Uncovering latent features in massive open online courses

Submitted for the MSc by Research in Engineering Science

New College



Supervisors:
Dr. Michael Osborne
Professor Stephen Roberts

Nabeel Gillani

Michaelmas 2014

Acknowledgements

This dissertation would not be possible without the invaluable contributions of so many people. First and foremost, I'd like to thank God, my mother, father, my brother Riaz, and brother from another mother Zafrin for investing so much care and love in me and my intellectual pursuits. Their support has enabled – and challenged – me to explore how we can build bridges between science and society, as this dissertation hopes to do.

Many thanks to my incredibly supportive supervisors, Dr. Michael Osborne and Professor Stephen Roberts, for their most helpful insights and advisory. They've taught me how to do well-informed, principled research while simultaneously encouraging me to seek knowledge and insights from people and places beyond the Machine Learning lab. Many thanks to Professor Michael Lenox and Kristin Palmer at the University of Virginia, as well as Coursera, for providing access to the datasets necessary for this effort. Thanks to Dr. Rebecca Eynon for her advisory over the past two years, and for building a cross-disciplinary team of wonderful collaborators – including Dr. Taha Yasseri and Dr. Isis Hjorth – to take this research to the next level. Without the diversity of thought and hard work put forth by each of these individuals, our explorations and insights would be severly limited.

Finally, thank you to all of my friends, family, and collaborators for engaging and inspiring me, listening to half-baked research ideas, proofreading manuscript drafts, and asking deep, piercing questions to improve the quality and clarity of this work. As we look to the future, I am inspired by the intelligence and abilities of these individuals, and I am excited to be a part of the world that they will build.

Publications

We have been most fortunate that a number of research communities have found the outputs of our investigations to be of interest. The following publications have resulted from this work:

N. Gillani, R. Eynon. Communication Patterns in Massively Open Online Courses (2014). *The Internet and Higher Education*, 23, 18-26.

This paper inspired much of the work included in this dissertation. For this paper, I helped formulate the research questions and methodology, conducted all data cleansing and computational analysis, synthesized results into key insights, and drafted the manuscript.

N. Gillani, T. Yasseri, R. Eynon, I. Hjorth (2014). Structural limitations of learning in a crowd: communication vulnerability and information diffusion in MOOCs. *Nature: Scientific Reports*, 4, 6447.

Chapter 3 is almost entirely represented by this paper. For this paper, I helped formulate the research questions and methodology, conducted data cleansing and computational analysis particularly as it pertained to the extraction of significant social networks and subsequent network vulnerability explorations, synthesized results into key insights, and drafted the manuscript.

N. Gillani, R. Eynon, M. Osborne, I. Hjorth, S. Roberts (2014). Communication Communities in MOOCs. Poster at the *2014 KDD Workshop on Data Mining for Educational Assessment and Feedback.*

Chapter 4 contains the majority of this paper and presents some additional analysis.

For this paper, I helped formulate the research questions and methodology, conducted data cleansing and computational analysis particularly as it pertained to manipulating content-labelled forum posts, synthesized results into key insights, and drafted the manuscript.

R. Eynon, I. Hjorth, N. Gillani, T. Yasseri (2014). 'Vote Me Up If You Like My Ideas!' Experiences of Learning in a MOOC. *2014 ICA Pre-conference working paper: Innovation in Higher Education: Building a Better Future?.*

Qualitative findings from this paper are discussed in Chapter 4. For this paper, I helped formulate the research questions and methodology, helped conduct relevant quantitative analysis, and provided comments on the manuscript draft.

Abstract

Massive Open Online Courses (MOOCs) bring together a global crowd of thousands of learners for several weeks or months. We use social network analysis and community detection to uncover the latent features of online discussions in MOOCs. We begin by using data from two successive instances of a popular business strategy MOOC to filter observed communication patterns and arrive at the "significant" interaction networks between learners. We then use complex network analysis to explore the vulnerability and information diffusion potential of the discussion forums. While network analysis offers a vibrant post-hoc analytical framework, it fails to answer a fundamental question: can we devise a model to represent the generation of the dataset at hand?

Moving from the structural properties of global-scale discussion to the discussion content itself, we employ existing educational theories to qualitatively content-analyse over 6,500 forum posts from a particular MOOC. We then use a generative model - Bayesian Non-negative Matrix Factorization (BNMF) - to extract communities of learners based on the nature of their online forum posts. We observe that the inferred communities are differentiated by the nature and topic of dialogue, as well as their composite students' demographic and course performance indicators. While qualitative analysis confirms these detected communities, additional quantitative sensitivity analysis shows that they are not crisply defined, illuminating key challenges of applying Machine Learning techniques to model noisy and incomplete learner data.

We conclude by discussing the key insights of this work for online education, namely, that different discussion topics and pedagogical practices promote varying levels of peer-to-peer engagement. Additional qualitative investigations reveal that many learners feel a sense of "content-overload" when deciding to participate in online discussions, often leading to their disengagement. These insights call for an interdisciplinary effort to help create relevant and personalized learning experiences in massive scale online settings.

Contents

1 Introduction **14**

2 Review **20**

2.1 Exploring Digital Education 20

2.2 Network Formulation and Analysis 24

2.2.1 Network Analysis in Education 25

2.2.2 Formulating the Network 26

2.2.3 Inferring Relevant Network Attributes 28

2.3 Latent Feature Models 32

2.3.1 Non-negative Matrix Factorization 33

2.3.2 Bayesian Latent Feature Models 34

2.4 Summary 40

3 Structural properties of learning in a crowd **42**

3.1 Introduction 42

3.2 Methodology 44

3.3 Results 46

3.3.1 Significant Interaction Networks 47

3.3.2 Communication vulnerability 52

3.3.3 Information Diffusion 56

3.4 Discussion 59

4 Content-inspired communication communities 62

4.1 Introduction 62

4.2 Dataset and Latent Features 63

4.2.1 Content Analysis of Forum Data 63

4.2.2 Probabilistic Generative Model 65

4.2.3 Inference and Cluster Assignment 66

4.3 Results 70

4.3.1 Model Benchmarking 70

4.3.2 Exploring Extracted Communities 73

4.3.3 Robustness and Sensitivity Analysis 81

4.4 Discussion 86

5 Conclusion 88

5.1 Key contributions 89

5.1.1 What are the structural and diffusion properties of communication networks in MOOCs? 89

5.1.2 How can the textual content of discussion forum posts be used to infer latent learner communities, and how crisp are these communities? 90

5.2 Improving massive-scale learning 92

5.2.1 Automating or crowdsourcing content analysis 92

5.2.2 Dealing with "content-overload" 92

References 95

Appendix 106

Supplementary Information for Communication Network Analysis 106

Supplementary figures 106

Robustness and Sensitivity Analysis for Communication Communities Extraction 110
Modularity scores for different content labels 110
Analysis Toolbox 113

List of Figures

3.1 The observed (a) and derived (b) communication networks for the *Study Groups* sub-forum. Here, we can see the impact of link filtration on network properties such as modularity score, which equals 0.62 and 0.80 for a and b, respectively. Colours correspond to the detected communities. 50

3.2 Network vulnerability for the different sub-forums in FOBS-1. Here, LCC refers to the largest connected component in each sub-forums communication network. 54

3.3 Vulnerability versus random removal for Cases sub-forum in FOBS-1. Here, LCC refers to the largest connected component in this sub-forums communication network. 54

3.4 (a) shows the percentage of infected nodes vs. simulation time for different networks. The solid lines show the results over the original network and the dashed lines for the degree-preserved shuffled network (configuration model), and (b) shows the value of e (i.e., the diffusion efficiency) for different sub-forums. 57

4.1 Content dimensions and labels used to qualitatively analyse discussion forum posts. 64

4.2 An example representation of our data matrix X. The entry u_2u_3, for example, is equal to 0, suggesting that users 2 and 3 had no similar content labels assigned to their posts in a particular sub-forum. . . . 66

4.3 Graphical model for the joint distribution over all variables in the BNMF model. Here, H=W because our similarity matrix is symmetric. 67

4.4 Community Detection using Bayesian NMF. As described in (Tan and Fèvotte, 2009), the algorithm uses an efficient multiplicative coordinate descent algorithm to minimize the KL-Divergence between the observed data X and its factor matrices, W and H, arriving at point estimates for the parameters of interest. To learn the full posterior parameter distributions, other methods (e.g. Gibbs sampling, variational inference) may be used, for example, as in (Schmidt et al., 2009). 69

4.5 Graphical models for BNMF (a) and LPGM (b). In addition to the explicit representation of the binary latent feature matrix in (b), one of the most notable discrepancies in both models is the absence of a parameter in the LPGM that ties together the rows and columns of the latent factor matrices. Highly-correlated values in the real-world dataset may warrant this common precision parameter, thereby limiting the predictive potential of LPGM. 72

4.6 Plot (a) illustrates the dialogue acts represented in the posts made by learners belonging to each community, and plot (b) depicts their geographic locations. 77

4.7 Plot (a) illustrates the discussion topics represented in the posts made by learners belonging to each community, and plot (b) depicts their course outcomes. 80

4.8 Each histogram depicts the number of learners with a maximum community membership belief score depicted by the values on the x-axis, renormalized over only those communities to which at least two learners were allocated (a membership proportion of 1 means that a particular learner belongs entirely to a single community). From these charts, we can infer that for many learners, community membership is highly entropic, suggesting inherently weak community structure. . 83

4.9 The x-axis of each plot represents the sub-network induced over the full similarity matrix of each sub-forum where only the top x% of learners with highest proportional commitment to their respective communities are considered, and the y-axis reports the corresponding Louvain modularity score for each sub-network (so, the x-value of 1 and corresponding y-value indicates the modularity score for the full similarity matrix). Plots (a) and (b) depict trends for the Cases and Final Project sub-forums, respectively. 84

1 Forum post activity over time in the Final Project sub-forum of FOBS-1. The large peak around the end of week 6 corresponds to students posting last-minute questions about the final project submission deadline. 107

2 Forum post activity over time in the Cases sub-forum of FOBS-1. Like many of the other sub-forums, participation decreases as the course progresses, but there are still peaks of activity each week corresponding to the weekly case discussions. 107

3 The number of views and posts per discussion threads across all sub-forums, in log-log scale, for FOBS-1. The charts suggest a fat-tailed distribution of views and posts across threads i.e., the vast majority of discussion threads have very small numbers of posts and views, with a few threads harbouring high posting and viewing behaviour. . 108

4 Comparison of posts and views for each thread in a particular sub-forum, denoted by the coloured circles shown here, for FOBS-1. The size of each circle indicates the Popularity time persistence of the corresponding thread, i.e., the amount of time that elapses before 90% of all posts are made to that thread (hence, small circles depict threads with very short lifespans). 108

5 Communication vulnerability in the different sub-forums of FOBS-2. These trends are similar to those observed in FOBS-1. 109

6 shows the percentage of infected nodes vs. simulation time for different networks in FOBS-2 (similar to those observed in FOBS-1). The solid lines show the results over the original network and the dashed lines for the degree-preserved shuffled network (configuration model). 109

List of Tables

3.1 Summary statistics for forum participation in FOBS-1 and FOBS-2. N_l, N_p, and N_p / N_l denote the number of learners, number of posts, and average posts per learner (along with standard deviations in brackets) in each sub-forum, respectively. 47

3.2 Observed and derived communication networks for the different sub-forums in FOBS-1 and FOBS-2. Here, $|E_o|$ indicates the number of edges originally in the denoted sub-forum's network; $|E_s|$ indicates the number of edges in the derived significant network; N indicates the number of learners connected to at least one other learner in the network; and Δ is the percentage decline in the number of edges after the link filtration. 51

3.3 Modularity scores of the original and derived significant networks. Here, M_o is the modularity of each original network; M_s is the modularity of each derived significant network; and NM indicates each network's normalized modularity (i.e., M_o/M_s). 59

4.1 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. 73

1 Modularity scores for the Cases sub-forum's similarity matrices, constructed by considering different subsets of content label dimensions. . 111

2 Modularity scores for the Final Project sub-forum's similarity matrices, constructed by considering different subsets of content label dimensions. 112

Chapter 1

Introduction

There has been no shortage of either hype or disillusionment around Massive Open Online Courses (MOOCs) over the past three years. Since their popularisation at the end of 2011, both proponents and critics have analysed how MOOCs can facilitate learning across geographic divides, prior levels of educational attainment, and socioeconomic contexts. While some researchers have indicated the potential of these courses to "democratize education" and extend access to marginalised populations (Mahraj, 2012), others have revealed that early MOOC participants are already well-educated and primarily reside in western countries (Ezekiel, 2013).

Despite this polarised attention, it is important to note that online education, as well as its underlying research, is not new. Since the launch of the public internet browser, some educationists have warned of the rise of "digital diploma mills" resulting from the mechanisation of education offered through online media (Nobel, 1998). Others have embraced the learning possibilities afforded by a boundary-less medium, calling for new education theories that build on constructivist pedagogy (Vygotsky, 1978) to explain how connections between people can be supported and amplified to facilitate "social learning" (Siemens, 2005).

These pedagogical frameworks have drawn inspiration from – and themselves

have further influenced – the realities of online learning. The launch of MIT's OpenCourseware (OCW) effort at the turn of the century (Abelson, 2008) offered hope for universal access to high-quality educational resources. MOOCs are, in some sense, a new incarnation of previous forms of distance education like OCW. They differ primarily in their scale and semi-synchronicity: with a defined start date and duration, they bring together hundreds to hundreds of thousands of learners from around the world for a period of time, during which learners can engage with content but also one another through online discussion forums. Moreover, MOOC platforms like Coursera and edX have implemented software to continuously capture granular digital trace data at the learner-level, recording millions of interactions per course that range from the full-text of discussion forum posts to the time and location of each mouse click.

Early MOOC research has exploited the availability of this granular data to try and understand broad trends in participation and engagement with course content (Breslow et al., 2013; Kizilcec et al., 2013). More recent work has begun to investigate the semi-synchronous nature of co-participation in MOOCs, studying online interactions and communication in these spaces (Brinton et al., 2013; Anderson et al., 2014). In order to infer common themes and structures underpinning the interactions of thousands of lifelong learners, some MOOC researchers have started to employ techniques in computational social science and machine learning, for example, to calibrate peer-assessment and predict drop-out rates (Piech et al., 2013; Rose, 2013; Taylor et al., 2014).

Given the importance of understanding how social interactions unfold in MOOCs, it is useful to explore how machine learning can aid in this analysis. Previous computational education research has used data mining and basic machine learning techniques to understand and model learning processes. However, few have assembled and used a hybrid toolbox of techniques that include complex network analysis

and principled Bayesian machine learning. Network analysis is a widely-used framework for exploring the interactions between actors in social, biological, and other settings (Wasserman and Faust, 1994; Haythornthwaite, 1996). It has been used across domains – including education (e.g. Vaquero and Cebrian, 2013), and more recently, MOOC research (Gillani and Eynon 2014; Kellogg et al., 2014) – largely because it provides in-built metrics for describing the density of connections between actors, correlations between actors and their characteristics, inherent community groupings/structures, and a host of other patterns of interest.

While network analysis offers a vibrant post-hoc analytical framework, it fails to answer a fundamental question: can we devise a model to represent the generation of the dataset at hand? Bayesian probabilistic generative models offer flexible frameworks for representing these datasets through the incorporation of prior information, and in nonparametric settings, infinite-dimensional objects that make minimal assumptions about the data in which they wish to infer patterns (Gershman and Blei, 2011). Prior work on recommender systems, for example, has used Bayesian generative formulations to extract perference information on actors in a particular setting, such as those rating movies on Netflix (e.g., Mackey et al., 2010). Indeed, much recent work has used Bayesian methods to better understand, and extract insights from, social networks in order to detect latent communities and quantify the likelihood of an underlying social tie in the absence of an observed connection between two or more actors (Schmidt and Mørup, 2013). Many of these methodologies, which have been applied to the social and life sciences in the past, are important to investigate for their relevance and usefulness to the education domain.

With large amounts of digital trace data from learners in massive online environments, we seek to explore the following research questions:

1. What are the structural and diffusion properties of communication networks in MOOCs?

2. How can the textual content of discussion forum posts be used to infer latent learner clusters, and how crisp are these clusters?

3. What roles can machine learning and computational social science play in communicating actionable insights into social learning patterns to help improve learning experiences online?

Given the breadth and complexity of these questions, a hybrid analytical framework that combines post-hoc social network analysis and principled generative Bayesian models that are well-suited to account for uncertainty in observed data is needed. However, simply employing a hybrid computational toolbox is not enough to deliver both technically and practically sound results to the education community; it is also important to cross-check these results with existing educational theory and domain experts. Indeed, this complete approach may help prevent the miscommunication or misinterpretation of results that could ultimately negatively impact the learning experiences or outcomes for millions of people around the world.

With these guiding approaches in-hand, **we hope to strengthen our understanding of how social learning occurs by identifying the key latent forces that characterize and influence online discussions between individuals in a "global classroom"**. The rest of this document proceeds as follows:

In **Chapter 2**, we present a Review of the literature, including the use of network analysis and Bayesian generative models in the educational data minining community. We explore probabilistic network formulations as mechanisms for more accurately depicting true underlying relationships in noisy social settings (e.g., online discussion forums). We also present a review of different latent feature models used by the Machine Learning community, with special emphasis placed on matrix factorization techniques.

In **Chapter 3**, we explore the structural properties of communication networks

from two successive instances of a business strategy MOOC with over 150,000 participants in total (Research Question 1). Here, we assume the observed communication network in a MOOC is a noisy sample from some true underlying network, and apply a link filtration algorithm used in ecology and other domains to derive this "significant" discussion network. We then explore two non-canonical network characteristics – communication vulnerability and information diffusion potential – to show how certain discussion topics in MOOCs tend to promote more engagement and inclusive dialogue than others.

In **Chapter 4**, we transition from a high-level look at the structural properties of learners' communciation networks to explore the actual content of online discussion and the latent learner groups that can be inferred based on this content (Research Question 2). We describe a content-labelling scheme used to hand-label over 6,500 posts and use Bayesian Non-negative Matrix Factorization (BNMF) to cluster learners based on their labelled forum posts. We see that the communities BNMF learns are constituted of individuals with significantly different socioeconomic and demographic backgrounds. While qualitative interviews appear to confirm the nature of these communities, additional quantitative sensitivity analysis shows that the communities are not crisply defined. This presents an important insight for those engaged in computational education research, particularly those focused on MOOCs: it can be difficult to find robust structure in noisy human-behaviour data, and doing so requires combining multiple approaches to validate any derived insights.

In **Chapter 5**, we conclude by reviewing the key contributions of this work, discussing implications for educators and instructional designers, and describing opportunities for further research in this nascent field.

Indeed, the scale and complexity of data in massive online learning settings requires multi-level analysis, exploring both the macro-trends in social exchanges and shared dialogue in discussion forums, but also micro-trends that characterize

learners and the discussion content that they produce. By leveraging a hybrid toolbox, we explore and model MOOC data at multiple levels of granularity in order to arrive at the following key insights: 1) different discussion topics and pedagogical practices promote varying levels of peer-to-peer engagement and inclusiveness, and 2) that many learners feel a sense of "content-overload" when deciding to participate in online discussions, often leading to their disengagement. These insights call for education practitioners, software engineers, and data scientists to come together to create new software and technical frameworks that make use of granular learner data in order to help create relevant, personalized learning experiences online.

Chapter 2

Review

2.1 Exploring Digital Education

The interdisciplinary field of educational data mining (EDM) has, in recent years, brought together qualitative and quantitative researchers to collaboratively explore how to better facilitate and support the learning processes of a wide range of students. As Romero and Ventura highlight in their 2010 review, a wide range of studies have leveraged hybrid computational toolboxes, including social network analysis and machine learning, in order to further explore relationships that form in learning settings, understand the effectiveness of adaptive tutoring environments, recommend course content to students, and aid in the planning and scheduling of future curricula, for example (Romero and Ventura, 2010).

Research into user modelling and implementing intelligent tutoring systems has been particularly widespread, leveraging non-probabilistic classification methods like support vector machines (SVMs) and neural networks in order to predict course outcomes (Hamalainen et al., 2006). Some have also explored Bayesian Knowledge Tracing (BKT) in this setting, using a hidden Markov model (HMM) to model a student's progression through educational content in order to infer the likelihood

of content mastery based on previous encounters and applications (Van De Sande, 2013). Indeed, BKT has been adapted by early MOOC researchers to account for non-linear pathways through course content (Pardos et al., 2013). In other instances, Bayesian networks have been used to learn the learning styles of students in web-based educational environments, where nodes represent particular actions within a course environment and inference is performed to determine the likelihood of certain actions based on those that have been observed (Garcia et al., 2005). In many cases, inference in these settings has been performed through models trained via maximum likelihood methods (Yudelson et al., 2013).

Still, computational research in educational settings has not been confined to developing models that characterize or predict learning processes. Many researchers, educators, and policy-makers have also expressed interest in the social networks that emerge in learning settings. Social network analysis is a powerful framework for exploring the relationships (edges) between a collection of actors (nodes) (Haythornthwaite, 1996). it has been used across disciplines (e.g., Buchanan and Calderalli, 2010) to help characterize connectivity and resource exchanges. In education, social network analysis has been used to explore communication patterns between students in a computer-supported collaborative environment (Palonen and Hakkarainen, 2000), and more recently, file exchanges between thousands of learners, where high performing students tended to form "rich clubs" with other high performing students in a digital learning environment (Vaquero and Cebrian, 2013).

In addition to the growing interest in learner/instructor social networks, researchers and practictioners have also explored how to group students together based on similar attributes or behaviours in order to inform course design, content delivery, and a number of other factors that could influence learning. One study compared agglomerative, k-means, and mixture-model clustering in order to group students according to similar skill profiles (Ayers, 2009). Others analysed learners' usage

of online discussion forums to construct time-series of recorded interactions, which they ultimately clustered using agglomerative hierarchical clustering (Cobo et al., 2011). More recent efforts have explored graph-based clustering, using spectral partitioning on bipartite learner-to-feature networks to infer latent groups based on the correlations between learners and their course performance features (Trivedi et al., 2012). In most cases, the clustering methods used have not specified a full probabilistic generative model of the data. In instances where probabilistic models (e.g., Markov models) have been employed, clustering has been conducted according to a "hard partition" approach, that is, assigning a particular learner exclusively to one group instead of informing a distribution or belief over their possibly-overlapping group participation.

For MOOCs, where observed data is often noisy due to sporadic learner participation and high attrition rates (Kizilcec et al., 2013), an understanding of both macro and micro-trends in these datasets is important to inform pedagogical design, practices, and when relevant, interventions. For social network analysis to serve as an appropriate macro-analytical tool, sufficient care must be given to formulating the network topology (namely, what nodes and edges represent), and determining networks metrics that have practical significance in explaining learner interaction patterns.

For a robust understanding of the data at a micro level – for example, to infer groupings and other latent structures that exist in an inherently noisy dataset – a probabilistic generative model of the data is favourable because it 1) inherently assumes the observed data is a noisy sample from some underlying generating force, 2) provides a framework for predicting values for missing or incomplete data, and 3) depending on the model, may inform a belief over object (i.e., learner) assignments to multiple groups. A Bayesian context further specifies prior and likelihood models for the data, which, given the diversity of the type of data collected in online course

settings, allows for model flexibility that is better-suited to represent, and ultimately extract insights from, observed trends. This flexibility also enables existing models to be extended and adapted to help answer new research questions.

Given these advantages of probabilistic generative models, recent MOOC studies have made efforts to leverage network analysis and clustering approaches to better understand student experiences in these courses. Some work has begun to explore how MOOCs can harbour the types of social environments needed to promote engagement and resilience and decrease attrition rates (Rose, 2013), leveraging an adaptation of the Mixed-Membership Stochastic Blockmodel (Airoldi et al., 2008) as a probabilistic community detection/clustering model. In order to characterize the content of MOOC discussion forums, some researchers have proposed generative topic models that indicate the relevance of communication to course content (Brinton et al., 2013). To our knowledge, Bayesian Nonparametric methods have yet to be used to explore MOOC data, although recent work employed Gaussian Process optimization to select optimal instructional policies for learners in non-MOOC educational settings (Lindsey et al., 2013).

There is an immense opportunity at hand to build upon early MOOC research to define and analyse social networks and leverage probabilistic generative models to infer latent groups in educational settings. With a hybrid analytical framework, we can begin to better understand and characterize social learning in large-scale, online contexts. We now turn to a discussion of social network analysis in education, in particular, how network formulation is a crucial step to deriving meaningful metrics and insights to inform decision-making.

2.2 Network Formulation and Analysis

Network analysis is the study of how actors (nodes) in some setting exchange resources with one another through specific relationships (edges) (Newman, 2010; Wasserman and Faust, 1994). Given this generality, network analysis has received much attention over the past years from a wide spectrum of researchers and practitioners, each using several different methods for defining network topologies and subsequently exploring their underlying (sometimes latent) properties like vulnerability, community structure, and diffusion potential (details provided later in this chapter). Open-source network analysis toolboxes like *NetworkX* and *Gephi*, for example, have provided standard, out-of-the-box algorithms and visualization packages to further promote these explorations. Beyond immediate applications, many mathematicians and statisticans have delved into a theoretical exploration of networks in order to explore how they may better represent time-evolving phenomena (Porter and Gleeson, 2014), or how they may be manipulated (sometimes drawing upon graph theory from computer science) in order to inform recommendations or predictions (Caron, 2012). This latter strand of investigation has been of particular interest to machine learning researchers, as the prevalence of network structures in real-world (particularly social) settings has given rise to new models for community detection (e.g., Airoldi et al., 2008) and edge prediction in the wake of noisy or incomplete datasets (e.g., Schmidt and Mørup, 2013).

Despite widespread interest in and applicability of network analysis, its application to specific domains and the subsequent extraction of meaningful insights is largely contingent upon *how* the network topology is defined, and *which* network attributes are explored. In this section, we begin with a brief review of network analysis in education and the network formulations adopted by these investigations. We then discuss two different approaches to formulating networks that have been

used across domains, followed by several post-hoc analytical methodologies, and how these tools can be leveraged to better understanding learner communications and interactions in MOOCs.

2.2.1 Network Analysis in Education

Network analysis in education research to date has primarily focused on modelling small-scale (e.g., classroom-level) social interactions between students. For example, (Sundararajan, 2010) used social network analysis with a directed network topology to model communication and information exchange between learners over the course of three semesters. Similarly, (Reffay and Chanier, 2003) explored a number of potential network topologies for different forms of learner-to-learner resource exchanges in computer-supported collaborative environments (Stahl, 2006). Reffay used directed networks to represent email correspondence or discussion forum participation, both of which define a link as representing a message initiated by learner A and opened by learner B, during some time period P. In the case of discussion forums, the author noted high network density, proceeding to analyse the resultant networks (e.g. computing learner co-participation in cliques as a measure of centrality) instead of exploring how the network formulation could be refined to only consider some subset of relevant learner interactions.

A number of other works have leveraged social network analysis in education contexts to explore communication in digital forums (e.g., Kepp and Schorr, 2009 [as cited in Vaquero and Cebrian, 2013]; Cho and Gay, 2007 [as cited in Vaquero and Cebrian, 2013]; Huerta-Quintanilla, 2013), but as recent researchers (e.g. Vaquero and Cebrian, 2013) point out, most of these studies have constructed networks that are "static snapshots" of small-scale communication dynamics at some point in time. Some early MOOC investigations have started to explore communication dynamics at scale, for example, by using network analysis and random graph models

to characterise interactions (Kellogg et al., 2014). While the sample size for these recent efforts has been much larger due to the massive-scale setting, little attention has been paid to formulating a network that accounts for the rapid decay of MOOC interactions over time, or perhaps more importantly, the underlying (latent) social relationships that form as learners interact during a course.

We seek to fill this gap in the literature by investigating how network analysis can be leveraged to infer the underlying strength of social connections between MOOC participants. We start by investigating alternative network formulation methodologies that can be adapted to the MOOC environment, after which we explore a subset of metrics to further analyse the macro-level structural properties of dialogue and social learning in MOOCs.

2.2.2 Formulating the Network

Despite the simple and general analytical framework offered by a networks perspective of resource exchange and its previous applications to education, several works have highlighted different methods for defining a particular network topology (e.g., Krause et al., 2009 [as cited in Psorakis, 2013]). Here, we draw inspiration from (Psorakis, 2013), which explored social network formulation based on spatio-temporal sensor data from wild bird interactions, as we discuss different network formulation methodologies. Our ultimate goal is to highlight models that will help us infer the latent social relationships of MOOC participants based on observed online communication tendencies.

The time-window approach

With growing interest in the temporal evolution of networks (including in education – e.g., Vaquero and Cebrian, 2013), some researchers have leveraged temporal information as they define the network topology before performing subsequent anal-

ysis. For example, as (Psorakis, 2013) notes, several works (e.g. Krings et al., 2012) have defined undirected connections between actors A and B if both co-participated in a given social setting during a particular time window Δt (e.g., $\Delta t = 7$ days). However, because varying this time window parameter often leads to drastically different topological realities (e.g., network density increases with Δt), without precise domain or expert-driven rationale to pre-select a particular value for t, the choice remains arbitrary, and any insights derived from analysis performed on the subsequently formed network are subject to change.

In the MOOC setting, where course content and assignments are often presented week-by-week, $\Delta t = 7$ days may be a reasonable starting point. Indeed, previous work (Gillani and Eynon, 2014) leveraged week-by-week temporal resolution to explore how participation in discussion threads evolved throughout a 6-week long MOOC. However, discussion threads have varying lifespans (see Appendix), and two learners that co-participated in a shared discussion space during weeks 1 and 6, respectively, may still have had an exchange that "counts" as a significant interaction. This suggests that a time-window based approach – and more broadly, network formulation methodologies in most education/MOOC research to date – are insufficient in depicting meaningful social interactions and relationships in noisy online learning environments where sporadic and irregular participation are the norm.

The statistical approach

The issues with Δt-inspired topological definitions extend beyond the arbitrariness of selecting t: they assume that all relationships (denoted by an edge between two actors) are of equal significance. In MOOCs, where thousands of learners are participating in discussion forums, and posts are often motivated by course announcements, assignments, pending deadlines, or other triggering events, not all co-participations in discussion threads are indicative of a meaningful social exchange – rather, many

happen coincidentally, and therefore, should not necessarily be represented when constructing a network that captures significant social connections. (Psorakis, 2013; Psorakis et al., 2011b) Addresses an analogous issue when modelling the social networks of wild bird populations. The author defines an observed bird-to-bird social network with an edge between any two birds representing co-participation in at least one "gathering event". The author then generates a set of independent sample bird-to-bird networks based on shuffled participation profiles to serve as a null model against which the statistical significance of each observed connection is tested (see Psorakis, 2013) for full details of this model). In this way, the resultant network depicts the social interactions that are indicative of an underlying *significant* exchange or relationship while filtering out those interactions that appear to have occurred by chance. A similar null-model / hypothesis testing methodology has been adopted in recent work on mobile communication networks (Li et al., 2014) and pattern recognition in the strategies of financial investors (Tumminello et al., 2012), suggesting the potential suitability of this method to infer latent structures in complex human behaviour data.

In the case of MOOCs, because most discussion threads are short-lived and many interactions are the result of coincidence, a key underlying objective is to uncover latent meaningful social relationships between learners. Indeed, Chapter 3 slightly extends the work of (Psorakis, 2013; Psorakis et al., 2011b) to better represent human interactions in massive-scale learning settings and presents details of the probablistic model and sampling procedures used to determine which connections in the observed social network appear to represent a meaningful social relationship.

2.2.3 Inferring Relevant Network Attributes

A robust network formulation methodology is an important first-step in network analysis. Many researchers across domains have chosen network analysis precisely

because of the wealth of metrics it offers to characterise the structural properties of resource exchanges (e.g, average node degree, network diameter, etc. (Newman, 2010)). Below, we briefly explore two particular metrics – node centrality and modularity – and intrepret their usefulness in understanding the macro-level structural attributes of learning in MOOCs.

Centrality

The notion of a given node's "centrality" in a network has long been regarded as an important metric across application domains – but several methods for measuring this centrality have emerged (Freeman, 1978). For example, a node's in-degree centrality (or more generally, degree centrality for undirected networks) is one of the simplest measures of relative placement and is computed by counting the number of edges that are incident upon that particular node (Costenbader and Valente, 2003). Some measures like the popular (Google) PageRank function define node a's centrality in terms of the number of edges that are incident upon a – giving additional weight to nodes that themselves are "more important" (e.g., have a higher centrality score). Indeed, PageRank is one special example of yet another centrality measure known as eigenvector centrality (Spizzirri, 2011).

In the context of social networks, one of the most popular centrality measures is "betweenness centrality" because it quantifies a node's tendency to be a "gatekeeper" of knowledge or other resources between network actors (Costenbader and Valente, 2003). A node's betweenness centrality measures the number of shortest paths between any node pair that travel through that particular node. More formally, the betweenness centrality of node a, given by bc_a, can be expressed as $bc_a = \sum_i \sum_j \frac{g_{iaj}}{g_{ij}}$ for $i \neq j \neq a$, where g_{ij} are all of the shortest paths between nodes i and j, and g_{iaj} are those shortest paths that also travel through a (Leydesdorff, 2007).

Recently, betweenness centrality has been used in the MOOC literature to ex-

plore correlations between network positioning and other characteristics (e.g. Sinha, 2014). For example, (Jiang et al., 2014) found that communication networks in two separate MOOCs (algebra and financial planning) both exhibited statistically significant node betweenness centrality values, suggesting higher levels of participation and influence among certain participants than what would be expected by chance. However, the authors also noted mixed results when exploring correlations between a given learner's betweenness centrality and final marks in each course, perhaps due to the popular tendency for MOOC participants to engage in discussion forums even if they have no intention of pursuing formal completion (Gillani and Eynon, 2014). Another paper (Yang et al., 2013) explored several network centrality measures as potential predictors of attrition. While results on betweenness centrality were not reported, results revealed that a similar network attribute – authority – was inversely correlated with the likelihood of dropout. While there is still no conclusive evidence of betweenness centrality's appropriateness in predicting MOOC outcomes, the metric itself may serve as a foundation for new simulations that reveal the role and "connectedness" of each learner that participates in forum discussions (as we explore further in Chapter 3).

Community Structure and Modularity

Identifying and evaluating community structures in networks has been an active area of research. As noted in (Newman and Girvan, 2004), a network community structure can be defined as a group of nodes within which there is a high density of edges, but a relatively low density of edges between nodes that are in different groups. Early community detection methods drew upon graph partitioning methods from computer science but were severely limited in that they required the number of implicit groups to be specified a priori. Agglomerative clustering methods were also used, but often tended to give more weight to densely connected nodes, thereby

neglecting those on the "periphery". Indeed, (Newman and Girvan, 2004) proposed a new class of divisive clustering methods, which removed edges based on a range of edge-related metrics that were iteratively recomputed (e.g. betweenness centrality values) until a pre-specificed stopping criteria was met.

Even with innovations in network community detection methodologies (e.g., Blondel, 2008), each algorithm's performance reveals little about the strength or robustness of the data's inherent community structure. Perhaps one of the most valuable contributions of Newman and Girvan's seminal work was the definition of a "fitness score" – which they termed *modularity* – to indicate the strength of community structure in a particular network. For a given network with K communities, its modularity score, Q, is defined as $Q = \sum_i (e_{ii} - a_i^2)$, where the (i, j)th entry of the $K \times K$ matrix $\mathbf{e}$ denotes the fraction of all network edges that connect communities i and j (diagonal entries indicate the number of edges that connect the nodes contained in a particular network), and $a_i = \sum_j e_{ij}$, i.e., the fraction of edges that connect to nodes in community i. With this definition, we can see that the modularity score Q is driven to 0 – indicating no inherent community structure – when the fraction of edges within a community is approximately the same as the fraction of edges that are connecting to nodes in that particular community (which is approximately equivalent to what one would expect from a network whose edges have been randomly rewired). When Q is close to 1 – indicating very strong inherent community structure – we see that the largest fraction of edges within the network lie within, instead of between, communities (Newman and Girvan, 2004).

Beyond providing a fitness score for the strength of a network's inherent community structure, modularity can also reveal several other latent network attributes – for example, the network's potential to facilitate information diffusion (e.g., Nematzadeh et al., 2014). To our knowledge, other than a brief discussion of the important insights computing modularity and information diffusion simulations in

MOOC environments could reveal about the nature of "social learning" in these settings (e.g. Sinha, 2014), these network attributes have not been explored empirically in the MOOC context. Therefore, there is an important opportunity at hand to leverage these methods and the broader network analysis toolbox for insights into the latent dynamics and structural properties that characterise communication in MOOCs.

2.3 Latent Feature Models

As the preceding section revealed, the formulation of a particular network topology is the basis upon which subsequent analysis is performed, ultimately informing the macro-trends that are inferred for any given application domain. There are, of course, several properties about learner communications in MOOCs that can be explored outside of a network analysis framework. In the past, latent feature analysis has been used for insights into the micro-trends characterising behaviours in educational settings. In this section, we explore several latent feature models from machine learning. We begin by investigating Non-negative Matrix Factorization (NMF) as a latent feature model, including its relationship with the popular k-means clustering scheme used in numerous previous clustering studies in education research. We then described how NMF can be formally situated in a principled Bayesian generative framework. We broaden our discussion of NMF to explore other parametric and nonparametric latent feature models. Our exploration of these latent feature models aims to provide a foundation upon which we can conduct more granular analysis of the forces that underlie communication and interaction in MOOCs.

2.3.1 Non-negative Matrix Factorization

Non-negative Matrix Factorization represents an N $\times$ D data matrix X, i.e. X $\approx$ WH, where W and H are N $\times$ K and K $\times$ D matrices, respectively (Lee and Seung, 1999). Unlike other latent feature / dimensionality reduction techniques like PCA and Vector Quanitization, the factor matrices W and H are constrained to have positive elements, leading to an additive reconstruction of X by these parts-based multiplicative factors. Indeed, W can be interpreted as indicating "how much" column (or cluster) k helps to explain a particular row n (or item) to be clustered, whereas H indicates the extent to which each cluster k is characterized by each data dimension, d. NMF has been used widely across application domains, ranging from face recognition to document topic modelling (Lee and Seung, 1999).

As (Ding at al., 2005) shows, NMF is actually equivalent to kernel k-means clustering (Dhillon et al., 2004), a generalization of the popular k-means algorithm which enables the clustering of non-linearly separable data. NMF essentially relaxes the orthogonality constraint of the solution to kernel k-means, which is represented by the N x K matrix H where $h_{nk} = 1$ implies item n belongs to only to cluster k (with 0's for all other values of k). This orthogonality implies a hard-partitioning of items into clusters. Under an NMF formulation, we assume our data X = HH$^\top$, where H is near-orthogonal. This near-orthogonality yields a soft-partioning of items into clusters, but still yields a cluster for each item to which it "most" belongs (Ding et al., 2005).

In addition to the connection to the popular k-means scheme, it is valuable to note that the original exposition of NMF by (Lee and Seung, 1999) used an objective function that is equivalent to a factorized Poisson likelihood, i.e. where the rows and columns of the data matrix are assumed to be independent (this point was also mentioned in Gopalan et al., 2013). Under this formulation of NMF, inference for the factor matrices W and H was performed under a maximum likelihood setting,

with no incorporation of prior information about either latent matrix.

Given the relationship between NMF and k-means, which has been widely used in computational education research to date, and NMF's original formulation as a basic probabilistic model, it serves as an excellent bridge between past and future research in education, as well as non-Bayesian and Bayesian models for latent feature analysis. Exploring Bayesian latent feature analysis techniques in the context of MOOCs offers hope for a more rigorous and robust understanding of how probabilistic models can be harnessed to represent and interpret noisy data from large-scale online learning environments.

2.3.2 Bayesian Latent Feature Models

We can begin by placing NMF in a Bayesian context, discussing different models that have emerged along this strand. Given the model-selection challenge characteristic of parametric latent feature models like some implementations of Bayesian NMF, we explore nonparametric variants that have emerged in recent years, including their advantages and shortcomings.

Bayesian Non-negative Matrix Factorization

A Bayesian extension to NMF (BNMF) was proposed by (Schmidt et al., 2008), which placed exponential priors on the factor matrices and leveraged a Gaussian data likelihood model, conducting inference via Gibbs sampling (Geman and Geman, 1984). Since then, there have been many model/inference schemes proposed for BNMF. (Cemgil, 2009) presented a Poisson observation model with gamma priors for conjugacy, using variational methods for inference (Jordan et al., 1999) – which have been chosen for their tractability over MCMC methods in other latent feature models (e.g., Airoldi et al., 2008). (Gopalan et al., 2013) also presented a variational inference scheme for Bayesian Poisson Factorization, introducing auxiliary

latent variables to better account for the inherent sparsity in adjacency matrices representing social or other interaction networks (e.g. those forming the basis of recommender systems).

In many of these cases, model-selection (Guyon et al., 2010) on the number of latent features has been used to determine the most likely number of latent features in the data. (Tan and Fèvotte, 2009) proposed a BNMF model with a Poisson likelihood that leveraged automatic relevance determination priors (MacKay, 1995) under a maximum a posteriori (MAP) inference scheme. These priors, which are indexed by a particular cluster k and are the inverse precision parameters for the half-normal distributions governing the entries of the factor matrices, tie together the columns of W and rows of H and are driven to large values (and low precision levels) when the corresponding cluster k is irrelevant in explaining the observed data (Tan and Fèvotte, 2009). Given its computational efficiency, a number of works across application domains have used this scheme for community detection in social networks, including (Psorakis et al., 2011a) and (Gillani et al., 2014).

The prevalance of count data in MOOC settings (such as the number of forum posts or views by each learner; the number of lecture downloads; etc.) makes Poisson BNMF models of particular interest for the application at hand. The Poisson likelihood, as a member of the exponential family and a maximum-entropy distribution, makes minimal assumptions about the data while simultaneously enabling properties like conjugacy to be exploited for computational ease. Below, we explore Bayesian latent feature models not necessarily formulated as standard matrix factorization problems.

Other Parametric Latent Feature models

Clustering is a well-studied problem in machine learning. A large number of probabilistic generative clustering schemes have been proposed and applied across do-

mains, including the social and life sciences (e.g. Psorakis et al., 2011b, Holmes et al., 2012). The Gaussian Mixture Model (GMM) is a popular probabilistic generative clustering framework and assumes that each observation x_i in some dataset X is drawn from a latent Gaussian distribution (Stauffer and Grimson, 1999). In other instances, tools like Naive Bayes has been used to infer hidden classes in text corpora (Baker and McCallum, 1998). Despite their widespread use and applicability, these simple approaches have made strong assumptions about the data, such as the Gaussianity of the latent structures or the independence of a particular observation's constituent features, that are not always appropriate (Xu et al., 2003).

In pursuit of a more flexible generative latent feature model, (Blei et al., 2003) proposed Latent Dirichlet Allocation (LDA). LDA is a hierarchical model that has been used extensively in document topic modelling. It represents a corpus as a set of documents, each of which in turn is comprised of a set of words. Documents can be characterised by multiple latent features (e.g. topics). The generative model first samples the distribution over possible topics from a Dirichlet. It then samples a topic from a Multinomial parameterized by the Dirichlet sample, and subsequently samples a word from a Multinomial that is conditioned on the selected topic and some prior distribution over words (Blei et al., 2003). The original LDA paper proposed a variational inference scheme, although extensions since then have introduced both uncollapsed and collapsed (e.g., Xiao and Stibor, 2010) Gibbs samplers.

The document metaphor, of course, can be generalized to many different application settings, as a number of authors have done (e.g. Perina et al., 2010). LDA's flexibility arises in large part from its ability to explicitly represent a number of different objects (e.g. documents, words, and topics) and allow for a soft-partitioning of documents across constituent topics. Some authors have extended LDA to alternative likelihood models, such as the Gamma Poisson topic model (Canny, 2004), which can be interpreted as a maximum likelihood instance of Poisson matrix fac-

torizaton (Gopalan et al., 2013). Indeed, (Gopalan et al., 2013) illustrated that LDA and Bayesian Poisson Matrix Factorization are equivalent, subject to certain conditioning and scaling requirements.

In social networks, however, interactions between people are often complex, and can be the result of many different underlying forces. To model these dynamics, the Mixed Membership Stochastic Blockmodel (MMSB - Airoldi et al., 2008) was proposed, lifting the assumption of previous mixed-membership generative models like LDA that the data are conditionally independent given their individual membership vectors over latent features. The MMSB generative model first draws a K-dimensional mixed membership vector for each node n in the network (e.g., the "intrinsic" membership vector of node n). For each node pair, it then samples membership indicator vectors for both the initiator n_i and receiver n_r nodes in a particular interaction. While the indicator vectors are parameterized by the intrinsic membership vector of each initiating/receiving node, they help model the context-dependent membership of each. Finally, a value for the interaction between n_i and n_r is sampled based on the membership indicator vectors, often a Bernoulli random variable to indicate the presence or absence of a particular interaction.

Given its sensitivity to the relationships in between particular nodes, the MMSB is a latent feature model for non-exchangeable relational data. When compared to matrix factorization methods, MMSB is advantageous in that it explicitly models how two users (or users and items in a classic recommendation-based matrix factorization setting) interact via their context-specific membership vectors; however, it fails to model the direct interactions between these objects, which has proven to be too restrictive of an assumption in some settings (Mackey et al., 2010). Given the nature of our data - where we do not yet consider complex social interactions and instead only similiarities between learners based on their participation in discussions - we plan to continue to leverage matrix factorization, although other ongoing

MOOC research is developing extensions to MMSB to model learner resilience in communication settings (Rose, 2013).

Bayesian Nonparametric Latent Feature Models

In the latent feature models described above, a key challenge has been one of model selection - namely, determining a priori the most likely number of clusters that exist in the data and using this as a parameter for each model. Model selection, however, can be computationally burdensome and prohibitive, particularly in cases where inference is non-trivial and must be conducted a number of times to select the "best" model to represent the data (Gopalan et al., 2014). Fortunately, we can turn to Bayesian Nonparametrics (BNP), which specifies a collection of methodologies that empower the data to determine the complexity of the model (Gershman and Blei, 2011).

Latent-feature extraction, like in parametric settings, has been well studied in the BNP literature (Gershman and Blei, 2011). One of the simplest nonparametric latent feature models is the Infinite Relational Model (IRM - Kemp et al., 2006). The IRM models a possibly unbounded number of latent features with a Chinese Restaurant Process prior, which mathematically favours a small number of clusters (Aldous, 1985). The IRM can simultaneously cluster both objects and their features – an innovation over previous nonparametric clustering approaches like the Infinite Mixture Model (Rasmussen, 2000). However, it is inflexible in that it induces a hard-partition over the data, allowing each object or feature to be assigned to only one cluster.

To achieve soft-partitioning of objects over their latent features, (Griffiths and Gharamani, 2005) introduced the Indian Buffet Process (IBP). The IBP is perhaps the most prominent infinite latent feature model, specifying a prior distribution over N $\times$ K binary matrices that indicate which objects $n \in$ N are characterized by

which latent features $k \in K$. In this way, the IBP is similar to LDA, MMSB, and NMF – but differs in that 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. Still, conjugacy in linear-Gaussian models has enabled inference over larger-scale datasets via Gibbs sampling. Additionally, different constructions of the IBP - such as the stick-breaking construction of (Teh et al., 2007) - have revealed theoretical connections with other generative processes that have enabled efficient inference procedures, like slice sampling, to be adapted and used (Teh et al., 2007). The development of variational inference schemes and efficient sampling procedures (Doshi-Velez, 2009) over the past 15 years have further made inference tractable for datasets with thousands of objects, exploiting conjugacy in the (often linear-Gaussian) likelihood models. More recent explorations have revealed how inference in the IBP can be scaled via submodular optimization (Reed and Ghahramani, 2013), although again, by exploiting the structure of the linear-Gaussian likelihood model. Interestingly, benchmark tests conducted in (Reed and Ghahramani, 2013) revealed the high relative accuracy and scalability of BNMF compared to a number of other IBP-based models.

Nonparametric variants of Bayesian Poisson NMF have also been explored. (Titsias, 2007) introduced the Infinite Gamma-Poisson feature model as an extension to previous IBP models by representing the N $\times$ K latent feature matrix Z as a draw from a Poisson, indicating the number of times object $n \in N$ contained feature $k \in K$. More recently, (Gupta et al., 2012) proposed a nonparametric model for count data – the Linear Poisson Gamma Model (LPGM) – which also used an IBP prior over latent features and explicitly learned the factorization error as a part of the model. Both of these methods used Gibbs sampling procedures with poor scala-

bility in the wake of thousands of data points. To address these scalability issues, a recent paper by (Gopalan et al., 2014) proposed a nonparametric Poisson factorization method for recommender systems based on a stick breaking construction of the Gamma Process, extending their previous parametric efforts to continue to use latent auxiliary variables to make inference in sparse datasets more efficient (Gopalan et al., 2013). Beyond Poisson factorization, recent efforts have been made to leverage online stochastic variational inference to discover latent structure in social networks with tens of thousands of nodes (Kim et al., 2013).

While many BNP models have proven impractical in real-world settings because they do not scale well, the emerging state of the art in nonparametric latent feature extraction offers new directions for exploration and application. Comparing the representational power and effiicency of BNP models with their parametric counterparts on large-scale MOOC data will be important to determine how to better understand and support these nascent digital learning environments.

2.4 Summary

The rich set of theoretical and methodological innovations in education, network analysis, and machine learning combine to lay a strong foundation for the analytical frameworks leveraged in this work. From the domain of network analysis, we have different perspectives on constructing a robust, representative network topology as a foundational step to performing subsequent analysis. From machine learning, we have the example of several parametric and non-parametric clustering/community detection models, many of which leverage a Bayesian framework to model the inherent noisiness and complexity of real-world datasets. As the empirical investigations and results of the subsequent chapters reveal, this hybrid toolbox is fundamental to capturing both macro and micro-level insights into the hidden patterns and struc-

tures of communication in MOOCs.

Chapter 3

Structural properties of learning in a crowd

3.1 Introduction

Over the last few years, millions of self-selected learners have enrolled in courses on large-scale learning platforms, typically co-participating with thousands of peers in courses of their choice (Littlejohn, 2013). Within the field of Computer Supported Collaborative Learning (CSCL), the structural properties of online group learning have been studied in detail (Stahl, 2006; Vaquero and Cebrian, 2013). However, these studies have been carried out in the contexts of more traditionally organized, smaller-scale online classrooms. The novelty of learning at scale has inspired a number of research studies that explore how participants interact with course content. Some studies have revealed that these courses are taken primarily by those already with college degrees (Ezekiel, 2013). Others have traced the behavior of MOOC participants through a course's lifespan by leveraging granular clickstream data (Breslow et al., 2013; Kizilcec et al., 2013). Still others have started to cluster learners based on their patterns of engagement in order to predict dropouts (Yang

et al., 2013).

Despite a growing body of research (Brinton et al., 2013; Anderson et al., 2014), many questions relating to the characteristics of group interactions and dialogue in these courses have largely been ignored. Contemporary socio-cultural learning theory emphasizes the role of group interaction for cognition, highlighting the need for understanding the degree to which MOOCs in practice allow for deep and meaningful learning through the facilitation of significant interactions and the spread of information between participants as they seek to acquire and generate new knowledge (Vygotsky, 1978; Siemens, 2005).

In this chapter, we aim to expand this area of research by examining group level interactions in MOOC forums. Regardless of the type of MOOC, learners are typically offered the opportunity to participate and collaborate with one another in online discussion forums to enhance their educational experience. The forum participants have a diverse range of backgrounds and motivations for taking the courses. There are few rules, and are best defined as non-formal learning spaces (Colley et al., 2002), where learners are free to pick and choose how and if they interact. The overall governance structure is relatively weak, set primarily by the educators' questions / assignments for the forums, the technical design of the forums (e.g. the division of forums into specific topics) and the roles participants themselves take on during the course (typically 8 weeks).

While theoretical perspectives and emphases differ in studies of online learning, it is recognised that understanding the learning process in online forums requires consideration of interactions at the individual and group level (Stahl, 2006; Goodman et al., 2011; Trognon and Batt, 2012). The interactions at the group level within these forums can be viewed as a kind of scaffold through which learning can occur (Tay, 2012), and therefore, is of significant practical concern when considering the future design and development of courses.

A key challenge in addressing group-level interaction in MOOCs is of methodological character: we need to determine what constitutes a significant interaction in these learning environments of unprecedented scope. In large-scale MOOC forums, with socio-culturally diverse learners, this is a non-trivial problem. Many of the tens of thousands of interactions in the forum may have little relevance to enhancing the learning process; thus, the first question we address asks: how can we determine the underlying social networks that depict significant interactions?

Once the significant interactions have been established, we ask two questions core to understanding group interactions for learning in MOOCs: What is the vulnerability of the communication networks in the forums, and how does information flow through the forums? The large scale and open nature of the forums means that the group has a vast set of resources available in the form of people with varied experiences and expertise. However, this kind of group membership, together with the non-formal and short-term nature of these courses, means that relatively weak interpersonal relationships are likely (Tay, 2012; Butler, 2001).

3.2 Methodology

We analysed data from two successive instances of a business MOOC (FOBS-1 and FOBS-2), offered on the Coursera platform in Spring and Autumn of 2013. Nearly 90,000 students registered for FOBS-1 and over 77,000 for FOBS-2. The courses lasted for six weeks each, during which 4,500 FOBS-1 and 3,300 FOBS-2 learners contributed over 15,000 and 14,000 posts or comments in the discussion forums, respectively. More than 15,000 and 11,500 learners viewed at least one discussion thread in both instances, contributing to 181,911 and 139,858 total discussion thread views in both instances of the course.

In both instances, the discussion forum was segmented into sub-forums: the

Final Project sub-forum facilitated questions, debates, and team formation for the final strategic analysis assignment; the *Cases* sub-forum facilitated weekly discussion about a particular company (e.g., Google, Apple, Disney, etc.) and its business strategy. This sub-forum was divided into additional sub-forums for each week's selected company, and each company's sub-forum was in turn divided into further sub-forums that asked a specific question about the company for that week (for example, questions about the companies competitive advantage, market position, etc.); students posted for both logistical and content clarifications in the *Questions for Professor* sub-forum; *Technical Feedback* and *Course Material Feedback* sub-forums provided channels for voicing gratitude and felicitations at the end of the course, and suggestions for future improvements; *Readings* and *Lectures* harboured discussion around core course content; and the *Study Groups* sub-forum enabled learners to form cohorts with their peers to experience the course together often interacting with others from similar timezones. Learners could create discussion threads within these sub-forums, which contained new posts or comments on existing posts.

In FOBS-1, students were not evaluated on their performance in the forums and over 4,500 individuals participated to contribute over 15,000 posts; in FOBS-2, 8% of students' final scores was derived from their forum participation as a function of the total number of upvotes they received on their posts or comments. We analysed the dynamics of sub-forum across both courses separately for a more granular understanding of learner behaviours. This approach was justified by low overlap between learners across sub-forums (less than 10% in all instances, and in most cases, below 3%). These numbers are revealing in their own right: most students participated infrequently in only a few sub-forums.

We used social network analysis to capture both broad trends in communication and the roles of individuals in facilitating discussions. Network nodes represented

learners that created at least one post or comment in a discussion thread; an edge connected two learners if they co-participated in at least one discussion thread. Formally, the communication network is represented as a graph $G = (V, E)$, where $V = \{s_i\}_{i=1}^{N}$ for the N participants s_i, and $E = \{e_{ij}\}_{i,j=1;i\neq j}$ where $e_{ij} = m$ implies students i and j co-participated m times in forum discussion threads.

The networks presented here only depict explicit participation (e.g. posting or commenting), although it is important to note that the prominence of lurking suggests the need for more robust network models that depict both viewing and posting trends. Indeed, an alternative network formulation may capture the viewing frequencies of each learner as a node-level attribute (e.g., setting the node size to be proportional to the number of thread views for each corresponding learner – which would in turn enable analysis of assortativity and other similar metrics). More granular data on the specific posts viewed by each learner could also be considered to create a directed network topology, where an edge from i to j would indicate that learner i posted in a thread after learner j, and the edge weight, $e_{ij} = m$, would reveal the number of times learner i viewed all posts made by learner j.

3.3 Results

Both posting and viewing in the course's sub-forums tended to occur in bursts, with most activity happening at the beginning of the course, and in some cases, triggered by recurring or final course milestones (see Appendix). The number of views and posts per discussion thread had a fat-tailed distribution (see Appendix) and the threads of different sub-forums tended to have different view lifespans; the time period that 90% of views occurred within (see Appendix). Given our interest in exploring explicit forum participation, we focus the remainder of this work on posting behaviour. Table 3.1 provides basic summary statistics for the different sub-forums

	FOBS-1			**FOBS-2**		
Sub-forum	N_l	N_p	N_p/N_l	N_l	N_p	N_p/N_l
Technical Feedback	250	519	2.08(2.8)	109	233	2.14(3.3)
Course Material Feedback	281	487	1.73(1.9)	289	537	1,86(2.5)
Study Groups	1,387	2,730	1.97(3.4)	2,022	4,573	2.26(3.9)
Readings	1,120	2,334	2.08(3.0)	137	209	1.53(1.0)
Lectures	638	1,262	1.98(2.8)	470	1,175	2.50(3.0)
Questions for Professor	305	610	2.00(3.1)	370	785	2.12(3.2)
Final Project	1,078	2,612	2.42(4.4)	630	2,224	3.53(6.6)
Cases	1,222	4,669	3.82(5.6)	987	4,397	4.45(7.8)

Table 3.1: Summary statistics for forum participation in FOBS-1 and FOBS-2. N_l, N_p, and N_p / N_l denote the number of learners, number of posts, and average posts per learner (along with standard deviations in brackets) in each sub-forum, respectively.

in the course (see the Methodology for an explanation of the sub-forums).

3.3.1 Significant Interaction Networks

We modelled communication between learners as a social network, where nodes represent learners that posted explicitly in the discussion forums and an edge between two learners indicates that they co-participated in at least one discussion thread (see Appendix for formal definition). However, not all links generated this way are equally important, and two learners' co-participation in a thread is not necessarily indicative of a meaningful social exchange. Drawing upon work from ecology and engineering (Psorakis et al., 2011b), as well as recent work in mobile communication networks (Li et al., 2014), we assumed that the observed communication network in each sub-forum was a noise-corrupted version of the true network i.e., one that depicts meaningful communication between students. Our task, then, was to derive this true "significant" network by filtering out links between learners thought to have been generated by random, or chance, encounters. Previous research has used interaction time windows to determine which links to keep and which to remove

in a network (e.g., Krings et al., 2012). Because we did not have reliable a priori knowledge about what a reasonable time window would be, we instead turned to a significant network extraction model used to infer social networks in ecological settings (Psorakis et al., 2011b) in order to inform our efforts.

The derivation of a significant social network proceeded as follows. For each sub-forum f, we first constructed a bipartite graph to represent the *learner-to-thread* adjacency matrix $B_f^{N_f \times T_f}$, where N_f represents the number of learners that explicitly posted and T_f the number of threads, both in sub-forum f. Each entry $b^{n,t} \in B_f^{N_f}$ was an integer greater than or equal to 0, denoting the number of times learner n participated in thread t. Next, we computed a standard weighted one-mode projection of B_f to recover the *learner-to-learner* network, $L_f^{N_f \times N_f}$ 3. Each entry $l^{i,j} \in L_f^{N_f \times N_f}$ was also an integer greater than or equal to 0, depicting the number of threads that learners i and j co-participated in.

With a learner to learner network L_f, we now sought to identify which edges in L_f (i.e., $l^{i,j} \in L_f$ s.t. $l^{i,j} \neq 0$) depicted a *significant* interaction between two learners. Our goal was to generate a family of M sample networks against which we could test the significance of each $l_{i,j} \in L_f$. We started by noting that the observed learner-to-thread network B_f depicted each learner n's participation in a particular thread t. We modelled this participation – i.e., each row n of B_f - as a draw from Multinomial(k_n, p_n), where $k_n = \sum_{t \in T_f} b_{n,t}$ and $p_n = (p_{n,1}, ..., p_{n,T_f})$ for $p_{n,t} = b_{n,t}/k_n$. It is important to note that $p = \{p_n\}_{i=1}^{N}$ represented the observed social relationships between learners as indicated by the likelihood of each student's participation in a particular thread.

If we wished to test the significance of the observed edges – i.e., the observed social interactions – we needed to determine a mechanism for generating possible social networks that do not possess the same social patterns as the observed one. In order to explore alternative social structures, we first defined a shuffling function

σ such that each row of the sth sample learner-to-thread network B_f^s is drawn from $Multinomial(k_n, \sigma(p_n))$ with k_n and p_n defined as above. We defined σ such that it preserved learner n's proportional participation in different threads (e.g., the entropy of each p_n), but accounted for the possibility of participation in alternate threads. As an extension to (Psorakis et al., 2011b), we constrained σ to only shuffle each entry of p_n with a location (e.g., thread) t that has popularity greater than or equal to the least popular thread that learner n participated in, where thread popularity was defined as the number of posts it contains. This constraint was meant to reflect which threads learners could have possibly participated in, since in many cases, discussion threads only had a single or very small number of posts, and therefore, it was unrealistic to assume that a learner who participated primarily in popular threads may have also participated in isolated ones. Without this constraint, the shuffling allocated participation probabilities to a larger set of threads, increasing the likelihood particularly for those individuals with low participation volumes but high proclivity to post in popular threads that these one-off interactions were deemed significant. Additionally, this constraint was informed by real-world discussions with participants from FOBS 1, some of who indicated that the popularity of a particular discussion thread often influenced their decisions to view or post. Therefore, the constrained shuffling more accurately captured learner behaviour and detected one-off participation in high-activity threads as insignificant (thereby, pruning more edges) when compared to its under-constrained counterpart.

With a sampling procedure in place, we generated each B_f^s and computed its one-mode projection to arrive at the set of sampled learner-to-learner networks, i.e., $G = \{G_f^s\}_{s=1}^M$. We then compared each entry $l_{i,j} \in L_f$ to $\frac{1}{M}\sum_s g_{i,j}$ for $g_{i,j} \in G_f^s$, computing the z-score and labelling as significant if the right-tailed p-value was less than 0.001 (we assumed a relatively small p-value threshold due to the sparsity of participation in the discussion threads). Our derived significant network was the

collection of $l_{i,j} \in L_f$ labeled as significant by this procedure.

As Figure 3.3.1 illustrates, filtering out insignificant edges impacted various network attributes, such as modularity (Blondel et al., 2008).

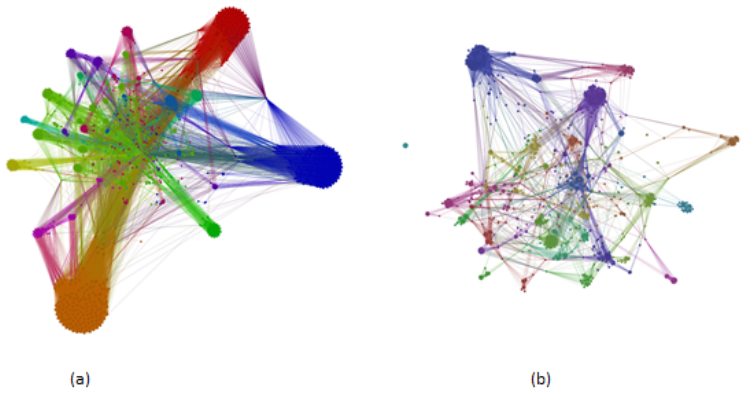


Figure 3.1: The observed (a) and derived (b) communication networks for the *Study Groups* sub-forum. Here, we can see the impact of link filtration on network properties such as modularity score, which equals 0.62 and 0.80 for a and b, respectively. Colours correspond to the detected communities.

Table 3.2 shows the proportion of edges pruned for the significant network derivation of each top-level sub-forum in FOBS-1 and FOBS-2. In both instances, the feedback forums had the greatest decrease in density, which was to be expected: most participation in these sub-forums was one-off, posting either specific technical questions/comments or a final thank you to the course staff at the end of the class. On the other hand, the Cases sub-forum lost the smallest proportion of edges at 44% and 39% in both instances, respectively. This is likely due to the type of discussion encouraged in this sub-forum: the weekly case discussions asked students to provide their thoughts and opinions on a series of open-ended questions. The sub-forums accommodated these questions and aimed to incite, and guide, group discussion. In many cases, students would read the analyses posted by their peers and comment with additional insights or critiques, leading to greater engagement and knowledge construction.

	FOBS-1		**FOBS-2**	
Sub-forum	$\|E_o\|$ (N)	$\|E_s\|$ (Δ)	$\|E_o\|$ (N)	$\|E_s\|$ (Δ)
Tech. Feedback	3,087 (231)	339 (89%)	767 (105)	83 (89%)
C. Mat. Feedback	2,752 (252)	729 (74%)	7,824 (279)	1,253 (84%)
Study Groups	41,819 (1,359)	11,609 (72%)	130,821(1,993)	38,548 (71%)
Readings	35,728(1,108)	11,259(68%)	1,873(137)	125(93%)
Lectures	12,644(617)	3,988(68%)	6,839(463)	3,472(49%)
Quest. for Prof.	2,758(284)	896(68%)	3,258(345)	1,358(58%)
Final Project	23,244(1,019)	12,557(46%)	16,116(611)	7,760(52%)
Cases	102,171(1,114)	57,490(44%)	45,257(925)	27,586(39%)

Table 3.2: Observed and derived communication networks for the different sub-forums in FOBS-1 and FOBS-2. Here, $|E_o|$ indicates the number of edges originally in the denoted sub-forum's network; $|E_s|$ indicates the number of edges in the derived significant network; N indicates the number of learners connected to at least one other learner in the network; and Δ is the percentage decline in the number of edges after the link filtration.

Interestingly, more than 2 out of 3 connections in the *Study Groups* sub-forum were considered insignificant in both instances of the course. Informal content analysis of this sub-forum revealed that most learners posted to introduce themselves early on in the course, particularly to their peers in similar geographic regions/time zones, then opting to move their discussions to other online platforms (e.g. Facebook). Largely, the *Study Groups* sub-forum served more as a meeting point for learners instead of an environment for sustained community-building and engagement.

Before proceeding, we reiterate that our definition of a significant link (e.g. a meaningful social exchange) is a function of the frequency of thread co-participation of any two learners, as well as the popularity of the threads in which they take part. Indeed, frequency of forum participation has been investigated by others as a means of determining engagement and evaluating performance in educational settings (Baldwin et al., 1997). For illustrative purposes: if two learners tend to only participate with one another a small amount in a very popular thread, the filtration

algorithm is likely to label this exchange as insignificant, because it has a high probability of also appearing in the sample networks (thereby, suggesting that it may have resulted from chance alone). This is motivated by the MOOC context, where highly popular threads – for example, threads where learners introduce themselves – tend to harbour increasingly noisy exchanges without observable follow-up discussions or deeper social engagement. Conversely, if learners tend to engage in smaller-group discussions with one another through repeated, iterative dialogue, this potentially exhibits a more meaningful exchange with a greater likelihood of facilitating knowledge construction among participants. This sort of exchange is likely to be labelled as "significant" by the algorithm, because it is unlikely to repeatedly manifest in the sample networks: it is a unique and robust exchange.

Of course, in a complex social setting such as an online course, the significance of communication is not entirely dependent on the frequency of co-participation or the popularity of a thread, but also on the nature of the content exchanged, and perhaps more importantly, the unique, individual motivations of the learners as they relate to the discussion content. Therefore, a key opportunity for future work is to build upon – or perhaps even re-envision – the notion of significance presented here to arrive at more robust and suitable filtration schemes.

3.3.2 Communication vulnerability

The vulnerability of networks has been studied across disciplines (Albert et al., 2000). For example, power systems engineers often ask which critical set of network components must be damaged in a functioning circuit in order to cut off the supply of electricity to the remaining nodes (Simonsen et al., 2008). We asked an analogous question from an educational perspective: which critical set of learners is responsible for potential information flow in a communication network - and what would happen to online discussions if the learners comprising this set were removed? We defined

vulnerability in the education context to be the proportion of nodes that must be disconnected from the network in order to degrade the relative size of the largest connected component to the total number of nodes. To our knowledge, the vulnerability of communication networks in educational settings has not been explored in previous research.

The vulnerability of MOOC discussion networks indicates how integrated and inclusive communication is. Discussion forums with fleeting participation tend to have a small proportion of very vocal participants comprise this set: removing these learners from the online discussions would rapidly eliminate the potential of discussion and information flow between the other participants. Conversely, forums that encourage repeated engagement and in-depth discussion among participants have a proportionally larger critical set, and discussion is distributed across a wide range of learners. By analysing vulnerability in different sub-forums, we sought to understand how group communication dynamics differed according to the topics being discussed.

To chart the vulnerability of each sub-forum, we executed the following algorithm: for each extracted network of significant interactions, iteratively find the node with the highest betweenness centrality (Newman, 2010), disconnect it from the network, and compute the resultant proportion of nodes in the networks largest connected component to the total number of nodes. We used betweenness centrality as our removal metric as its definition, particularly for undirected graphs, indicated each node's potentiality as a conduit for the spread of information or ideas. We compared results from this removal strategy to one where nodes are removed at random, which has served as a baseline evaluative mechanism for similar analyses in other application domains (Albert et al., 2000).

Figure 3.3.2 depicts the results for all of the sub-forums in FOBS-1 (with similar results for FOBS-2 illustrated in the Appendix). Figure 3.3.2 zooms into the *Cases*

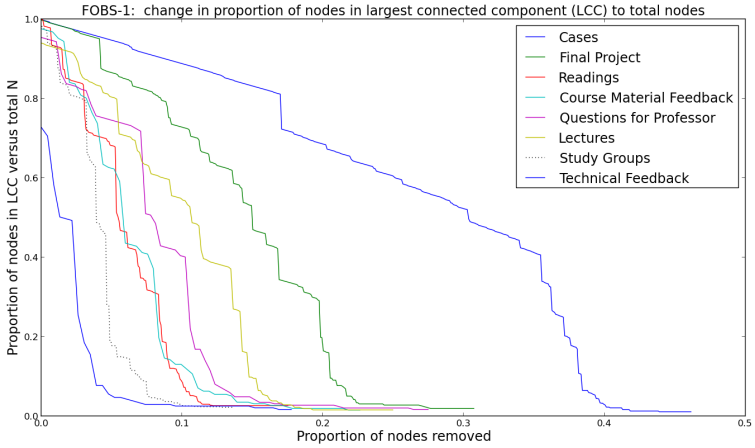


Figure 3.2: Network vulnerability for the different sub-forums in FOBS-1. Here, LCC refers to the largest connected component in each sub-forums communication network.

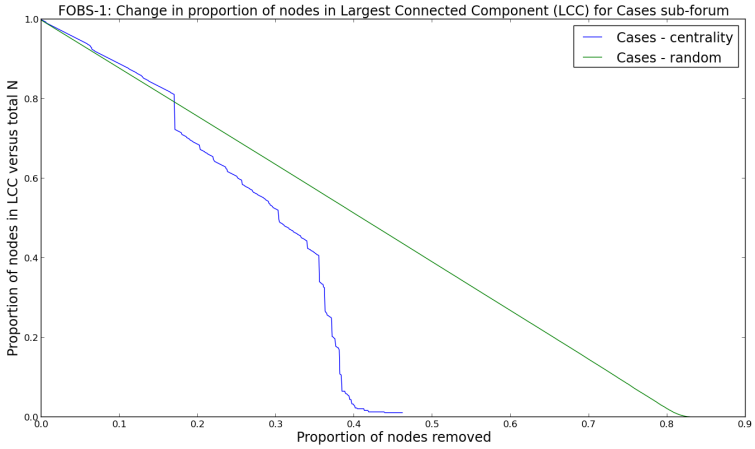


Figure 3.3: Vulnerability versus random removal for Cases sub-forum in FOBS-1. Here, LCC refers to the largest connected component in this sub-forums communication network.

sub-forum from FOBS-1 to depict the degradation resulting from iteratively removing the node with the highest betweenness centrality to a random removal strategy. From Figure 3.3.2, it is clear that different sub-forums had different vulnerability thresholds. For example, in both instances of the course, the *Cases* and *Final Project* sub-forum were the least vulnerable, as determined by the relatively higher proportion of nodes requiring removal before the relative size of the largest connect component was driven close to zero. This likely resulted from the high levels of iterative dialogue and knowledge construction characteristic of both sub-forums. Indeed, learners tended to use these forums to share ideas and insights as they related to the weekly case questions, their final strategic analyses, or questions about course outcomes and the peer review process (Gillani et al., 2014). These trends, which could be interpreted as conducive to promoting learners participation in multiple discussions with many other learners, were largely absent in other sub-forums. For example, in the *Study Groups* sub-forum - one of the most vulnerable - the majority of individuals posted once to introduce themselves to a particular group of students and then proceeded to move their discussions to other platforms, or perhaps, cease engagement altogether. Indeed, it is interesting to note differences in sub-forum vulnerability across both FOBS-1 and FOBS-2. The proportion of nodes required to rapidly degrade both the *Cases* and *Final Project* sub-forums is higher in FOBS-2 than in FOBS-1, suggesting less communication vulnerability. This may have been the result of an additional evaluation criteria in FOBS-2, where 8% of students final scores was computed as a function of the total number of upvotes they received on their posts or comments in the discussion forum.

The different vulnerability thresholds across sub-forums and course iterations suggest that the different topics being discussed, and perhaps, different incentives for participation promoted different levels of inclusiveness and engagement among learners.

3.3.3 Information Diffusion

From analyzing network vulnerability, it is clear that different sub-forums have different critical sets of participants that characterize the inclusiveness of discussions. Still, it is important to explore how information spreads in these networks, as doing so may ultimately reveal how forum participation promotes knowledge construction. Therefore, we ask: *how does information flow through the forums?*.

To investigate this, we simulated an information diffusion model similar to the SI (Susceptible-Infected) model of contagion (Kermack and McKendrick, 1972; Anderson and May, 1992) which has been extensively used in previous work to model social contagion (Onella et al., 2007; Karsai et al., 2011). The simulation first infected one node randomly, which then subsequently infected all neighboring nodes with the probability p; this process was repeated for all infected nodes. The outreach was measured as the ratio of the infected nodes to the total number of nodes over time. This number was averaged over a sample of 400 realizations with different initial seeds. The value of p was chosen to be 0.01; this choice did not impact the generality of the results (the same trends were obtained for different values of p).

Although very simplistic, the SI model is very useful in analyzing the topological and temporal effects on networked communication systems. The SI model does not take into account effects such as decaying interest over time, the influence of peers, and more sophisticated mechanisms of social contagion, but it adequately determines the upper limit of the contagion rate based on the topology and the connectedness of the interaction network. It also enables us to compare different topologies and their efficiency in information spread within a quantitative framework.

As a benchmark, we performed the same diffusion simulation on a randomized network, where each node maintained its degree but had a different set of neighbors than those observed in the significant network (i.e., a configuration model (Catanzaro et al., 2005)). The purpose of shuffling neighbors in the randomized network is

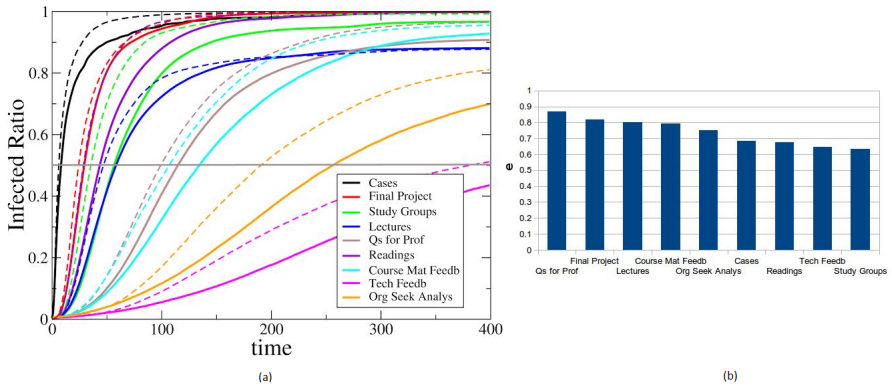


Figure 3.4: (a) shows the percentage of infected nodes vs. simulation time for different networks. The solid lines show the results over the original network and the dashed lines for the degree-preserved shuffled network (configuration model), and (b) shows the value of e (i.e., the diffusion efficiency) for different sub-forums.

to present the diffusion potential of the corresponding sub-forum without inherent modularity i.e., the benchmark provides a baseline of how well information would flow between participants in the forum if the observed community structures did not exist. The time evolution of the percentage of reached nodes both for the original and randomized networks is depicted in Figure 3.4(a). It is evident from this figure that the spread is uniformly faster in the randomized networks, and throughout the process starting from a single infected node until the whole system is infected. To quantify the difference between the randomized and original networks, we computed the time it took for a simulated information packet to come into contact with half of the network's nodes (i.e., participants), T_{half}. It is possible to consider a different threshold for prevalence (e.g., some have used 20% (Karsai et al., 2011)), but as evident from Figure 3.4(a) the choice of this threshold does not change the overall patterns.

We defined an information diffusion efficiency e of a sub-forum as $e = T_{half}$ (random) / T_{half} (original). Values of e that were larger/smaller than 1 indicated that the structure of the discussion network correlated with enhanced/diminished infor-

mation diffusion compared to the randomized benchmark. Figure 3.4(b) illustrates the value of e for the different sub-forums of FOBS-1. Similar trends have been observed for FOBS-2 (depicted in the Appendix).

Across all sub-forums, information spreads more slowly in the original networks compared to the randomized benchmarks. This is likely explained by the existence of local community structures in the derived significant networks. It has recently been shown that a dense cluster of nodes with a large number of connections could impede diffusion processes (Onella et al., 2007). The *Questions for Professors* sub-forum was the most efficient at facilitating information spread; conversely, the highly modular Study Groups sub-forum depicted the least diffusion efficiency (the Methodology section explains the intended function of each of the depicted sub-forums).

To further support this finding, we calculated the modularity score (Newman, 2006) of the original and randomized networks of each sub-forum , using the Louvain community detection method (Blondel et al., 2008) with resolution 1 (Lambiotte et al., 2009). We then computed the normalized modularity score for each sub-forum by taking the ratio of the original networks modularity score to the modularity score of its shuffled counterpart. Table 3.3 presents these modularity metrics for each sub-forum. When compared to the efficiency scores from Figure 3.4(b), it is clear that the most modular sub-forums also tended to have the lowest diffusion efficiency (e.g. *Readings* and *Study Groups*) and vice versa (e.g. *Questions for Professor* and *Course Material Feedback*). Despite these trends, it is important to note the Technical Feedback forum, which had both low normalized modularity and diffusion efficiency (this sub-forum also had the lowest amount of participation and aimed to capture technical platform glitches instead of facilitating discussion about course content). Given the limited number of sub-forums over both MOOC instances, we refrain from reporting a Pearson r-value to quantify the anti-correlation between normalized modularity and diffusion efficiency. Indeed successive iterations of this MOOC will

	FOBS-1			**FOBS-2**		
Sub-forum	M_o	M_s	NM	M_o	M_s	NM
Tech. Feedback	0.78	0.55	1.42	0.86	0.72	1.19
C. Mat. Feedback	0.66	0.36	1.83	0.47	0.26	1.81
Study Groups	0.80	0.19	4.21	0.71	0.12	5.92
Readings	0.70	0.17	4.11	0.43	0.15	2.87
Lectures	0.61	0.22	2.77	0.51	0.22	2.32
Quest. for Prof.	0.65	0.36	1.81	0.62	0.32	1.94
Final Project	0.58	0.15	3.87	0.42	0.15	2.80
Cases	0.27	0.10	2.70	0.34	0.10	3.40

Table 3.3: Modularity scores of the original and derived significant networks. Here, M_o is the modularity of each original network; M_s is the modularity of each derived significant network; and NM indicates each network's normalized modularity (i.e., M_o/M_s).

offer additional opportunities and data to verify if the inverse relationships between modularity and efficiency scores observed in this investigation continue to hold.

Overall, these results reveal an important characteristic of discussion in MOOCs: when it comes to significant communication between learners, there is simply too many discussion topics and too pronounced heterogeneity characterizing participation to realize truly global-scale discussion. Instead, most information exchange, and by extension, any knowledge construction in the discussion forums occurs in small, short-lived groups.

3.4 Discussion

Our research reveals that forum participation is heterogeneously distributed both temporally and according to the different discussion topics. User participation is driven by course milestones (for example, course launch and final project due date). Additionally, group dynamics namely, the significance and vulnerability of communication vary according to what is being discussed and how the forums are used

by course staff to encourage participation. Finally, modularity in MOOC forum networks appears to "trap" information in small learner groups. This finding is important as it highlights structural limitations that may adversely impact the ability of MOOCs to facilitate communication amongst learners that look to learn "in the crowd".

These insights into the communicative dynamics at play motivate a number of important questions about how social learning can be better supported, and facilitated, in MOOCs. Recent work on a subset of this data employed qualitative content analysis - combined with community detection schemes from machine learning - to infer latent learner communities according to the content of their forum posts. Interestingly, for the *Cases* and *Final Project* sub-forums, the inferred communities had statistically significant differences in the geographic and prior educational experiences of constituent learners, as well as their final course performance and overall engagement in the discussion forums (Gillani et al., 2014). Moreover, ongoing semi-structured interviews of FOBS-1 forum participants revealed different motivations for engaging in discussions and different strategies used to manage "content overload" within the forums. These insights - coupled with those presented above - suggests the necessity of large-scale online learning platforms to increasingly employ intelligent machine learning algorithms to support the needs of crowd-based learners. Such systems might, for example, detect different types of discussion and patterns of engagement during the runtime of a course to ultimately help students identify and engage in conversations that promote learning in accordance with individual objectives and learning preferences. Without such interventions - and the technical and pedagogical governance structures for online discussions that may result - the current structural limitations of social learning in MOOCs may prevent the realization of a truly global classroom.

In the next chapter, we explore how looking beyond the structure of discus-

sion forum interactions and directly at communication content can reveal additional insights to help support learners in massive online contexts.

Chapter 4

Content-inspired communication communities

4.1 Introduction

Most early attempts at understanding social learning in MOOCs have largely relied on post-hoc descriptive analysis, in part, because the sheer volume of thousands of discussion forum posts has prevented full-scale content analysis to model the *content* and *context* of communication. In this chapter, we build upon our macro-level insights into forum engagement patterns and use latent feature analysis to further explore communication in MOOCs. In particular, we: 1) introduce a new content-analysed dataset of MOOC forum data; 2) explore and validate Bayesian Non-negative Matrix Factorization as a probablistic generative model for content clusters that emerge in online discussions; 3) explore the sensitivity of the inferred clusters using both qualitative and quantitative indicators; and 4) reflect on the implications of our findings for the future of digital learning.

4.2 Dataset and Latent Features

4.2.1 Content Analysis of Forum Data

Similar to the preceding chapter, we analysed data from a business strategy MOOC offered on the Coursera platform in Spring 2013, but focused on only one instance of this course. 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. As before, 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, classifying 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

Learning	Communication	Affect	Topic	Relevance
Statement of Observation or Opinion	Argumentative	Positive activating	Lectures or videos	High relevance – content corresponds to sub-forum and thread topic
Confirm or agree or support examples by statements given by others	Responsive	Negative activating	Quizzes	Low relevance – content relates to course activities, but wrong forum or thread
Asking or answering questions to clarify details of statements	Informative	Positive de-activating	Readings	No relevance – content seems out of context or is inappropriate
Definition, description or identification of a problem	Elicitative	Negative de-activating	Arrange offline meet-ups	
Discovery and exploration of dissonance or inconsistency among participants	Imperative	Neutral	Arrange virtual meet-ups	
Negotiation of meaning or knowledge co-construction			Final reports or projects	
Testing and modification			Peer review	
Phrasing of agreement and applications of newly constructed meaning			Course outcome, grades, certificates	
No learning			Cases	
			Introductions	
			Other	

Figure 4.1: Content dimensions and labels used to qualitatively analyse discussion forum posts.

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. Figure 4.1 shows the possible content labels for each dimension, where a content label is defined as a category within a particular dimension (a post, then, is characterised by a set of five content labels – one assigned for each dimension).

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.

4.2.2 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 (represented by the rows in Figure 4.1). So, for example, one column may correspond to the content label "no learning" (falling within the Knowledge Construction dimension); another may correspond to the content label titled "readings" (falling within the Topic dimension); etc. 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 $N \times N$ *similarity 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 similarity matrix do *not* necessarily depict communication between them (as opposed to the preceding chapter, where the interaction matrix depicted discussion thread co-participation); instead, they indicate similar discussion content (in some sense, "content co-participation"). Figure 4.2 presents an example representation of this data.

We assume that the pairwise similarities described by X are drawn from a Poisson distribution with rate $\hat{\mathrm{X}} = \mathrm{WH}$, i.e. $x_{ij} \sim \text{Poisson}(\sum_{k=1}^{K} w_{ik}h_{kj})$, where the inner rank K denotes the unknown number of clusters 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

$$\mathrm{X} = \begin{array}{c|cccc} & u_1 & u_2 & u_3 & u_4 \\ \hline u_1 & 4 & 3 & 2 & 2 \\ u_2 & 3 & 3 & 0 & 0 \\ u_3 & 2 & 0 & 3 & 4 \\ u_4 & 2 & 0 & 4 & 10 \end{array}$$

Figure 4.2: An example representation of our data matrix X. The entry u_2u_3, for example, is equal to 0, suggesting that users 2 and 3 had no similar content labels assigned to their posts in a particular sub-forum.

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:

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

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

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

Figure 4.2.2 presents a graphical model for equation (1).

4.2.3 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:

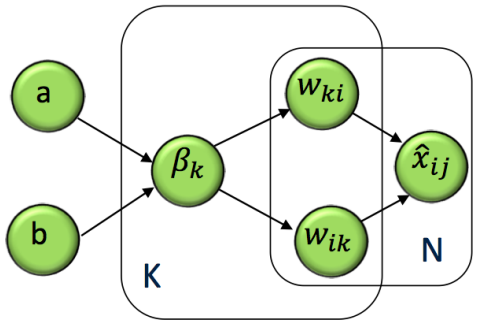


Figure 4.3: Graphical model for the joint distribution over all variables in the BNMF model. Here, H=W because our similarity matrix is symmetric.

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

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:

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

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.

The negative log priors are:

$$
\begin{aligned}
-\log p(\mathrm{W}|\beta) &= -\sum_{i=1}^{N}\sum_{k=1}^{K}\log \mathcal{HN}(w_{ik};0,\beta_k^{-1}) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\left(\tfrac{1}{2}\beta_k w_{ik}^2\right) - \tfrac{N}{2}\log\beta_k \\
&\quad + \text{const.} \qquad (5)
\end{aligned}
$$

$$
\begin{aligned}
-\log p(\mathrm{H}|\beta) &= -\sum_{k=1}^{K}\sum_{i=1}^{N}\log \mathcal{HN}(h_{ki};0,\beta_k^{-1}) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{N}\left(\tfrac{1}{2}\beta_k h_{ki}^2\right) - \tfrac{N}{2}\log\beta_k \\
&\quad + \text{const.} \qquad (6)
\end{aligned}
$$

Further to (Psorakis et al. 2011a), we assume the β_k are independent and place a standard Gamma distribution over them, yielding the following negative log hyper-priors:

$$
\begin{aligned}
\text{-}\log p(\beta) &= -\sum_{k=1}^{K}\log \mathcal{G}(\beta_k|a,b) \\
&= \sum_{k=1}^{K}\left(\beta_k b - (a-1)\log\beta_k\right) \\
&\quad + \text{const.} \qquad (7)
\end{aligned}
$$

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 4.4 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 insight

CD-BNMF(X; K_0; a; b)

1 **for** $i = 1$ *to* n_{iter} **do**

2 $\quad \mathrm{H} \leftarrow \left(\frac{\mathrm{H}}{\mathrm{W}^\top 1 + \beta \mathrm{H}}\right) \mathrm{W}^\top \left(\frac{\mathrm{X}}{\mathrm{WH}}\right)$

3 $\quad \mathrm{W} \leftarrow \left(\frac{\mathrm{W}}{1\mathrm{H}^\top + \mathrm{W}\beta}\right) \left(\frac{X}{\mathrm{WH}}\right) \mathrm{H}^\top$

4 $\quad \beta_k \leftarrow \frac{N+a-1}{\frac{1}{2}\left(\sum_i w_{ik}^2 + \sum_j h_{kj}^2\right)+b}$

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 4.4: Community Detection using Bayesian NMF. As described in (Tan and Fèvotte, 2009), the algorithm uses an efficient multiplicative coordinate descent algorithm to minimize the KL-Divergence between the observed data X and its factor matrices, W and H, arriving at point estimates for the parameters of interest. To learn the full posterior parameter distributions, other methods (e.g. Gibbs sampling, variational inference) may be used, for example, as in (Schmidt et al., 2009).

into a node's cluster membership, which we can model and explore explicitly if desired.

4.3 Results

4.3.1 Model Benchmarking

In order to evaluate the BNMF model and inference scheme's ability to represent the data, we tested it 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, using an Indian Buffet Process prior over a infinite-dimensional binary hidden feature matrix. The LPGM model is:

$$
\begin{aligned}
\mathrm{X} &= (\mathrm{Z} \circ \mathrm{F})\mathrm{W} + \mathrm{E} \\
\\
\mathrm{Z} &\sim \text{StickIBP}(\alpha) \\
\mathrm{W}_{kj} &\sim \text{Gamma(a,b)} \\
\mathrm{F}_{ik} &\sim \text{Gamma(c,d)} \\
\mathrm{X}_{i,:} | \mathrm{W}, \mathrm{Z}_{i,:}, \mathrm{F}_{i,:} &\sim \text{Poisson}\big((\mathrm{Z}_{i,:} \circ \mathrm{F}_{i,:})\mathrm{W} + \lambda\big)
\end{aligned}
$$

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; W 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 $\mathrm{E}_{ij} \sim \text{Poisson}(\lambda)$; and $\circ$ is the Hadamard product.

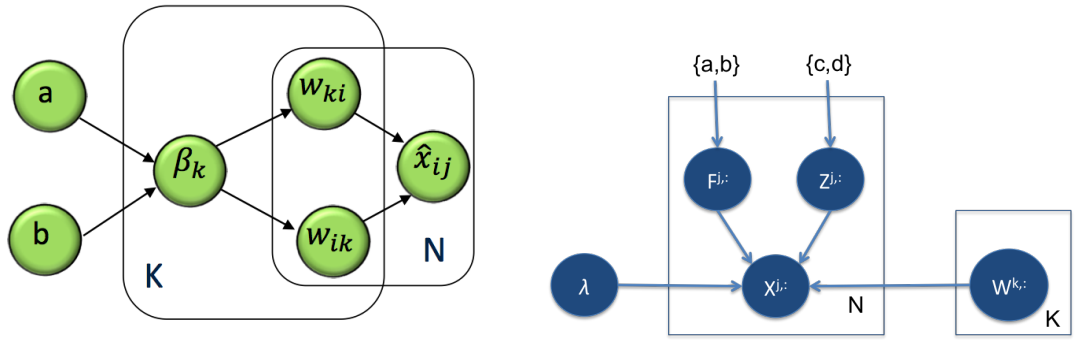
Due to poor scalability, it was infeasible to evaluate the LPGM on matrices with more than a few hundred rows and columns. We therefore evaluated both BNMF and LPGM on 20 randomly-selected 50×50 subsets of real-world data by comparing the root mean squared error (RMSE) and negative log likelihood (NLL) of held-out test data. Both the RMSE and NLL measured each probabilistic model's ability to represent the data: RMSE quantified predictive potential, while NLL illuminated

how "surprised" each model was to see the held-out data points. A value of 0 for both RMSE and NLL, therefore, would have implied perfect correspondence between the data and each model.

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 4.1 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].

We can review the graphical models of both BNMF and LPGM as exhibited in Figure 4.5 for insights into the performance discrepancies between both models. Firstly, it is worth noting that BNMF assumes that the factor matrices W and H are drawn from independent half-normal distributions while LPGM assumes independent gamma distributions over the same quantities. Next, the representation of the factor matrices is slightly different in both models: while BNMF seeks to learn two factor matrices, the product of which is an approximate reconstruction of our original data matrix X, the LPGM model learns three since the IBP requires inference of a binary feature matrix that indicates which learners exhibit which features

[1]the values presented in the table are generated by taking the arithmetic mean of the RMSE and NLL, computed for each of the 20 different subsets, with different data held-out for each subset.



(a) (b)

Figure 4.5: Graphical models for BNMF (a) and LPGM (b). In addition to the explicit representation of the binary latent feature matrix in (b), one of the most notable discrepancies in both models is the absence of a parameter in the LPGM that ties together the rows and columns of the latent factor matrices. Highly-correlated values in the real-world dataset may warrant this common precision parameter, thereby limiting the predictive potential of LPGM.

(or in our case, which learners are characterized by which latent clusters). Perhaps most obvious from the graphical models, however, is the absence of the prior precision parameters β_k in the LPGM to tie together the columns and rows of the individual factor matrices ($Z_{i,:} \circ F_{i,:}$) and W. The common precision parameters in the BNMF model may be favourable because of inherent correlations in the content-labelled data. The benefits of this tying behaviour in uncovering the true underlying clustering have also been validated empiricaly by (Tan and Fèvotte, 2009).

While the LPGM may be a more general model in that it makes minimal assumptions about dependencies in the dataset, for this particular application, BNMF appears to the superior model – not only because of its higher predictive accuracy, but also, its computational tractability (taking seconds to run on the full dataset, versus days for the LPGM).

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

Table 4.1: 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.

4.3.2 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. Before proceeding, it is important to note that the terms "community" and "cluster" are used interchangeably in this work, and that the word "community" does not necessarily imply a group of learners interacting with one another; rather, it denotes a latent group of learners detected due to similarly content-labelled forum contributions. Additionally, given the soft-partitioning scheme of BNMF, learners with varying probabilities of cluster participation were allocated to their respective groups, and therefore, the names used to categorize the inferred groups are more characteristic of *average* rather than nuanced *individual* behaviours.

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%). Interestingly, 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. Indeed, the more elicitative na-

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

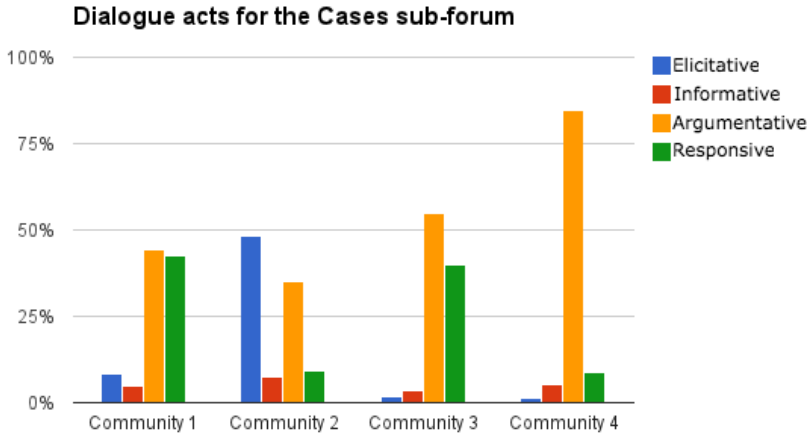
[3]Learners' locations were determined by leveraging a web-service to learn which country they last accessed course content from, based on a recorded IP address.

ture of dialogue in this group may suggest cultural differences in interpretations of, or responses to, various conversation topics.

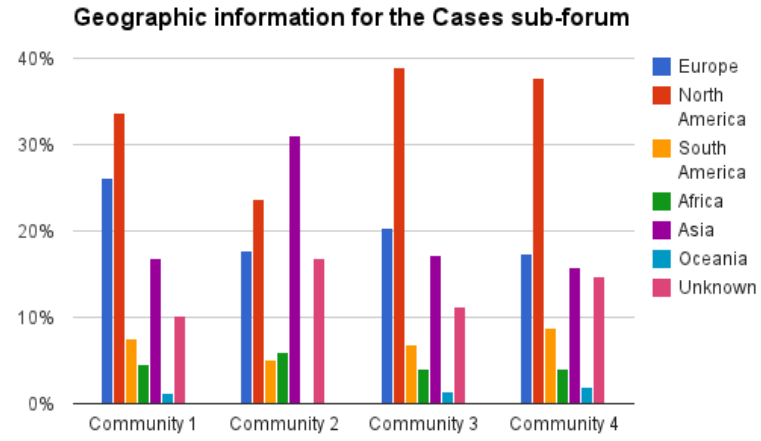
Community 3 (strategists). In many ways, people in this group were similar to group 1. They had similar levels of higher-order learning and tended to be responsive to others' comments. However, they had a greater proportion of argumentative statements (55%) and rarely had posts that reflected no learning (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 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 4.6 shows the dialogue acts and geographic locations of the members of each group.



(a)



(b)

Figure 4.6: Plot (a) illustrates the dialogue acts represented in the posts made by learners belonging to each community, and plot (b) depicts their geographic locations.

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%, $p\leq0.05$). 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.

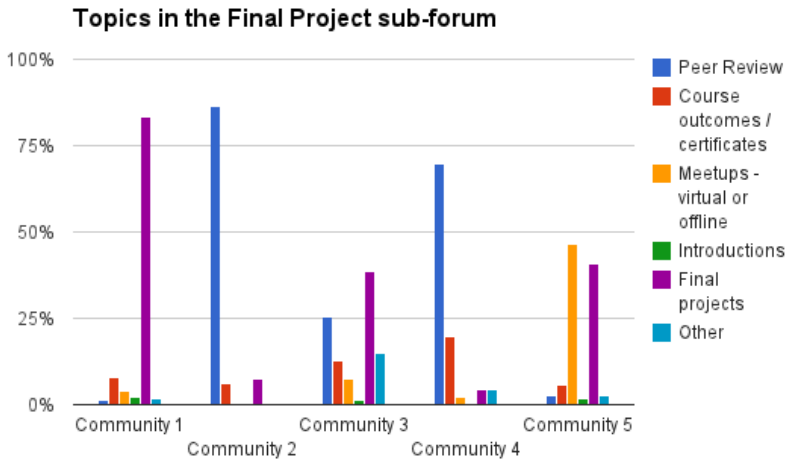
Community 2 (careful assessors). Participants in this group had the highest proportion of elicitative dialogue acts out of all of the groups (71%), but in contrast to group 1, the focus of their posts was about the peer review process (87%). They viewed more posts on average than groups 1, 4, and 5, 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

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

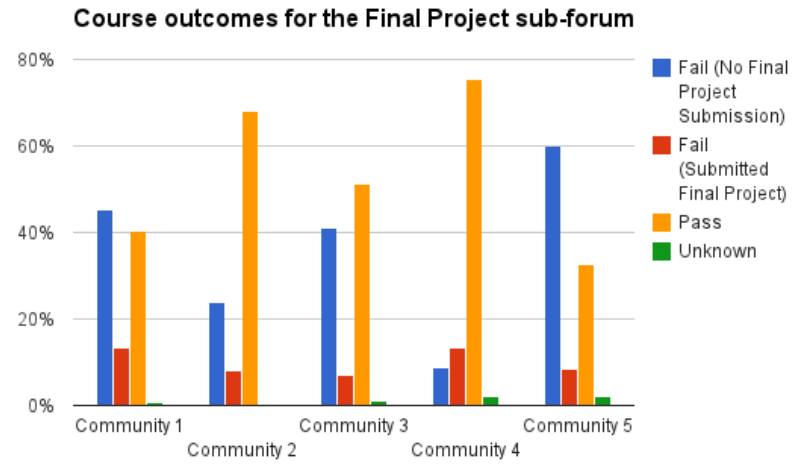
again if necessary. Like group 4, a high proportion of learners passed the course, compared to groups 1, 3, and 5 ($p<0.05$). These patterns suggest that this group needed to know more about the peer assessment process, but that its members were very strategic in their use of this sub-forum to obtain 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 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)



(b)

Figure 4.7: Plot (a) illustrates the discussion topics represented in the posts made by learners belonging to each community, and plot (b) depicts their course outcomes.

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

4.3.3 Robustness and Sensitivity Analysis

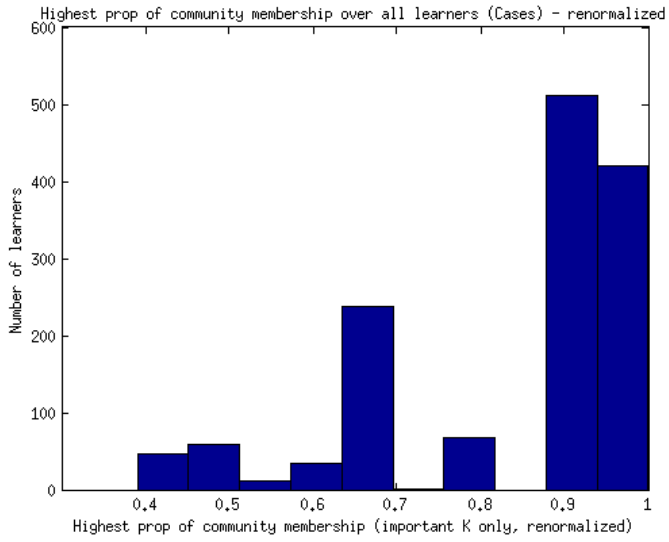
Through 28 in-depth interviews of learners from the MOOC, we were able to qualitatively classify 23 participants into one of the communities described in the preceeding sections (Eynon et al., 2014). This reinforced the practical significance of the detected communities and their appropriateness in characterising the behaviours of a number of interview participants. However, it is also important to note that in

some cases, learners played quite different roles – exhibiting different characteristics – across multiple different sub-forums (for example, two of the interview participants were classified as Individualists in the Cases sub-forum, but Community Builders in the Final Project sub-forum). These qualitative findings reveal the dynamicity of learners in different interaction settings, and imply the dangers of relying on solely quantitative methods to infer latent learner behaviour patterns.

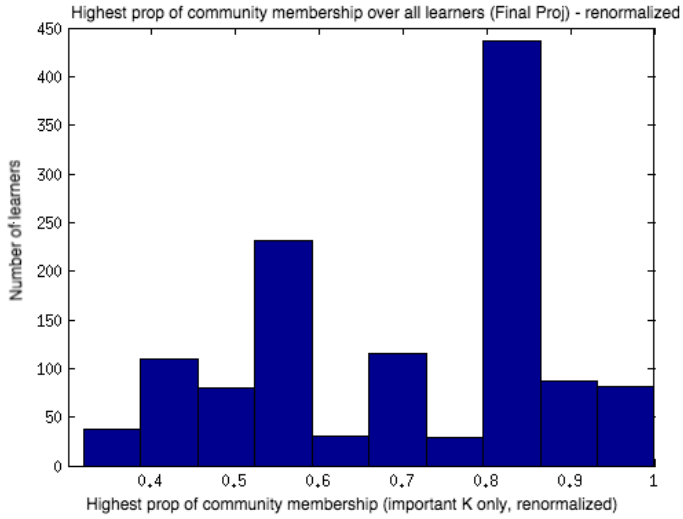
To explore the crispness of the detected community structure in the face of complex, dynamic human behaviour, we took advantage of the fact that our community detection algorithm – BNMF – yielded a soft-partition of learners to clusters to ultimately inform *how much* each learner belonged to each community. We started by computing histograms (Figure 4.8) that depicted this degree of learner-to-community membership in the Cases and Final Project sub-forums. From this figure, we can see that a non-trivial number of learners in each community were only weakly committed to their respective communities.

With the membership proportions of each learner in hand, we then computed the modularity of our content-similarity networks to explore if, and how, community structure and composition changes as a function of each learner's community membership strength. The modularity scores of the Cases and Final Project sub-forum networks including all community members (e.g., the full similarity matrices inputted into BNMF), were 0.06 and 0.19, respectively – indicative of very weak community structure. The charts in Figure 4.9 illustrates the trends in modularity scores for sub-networks induced according to different community membership strengths. It is clear that while modularity varies when considering different subsets of learners based on their community membership levels, overall, modularity scores are consistently low, and the network's tendency to yield crisp communities of learners based on their forum content is quite weak.

Perhaps more important than the crispness of the communities, however, is the

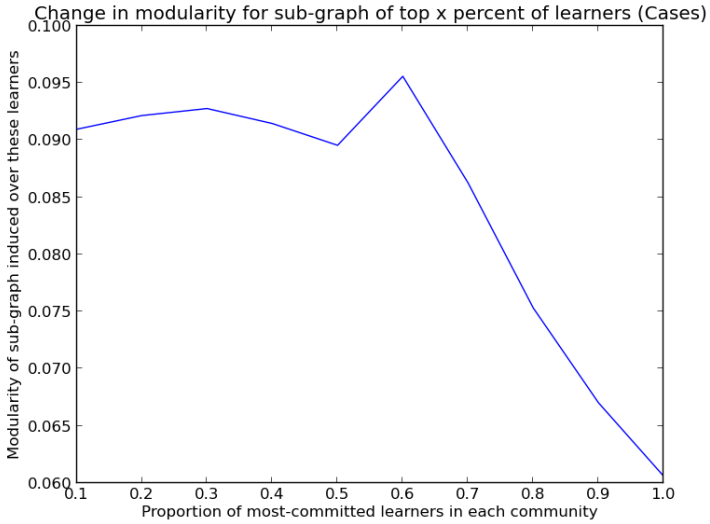


(a)

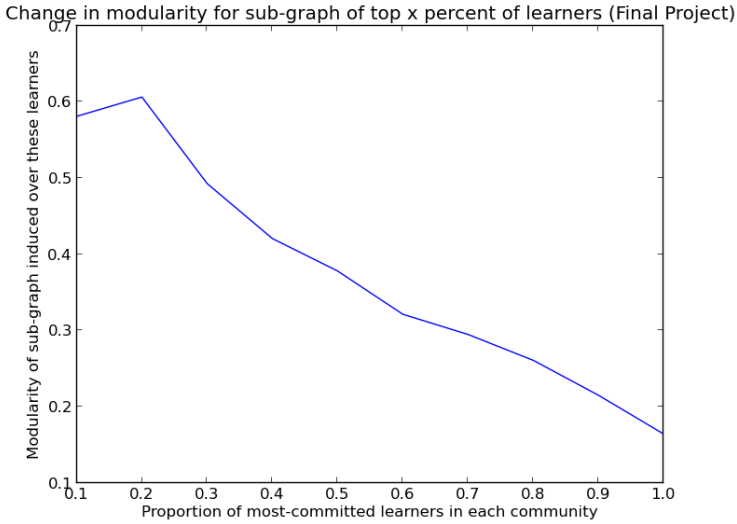


(b)

Figure 4.8: Each histogram depicts the number of learners with a maximum community membership belief score depicted by the values on the x-axis, renormalized over only those communities to which at least two learners were allocated (a membership proportion of 1 means that a particular learner belongs entirely to a single community). From these charts, we can infer that for many learners, community membership is highly entropic, suggesting inherently weak community structure.



(a)



(b)

Figure 4.9: The x-axis of each plot represents the sub-network induced over the full similarity matrix of each sub-forum where only the top x% of learners with highest proportional commitment to their respective communities are considered, and the y-axis reports the corresponding Louvain modularity score for each sub-network (so, the x-value of 1 and corresponding y-value indicates the modularity score for the full similarity matrix). Plots (a) and (b) depict trends for the Cases and Final Project sub-forums, respectively.

robustness of their defining characteristics as we consider alternative subsets of constituent learners. The hope is that the relative demographic and course performance characteristics of the inferred communities are consistent across different subsets of learners – i.e., that whether we consider the top 10% most committed learners or all learners for a particular community, the demographics, participation characteristics, and other descriptors of that community should be consistent.

To explore this empirically, we performed a post-hoc analysis over the 20% most committed learners in each of the above inferred communities (similar to the analysis presented in the results section above for both sub-forums). On the whole, while there were some similarities in pass rates characterising the subset and full communities, the demographic characteristics (geography, previous educational background) and course participation characteristics (average number of thread views and posts) were quite different. Due to the "label-switching" problem that makes community identification challenging across multiple runs of BNMF, we have omitted visuals that further portray these results.

In some sense, the largely inconsistent nature of both the full and learner-subset communities is to be expected given the weak modularity scores, high entropy in community membership as presented by BNMF's soft-membership output, and overall weak inherent community structure in the networks. By allocating learners with lower membership proportions to particular communities, the attributes that characterize each cluster are biased from those that characterize clusters with only those learners with high membership probabilities. Still, all hope is not lost: while the variability in results suggest that these methods may not be well-suited to help inform educators' targeted pedagogical interventions for specific learners, they may still be valid in providing a coarse metric for learner similarity that could help mitigate a key force of disengagement in MOOC discussion forums: feelings of "content-overload".

4.4 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 learners may influence how they choose to engage with the discussion forums.

As the above sensitivity analysis revealed, the similarity matrices for the different sub-forums used in community detection exhibit inherently weak community structure and inconsistent characteristics for different learner subsets. Therefore, to use these communities as infallible tools to aid educators' pedagogical interventions or assessment practices may compromise the learner's experiences and outcomes in massive-scale online course environments. Still, community detection in noisy MOOC data may be a valuable first-step towards developing automated content-recommendation methodologies – which, in the future, may be embedded directly within digital learning platforms. Understanding the similarities in different learners' participation preferences and strategies (for example, by using metrics like posterior community membership belief returned by BNMF) may enable the intelligent online learning platforms of the future to infer which discussion threads and

other course content may be more relevant and interesting to learners than others. Therefore, as large-scale social learning spaces proliferate, understanding and characterising latent "communication communities" within MOOCs may play a fundamental role in personalizing educational experiences by connecting learners with more engaging and personally-meaningful content.

Chapter 5

Conclusion

The rise in popularity of massive open online courses over the past few years has provided new challenges – and opportunities – for conducting interdisciplinary research into how people learn at scale. As MOOCs offer an unprecedented opportunity for "social learning", we adopted a hybrid analytical toolbox to explore learner communications in these educational settings. Through network analysis, we extracted insights into the structural properties and limitations of "learning in a crowd", identifying which types of discussion topics tended to promote participation and engagement. Based on these insights, we then conducted targeted latent feature extraction on content-analysed forum posts – using Bayesian nonnegative matrix factorization for community detection – in order to explore the relationships between learner demographics/course participation tendencies and the nature of their discussion contributions. By drawing upon parallel qualitative efforts (e.g., in-depth interviews) and additional robustness/sensitivity analysis, we explored the extent to which our empirical findings were grounded in existing educational theories and literature, and ultimately uncovered the inherent complexity of making sense of noisy and incomplete data capturing depicting human behaviours in massive online learning settings.

In the sub-sections that follow, we reflect on the key contributions of this work and consider future opportunities for course platforms and educators to better support learning in MOOCs.

5.1 Key contributions

With MOOC research in its infancy and minimal efforts made to date to explore, more specifically, communication within massive scale digital learning environments, we have presented a number of findings that may be of interest to both educators and practictioners. We summarize these findings according to our first two research questions.

5.1.1 What are the structural and diffusion properties of communication networks in MOOCs?

As Chapter 3 illuminated, most connections formed between learners through discussion thread co-participation in MOOCs are not indicative of strong underlying social relationships. Additionally, discussions in most sub-forums are highly "vulnerable" – that is, if we were to remove a small fraction of the most-connected learners (less than 40% of learners in most cases) from their respective discussion networks, these networks – and the potential for information diffusion – would rapidly degrade. Indeed, the online forums that promoted the most open-ended discussions – those focused on case analyses of real companies, or interactions surrounding the final strategic analysis project – tended to be the least vulnerable, suggesting greater engagement and participation from contributing learners.

Through information diffusion simulations, we found that the filtered interaction networks that emerge from forum participation are generally not conducive to the spread of information when compared to randomised network counterparts that

"break" implicit community structures. With respect to this finding, it is important to note that this inefficiency certainly highlights a structural attribute of MOOC communications, but is not necessarily an unfavourable attribute, as community structures and group interactions that seem to impede global information diffusion may indeed lead to improved learning experiences and outcomes. We hope that future researchers will expand upon this work by considering variants of diffusion analysis – including models that assume a higher likelihood of information propagation within communities versus between them, and learner engagement across sub-forums instead of simpler per sub-forum analyses as presented here. Doing so may enable instructional designers and educators to derive insights that help improve the design and implementation of design social learning experiences.

5.1.2 How can the textual content of discussion forum posts be used to infer latent learner communities, and how crisp are these communities?

Chapter 4 leveraged macro-level insights about the structural properties of MOOC communication to inform the subsequent content analysis of over 6,500 discussion posts, defining content dimensions based upon existing education and sociology literature. We sought a micro-level exploration of if, and how, latent learner groups could be inferred based on their contributions to the forums. As such, we built a *learner-by-learner* similarity matrix where entries denoted the number of shared category labels over all posts of any two learners and used Bayesian Non-negative Matrix Factorization to infer latent communities. We found that across the course's Cases and Final Project sub-forums, latent groups could be characterized by the nature and extent of their constituent learners' forum participations: while some chose to actively engage and participate to build communities and learn socially,

others simply posted to advance individual learning goals or final course achievement objectives. We also found significant differences in key demographic and course performance indicators across the learners in these different communities.

Qualitative interviews revealed that the inferred communities were, in fact, adept at capturing the different learner dynamics and preferences at play, and that the characteristics of these communities were well-grounded in education theory. However, the qualitative investigation also revealed that learners were dynamic, and that they often participated in the forums quite differently, depending on the discussion topic and their short and long-term learning goals. This dynamicity surfaced in subsequent quantitative robustness and sensitivity analysis, which revealed the low modularity scores for the networks depicted by the similarity matrix inputted into the community detection scheme. Further analyses showed that despite statistically significant differences in learner characteristics, the inferred communities and the attributes characterising constituent learners were not robust when considering different sub-networks/learner subsets.

These investigations revealed a crucial point for future MOOC researchers: when investigating large-scale, noisy social learning environments where complex human beings are sometimes experimenting with different participation styles, it is unsafe to assume that a single computational tool will be adequate in capturing and characterising the complex dynamics at play. Indeed, if researchers and practitioners simply trust these tools to generate meaningful, robust insights without additional quantitative and qualitative validation, they may adversely affect the future learning experiences of millions of individuals worldwide.

Therefore, beyond the insights into learner communication patterns, a key contribution of this research effort is the example it provides of the importance – and in some regards, necessity – of interdisciplinary and multi-faceted research when exploring inherently difficult-to-characterise human behaviour in online learning settings.

5.2 Improving massive-scale learning

In this section, we reflect on our third research question: **What roles can machine learning and computational social science play in communicating actionable insights into social learning patterns to help improve learning experiences online?**.

5.2.1 Automating or crowdsourcing content analysis

A key constraint to replicating the content analysis in Chapter 4 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 the course participants themselves, and in particular, both active and passive discussion forum 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.

5.2.2 Dealing with "content-overload"

Perhaps one of the biggest takeaways from this work is that many MOOC participants choose not to participate in the discussion forums simply because they feel overwhelmed by the volume of content. As participant interviews revealed,

with over 2,500 discussion threads and 15,000 posts, many learners simply felt lost when it came to course discussion forums – a trend that was supported by the low proportional participation and prevalent "lurking" behaviour, minimal significant interactions, and low levels of higher order knowledge construction. While some learners prefer not to learn socially, there appears to be a more fundamental force of fragmentation at play. Without the appropriate pedagogical design and technological tools (including the user interface implementation of online learning platforms), future learners may continue to miss out on opportunities to learn socially simply because they don't know where, or how, to engage.

Fortunately, the methods and results presented in this work offer a strong foundation for building "intelligent" MOOC platforms. Matrix Factorization techniques have long been used in recommender-systems, using social preference information in order to recommend new products or other often consumer-focused resources (e.g., Mackey et al., 2010). Building in recommender systems that operate in real-time and are based on collaborative filtering, community detection, or even topic-modelling technologies may help suggest a subset of relevant and interesting forum discussions to participants. As more courses are offered, data-pooling across courses and course platforms could enable these systems to grow "smarter" with time.

To deal with issues of data sparsity, particularly as they pertain to data generated through explicit course participation (e.g., posting in discussion forums), modelling user behaviour patterns by harnessing data that *is* readily available (e.g., clickstream data) will be an important first step for content recommendations based on "social profiling". For this, the literature on point processes with stochastic intensity functions (e.g., Møller et al., 1998) may provide a basis for exploring how constantly varying and seemingly unpredictable learner behaviours can be characterised. Explicitly factoring in learner motivations for course participation – and how these motivations influence learner behaviours throughout the course – may further en-

able robust personalization tools. It is our hope that these efforts to combat feelings of content-overload will provide learners with the support and guidance necessary to help them more deeply engage in the ways that they hope and intend to.

Still, it is important to note that recommendations in learning settings may, and perhaps should, be informed by different objectives than those in consumer domains. Many times, learning happens precisely when we are exposed to resources or people that fall outside of our discernable set of preferences. Therefore, as we expand the functionality and opportunities offered by digital learning environments, technologists, researchers, and education practitioners will need to work closely together to ensure that the appropriate tools and frameworks are applied to champion the learner and his/her educational ambitions. The challenges posed by massive-scale education research are too nuanced - and the potential to negatively affect learning through misinformed or incomplete approaches too grave - to fall short in this matter.

If MOOCs are to help promote more "democratic" educational opportunities, we must understand how to design and implement them in order to facilitate greater participation and personalised support for thousands of learners with unique pedagogical needs. Never before have educational resources that help facilitate global-scale social learning been so widely available. We hope researchers and practitioners will build upon this work to design new courseware, technologies, and pedagogies that help promote educational engagement and learning for all. The opportunities for envisioning, and building, the future of education are ripe.

References

Abelson, H. (2008). The Creation of OpenCourseWare at MIT. *Journal of Science, Education, and Technology*, 17(2): 164-174.

Adams, R. P., I. Murray, D. J. C. MacKay (2009). Tractable Nonparametric Bayesian Inference in Poisson Processes with Gaussian Process Intensities. *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada.

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

Albert, R., H. Jeong, A. L. Barabasi (2000). Error and attack tolerance of complex networks. *Letters to Nature*, 406, 378-382.

Aldous, D. (1985). Exchangeability and Related Topics. *École d'Été de Probabilités de Saint-Flour XIII-1983*, Springer, Berlin, 1-198.

Anderson, A., 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.

Anderson, R. M., R. M. May (1992). *Infectious Diseases of Humans: Dynamics and Control*, Oxford Science Publications, Oxford.

Ayers, E., R. Nugent, N. Dean (2009). A Comparison of Student Skill Knowledge Estimates. *International Conference on Educational Data Mining*, Cordoba, Spain: 1-10.

Baker, L. D., A. K. McCallum (1998). Distributional clustering of words for text

classification. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 96-103. doi:10.1145/290941.290970.

Baldwin, T., M. D. Bedell, J. Johnson (1997). The Social Fabric of a Team-Based M.B.A. Program: Network Effects on Student Satisfaction and Performance. *The Academy of Management Journal*, 40(6): 1369-1397.

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

Blondel, V. D., J. L. Guillaume, R. Lambiotte, E. Lefebvre (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theor. Exp.*, 10.

Breslow, L, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, D. T. Seaton (2013). Studying Learning in the Worldwide Classroom: Research into edX's First MOOC. *Research & Practice in Assessment*, 8: 13-25.

Brinton, C. G., 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.

Buchanan, M., G. Calderalli (2010). A networked world. *Physics World*, 23(2): 22-24.

Butler, B. S. (2001). Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Inf. Syst. Res.*, 12, 346-362.

Canny, J. (2004). GaP: A Factor Model for Discrete Data. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, Sheffield, South Yorkshire, UK.

Caron, F. (2012). Bayesian nonparametric models for bipartite graphs. *Advances in Neural Information Processing Systems 25*, Lake Tahoe, NV, USA: 2051-2059.

Catanzaro, M., M. Bogulna, R. Pastor-Satorras (2005). Generation of uncorrelated random scale-free networks. *Phys. Rev. E*, 71, 027103.

Cemgil, A. (2009). Bayesian inference for nonnegative matrix factorization models. *Computational Intelligence and Neuroscience*.

Cho, H., G. Gay, B. Davidson, A. Ingraffea (2007). Social networks, communication styles, and learning performance in a CSCL community. *Computers and Education*, 49, 309329.

Cobo, G., D. Garcia, E. Santamaria, J. A. Moran, J. Melenchon, C. Monzo (2011). Modeling students' activity in online discussion forums: a strategy based on time series and agglomerative hierarchical clustering. *4th International Conference on Educational Data Mining*, Eindhoven, the Netherlands: 253-258.

Colley, H., P. Hodkinson, J. Malcolm (2002). Non-formal learning: mapping the conceptual terrain. A consultation report. Leeds: University of Leeds Lifelong Learning Institute.

Costenbader, E., T. W. Valente (2003). The stability of centrality measures when networks are sampled. *Social Networks*, 25, 283-307.

Danon, L., A. Diaz-Guilera, J. Duch, A. Arenas (2005). Comparing community structure identification. *Journal of Statistical Mechanics*, 2005(09): P09008.

Ding, C., X. He, H. D. Simon (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. *SIAM International Conference on Data Mining*.

Dhillon, I. S., Y. Guan, B. Kulis (2004). Kernel k-means: spectral clustering and normalized cuts. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 551-556.

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

Eynon, R., I. Hjorth, N. Gillani, T. Yasseri (2014). 'Vote Me Up If You Like My Ideas!' Experiences of Learning in a MOOC. *2014 ICA Pre-conference working paper: Innovation in Higher Education: Building a Better Future?*.

Ezekiel, E. J. (2013). Online education: MOOCs taken by the educated few. *Nature*, 503: 342.

Freeman, L. (1978). Centrality in Social Networks Conceptual Clarification. *Social Networks*, 1, 215-239.

Garcia, P., A. Amandi, s. Schiaffino, M. Campo (2005). Using Bayesian Networks to Detect Students' Learning Styles in a Web-based Education System. *Proceedings of ASAI*, Rosario: 115-126.

Geman, S., D. Geman (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6): 721-741.

Gershman, S. J., D. Blei (2011). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56: 1-12.

Gillani, N., R. Eynon. Communication Patterns in Massively Open Online Courses (2014). *The Internet and Higher Education*, 23, 18-26.

Gillani, N., R. Eynon, M. Osborne, I. Hjorth, S. J. Roberts (2014). Communication Communities in MOOCs. *arXiv:1403.4640*.

Goodman, P. S., L. A. Dabbish (2011). Methodological Issues in Measuring Group Learning. *Small Group Res.*, 42, 379-404.

Gopalan, P., F. J. R. Ruiz, R. Ranganath, D. M. Blei (2014). Bayesian nonparametric Poisson factorization for recommendation systems. *The 17th International Conference on Artificial Intelligence and Statistics*, Reykjavik, Iceland.

Gopalan, P., J. Hofman, D. M. Blei (2013). Scalable Recommendation with Poisson factorization. *arXiv:1311.1704*.

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

Gupta, S. K., 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.

Guyon, I., A. Saffari, G. Dror, G. Cawley (2010). Model Selection: Beyond the Bayesian/Frequentist Divide. *Journal of Machine Learning Research*, 11: 61-87.

Hamalainen, W., M. Vinni (2006). Comparison of Machine Learning Methods for Intelligent Tutoring Systems. *Intelligent tutoring Systems*, 4053: 525-534.

Haythornthwaite, C. (1996). Social Network Analysis: An Approach and Technique for the Study of Information Exchange. *Library and Information Science Research*, 18(4): 323-342.

Holmes, I., K. Harris, C. Quince (2012) Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. PLoS ONE 7(2): e30126.

Huerta-Quintanilla, R., E. Canto-Lugo, D. Viga-de Alva (2013). Modeling Social Network Topologies in Elementary Schools. *PLOS One*, 8(2): 1-9.

Jiang, S., S. M. Fitzhugh, M. Warschauer (2014). Social Positioning and Performance in MOOCs. *Unpublished.*

Jordan, M., Z. Ghahramani, T. Jaakkola, L. Saul (1999). Introduction to variational methods for graphical models. *Machine Learning*, 37: 183-233.

Karsai, M. M., M. Kivela, R. K. Pan, K. Kaski, J. Kertesz, A. L. Barabasi, J. Saramaki (2011). Small but slow world: How network topology and burstiness slow down spreading. *Phys. Rev. E*, 83, 025102.

Kellogg, S., S. Booth, K. M. Oliver (2014). A social network perspective on peer support learning in MOOCs for Educators. *International Review of Research in Open and Distance Learning (in press).*

Kemp, C., J. B. Tenenbaum, T. L. Griffiths, T. Yamada, N. Ueda (2006). Learning systems of concepts with an infinite relational model. *Proceedings of the 21st National Conference on Artificial Intelligence.*

Kepp, S. J., H. Schorr (2009). Analyzing collaborative learning activities in

wikis using social network analysis. *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems*, New York, 42014206.

Kermack, W. O., A. G. McKendrick (1972). A Contribution to the Mathematical Theory of Epidemics. *Roy. Soc. Lond. A. Matt. P.*, 115, 700-721.

Kim, D., P. Gopalan, D. M. Blei, E. B. Sudderth (2013). Efficient Online Inference for Bayesian Nonparametric Relational Models. *Advances in Neural Information Processing Systems 26*, Tahoe, NV, USA: 962-970.

Kizilcec, R., C. Piece, Schneider, E. (2013). Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. *The 3rd Proceedings of the Learning Analytics and Knowledge Conference*, Leuven, Belgium.

Krause, J., D. Lusseau, R. James (2009). Animal social networks: an introduction. *Behavioral Ecology and Sociobiology*, 63:967973.

Krings, G., M. Karsai, S. Bernhardsson, V. D. Blondel, J. Saramaki (2012). Effects of time window size and placement on the structure of an aggregated communication network. *European Physical Journal Data Science*, 1(1):116.

Lambiotte, R., J. C. Delvenne, M. Barahona (2009). Laplacian Dynamics and Multiscale Modular Structure in Networks. *arXiv:0812.1770*.

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

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, 58: 13031319.

Li, M. X., et al. (2014). Statistically Validated mobile communication networks: Evolution of motifs in European and Chinese data. *arXiv:1403.3785*.

Littlejohn, A. (2013). Understanding Massive Open Online Courses. *CECMA EdTech Notes*.

Lindsey, R., M. Mozer, W. J. Huggins, H. Pashler (2013). Optimizing Instruc-

tional Policies. *Advances in Neural Information Processing Systems 26*, Tahoe, NV, USA: 2778 - 2786.

Møller, J., A. R. Syversveen, R. P. Waagepetersen (1998). Log Gaussian Cox Processes. *Scandinavian Journal of Statistics*, 25(3): 451-482.

Mackay, D. J. C. (1995). Probable networks and plausible predictions a review of practical Bayesian models for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469505.

Mackey, L., D. Weiss, M. I. Jordan (2010). Mixed membership matrix factorization. *International Conference on Machine Learning*, Haifa, Israel.

Mahraj, K. (2012). Using Information Expertise to Enhance Massive Open Online Courses. *Public Services Quarterly*, 8(4): 359-368.

Nematzadeh, A., E. Ferrara, A. Flammini, Y. Ahn (2014). Optimal Network Modularity for Information Diffusion. *Phys. Rev. Lett.*, 113(8):088701.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. of the Natl. Acad. Sciences*, 103, 8577-8582.

Newman, M. E. J. (2010). *Networks: an introduction*. Oxford University Press, Oxford.

Newman, M. E. J., M. Girvan (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2): 026113.

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

Onella, J. P., et al. (2007). Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U.S.A.*, 104, 7332-7336.

Palonen, T., K. Hakkarainen (2000). Patterns of Interaction in Computer-Supported Learning: A Social Network Analysis. In: B. Fishman, S. O'Conner-Divelbiss, eds. *Fourth International Conference of the Learning Sciences*. Mahwah, NJ: Erlbaum: 334-339.

Pardos, Z. A., Y. Bergner, D. Seaton, D. E. Pritchard. Adapting Bayesian Knowledge Tracing to a Massive Open Online Course in edX. *Proceedings of the 6th International Conference on Educational Data Mining*, Memphis, TN, USA.

Perina, A., P. Lovato, V. Murino, M. Bicego (2010). Biologically-aware Latent Dirichlet Allocation (BaLDA) for the Classification of Expression Microarray. *Pattern Recognition in Bioinformatics*, 6282: 230-241.

Piech, C., 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.

Porter, M. A., J. P. Gleeson (2014). Dynamical Systems on Networks: A Tutorial. *arXiv:1403.7663v1*.

Preece, J., B. Nonnecke, D. Andrews (2004). The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior*, 20(2): 201-223.

Psorakis, I. (2013). Probabilistic inference in ecological networks; graph discovery, community detection and modelling dynamic sociality. *Doctoral Dissertation, University of Oxford*.

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

Psorakis, I., 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.

Rao, V., Y. W. Teh (2011). Gaussian process modulated renewal processes. Advances in neural Information Processing Systems 24, 2474-2482.

Rasmussen, C. E. (2000). The Infinite Gaussian Mixture Model. *Advances in Neural Information Processing Systems 12*, 554-560.

Reed, C., Z. Ghahramani (2013). Scaling the Indian Buffet Process via Submod-

ular Maximization. arXiv:1304.3285v4.

Reffay, C., T. Chanier (2003). How Social Network Analysis can help to Measure Cohesion in Collaborative Distance-Learning. in *Designing for Change in Networked Learning Environments*, B. Wasson, S. Ludvigsen, and U. Hoppe, eds. Springer Netherlands: 343-352.

Romero, C., S. Ventura (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, 40(6): 601-617.

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

Schmidt, M. N., M. Mørup (2013). Nonparametric Bayesian modeling of complex networks: an introduction. *Signal Processing Magazine, IEEE*, 30(3): 110-128.

Schmidt, M. N., O. Winther, L. K. Hansen (2009). Bayesian Non-negative Matrix Factorization. *Independent Component Analysis and Signal Separation Lecture Notes in Computer Science*, 5441: 540-547.

Siemens, G. (2005). Connectivism: A Learning Theory for a Digital Age. *International Journal of Instructional Technology and Distance Learning*, 2(1): 3-10.

Simonsen, I., L. Buzna, K. Peters, S. Bornholdt, D. Helbing (2008). Transient Dynamics Increasing Network Vulnerability to Cascading Failures. *Phys. Rev. Lett.*, 100, 218701.

Simpson, E., 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.

Sinha, T. (2014). Together we stand, Together we fall, Together we win: Dynamic Team Formation in Massive Open Online Courses. *arXiv:1404.5521v1.*

Spizzirri, L. (2011). Justification and Application of Eigenvector Centrality.

Unpublished.

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

Stauffer, C., W. E. L. Grimson (1999). Adaptive background mixture models for real-time tracking. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference*, 2: 252.

Sundararajan, B. (2010). Emergence of the most knowledgeable other (mko): Social network analysis of chat and bulletin board conversations in a CSCL system. *Electronic Journal of e-Learning*, 8:191-208.

Tan, V., C. Fèvotte (2009). Automatic relevance determination in nonnegative matrix factorization. *SPARS09 - Signal Processing with Adaptive Sparse Structured Representations*, 119.

Tay, W. Y. (2012). Conceptualising learning in social and virtual worlds: an ethnography of three groups in Second Life. *Doctoral Dissertation, University of Oxford*.

Taylor, C., K. Veeramachaneni, U. O'Reilly (2014). Likely to stop? Predicting Stopout in Massive Open Online Courses. *arXiv:1408.3382*.

Teh, Y. W., D. Gorur, Z. Ghahramani (2007). Stick-breaking construction for the Indian buffet process. *Journal of Machine Learning Research - Proceedings Track*, 2:556-563.

Titsias, M. K. (2007). The Infinite Gamma-Poisson Feature Model. *Advances in Neural Information Processing Systems 20*, 1513-1520. Vancouver, B.C.

Trivedi, S., Z. A. Pardos, G. N. Sarkozy, N. T. Heffernan (2012). Co-clustering by bipartite spectral graph partitioning for out-of-tutor prediction. Proceedings of the 5th International Conference on Educational Data Mining, Chania, Greece: 33-40.

Trognon, A. M., M. Batt (2012). Group Dynamics and Learning. Published in *The Encyclopedia of the Sciences of Learning*, edited by N. M. Seel, 1388-1391.

Tumminello, M., F. Lillo, J. Piilo, R. N. Mantegna (2012). Identification of clusters of investors from their real trading activity in a financial market. *New. J. Phys.*, 14:013041.

Van De Sande, B. (2013). Properties of the Bayesian Knowledge Tracing Model. *Journal of Educational Data Mining*, 5(2): 1-10.

Vaquero, L., M. Cebrian (2013). The rich club phenomenon in the classroom. *Sci. Rep.*, 3, 1174.

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

Wasserman, S., K. Faust (1994). *Social Network Analysis.* Cambridge: Cambridge University Press.

Xiao, H., T. Stibor (2010). Efficient Collapsed Gibbs Sampling For Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 13: 63-78.

Xu, W., X. Liu, Y. Gong (2003). Document clustering based on non-negative matrix factorization. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 267-273. doi:10.1145/860435.860485.

Yang, D., T. Sinha, D. Adamson, C. P. Rose (2013). "Turn on, Tune in, Drop out": Anticipating student dropouts in Massive Open Online Courses. Paper presented at *Neural Information Processing Systems: Workshop on Data Driven Education*, Tahoe, NV.

Yudelson, M. V., K. R. Koedinger, G. J. Gordon (2013). Individualized bayesian knowledge tracing models. *Artificial Intelligence in Education*, 7926: 171-180.

Appendix

Supplementary Information for Communication Network Analysis

Supplementary figures

The figures below depict additional trends from FOBS-1 and FOBS-2. Additionally, the video located at `https://www.dropbox.com/s/rvkd18dnuiyd02v/finalprojects.avi` depicts the network vulnerability simulation for the Final Project sub-forum from FOBS-1. Each frame corresponds to one step in the algorithm, at which the node with the highest betweenness centrality is computed and disconnected from the graph.

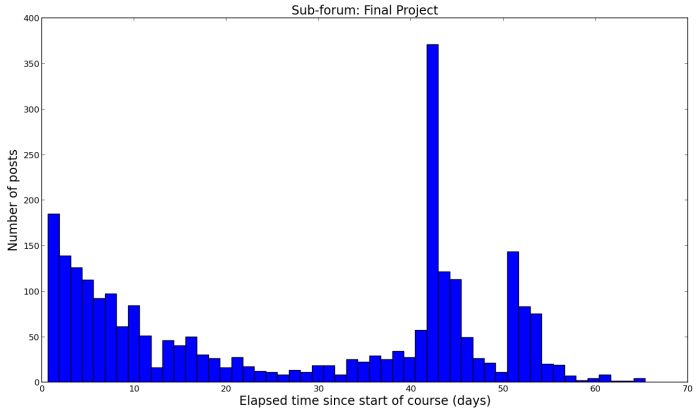


Figure 1: Forum post activity over time in the Final Project sub-forum of FOBS-1. The large peak around the end of week 6 corresponds to students posting last-minute questions about the final project submission deadline.

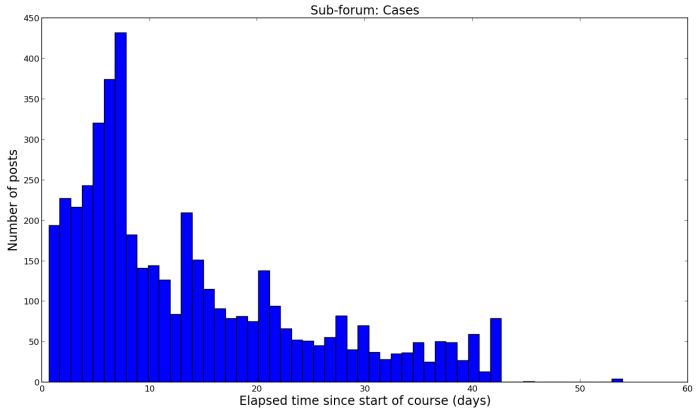


Figure 2: Forum post activity over time in the Cases sub-forum of FOBS-1. Like many of the other sub-forums, participation decreases as the course progresses, but there are still peaks of activity each week corresponding to the weekly case discussions.

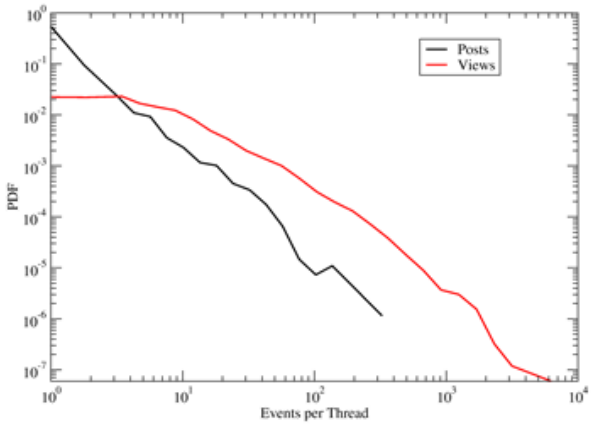


Figure 3: The number of views and posts per discussion threads across all sub-forums, in log-log scale, for FOBS-1. The charts suggest a fat-tailed distribution of views and posts across threads i.e., the vast majority of discussion threads have very small numbers of posts and views, with a few threads harbouring high posting and viewing behaviour.

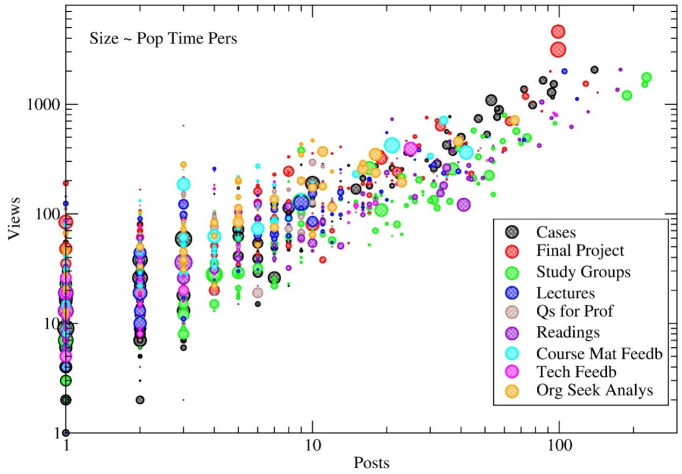


Figure 4: Comparison of posts and views for each thread in a particular sub-forum, denoted by the coloured circles shown here, for FOBS-1. The size of each circle indicates the Popularity time persistence of the corresponding thread, i.e., the amount of time that elapses before 90% of all posts are made to that thread (hence, small circles depict threads with very short lifespans).

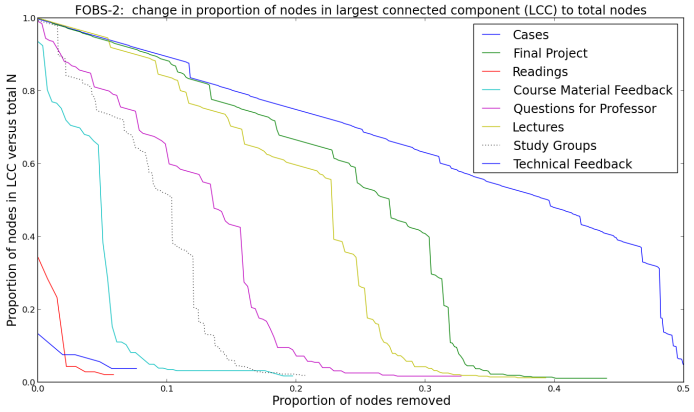


Figure 5: Communication vulnerability in the different sub-forums of FOBS-2. These trends are similar to those observed in FOBS-1.

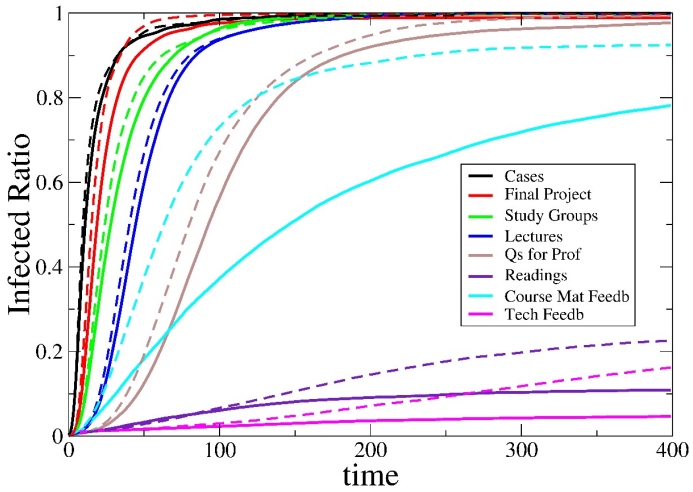


Figure 6: shows the percentage of infected nodes vs. simulation time for different networks in FOBS-2 (similar to those observed in FOBS-1). The solid lines show the results over the original network and the dashed lines for the degree-preserved shuffled network (configuration model).

Robustness and Sensitivity Analysis for Communication Communities Extraction

Modularity scores for different content labels

As a network's modularity is inherently tied to the manner in which it is formulated, we explored how modularity scores changes as a result of changing the content lables considered when formulating the similarity matrix described in Chapter 4. As there were 5 content labels, we considered all 31 (i.e., 2^5 - 1) possible subsets of labels. Tables 1 and 2 reveal the sorted modularity scores. It is important to note that for certain labels, content was largely homogeneous (e.g. most posts across sub-forums were labeled as "neutral" for affect), leading to a dense similarity matrix and understandably low modularity scores. These results suggest that regardless of which labels are used to formulate the similarity matrix, both the Cases and Final Project sub-forums exhibit weak inherent community structure.

Content dimension	**Modularity**
Communication	0.2103
Learning	0.1407
Learning, Communication	0.1352
Relevance, Communication	0.0927
Topic, Communication	0.0909
Communication, Affect	0.0865
Learning, Relevance, Communication	0.0837
Topic, Learning, Communication	0.0806
Learning, Communication, Affect	0.0791
Learning, Relevance	0.0612
Topic, Relevance, Communication	0.0611
Topic, Learning, Relevance, Communication	0.0606
Relevance, Communication, Affect	0.0606
Learning, Relevance, Communication, Affect	0.0600
Learning, Affect	0.0593
Topic, Learning	0.0588
Topic, Learning, Communication, Affect	0.0586
Topic, Communication, Affect	0.0565
Topic, Learning, Relevance, Communication, Affect	0.0480
Topic, Relevance, Communication, Affect	0.0458
Topic, Learning, Relevance	0.0430
Learning, Relevance, Affect	0.0428
Topic, Learning, Affect	0.0394
Topic, Learning, Relevance, Affect	0.0333
Relevance	0.0263
Relevance, Affect	0.0145
Topic, Relevance	0.0144
Topic, Relevance, Affect	0.0115
Topic, Affect	0.0018
Topic	0.0004
Affect	0.0002

Table 1: Modularity scores for the Cases sub-forum's similarity matrices, constructed by considering different subsets of content label dimensions.

Content dimension	**Modularity**
Communication	0.3103
Relevance	0.3010
Topic	0.2998
Topic, Relevance	0.2159
Relevance, Communication	0.1717
Topic, Communication	0.1641
Topic, Relevance, Communication	0.1561
Relevance, Affect	0.1307
Learning, Relevance	0.1258
Topic, Relevance, Affect	0.1206
Topic, Learning, Relevance	0.1177
Communication, Affect	0.1103
Learning, Communication	0.1079
Topic, Affect	0.1043
Topic, Relevance, Communication, Affect	0.1025
Topic, Learning	0.1023
Topic, Learning, Relevance, Communication	0.1009
Learning, Relevance, Communication	0.0995
Relevance, Communication, Affect	0.0990
Topic, Communication, Affect	0.0929
Topic, Learning, Communication	0.0911
Topic, Learning, Relevance, Affect	0.0841
Learning, Relevance, Affect	0.0831
Topic, Learning, Relevance, Communication, Affect	0.0774
Learning, Relevance, Communication, Affect	0.0734
Learning, Communication, Affect	0.0698
Topic, Learning, Communication, Affect	0.0661
Topic, Learning, Affect	0.0652
Affect	0.0194
Learning	0.0189
Learning, Affect	0.0175

Table 2: Modularity scores for the Final Project sub-forum's similarity matrices, constructed by considering different subsets of content label dimensions.

Analysis Toolbox

Over 12,500 lines of Python code have been written to process clickstream and relational database (MySQL) data from multiple MOOCs; compute basic statistics on discussion patterns; perform network analysis; and create visualizations. This code has made extensive use of a number of support packages, including *NumPy*, *SciPy*, *PyPlot*, and *NetworkX*. Additionally, nearly 1,000 lines of MATLAB code have been written as wrappers and other utilities to leverage existing implementations of MAP and Variational Bayes BNMF (used for benchmarking), as well as the IBP-based latent feature model presented in Chatper 4.

Please email nabeel.gillani90@gmail.com for more information.