Communication Communities in MOOCs

Nabeel Gillani
Machine Learning Group
Department of Engineering
University of Oxford
nabeel@robots.ox.ac.uk

Rebecca Eynon
Oxford Internet Institute
Department of Education
University of Oxford
rebecca.eynon@oii.ox.ac.uk

Michael Osborne
Machine Learning Group
Department of Engineering
University of Oxford
mosb@robots.ox.ac.uk

Isis Hjorth
Oxford Internet Institute
Department of Education
University of Oxford
isis.hjorth@oii.ox.ac.uk

Stephen Roberts
Machine Learning Group
Department of Engineering
University of Oxford
sjrob@robots.ox.ac.uk

ABSTRACT

The rise of Massive Open Online Courses (MOOCs) has brought together thousands of people from different geographies and demographic backgrounds – but to date, little is known about how they interact and learn. We introduce a new content-analysed MOOC dataset and use Bayesian Non-negative Matrix Factorization (BNMF) to extract communities of learners based on the nature of their online forum posts. We see that the communities BNMF learns are differentiated by their composite students' demographic and course performance indicators. We conclude with a discussion of how computationally efficient probabilistic generative models can be leveraged in conjunction with automated or crowdsourced content analysis to inform educators and learners about latent communication tendencies.

1. INTRODUCTION

There has been no shortage of polarised media hype around Massively Open Online Courses (MOOCs) – and yet, grounded research on their pedagogical effectiveness and potential is in its infancy. Research in online education is not new: proponents have lauded its potential and critics have warned of an imminent "mechanization of education" since the popularization of the internet nearly two decades ago (Nobel, 1998). Moreover, much education literature has focused on the way learners use online discussion forums, drawing upon social constructivist (Vygotsky, 1978) conceptions of knowledge creation and sharing. The rise of MOOCs has provided a new opportunity to explore how communication unfolds in a global, semi-synchronous classroom. A deeper understanding of how learners communicate in these social learning spaces, combined with new efforts in data visualization and software engineering, could influence how practitioners develop, and learners access, course content.

Some early research efforts have taken only a cursory look at the nature of MOOC communication, for example, by detecting and modelling the prevalence of chatter irrelevant to the course (Brinton et al., 2013) and, in some cases, exploring the frequency of words used by those that pass or fail (Anderson et al., 2014). Others have begun to explore how latent feature models from Machine Learning like the mixed-membership stochastic block model (Airoldi et al., 2008) can be used to explain forum participation and dropout rates (Rose, 2013).

In most of these first attempts to make sense of the nature of interaction and learning in MOOC forums, the sheer volume of thousands of discussion forum posts has prevented latent feature analysis from modelling the *content* and *context* of communication. Our efforts aim to explore how insights can be derived from global-scale dialogue in learning contexts by 1) introducing a new content-analysed dataset of MOOC forum data; 2) leveraging and validating Bayesian Non-negative Matrix Factorization as a robust probablistic generative model for online discussions; 3) indicate how future online course environments may incorporate efficient content-labelling and generative modelling schemes in order to inform feedback and assessment as it pertains to learner communications.

2. DATASET AND LATENT FEATURES

2.1 Content Analysis of Forum Data

We analysed data from a business strategy MOOC offered on the Coursera platform in Spring 2013. Nearly 90,000 students registered for the course, which lasted for six weeks and assessed students through a combination of weekly quizzes and a final project. The online discussion forum was comprised of a number of sub-forums, which in turn had their own sub-forums or user-generated discussion threads that contained posts or comments. There were over 15,600 posts or comments in the discussion forum, generated by nearly 4,500 learners. Over 15,000 learners viewed at least one discussion thread in both instances, contributing to 181,911 total discussion thread views.

We conducted qualitative content analysis on nearly 6,500 posts from this course – to our knowledge, an unprecedented undertaking to date in MOOC research. Content analyses have sometimes been used in online learning research, yet at

much smaller scales than presented here (e.g. De Weaver et al. 2006). The content analysis scheme for the present study was developed based on both existing academic literature and preliminary observations of online course discussions.

We selected five dimensions to capture key aspects of interaction and learning processes. The first dimension (learning) was used to collect data about the extent to which knowledge construction occurred through discussions, categorising each post using one of nine categories, ranging from no learning, through to four types of sharing and comparing of information, to more advanced stages of knowledge construction such as negotiation of meaning (Gunawardena et al., 1997). The second dimension identified communicative intent in the forums, selecting from five categories: argumentative, responsive, informative, elicitative and imperative (Clark et al., 2007; Erkens et al., 2008). The third dimension – affect – gauged levels and relative distributions of emotion in discourse, using five codes: positive / negative activating, positive / negative deactivating, and neutral (Pekrun et al., 2002). Based on our own observations of the forums, we also developed two more dimensions: one related to topic, which had 11 categories that reflected all course related topics (e.g. cases, quizzes, readings, arrange offline meet-ups, introductions); and the other a rating of relevance of the post to its containing thread and sub-forum. Relevance was rated on a three point scale: high relevance, low relevance and no relevance. In total, there were 7,425 possible label combinations for each post.

For simplicity (given the size of the dataset), the unit of analysis selected was the post. The qualitative analysis software NVivo was used for labelling content. Coding was conducted by four individuals who trained together over the course of two sessions and pilot tested the instrument together to enhance reliability.

2.2 Inferring Latent Features

We are interested in probabilistic generative models to infer hidden features in content-analysed MOOC forum data primarily because: 1) in a principled Bayesian setting, they enable the use of tools – e.g. priors and likelihoods – that are intuitively appropriate and fitting for the specific data at hand and 2) they enable data simulation, which, given the "noisiness" of data generated in online course environments, may be useful as educators and learners make inferences about how communication unfolds and the nature of their participation relative to their peers.

A variety of latent feature models have been proposed in the machine learning community. Some of the more popular parametric models include Latent Dirichlet Allocation (LDA – Blei et al., 2003) and the Mixed Membership Stochastic Block (MMB – Airoldi et al., 2008), which have been applied to extracting latent topics in documents and community structures in social networks. These models assume that the number of hidden features is known a priori before inferring how observed data relate to these features. Bayesian nonparametric models have lifted this assumption, often by treating the number of hidden features as a variable to be learned during inference. The Indian Buffet Process (IBP – Griffiths and Gharamani, 2005) is perhaps the most prominent nonparametric latent feature model, specifying a prior distribution over binary matrices that denote the latent features characterizing a particular dataset. Similar to LDA and MMB, the IBP enables an observation to be characterized by multiple features – but it treats the number of latent features as an additional variable to be learned as a part of inference. Unfortunately, given the large (exponentially-sized) support over the distribution of possible binary matrices for any fixed number of features, the IBP has proven intractable for large-scale datasets. While recent innovations have helped it scale (Doshi-Velez, 2009; Reed and Ghahramani, 2013), they have primarily exploited the structure of linear-Gaussian likelihood models.

Non-negative Matrix Factorization (Lee and Seung, 1999) is a computationally efficient parametric method that has been used to produce a parts-based representation of an $N \times D$ data matrix X, i.e. $\mathrm{X} \approx \mathrm{WH}$, where W and H are $N \times K$ and $K \times D$ matrices, respectively. More recently, Bayesian extensions to NMF (BNMF) have been proposed, leveraging both fully Bayesian as well as MAP inference schemes and providing an intuitive foundation for detecting hidden parts, or clusters, large-scale datasets.

3. BAYESIAN NMF

In this work, we apply BNMF for community detection, as proposed by (Psorakis et al. 2011a) for social networks, to understand how latent communities can be inferred among MOOC users based on the content of their forum posts. The model informs a *belief* about each individual's community membership by presenting a membership distribution over possible communities.

3.1 Probabilistic Generative Model

Our data can be represented as an $N \times D$ matrix C where each row represents a learner n that has posted at least once in the course's online forums and each column d represents a particular content label for each of the five dimensions described in the previous section (for example, one category in the Knowledge Construction dimension is "statement of observation or opinion"). Each entry of C, c_{nd}, is 1 if learner n has made at least one post assigned a content label of d, and 0 otherwise. Hence, C is a binary matrix (our content analysis scheme allowed each post to be labelled with only one category per dimension – however, users with multiple posts may be characterized by many different categories).

C depicts a bipartite *learner-to-category* network. Given our interest in uncovering latent groups of learners based on the category labels of their posts, we adopt the convention from (Psorakis et al. 2011b) and compute a standard weighted one-mode projection of C onto the set of nodes representing learners, i.e. $\{n_i\}_{i=1}^{N}$. The resultant $N \times N$ matrix X has entries $x_{ij} = \sum_{d=1}^{D} c_{id} c_{jd}$, i.e., the total number of shared categories across all posts made by learners i and j. It is important to note that connections between learners i and j in this adjacency matrix do *not* necessarily depict communication between them; instead, they indicate similar discussion content.

We assume that the pairwise similarities described by X are drawn from a Poisson distribution with rate $\hat{\mathrm{X}} = \mathrm{WH}$, i.e. $x_{ij} \sim \mathrm{Poisson}(\sum_{k=1}^{K} w_{ik} h_{kj})$, where the inner rank K denotes the unknown number of communities and each element k for a particular row i of W and column j of H indicates the extent to which a single community contributes to $\hat{x}_{i,j}$. In other words, the expected number of categories that two individuals i,j share across their posts, $\hat{x}_{i,j}$, is a result of the degree to which they produce similar discussion content. To address the fact that the number of communities K is not

initially known, we place *automatic relevance determination* (MacKay, 1995) priors β_k on the latent variables w_{ik} and h_{kj}, similar to (Tan and Fèvotte, 2009), which helps ensure that irrelevant communities do not contribute to explaining the similarities encoded in X.

The joint distribution over all the model variables is:

$$p(\mathrm{X},\mathrm{W},\mathrm{H},\beta) = p(\mathrm{X}|\mathrm{W},\mathrm{H})p(\mathrm{W}|\beta)p(\mathrm{H}|\beta)p(\beta) \quad (1)$$

And the posterior distribution over the model parameters given the data X is:

$$p(\mathrm{W},\mathrm{H},\beta|\mathrm{X}) = \frac{p(\mathrm{X}|\mathrm{W},\mathrm{H})p(\mathrm{W}|\beta)p(\mathrm{H}|\beta)p(\beta)}{p(\mathrm{X})} \quad (2)$$

3.2 Inference and Cluster Assignment

Our objective is to maximize the model posterior given the data X, which is equivalent to minimising the negative log posterior (i.e., the numerator of equation (2)) since $p(\mathrm{X})$ is not a random quantity. Like (Psorakis et al. 2011a), we represent the negative log posterior as an energy function U:

$$\begin{aligned} \mathrm{U} &= -\log p(\mathrm{X}|\mathrm{W},\mathrm{H}) - \log p(\mathrm{W}|\beta) \\ &\quad -\log p(\mathrm{H}|\beta) - \log p(\beta) \end{aligned} \quad (3)$$

The first term of U is the log-likelihood of the data, $p(\mathrm{X}|\mathrm{W},\mathrm{H}) = p(\mathrm{X}|\hat{X})$, which represents the probability of observing similar post content between two users i and j represented by x_{ij}, given an expected (or Poisson rate) of $\hat{x}_{ij}$. The negative log-likelihood is given by:

$$\begin{aligned} -\log p(\mathrm{X}|\hat{\mathrm{X}}) &= -\sum_{i=1}^{N}\sum_{j=1}^{N}\log p(x_{ij}|\hat{x}_{ij}) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\big(x_{ij}\log\tfrac{x_{ij}}{\hat{x}_{ij}} + \hat{x}_{ij} \\ &\quad -x_{ij} + \tfrac{1}{2}\log(2\pi x_{ij})\big) + \text{const.} \end{aligned} \quad (4)$$

Following (Tan and Fèvotte, 2009; Psorakis et al. 2011a), we place independent half-normal priors over the columns of W and rows of H with zero mean and precision (inverse variance) parameters $\beta \in \mathbb{R}^K = [\beta_1,...,\beta_K]$. Each β_k controls the importance of community k in explaining the observed interactions; large values of β_k denote that the elements of column k of W and row k of H lie close to zero and therefore represent irrelevant communities.

To optimize for W, X, and β, we use the fast fixed-point algorithm presented in (Tan and Fèvotte, 2009; Psorakis et al., 2011a) to optimize the objective function U. The algorithm has complexity $\mathrm{O}(NK)$, which involves consecutive updates of W, H, β until convergence (e.g. a maximum number of iterations) has been satisfied. Figure 1 presents pseudocode for this procedure.

In the case of our application, $\mathrm{W}_* = \mathrm{H}_*^\top$ since X is symmetric. Each element w_{ik}^*, or h_{ki}^* denotes the *degree of participation* of individual i in cluster k while each normalized row of W_* (or column of H_*) expresses each node's *soft-membership* distribution over the possible clusters. This soft-membership provides more context to our *belief* about a node's cluster membership, which we can model and explore explicitly if desired.

4. RESULTS

CD-BNMF(X; K_0; a; b)
1 **for** $i = 1$ *to* n_{iter} **do**
2 $\quad \mathrm{H} \leftarrow (\frac{\mathrm{H}}{\mathrm{W}^\top 1 + \beta\mathrm{H}})\mathrm{W}^\top(\frac{\mathrm{X}}{\mathrm{WH}})$
3 $\quad \mathrm{W} \leftarrow (\frac{\mathrm{W}}{1\mathrm{H}^\top 1 + \mathrm{W}\beta})(\frac{X}{\mathrm{WH}})\mathrm{H}^\top$
4 $\quad \beta_k \leftarrow \frac{N+a-1}{\frac{1}{2}(\sum_i w_{ik}^2 + \sum_j h_{kj}^2)+b}$
5 **end for**
6 $K_* \leftarrow$ # nonzero columns of W or rows of H
7 **return** $\mathrm{W}_* \in \mathbb{R}_+^{N\times K_*}, \mathrm{H}_* \in \mathbb{R}_+^{K_*\times N}$

Figure 1: Community Detection using Bayesian NMF.

4.1 Model Benchmarking

In order to evaluate the BNMF model and inference scheme's ability to represent the data, we computed the negative log likelihood (NLL) and root mean-squared error (RMSE) of held-out test data against the nonparametric linear Poisson gamma model (LPGM) of (Gupta et al., 2012). The LPGM is a latent feature model that treats the number of hidden clusters as a variable to be learned during inference, leveraging an Indian Buffet Process prior over a infinite-dimensional binary hidden feature matrix. The LPGM model is:

$$\mathrm{X} = (\mathrm{Z}\circ\mathrm{F})\mathrm{T} + \mathrm{E}$$

Here, X is the $N \times D$ matrix of observations; Z is the $N \times K$ binary latent feature matrix with entry $i,k = 1$ iff feature k is represented in datum (i.e., learner) i; T is a non-negative $K \times D$ matrix illustrating the represenation of each dimension $d \in D$ in feature $k \in K$; F is an $N \times K$ non-negative matrix indicating the strength of participation of i in feature k; E is the reconstruction error with rate λ such that $\mathrm{E}_{ij} \sim \mathrm{Poisson}(\lambda)$; and $\circ$ is the Hadamard product.

Given the scalability issues of the IBP, we evaluated both BNMF and LPGM on 20 randomly-selected 50×50 subsets of real-world data by comparing the root mean squared error (RMSE) and negative log likelihood (NLL) of held-out test data. Both the RMSE and NLL measured each probabilistic model's ability to represent the data. Each 50×50 subset was determined by randomly sampling the rows and corresponding columns of full content-analysed forum dataset used in this study, and for each row, 10% of entries were randomly selected for hold-out, with the remainder used for training. Inference on the infinite model was performed via 5000 iterations of Gibbs sampling, with no samples discarded for burn-in. For comparison purposes, we computed the RMSE and NLL of a naïve model (Pred-Avg) that predicts the arithmetic mean of the training data for each held-out data point, as well as the RMSE for a naïve benchmark (Pred-0) that always predicts 0 for all held-out data (because of the 0 prediction, computing the NLL for this naïve model would involve repeatedly evaluating a Poisson likelihood with rate $\lambda = 0$, ultimately yielding ∞). Table 1 summarizes the results, which reveal that the proposed BNMF model and inference scheme has greater predictive accuracy than its nonparametric IBP counterpart (confined to a finite number of sampling iterations) and both naïve approaches[1]. Moreover, given its computational tractability

[1]the values presented in the table are generated by taking the arithmetic mean of the RMSE and NLL, computed for

	BNMF	LPGM	Pred-Avg	Pred-0
RMSE	**0.4647**	1.2199	0.9330	1.5823
NLL	**251.97**	355.22	324.74	–

Table 1: RMSE and NLL results for the BNMF, LPGM, Pred-Avg, and Pred-0 models. Bold values indicate the strongest predictive performance on held-out test data.

(taking seconds to run on the full dataset, versus days for the LPGM), BNMF offers a favourable generative model for extracting latent features from online discussion forum data.

4.2 Exploring Extracted Communities

Students in the business strategy course were encouraged to interact through its online discussion forum, which was segmented into multiple sub-forums. Two sub-forums aimed at promoting learner engagement and interactions were content-analysed and explored for latent features: Cases and Final Projects. The Cases sub-forum facilitated weekly discussions about a real company and its business challenges / opportunities (for example, in week 1, the selected company was Google, and one of the questions was "Do you think Google's industry is a competitive market, in the technical sense? Does Google have a sustainable competitive advantage in internet search?"). The Final Project sub-forum facilitated questions, debates, and team formation for the final strategic analysis assignment. The remaining sub-forums were: Questions for Professor, Technical Feedback, Course Material Feedback, Readings, Lectures, and Study Groups.

Since participation in multiple sub-forums was minimal (in most cases, no more than 10% of participants in one sub-forum participated in another), we explored the latent features of communication and the characteristics of these underlying communities in both sub-forums independently of one another.

Learners were assigned to inferred communities by computing maximum a-posteriori (MAP) estimates for W and H as described in section 2 and greedily assigning each learner i to the community k_i to which it "most" belongs, i.e. $k_i = \text{argmax}_{k \in K} w_{ik}$. In repeated executions of the BNMF procedure, different community assignments were computed for some learners due to random initializations of W and H as well as numerical precision issues that affected the group allocation step. To mitigate this, we ran the algorithm 100 times and used the factor matrices with the highest data likelihood to compute the final allocations.

Our analysis of extracted communities sought to understand the demographics, course outcomes, broader forum behaviours and types of posts for each of its constituent learners.

4.2.1 Cases sub-forum

The Cases sub-forum had 1387 unique participants that created nearly 4,100 posts or comments. We used BNMF to detect latent communities based on the learning and dialogue acts reflected in these posts, as this particular sub-forum was set up for participants to practice the tools and frameworks they learned in the course, and so, the learning and dialogue dimensions were selected to reveal the ways in which people used the forums to engage with one another and construct knowledge. Four learner communities emerged, containing 238, 118, 500, and 531 people, respectively. We describe these communities as committed crowd engagers, discussion initiators, strategists, and individuals, respectively.

Community 1 (committed crowd engagers). Participants in this group tended to engage with others in the forum. Of all the groups they contributed the most responsive dialogue acts at 43% of total posts, and the second highest number of informative (8%) and elecitive (5%) statements. In terms of learning, they tended to achieve quite similar levels of higher-order knowledge construction to groups 2 and 3. These participants read and posted the most of all four groups. 45% of the group's participants passed the course – significantly more than any other group ($p<0.05$ [2]). Interestingly, members of this group were likely to be from Western continents[3], with a larger proportion of Europeans (26.1%), albeit only significantly greater than the other groups at the $p<0.1$ level. Nearly 31% had at least a Master's degree – similar to group 3. It is reasonable to suggest that this group found the forums an important part of their learning and used it as they sought to formally pass the course.

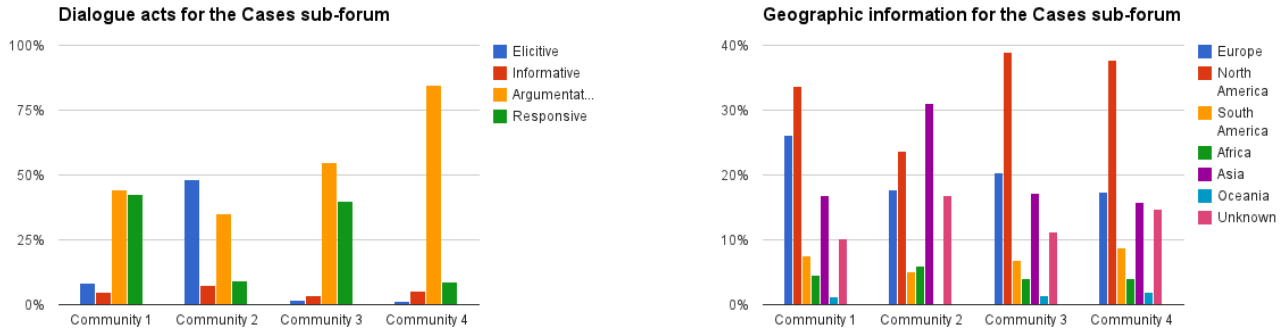
Community 2 (discussion initiators). Most notable for this group was its level of elicitative dialogue acts – which characterized over 48% of its participants' posts. Morever, 24% of their posts did not involve learning, a significantly greater proportion than the other groups ($p<0.07$ compared to group 1; $p\approx0$ compared to groups 3 and 4). Still, members of this group had a larger proportion of posts reflecting higher-order learning than the other groups (8.0%).ÂăInterestingly, this group had a significantly lower pass rate than groups 1 and 3 (25%, $p<0.05$), but this could be explained to a large extent by the high number of people who did not submit a final project (67%, similar to group 4). Members of this group viewed fewer discussion threads and contributed fewer posts than groups 1 and 3. Geographically speaking, a significantly higher proportion of this group's members were located in Asia in comparison to the other three (31%, $p<0.01$). This could suggest that geography played an important role in motivating discussion. ÂăIndeed, the more elicitative nature of dialogue in this group may suggest cultural differences in interpretations of, or responses to, various conversation topics.

Community 3 (strategists). In many ways, people in this group were similar to group 1. They had similar levels of higher-order learning and tended to be responsive to others' comments. However, they had a greater proportion of argumentative statements (55%) and rarely had posts that reflected no learning (1.6%). People in this group were second most likely to pass the final project (36.2%) and second most likely to try to pass, but ultimately fail (6.4%). They tended to be similarly educated to those in group 1 – with nearly 30% receiving at least a Master's degree. They viewed and contributed to the forums the second most number of times, but this was still significantly less than group

each of the 20 different subsets, with different data held-out for each subset.

[2]We used the nonparametric Kruskal-Wallis one-way analysis of variance to test for statistical significance.

[3]Learners' locations were determined by leveraging a web-service to learn which country they last accessed course content from, based on a recorded IP address.



(a) (b)

Figure 2: Plot (a) illustrates the dialogue acts represented in the posts made by learners belonging to each community, and plot (b) depicts their geographic locations.

1 ($p \approx 0$). Combined, these characteristics suggest that students in group 3 were more strategic in their approaches, using the Cases sub-forum only as needed to achieve their learning goals.

Community 4 (individualists). People in group 4 were highly distinctive in their large proportion of argumentative statements (85%). They had a smaller proportion of posts featuring higher-order learning (3.7%) compared to groups 1 - 3. They read and posted in the forums less than any other group (significant at $p \approx 0$ compared to groups 1 and 3). They were the most likely to not submit a final project (68%) - a similar number to group 2. Of all the groups, participants in this group had the smallest proportion of people attain at least a Master's degree (23.2%, $p<0.05$ compared to groups 1 and 3). These indicators may suggest a number of possibilities: that members of this group were the most likely to drop out of the course of all four groups, may have had limited experience of using forums to construct their knowledge, or simply preferred to learn individually.

Figure 2 shows the dialogue acts and geographic locations of the members of each group.

4.2.2 *Final Project sub-forum*

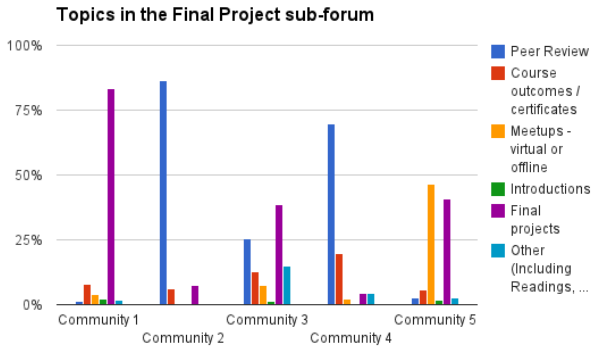
The Final Projects sub-forum had 1256 unique participants creating nearly 2,400 posts or comments. We selected the communication and topic labels as inputs into BNMF because of the nature of the sub-forum: it was a place for participants to find others to discuss their individual final projects with prior to the submission. Therefore, how people engaged with each other and the topics of their engagements were central to this setting. We detected 5 communities with 296, 50, 611, 45, and 237 individuals, which we characterised as: instrumental help seekers, careful assessors, community builders, focused achievers, and project support seekers, respectively. [4]

Community 1 (instrumental help seekers). Participants in this group had a high proportion of elicitative dialogue acts (64%) and primarily discussed the final project (83%). On average, they posted more than groups 2 and 4 and their amount of views of the forum were relatively low (similar to groups 4 and 5). The proportion of people who passed was significantly lower than in groups 2, 3 and 4 (41%, $p<0.01$). People in this group were also more likely to submit and fail the final project than clusters 3 and 5 (14%, $p \leq 0.05$). There were fewer people with postgraduate qualifications compared to groups 3 and 5 (20%, $p<0.01$). These trends suggest that members of this community sought help by asking questions and discussing the final project with their peers, but still did not pass the course.

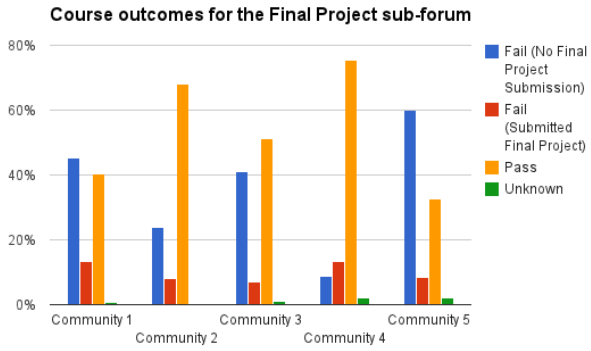
Community 2 (careful assessors). Participants in this group had the highest proportion of elicitative dialogue acts out of all of the groups (71%), but in contrast to group 1, the focus of their posts was about the peer review process (87%). They viewed more posts on average than groups 1, 4, and 5, but only groups 4 posted fewer comments on average. Thus, it seems that participants in this group used the forums to look for answers to questions they had about peer review, and only posted again if necessary. Like group 4, a high proportion of learners passed the course, compared to groups 1, 3, and 5 ($p<0.05$). These patterns suggest that this group needed to know more about the peer assessment process, but that its members were very strategic in their use of this sub-forum to obtain necessary information.

Community 3 (community builders). Participants in this group were distinctive in the proportion of posts that were responsive to others (55%). In contrast to the other groups, the focus of their discussions were spread across final projects and peer review. Interestingly this group seemed the most engaged of groups in the forum, being the most likely to view and post in this sub-forum of all participants in other groups ($p<0.05$, $p<0.001$, respectively). Likewise, the average length of posts submitted by supporters (712 words) was markedly higher than in any other group (Group 1 had the 2nd highest average of 382 words – $p<0.001$). Their pass rate (51%) was higher than clusters 1 and 5 ($p<0.01$), but lower than 2 and 4 ($p<0.05$), partly due to the high proportion of learners (41%) that did not submit a final project. This suggests that participants in group 3 were more interested in exchanging ideas with others as opposed to receiving formal acknowledgement or recognition for passing the

[4] 17 individuals were assigned to their own groups; for the purposes of analysis, we only investigated clusters with at least two members.



(a)



(b)

Figure 3: Plot (a) illustrates the discussion topoics represented in the posts made by learners belonging to each community, and plot (b) depicts their course outcomes.

course.

Community 4 (focused achievers). Participants in this group were distinctive as they had a higher proportion of argumentative dialogue acts (68%). While most focus was on peer review (70%), many posts also discussed course outcomes and certificates (20%). They had the highest proportion of posts that evidenced some form of learning (32%). They posted the least (on average, 2.5 times), and had the smallest average post size (146 words) and number of thread views views (38), both statistically significant only when compared to group 3 ($p<0.05$). Interestingly, they had the highest proportion of participants submit a final project and pass the course (76%, $p<0.01$ compared to groups 1, 3, and 5), yet a similar proportion to group 1 who submitted but still failed (13%). Furthermore, they comprised a group that showed the most emotion in their posts (20%) of all the groups. These patterns suggest a very focused group of participants who only used the forums when necessary to achieve their goals – and to express both joy and unhappiness with their own course outcomes.

Community 5 (project support seekers). Participants in this group were similar to those in group 1, although they were distinguished by a high proportion of imperative dialogue acts (50%) and organizing virtual meet-ups (45%). The average number of discussion thread views was relatively low (40 - similar to groups 1 and 4); moreover, participants made posts more often than groups 2 and 4, albeit not with statistical significance. This pattern suggests that participants in this group were seeking support and opportunities for collaboration on the final project. Interestingly, this was the only group where significant differences were found in geographic region: there were more people from South America in this group compared to 3 ($p<0.01$), which may indicate a wish for people from the same part of the world to collaborate. While this group had a higher number of participants with postgraduate degrees than group 1 (29%, $p<0.05$), they had the lowest pass rate out of all other groups (32%, $p<0.05$), partly explained by having the highest proportion of participants who did not submit a final project (57%, $p<0.05$).

Figure 3 shows the discussion topics and course outcomes of the members of each group.

5. DISCUSSION

Using BNMF to extract latent features from the dataset and subsequently exploring the composition of these features reveals that the different sub-forums in MOOCs offer participants different ways to engage with course content – and each other. As the latent learner groups suggest, discussion forum post content can provide insights into the background characteristics, course performance, and overall engagement levels of learners. These clusters may then be used for a variety of practices that relate to assessment and feedback. Forum participants, to a large extent, choose for themselves how much they wish to use the forums to construct knowledge together. Some adopt a more socio-cultural approach to learning and others use the forums as a way to reflect on their own ideas, more in-line with cognitive and social constructivist theories (Stahl, 2006). Detecting these different modes of social learning and presenting them to instructors or the learners themselves may influence how certain content is communicated, and how the discussion forums are leveraged in order to advance pedagogical objectives.

While providing information on latent communication communities to both educators and learners may serve as a valuable feedback mechanism, it is important to note the sensitivity of detected clusters to a number of factors. For one, the composition of communities could change depending on which content-labelling dimensions are considered when building the post-similarity adjacency matrix over all users. Additionally, we may leverage the fact that BNMF returns a posterior belief (soft-partitioning) of each learner's membership in each cluster in order to explore how the demographic and other characteristics of each cluster change as we assign learners to clusters with varying degrees of belief. Therefore, presenting tools to educators and learners that enable them to use their own theoretical and practical conceptions of learning to explore the sensitivity of clusters, for example, by changing the content dimensions or participants considered when computing the partition, may help provide more context about the trends observed within these hidden groups.

As we consider how to automate feedback that pertains to communication communities in large-scale learning set-

tings, a key constraint to replicating this study and embedding generative modelling of discussion forum content in online environments is the time and effort required to content-analyse thousands of forum posts. We can envision a future scenario where these content labels are inferred automatically through natural language processing software. Another possibility is crowdsourcing – not necessarily through traditional channels like Amazon Mechanical Turk, but rather, through forum viewers and other course participants. Existing methods for calibrating responses in crowdsourcing efforts (Simpson et al., 2013; Piech et al., 2013) could then be adapted in order to infer the bias and reliability of contributing individuals. Indeed, creating a user experience around post-labelling in an online course environment and making this one component of coursework could doubly serve as an opportunity to track engagement-levels of forum lurkers, since presently, clickstream data on discussion forum and discussion thread views fail to reveal the extent to which a learner actually reads and absorbs posted content.

As large-scale social learning spaces proliferate, understanding the different ways diverse groups of individuals contribute and engage in discussions and exploring scalable mechanisms for data collection may play a fundamental role in how we assess learners and provide feedback to both learners and educators around the world.

Acknowledgements

The authors would like to thank the MOOC Research Initiative for enabling this interdisciplinary work, as well as the University of Virginia and Coursera for their support in accessing relevant datasets. Thanks also to Chris Davies and Bav Radia in Oxford's Department of Education for their help in content analysis, and Taha Yasseri, Ioannis Psorakis, Rory Beard, Tom Gunter, and Chris Lloyd for their insights. Finally, thanks to Sunil Gupta at the University of Deakin for sharing code for the LPGM.

References

A. Anderson, D. Huttenlocher, J. Kleinberg, J. Leskovec (2014). Engaging with Massive Open Online Courses. *Proceedings of the 14th International World Wide Web Conference Committee*. Seoul, Korea.

B. De Weaver, T. Schellens, M. Valcke, and H. Van Keer (2006) Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers and Education*, 46(1): 6-28.

C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, F. M. F. Wong (2013). Learning about social learning in MOOCs: From statistical analysis to generative model. arXiv:1312.2159v2.

C. N. Gunawardena, C. A. Lowe, and T. Anderson (1997). Analysis of a Global Online Debate and the Development of an Interaction Analysis Model for Examining Social Construction of Knowledge in Computer Conferencing. *Journal of Educational Computing Research*, 17(4): 397âĂŞ431.

C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, D. Koller (2013). Tuned Models of Peer Assessment in MOOCs. *6th International Conference on Educational Data Mining*, memphis, Tennessee, USA.

C. Reed and Z. Ghahramani (2013). Scaling the Indian Buffet Process via Submodular Maximization. arXiv:1304.3285v4.

C. Rose (2013). Enabling Resilient Massive Scale Open Online Learning Communities through Models of Social Emergence. *MOOC Research Initiative Conference*, Arlington, TX.

D. B. Clark, V. Sampson, A. Weinberger, and G. Erkens (2007). Analytic Frameworks for Assessing Dialogic Argumentation in Online Learning Environments. *Educational Psychology Review*, 19(3): 343-374.

D. Blei, A. Ng, M. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993-1022.

D. D. Lee, H. S. Seung (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401: 788-791. doi:10.1038/44565.

D. F. Nobel (1998). Digital diploma mills: the automation of higher education. *Science as Culture*, 7(3): 355-368.

D. J. C. Mackay (1995). Probable networks and plausible predictions a review of practical Bayesian models for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469-505.

E. Airoldi, D. M. Blei, S. E. Feinberg, E. P. Xing (2008). Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, 9: 1981-2014.

E. Simpson, S. Roberts, I. Psorakis and A. Smith (2013). Dynamic Bayesian Combination of Multiple Imperfect Classifiers Decision Making and Imperfection. Intelligent Systems Reference Library series Vol 474. Springer.

F. Doshi-Velez (2009). *The Indian Buffet Process: Scalable Inference and Extensions* (Master's Dissertation). University of Cambridge, Cambridge, UK.

G. Erkens, and J. Janssen (2008). Automatic coding of dialogue acts in collaboration protocols. *International Journal of Computer-Supported Collaborative Learning*, 3(4): 447âĂŞ470. doi:10.1007/s11412-008-9052-6

G. Stahl, T. Koschmann and D. Suthers (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer, ed. *Cambridge handbook of the learning sciences.* Cambridge, UK: Cambridge University Press, 409-426.

I. Psorakis, S. Roberts, M. Ebden, B. Sheldon (2011a). Overlapping Community Detection using Nonnegative Matrix Factorization. *Physical Review E*, 83, 066114.

I. Psorakis, S. Roberts, I. Rezek and B. Sheldon (2011b). Inferring social network structure in ecological systems from spatio-temporal data streams. *Journal of the Royal Society Interface*, 9(76): 3055-3066.

L. Vygotsky (1978). *Mind in Society.* Harvard University Press, Cambridge, MA.

R. Pekrun, T. Goetz, W. Titz, and R. P. Perry (2002). Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program of Qualitative and Quantitative Research. *Educational Psychologist*, 37(2), 91âĂŞ105.

S. K. Gupta, D. Phung, S. Venkatesh (2012). A Nonparametric Bayesian Poisson Gamma Model for Count Data. *Proceedings of 21st International Conference on Pattern Recognition*, 1815-1818. Tsubka Science City, Japan.

T. Griffiths and Z. Ghahramani (2005). Infinite latent feature models and the Indian buffet process. *Technical Report 2005-001, Gatsby Computational Neuroscience Unit.*

V. Tan and C. Fèvotte (2009). Automatic relevance determination in nonnegative matrix factorization. *SPARS09 - Signal Processing with Adaptive Sparse Structured Representations*, 1-19.