manner of participation that contributes to that mixture of subcommunities via the practices that are displayed in their discussion behavior. Within each thread *t*, each user *u* is considered to have a probabilistic association with multiple subcommunities $c_1 \ldots c_n$ based on who he spends time talking with and the way he talks.

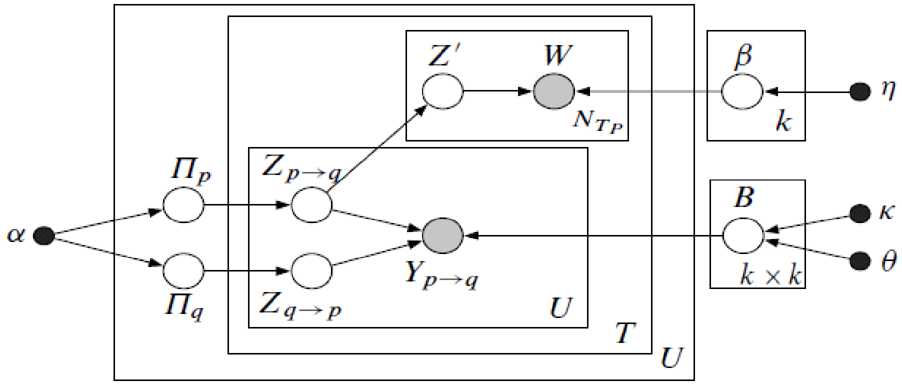

Figure 1. Graphical representation of the integrated LDA-MMSB model.

More technical readers may refer to the graphical representation of the model in plate notation in Figure 1. With reference to this representation we can state more formally as represented within the inner U plate that for each pair of users within a thread, which we may refer to as user p and user q, the distribution of subcommunities drawn for user p that reflects p addressing q is represented in the plate notation as $Z_{p\text{->}q}$, and likewise the distribution drawn for q is represented as $Z_{q\text{->}p}$. In addition to each thread specific distribution of subcommunity associations, users each have an overall distribution that represents their average tendency across all of the threads they have participated in. This is represented within the plate notation as Π_p and Π_q. This enables the model to prefer some consistency of user behavior across threads. The influence users p and q exert on one another's behavior arises from the MMSB portion of the model, which comprises a dirichlet prior (i.e., α, initialized with an assumed number of topics), from which are drawn the prior probability distribution over subcommunities associated with each user (i.e., Π_p and Π_q), and the inner U plate already described. As represented within the T plate, the LDA portion of the model reflects $Z_{p\text{->}q}$ as a mixture of word distributions, where each Z' represents a word distribution reflecting that of users when they are speaking as members of the subcommunity associated with Z'. A more extensive discussion of the technical details related to the model along with its parallelized approximate inference approach are published separately (Kumar et al., 2014).

Reflecting on the model from a conceptual standpoint, consistent with theories of social emergence (Sawyer, 2005), it is important to note that influence works both top-down, from the norms of the group to the behavior of the students within the group, and bottom-up, from the behavior of the student to the emerging norms of the group within discussions. Specifically, when users talk together on a thread, each user exerts some influence on the distribution of subcommunities whose values and goals are ultimately reflected in that conversation. However, each user is interacting with and responding to the other users on the thread. As a result, the set of users cumulatively exert some influence over the stance taken by each participant within the discussion. Thus, within a specific context, the distribution of subcommunities reflected in a participating user's behavior will be related both to the user's own tendencies and also to the tendencies of the other participants in that discussion. More formally, the cumulative reflected association of subcommunities within a thread t will emerge from the interaction of the set of users $u_1 \ldots u_n$ who are participating on t. And for each user u on thread t, his behavior on that thread will reflect each subcommunity c to the extent that it is associated with that user's own stance within that thread t. Because of this two way influence, it is reasonable to consider that subcommunity structure arises both from the pattern of connections embedded within the network constructed from the threaded reply structure and from the behaviors reflected through the text contributed within that structure. From a technical perspective, the interests and values of subcommunity c are reflected through an associated word distribution computed from the set of texts uttered by participants in subcommunity c. But they are also reflected through an association between nodes within the social network graph and subcommunity c. Thus, the representation of latent subcommunities $c_1 \ldots c_n$ mediates the network and the text.

Our model formulation integrates these two complementary views of subcommunity structure in one jointly estimated probabilistic model. This two-way influence may be modeled within this probabilistic framework through the iterative manner in which the model is estimated, which gives it a representational advantage over earlier multi-agent approaches to modeling social emergence (Hedtröm, 2005). In particular, as reflected in the structure of the plate notation, the model is estimated over the whole data set, but it is done by iterating over threads. On each thread iteration, the estimation algorithm iterates over the pairs of users who participate on the thread. And for each pair of users, it alternates between holding the LDA portion of the model constant while estimating the MMSB portion, and then holding the estimated MMSB portion constant while estimating the LDA portion.

The probabilistic formulation also has another advantage from a representation standpoint. In our model, a separate link structure is constructed for each thread. However, since each thread is associated with a distribution of subcommunities, and each subcommunity is associated with multiple threads, the text and network structures are conceptually linked. Most importantly, this probabilistic formulation enables us to represent the fact that participants reflect their connection to different subcommunities at different times depending on who they are talking with and what they are talking about. This novel approach contrasts with existing techniques built on a simple aggregation of reply networks into a single graph and user text across subforums into a single document per user and a hard partitioning of the network structure such that each user is treated as belonging only to one partition (e.g., Karypis & Kumar, 1995). This simplistic approach makes an invalid assumption about consistency of user behavior and can thus cause a severe loss of information in the resulting model, as demonstrated in our earlier work (Kumar et al., 2014).

Model reflections.

Our modeling approach integrates two types of probabilistic graphical models. First, in order to obtain a soft partitioning of the social network of the discussion forums, we used a mixed membership stochastic blockmodel (MMSB) (Airoldi et al., 2008). The advantage of MMSB over other graph partitioning methods is that it does not force assignment of students solely to one subcommunity. The model can track the way students move between subcommunities during their participation.

We made several extensions to the basic MMSB model. First, while the original model could only accommodate binary links that signal either that a pair of participants have interacted or not, we were able to make the representation of connections between nodes more nuanced by enabling them to be counts rather than strictly binary. Thus, the frequency of interaction can be taken into account. Secondly, we have linked the community structure that is discovered by the model with a probabilistic topic model, so that for each person a distribution of identified communicative themes is estimated that mirrors the distribution across subcommunities. By integrating these two modeling approaches so that the representations learned by each are pressured to mirror one another, we are able to learn structure within the text portion of the model that helps identify the characteristics of within-subcommunity communication that distinguish various subcommunities from one another. A well known approach is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which is a generative model and is effective for uncovering the thematic structure of a document collection.

LDA works by associating words together within a latent word class that frequently occur together within the same document. The learned structure in LDA is more complex than traditional latent class models, where the latent structure is a probabilistic assignment of each whole data point to a single latent class (Collins and Lanza, 2010). An additional layer of structure is included in an LDA model such that words within documents are probabilistically assigned to latent classes in such a way that data points can be viewed as mixtures of latent classes. By allowing the representation of documents as arbitrary mixtures of latent word classes, it is possible then to keep the number of latent classes down to a manageable size while still capturing the flexible way themes can be blended within individual documents.

Modeling Attrition

In order to evaluate the impact of social factors on continued participation within the MOOC context, we used a survival model, as in prior work modeling attrition over time (Wang, Kraut, & Levine, 2012; Yang et al., 2013). Survival analysis (Skrondal & Rabe-

Hesketh, 2004) is known to provide less biased estimates than simpler techniques (e.g., standard least squares linear regression) that do not take into account the potentially truncated nature of time-to-event data (e.g., users who had not yet left the community at the time of the analysis but might at some point subsequently). From a more technical perspective, a survival model is a form of proportional odds logistic regression, where a prediction about the likelihood of a failure occurring is made at each time point based on the presence of some set of predictors. The estimated weights on the predictors are referred to as hazard ratios. The hazard ratio of a predictor indicates how the relative likelihood of the failure occurring increases or decreases with an increase or decrease in the associated predictor.

Results

Quantitative Analysis

Identifying subcommunity structure as it emerges is interesting for a variety of reasons outlined earlier in this article. As just one example of its possible use, in this quantitative analysis we specifically illustrate how our integrated modeling framework can be used to measure the impact of subcommunity participation on attrition using a survival analysis. This enables us to validate the importance of the identified structure in an objective measure that is known to be important in this MOOC context.

As discussed above, we apply our modeling framework to discussion data from each of three different Coursera MOOCs, namely Accountable Talk, Fantasy, and Python. An important parameter that must be set prior to application of the modeling framework is the number of subcommunities to identify. In this set of experiments, we set the number to twenty for each MOOC based on intuition in order to enable the models to identify a diverse set of subcommunities reflecting different compositions in terms of content focus, participation goals, and time of initiating active participation. The trained model identifies a distribution of subcommunity participation scores across the twenty subcommunities for each student on each thread. Thus we are able to construct a subcommunity distribution for each student for each week of active participation in the discussion forums by averaging the subcommunity distributions for that student on each thread that student participated in that week. In the qualitative analysis we will interpret these variables in terms of the associated thematic structure via the text portion of the model. Thus, for consistency, we refer to these twenty variables as $\text{Topic}_1 \ldots \text{Topic}_{20}$. Note that the meaning of each of these topic variables is specific to the MOOC data set the model was estimated on.

We assess the impact of subcommunity participation on attrition using a survival model, specified as follows.

Dependent variable.

Drop: We treat commitment to the course as a success measure. Thus, the binary dependent variable is treated as having a value of 1, indicating failure, for a time point if that week was the last week in which a student participated in the course according to the data we have, which for the three MOOCs we discuss in this paper only includes the forum data. For all other time points, the variable is treated as having a value of 0.

Independent variables.

Topic$_1$...Topic$_{20}$: The numeric value of each topic variable represents the percentage of time during the time point (i.e., week of active participation) the student is identified by the model as participating in the associated subcommunity.

For each student in each MOOC we construct one observation for each week of their active participation. Weeks of no discussion participation were treated as missing data. The values of the independent variables were standardized with mean 0 and standard deviation 1 prior to computation of the survival analysis in order to make the hazard ratios interpretable. The survival models were estimated using the STATA statistical analysis package (Skrondal & Rabe-Hesketh, 2004), assuming a Weibull distribution. For each independent variable, a hazard ratio is estimated along with its statistical significance. The hazard ratio indicates how likelihood of dropping out at the next time point varies as the associated independent variable varies.

If subcommunity structure had a random association with attrition, we might expect one subcommunity variable to show up as significant in the analysis by chance. However, in our analysis, across the three courses, a minimum of two and a maximum of four were determined to be significant, which supports the assertion that subcommunity structure has a non-random association with attrition in this data. Hazard ratios for subcommunity topics identified to have a significant association with attrition over time in the survival model for the Fantasy course, the Accountable Talk course, and the Python course are displayed in Tables 1-3 respectively. For these analyses we removed the variables that corresponded to topics that did not have a significant effect in the model. For each subcommunity topic identified as associated with significantly higher or lower attrition, the associated effect was between 5% and 12%. The strongest effects were seen in the Fantasy course.

A hazard ratio greater than 1 signifies that higher than average participation in the associated subcommunity is predictive of higher than average dropout at the next time point. In particular, by subtracting 1 from the hazard ratio, the result indicates what percentage more likely to drop out at the next time point a participant is estimated to be if the value of the associated independent variable is 1 standard deviation higher than average. For example, a hazard ratio of 2 indicates a doubling of probability. As illustrated in Table 1, the four identified subcommunity topics have hazard ratios of 1.07, 1.12, 1.06, and 1.07 respectively, which correspond to a 7%, 12%, 6%, and 7%

higher probability of dropout than average for students participating in the associated subcommunities with a standard deviation higher than average intensity. Table 3 also presents two subcommunity topics associated with higher than average attrition.

A hazard ratio between 0 and 1 signifies that higher than average participation in the associated subcommunity is predictive of lower than average dropout at the next time point. In particular, if the hazard ratio is .3, then a participant is 70% less likely to drop out at the next time point if the value of the associated independent variable is 1 standard deviation higher than average for that student. As illustrated in Table 2, the two identified subcommunities have hazard ratios of .93, which indicates a 7% lower probability of dropout than average for students participating in the associated subcommunities with a standard deviation higher than average intensity. Table 3 also presents two subcommunity topics associated with lower than average attrition.

Survival curves that illustrate probability of dropout over time within the three courses as a visual interpretation of these hazard ratios is displayed in Figure 2. Again we see the most dramatic effect in the Fantasy MOOC.

Table 1

Hazard Ratios for Four Different Subcommunity Topics in the Fantasy Course

Independent variable	Hazard ratio	Standard error	P value
Topic5	1.07	.04	P < .05
Topic8	1.12	.05	P < .01
Topic11	1.06	.03	P < .05
Topic13	1.07	.03	P < .05

Note. Each is associated with higher than average attrition, which can be observed in that the hazard ratios are all greater than 1.

Table 2

Hazard Ratios for Two Different Subcommunity Topics in the Accountable Talk Course

Independent variable	Hazard ratio	Standard rrror	P value
Topic8	.93	.03	P < .05
Topic12	.93	.03	P < .05

Note. Each is associated with lower than average attrition, which can be observed in that the hazard ratios are all less than 1.

Table 3

Hazard Ratios for Four Different Subcommunity Topics in the Python Course

Independent variable	Hazard ratio	Standard error	P value
Topic9	1.06	.01	P < .01
Topic13	.95	.02	P < .05
Topic17	1.09	.01	P < .01
Topic18	.95	.02	P < .01

Note. Two are associated with higher than average attrition, and two are associated with lower than average attrition.

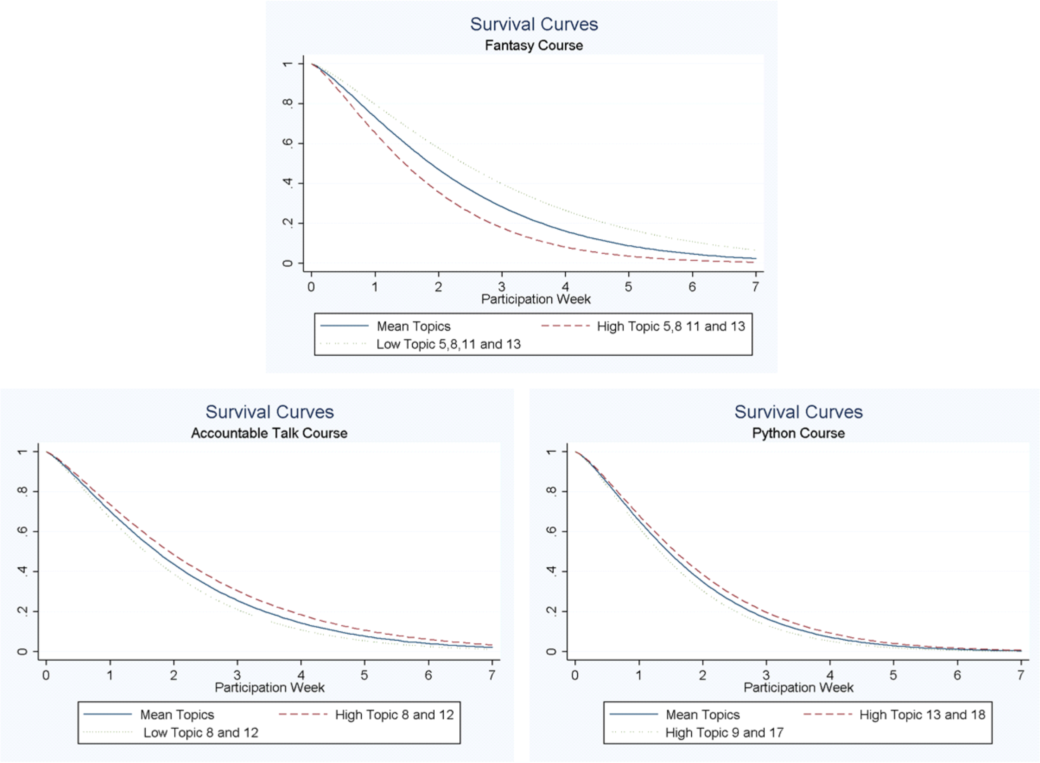

Figure 2. Survival curves for significant topics in three MOOCs.

Qualitative Analysis

In our qualitative analysis we compare topics that predict more or less attrition across MOOCs in order to demonstrate that the findings have some generality. We discuss here in detail all of the topics that were associated with significant effects on attrition in the survival models. One interesting finding is that we see consistency in the nature of topics that predicted more or less attrition across the three MOOCs.

In our analysis, we refer to student-weeks because for each student, for each week of their active participation in the discussion forum, we have one observational vector that we used in our survival analysis. The text associated with that student-week contains all of the messages posted by that student during that week. We will use our integrated model to identify themes in these student-weeks by examining the student-weeks that have high scores for the topics that showed significantly higher or lower than average attrition in the quantitative analysis.

When an LDA model is trained, the most visible output that represents that trained model is a set of word distributions, one associated with each topic. That distribution specifies a probabilistic association between each word in the vocabulary of the model and the associated topic. Top ranking words are most characteristic of the topic, and lowest ranking words are hardly representative of the topic at all. Typically when LDA models are used in research such as presented in this paper, a table is offered that lists associations between topics and top ranking words, sometimes dropping words from the list that don't form a coherent set in connection with the other top ranking words. The set of words is then used to identify a theme. In our methodology, we did not interpret the word lists out of the context of the textual data that was used to induce them. Instead, we used the model to retrieve messages that fit each of the identified topics using a maximum likelihood measure and then assigned an interpretation to each topic based on the association between topics and texts rather than directly to the word lists. Word lists on their own can be misleading, especially with an integrated model like our own where a student may get a high score for a topic within a week more because of who he was talking to than for what he was saying. We will see that at best, the lists of top ranking words bore an indirect connection with the texts in top ranking student-weeks. However, we do see that the texts themselves that were associated with top ranking student-weeks were nevertheless thematically coherent.

Because LDA is an unsupervised language processing technique, it would not be reasonable to expect that the identified themes would exactly match human intuition about organization of topic themes, and yet as a technique that models word co-occurrence associations, it can be expected to identify some things that would make sense as thematically associated. In this light, we examine sets of posts that the model identifies as strongly associated with each of the topics identified as predicting significantly more or less drop out in the survival analysis, and then for each one, identify a coherent theme. Apart from the insights we gain about reasons for attrition from the qualitative analysis, what we learn at a methodological level is that this new integrated model identifies coherent themes in the data, in the spirit of what is intended for LDA, and yet the themes may not be represented strictly in word co-occurrences.

Fantasy and Science Fiction course.

A common pattern we found among the topics that each predicted significantly higher attrition in the survival analysis for the Fantasy course was that they expressed confusion with course procedures or a lack of engagement with the course material. In many cases, these students appeared to be excited about the general topic of Fantasy and Science Fiction, but not necessarily excited about this particular course's content. Thus, the specific focus of this course may not have been a good enough fit to keep them engaged. We see students engaged in positive interactions with one another, but not in a way that encouraged them to make a personal connection with the course.

Topic5 [more attrition]. The top ranking words from the model included Philippines, looked, thank, reads, building, seem, intimidating, lot, shortfall, and weirdness. When we examined the texts in the top ranking student-weeks for this topic, the texts did not include many of these or even words found in the list of top 50 ranking words. What this means is that this topic assignment was influenced more by the network connections than by similarities at the word level. When we compared the posts in the top ranking student-weeks, there was indeed a lot of commonality at a more abstract level that might not be visible strictly based on word co-occurrences or word overlap between student-weeks. Many of the posts were introductions and discussions about confusions about course procedures, such as "I've got the same problem here. I think I'll do the assignment anyway; based on what I've read until tomorrow". At first the word *Philippines* seemed puzzling as a top ranking word, however, it was mentioned in several introductions. The word *thank* came up several times when students received a helpful response to their confusion, as in "Thanks; I've sent a request to join". An overwhelming number of these messages about confusion were from the initial week of student participation. Overall, this appears to be a topic that signifies getting oriented to the course and figuring out course procedures. The association with higher attrition is not surprising in that it would be reasonable to expect students to be vulnerable to dropout before they feel settled in a course.

Topic8 [more attrition]. Top ranking words in this topic included https, imaging, regarded, connections, building, course book, hard code, unnoticeable, arises, and staying. The most common connection between top ranking words and the texts in the top ranking student-weeks was discussion about the course books, but also other books the students were interested in, and even a comment indicating more interest in these other books than the ones that were assigned: "I should really start the course books lol". Students talked about their interpretation of symbolism in books and connections in usage of symbolism across books, but again, not necessarily the assigned books. There was some discussion of books versus movies. Overall, the discussion appeared to be lively and engaging, but not necessarily engaged with the assigned content. There was a lot of story telling about the students' own lives and experiences with books from their own countries.

Topic11 [more attrition]. Top ranking words in this topic included childbearing, hitch, range, looks, intimidating, beginning, thanks, behalf, somebody, and feelings. Similar to the other topics we have discussed, in the case of many of the top ranking words, we don't see those exact words showing up in the top ranking student-weeks, but we see words related to them conceptually. For example, the word "feelings" did not show up, but lots of emotional language describing student feelings about books, places, experiences was included. The content in many of the top ranking student-weeks appeared to focus on recommendations passed back and forth between students, sometimes recommending books they had written themselves. Examples include "You might be interested in this.ttp://irishgothichorrorjournal.homestead.com/maria.html" or "Thanks for telling us about it. I can't wait to read it." Similar to topic8, the recommendations were not necessarily for readings that were formally part of the course. Thus, we see students engaged in active exchange with one another, but not necessarily with the curriculum of the course.

Topic13 [more attrition]. Top ranking words include oneself, excitedly, experienced, releasing, somewhere, penalized, commentaries, somehow, thank, and released. Top ranking student-weeks included posts with a lot of troubles talk about the course such as “Well; I'm actually not happy with the essay I submitted. I find it soooooooo very hard to express my thoughts in English”, “All is in the eye of the peer reviewing” or “As I am writing the first 'essay'; I have a feeling that I am not writing the essay but a note. I wonder what makes an essay and a note.” Students also expressed some explicit disappointment with the readings, as in “i was expecting a real bang up ending. But it just sort of...ends. Oh well.” There was also some indication that students came to the course with different expectations than what may have been warranted, as in

> I came to the Discussion Forum looking for answers and the first one I read (yours) dissipates all my doubts. It is true; the name of the course indicates the POV from which we are suppose to be looking at the stories: a neverending interpretation of our modern world and how we explain our existence and that of others in it.

These texts had little overlap with the top ranking word list, but as with earlier topics, we see some conceptual links, such as “oneself” and “our existence”. Some student-weeks further down in the rankings that did overlap with the top ranking words were from new students just starting the course, as in “I am so excited to get back to learning.”

Accountable Talk course.

Although it was true that the connection between top ranking words and the content of the posts in the topics we examined for the Fantasy course were indirect, they were even more remote in the Accountable Talk course. In fact, we will see that the two identified topics were thematically coherent, but not in terms of word overlap. In both cases we see evidence of strong motivation for students to grapple with the course material and apply it in their own lives, which might explain why these topics were both associated with lower attrition.

Topic8 [less attrition]. Top ranking words include coast, joins, preach, thanks, hello, changed, unsurprised, giver, other, and centered. The top ranking student-weeks had very little overlap with the top ranking words. But the texts within that set were very thematically related with each other nevertheless. The bulk of top ranking student-weeks were focused on discussion about a video in the course called “The Singing Man”. Students talked about how inspired they were by the video and how they hoped to be able to achieve these effects in their own teaching, as in “I can't wait to see what explicit training the students received. They were clearly trained to respond to each other and to back up their ideas with the text.” There was some troubles talk where participants talked about why they thought this might be hard or where they have struggled in the past in their own teaching. Some non-teacher participants did the equivalent for their own “world”, such as parents who talked about their issues with communicating with their children.

Topic12 [less attrition]. Top ranking words included well, partaking, excitedly, fanatic, useable, implemented, naysayer, somebody, applying, and modeled. Many of the top ranking student-weeks for this topic were about questions that students had or things they were wondering about, such as "I was very curious to know how this course could help me in my job as an educator and what type of relationship there is between motivational interviewing and accountable talks", "Hey G; How nice to see you here with my favorite topic: Quote of the day. So what is your most favorite quote. I am keen to know. Thanks!", "For the life of me I cannot find where this pointer is available. I would greatly appreciate your time and consideration in helping me discover this content." These students expressed eagerness to learn or find specific things in this course and to hear the perspectives of others in the course.

Python course.

What is interesting about the Python course is that we have topics within the same course, some of which predict higher attrition and others that predict lower attrition, so we can compare them to see what is different in their nature. Similar to the Accountable Talk course, the connection between the top ranking words in each topic and the topic themes as identified from top ranking student-weeks bore little connection to one another, although we see some inklings of connection at an abstract level. Similar to the Fantasy course, topics that signified higher than average attrition were more related to getting set up for the course, and possibly indicating confusion with course procedures. Like in the Accountable Talk course, topics that signaled lower than average attrition were ones where students were deeply engaged with the content of the course, working together towards solutions. Similar to the findings in the Fantasy course, the interactions between students in the discussions associated with higher attrition were not particularly dysfunctional as discussions, they simply lacked a mentoring component that might have helped the struggling students to get past their initial hurdles and make a personal connection with the substantive course material.

Topic9 [more attrition]. Top ranking words included keyword, trying, python, formulate, toolbox, workings, coursera, vids, seed, and tries. The top ranking student-weeks contained lots of requests to be added to study groups. But in virtually all of these cases, that was the last message posted by the student that week. Similarly, a large number of these student-weeks included an introduction and no other text. What appears to unify these student-weeks is that these are students who came in to the course, made an appearance, but were not very quick to engage in discussions about the material. Some exceptions within the top ranking student-weeks were requests for help with course procedures. This topic appears to be similar in function to Topic5 from the Fantasy course, which was also associated with higher than average attrition.

Topic13 [less attrition]. Top ranking words include name error, uses, mayor, telly, setattr, hereby, gets, could be, every time, and adviseable. In contrast to Topic9, this topic contained many top ranking student-weeks with substantial discussion about course content. We see students discussing their struggles with the assignment, but not just complaining about confusion. Rather, we see students reasoning out solutions together. For example, “So 'parameter' is just another word for 'variable;' and an 'argument' is a specific value given to the variable. Okay; this makes a lot more sense now” or “For update_score(): Why append? are you adding a new element to a list? You should just update the score value.”

Topic17 [more attrition]. Top ranking words include was beginner, amalgamate, thinking, defaultdef, less, Canada, locating, fundamentalist, only accountable, and English. Like Topic 9, this topic contains many top ranking student-weeks with requests to join study groups as the only text for the week. The substantive technical discussion was mainly related to getting set up for the course rather than about Python programming per se, for example “Hi;I am using ubuntu 12.04. I have installed python 3.2.3 Now my ubuntu12.04 has two version of python. How can I set default version of 3.2.3Please reply” or “For Windows 8 which version should I download ?Downloaded Python 3.3.2 Windows x86 MSI Installer?and I got the .exe file with the prompter ... but no IDLE application”.

Topic18 [less attrition]. Top ranking words include one contribution, accidental, workable, instance, toolbox, wowed, meant, giveaway, patient, and will accept. Like topic13, we see a great deal of talk related to problem solving, for example “i typed s1.find(s2;s1.find(s2)+1;len(s1)) and i can't get why it tells me it's wrong? do not use am or pm.... 3am=03:00 ; 3pm=15:00”, or “I don't see why last choice doesn't work. It is basically the same as the 3rd choice. got it! the loop continues once it finds v. I mistakenly thought it breaks once it finds v. thanks!”. The focus was on getting code to work. Perhaps “workable” is the most representative of the top ranking words.

Discussion/Conclusion

In this paper, we have developed a novel computational modeling methodology that provides a view into the evolving social structure within a massive open online course (MOOC). In applying this integrated approach that brings together a view of the data from a social network perspective with a complementary view from text contributed by students in their threaded discussions, we illustrated how we are able to identify emergent subcommunity structure that enables us to identify subcommunities that represent behavior that is coordinated both in terms of who is talking to who at what time and how they are using language to represent their ideas.

In this paper, we have illustrated that this identified subcommunity structure is associated with differential rates of attrition. A qualitative posthoc analysis suggests that subcommunities associated with higher attrition demonstrate lower comfort with course procedures and lower expressed motivation and cognitive engagement with the course materials, which in itself is not surprising. However, the real value in such a model is that it offers a bird's eye view of the discussion themes within the course. The nature of the themes identified is qualitatively different from those identified using LDA alone because of the influence of the network on the topic structure. As pointed out in the qualitative analysis, the semantic connection between high ranking words within a topic is far more indirect than what is achieved through LDA, where word co-occurence alone provides the signal used to reduce the dimensionality of the data. The meaning of the topics is more abstract, and possibly richer, since it represents the collection of themes and values that emerge when a specific group of students are talking with each other.

The purpose of exploratory models such as this probabilistic graphical model is to identify emergent themes and structure in the data. It can be used as part of a sensemaking process, but it is not meant to test a hypothesis. In the case of the analysis presented in this paper, the findings about the association between low engagement and attrition might suggest that it would be worth the effort to formalize the structure so that a more rigorous analysis of the issue could be conducted. Along these lines, in some of our prior work where we have explicitly and directly modeled motivation and cognitive engagement as it is expressed in text only, we have also found evidence that higher expressed motivation and cognitive engagement are associated with lower attrition (Wen et al., 2014). In that work, the effect was much stronger, but it took an investment of time and effort to do the analysis. An exploratory analysis that suggests which issues would be worth investing time to pursue more rigorously within a data set could be valuable from the stand point of being strategic about the investment of research resources, especially when one considers the broad range of research questions that analysis of interaction data affords. The take home message is that exploratory models such as this could usefully be used for hypothesis formation, followed by more careful, direct modeling approach.

A limitation of selecting a probabilistic graphical modeling approach, as with any unsupervised clustering approach, is that the number of topics must be specified before the model is inferred. In our work, we selected a number based on intuition. It should be noted that one can tune the number of features using measures of model fit to determine which number to use. This approach might be especially useful for researchers who prefer not to make an ad hoc choice.

Our long term vision is to use insights into emerging social structure to suggest design innovations that would enable the creation of more socially conducive MOOCs of the future. The lesson we learn from the qualitative analysis presented in this paper is that students are vulnerable to dropout when they have not yet found a personal connection between their interests and goals and the specific content provided by the course. Mentors present within the discussions to coach students to find such personal connections might serve to keep students motivated until they have made it past initial confusions and have settled more comfortably into the course. On average, it is the more motivated students who participate in the discussions at all. However, the analysis presented here reveals that even among those students, we can identify ones that are vulnerable. Real time analysis of the texts could enable triggering interventions, such as alerting a human mentor of an opportunity to step in and provide support to a student who is motivated, but nevertheless does not possess quite enough of what it takes to make it in the course without support. Real time analysis of discussions for triggering supportive interventions that lead to increased learning are more common in the field of computer supported collaborative learning (Kumar & Rosé, 2011; Adamson et al., 2014), and such approaches could potentially be adapted for use in a MOOC context.

The current modeling approach has been applied successfully to Coursera MOOCs in this paper. However, cMOOCs provide a richer and more intricate social structure where students interact with one another not only in threaded discussions, but in a variety of different social settings including microblogs, synchronous chats, and email. Just as the current modeling approach integrates two complementary representations, namely network and text, in future work we will extend the approach to integrate across multiple networks in addition to the text so that each of these social interaction environments can be taken into account. The challenge is that a model of that complexity requires much more data in order to properly estimate all of the parameters. Thus it will likely require jointly estimating a model over multiple courses simultaneously using a hierarchical modeling approach that properly treats within course dependencies within the heterogeneous dataset.

Acknowledgements

This research was supported in part by the Gates Foundation Grant: The MOOC Research Initiative and NSF Grants OMA-0836012, IIS-1320064, and IIS-1302522.

References

Adamson, D., Dyke, G., Jang, H. J., Rosé, C. P. (2014). Towards an agile approach to adapting dynamic collaboration support to student needs. *International Journal of AI in Education, 24*(1), 91-121.

Airoldi, E., Blei, D., Fienberg, S. & Xing, E. P. (2008). Mixed membership stochastic blockmodel. *Journal of Machine Learning Research, 9*(Sep), 1981—2014.

Bakhtin, M. (1981). Discourse in the novel. In Holquist, M. (Ed.). *The dialogic imagination: Four essays by M. M. Bakhtin*. University of Texas Press: Austin.

Blei, D., Ng, A. and Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research* (3), 993-1022.

Collins, L. M., & Lanza, S. T., (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New York: Wiley.

Hedström, P. (2005). *Dissecting the social: On the principles of analytical sociology*. Cambridge University Press.

Jordan, K. (2013). MOOC completion rates: The data. Retrieved 23 April 2013.

Karypis, G. & Kumar, V. (1995). Multilevel graph partitioning schemes. In *ICPP* (3), 113-122.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 170-179). ACM.

Koller, D., Ng, A., Do, C. and Chen, Z. (2013). Retention and intention in massive open online courses. In Depth. Educause. 3.

Koller, D. & Friedman, N. (2009). Probabilistic graphical models: principles and techniques. The MIT Press

Kumar, R. & Rosé, C. P. (2011). Architecture for building conversational agents that support collaborative learning. *IEEE Transactions on Learning Technologies, 4*(1), pp 21-34

Kumar, A., Yang, D., Wen, M., Xing, E., & Rosé, C. P. (2014). *Large scale structure aware community discovery in online communities*. Language Technologies Institute Technical Report.

Rosé, C. P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P. & Sherer, J. (2014). Social factors that contribute to attrition in MOOCs. *Proceedings of the First ACM Conference on Learning @ Scale.*

Sawyer, K. (2005). *Social emergence: Societies as complex systems*. Cambridge University Press.

Siemens, G. (2005). Connectivism: A learning theory for a digital age. *International Journal of Instructional Technology and Distance Learning*, *2*(1).

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modleing: Multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC.

Smith, B. & Eng, M. (2013). MOOCs: A learning journey: Two continuing education practitioners investigate and compare cMOOC and xMOOC learning models and experience. In *Proceedings of the 6th International Conference on Hybrid Learning and Continuing Education* (pp. 244-255).

Wang Y, Kraut R, and Levine J. M. (2012) To stay or leave? The relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of Computer Supported Cooperative Work* (pp. 833-842).

Wen, M., Yang, D., Rosé, D. (2014). Linguistic reflections of student engagement in massive open online courses. In *Proceedings of the International Conference on Weblogs and Social Media.*

Yang, D., Sinha, T., Adamson, D., & Rosé, C. P. (2013). *Turn on, tune in, drop out: anticipating student dropout in massive open online courses*. NIPS Data-Driven Education Workshop.