ABSTRACT

Title of dissertation: A Probabilistic Approach to
Modeling Socio-Behavioral Interactions

Arti Ramesh,
Doctor of Philosophy, 2016

Dissertation directed by: Professor Lise Getoor
Department of Computer Science

In our ever-increasingly connected world, it is essential to build computational models that represent, reason, and model the underlying characteristics of real-world networks. Data generated from these networks are often heterogeneous, interlinked, and exhibit rich multi-relational graph structures having unobserved latent characteristics. My work focuses on building computational models for representing and reasoning about rich, heterogeneous, interlinked graph data. In my research, I model socio-behavioral interactions and predict user behavior patterns in two important online interaction platforms: online courses and online professional networks. Structured data from these interaction platforms contain rich behavioral and interaction data, and provide an opportunity to design machine learning methods for understanding and interpreting user behavior. The data also contains unstructured data, such as natural language text from forum posts and other online discussions. My research aims at constructing a family of probabilistic models for modeling social interactions involving both structured and unstructured data. In the early part of this thesis, I present a family of probabilistic models for online courses

for: 1) modeling student engagement, 2) predicting student completion and dropouts, 3) modeling student sentiment toward various course aspects (e.g., content vs. logistics), 4) detecting coarse and fine-grained course aspects (e.g., grading, video, content), and 5) modeling evolution of topics in repeated offerings of online courses. These methods have the potential to improve student experience and focus limited instructor resources in ways that will have the most impact. In the latter part of this thesis, I present methods to model multi-relational influence in online professional networks. I test the effectiveness of this model via experimentation on the professional network, LinkedIn. My models can potentially be adapted to address a wide range of problems in real-world networks including predicting user interests, user retention, personalization, and making recommendations.

A Probabilistic Approach to Modeling Socio-Behavioral Interactions

by

Arti Ramesh

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:

Prof. Lise Getoor, Univ. of Maryland, College Park	Chair/Advisor
Prof. Jennifer Golbeck, Univ. of Maryland, College Park	Dean's Representative
Prof. Hal Daumé III, Univ. of Maryland, College Park	Committee Member
Prof. Amol Deshpande, Univ. of Maryland, College Park	Committee Member
Prof. Dan Goldwasser, Purdue University	Committee Member

To Anand, for all his love and support!

Acknowledgments

This PhD journey has been nothing short of a roller coaster ride and a page is insufficient to thank everyone who were part of this wonderful adventure. First and foremost, I would like to thank my advisor Lise Getoor for her effort and care in shaping my research. She always had my back when things did not go as planned, which usually happened quite often in research! It is amazing how much Lise cares for her students and shows incredible patience. I truly cannot imagine a better advisor. So, thank you Lise, for believing in me! In a field where there are very few women professors, I am also lucky and very happy to have found the perfect person to learn from and emulate!

I am also very grateful to Hal Daumé III for helping me shape my research ideas. I worked with Hal for my first couple of projects, where he helped me with insights on combining computational linguistics and structured prediction and I have ever since been attracted to models that reason about language in a networked setting. I consider myself most fortunate that I had the pleasure of working with Dan Goldwasser and Bert Huang, who were both postdoctoral researchers at the time at UMD. I learnt a tremendous amount working closely with them as they helped me translate research ideas and results into words the scientific community understood. I attribute most of my "how to do research" training to them. Some of the best memories from my PhD are working with them on my first research project, which I will cherish forever. I am also thankful to my committee members Amol Deshpande and Jen Golbeck for their insightful comments.

Being an international student, LINQS gave me the ideal home away from home, providing me with much needed emotional support. I am very blessed to have the best lab mates; they made navigating tough grad-school situations such as pressing paper deadlines, job search, moving to Santa Cruz less stressful and sometimes even fun! I am very grateful that Lise recruited Shobeir Fakhraei along with me, with whom I have navigated the ups and downs of grad school together—be it taking classes, discussing what to write on weekly reports, or figuring out the move to Santa Cruz! When Lise moved to UCSC, I was so distraught at the idea of moving schools and working from UCSC toward a PhD at UMD. But in retrospect, it gave me twice the amount of fun and camaraderie—Steve Bach, Shobeir Fakhraei, Ben London, Alex Memory, Hui Miao, Walaa Mustafa, Jay Pujara, and Theodoros Rekatsinas at UMD and Jimmy Foulds, Pigi Kouki, Shachi Kumar, Dhanya Sridhar, and Sabina Tomkins at UCSC. I have learnt a tremendous amount from each of you, and I thank you for being part of my PhD experience! Thanks are also due to my numerous friends at both UMD and UCSC for their help and support.

My graduate career started at UMass where I was a Masters student and had the pleasure of working with Andrew McCallum in the IESL lab. Andrew introduced me to machine learning research and also helped me pick the perfect PhD advisor, which I wouldn't have been able to accomplish on my own. I am also thankful to IESL lab mates—Pallika Kanani, Sameer Singh, Limin Yao, and Anton Bakalov for stimulating research discussions that inspired me to pursue a PhD. I am also grateful to letter writers at UMass—Prof. Victor Lesser, Prof. Ramesh Seetharaman, and Prof. Daeyoung Kim and

undergraduate letter writers—Dr. K. Sarulada, Dr. S. Saraswathi, and Dr. D. Loganathan, for their thoughtful letters, which helped me secure admissions for graduate study.

Finally, I thank my family for their love, support, and care, which made this PhD possible. My mother has always been my supreme source of inspiration and is one of the strongest women I have known in my life. I am highly indebted to her for her unconditional love, care, and encouragement whenever I needed them. I am also truly grateful to have a working mother, which not only inspired me always to go the extra mile but also shaped my perspectives on life in general. So thank you amma, for being you! I am also very thankful for my father's keen insight that led me to pursue a career in computer science. I could not have found a better suiting profession myself. His profound wisdom and principled life have been my guiding light throughout my school years. I am also very grateful to both of them for striving constantly to give me a better life by investing time and money in my health, education and well-being, often comprising on their comforts! I am also thankful to my learned grandmothers, who continually fought societal and familial restrictions to self-educate themselves. Whenever I felt bogged down in my PhD, I often thought of their struggles and it gave me the courage to push harder! Both my grandmothers passed away during the course of my PhD, but their spirit will always be alive in my work. I am also immensely thankful to my mother-in-law and father-in-law for their prayers and words of encouragement, which helped me especially during the final phase of my PhD. Special thanks are also due to my sister-in-law Sudha and brother-in-law Sougata for their well wishes. They understood several tricky life situations (such as two-body issues) more easily than others, filling the void of not having siblings of my own. Thanks are also due to my cute nephew Kuttan, whose sweet words served as a welcome relaxation on stressful days. Special thanks are also due to my uncle and aunt who helped me settle in the United States when I initially arrived.

My graduate expedition gave me the best things in life, including my husband Anand! I met Anand at UMass at a crucial juncture in my career when I was trying to decide between a job and PhD and I would be hardly exaggerating if I say it was primarily Anand's influence that led me to alter my career path and pursue a PhD. I am highly indebted to him for all the sacrifices he has made and ordeals he has undertaken for making this possible, and above all being very understanding and forgiving when I had to prioritize research over spending time together! I have benefitted tremendously from his "been there, done that" advice (though I often did not believe him immediately, leading to "I told you so" scenarios)! I am incredibly grateful for his continuous reassurance, often believing in me more than I believed in myself! I am also very lucky to have a spouse from the same field, who understands my research and I am immensely thankful for the countless dinner-time research discussions that made this PhD experience fun, enjoyable, and more memorable!

Table of Contents

Chapter 1: Introduction

The rapid growth and reach of internet and social media in the recent years has led to increase in avenues for socio-behavioral interactions. Various online platforms exist ranging from popular social networks such as *Facebook*, *Twitter*, *LinkedIn*, and *LivingSocial*, to countless other online discussion forums, such as *StackExchange*, and *Quora*, to name a few. The ever-increasing number of online interactions has lead to a growing interest to understand and interpret online interactions to enhance user experience. This includes personalization, user retention, and product and friend recommendations.

In this thesis, I present models for rich socio-behavioral interactions in two popular online interaction networks: online courses, and online professional networks. I focus on online courses in the early part, moving to online professional networks towards the end. Various forms of distance education are emerging; they extend high quality education from top universities to nooks and corners of the world, transforming lives and inspiring future generations. Of particular interest are massive open online courses (MOOCs)—online courses hosted by education companies such as Coursera, EdX, and Khan Academy, that are available to people around the world for free or limited cost. MOOCs are redefining the education system and transcending boundaries posed by traditional courses. The open nature of these online courses attract a wide range of students

from different nationalities, ethnicities, and education backgrounds. Structured data from these courses containing behavioral and interaction data from participants provides a tantalizing opportunity to study user interaction and develop methods to improve teaching and learning experience. Previous research in the field of education has been focused primarily on classroom settings involving small populations. With the rise of MOOCs, the opportunity is ripe for developing data-driven models for student behavior and interaction, extending existing research to large scale populations in MOOCs. I identify challenges and opportunities presented by MOOCs, and develop statistical relational learning based methods to improve the teaching and learning experience for MOOC participants.

Online professional networks are specialized social networks for professional networking that connect people to potential job opportunities, business partners, and industry experts. These networks have richer behavioral information from additional entities such as companies, and a wider range of user actions such as moving jobs, and adding skills. I develop methods to represent and reason about various user actions in these specialized networks and quantify the influence these actions have on the users' connections.

1.1 Thesis Contributions

The contributions of this thesis are as follows: 1) First, I demonstrate how to understand student engagement in MOOCs by creating a data-driven formulation for student engagement using latent variables and how to use the latent engagement models for predicting student success in MOOCs, 2) Second, I demonstrate the utility of content analysis of discussion forums by using it to predict student course completion, 3) Third, I develop

methods for predicting fine-grained issues and student opinion in discussion forums, 4) Then, I perform a temporal analysis to understand evolution of topics in MOOC discussion forums across multiple iterations of two long-running online courses, and 5) Finally, I present multi-relational models of influence for large online professional networks.

I focus on understanding and defining engagement in the context of online courses in Chapter 3. Maintaining and cultivating student engagement is critical for learning. Understanding factors affecting student engagement will help in designing better courses and improving student retention. The large number of participants in massive open online courses (MOOCs) and data collected from their interaction with the MOOC open up avenues for studying student engagement at scale. To this end, I develop a data-driven model for student engagement using latent variables and demonstrate that formulating engagement is helpful in predicting student success in MOOCs. My first contribution is the abstraction of student engagement types using latent representations and using that in a probabilistic model to connect student behavior with course success indicators. I identify two important course success indicators in MOOCs—earning a certificate (performance), and staying with the course till its completion (survival) and demonstrate that the latent formulation for engagement helps in predicting student success across three MOOCs. Next, in order to initiate better instructor interventions, I need to be able to predict student survival early in the course. I demonstrate that I can predict student survival early in the course reliably using the latent model. Finally, I perform a closer quantitative analysis of user interaction with the MOOC and identify student activities that are good indicators of student success at different points in the course.

Discussion forums serve as a platform for student discussions in massive open on-

line courses (MOOCs). With the increase in popularity of MOOCs, there is a corresponding increase in need to understand and interpret the communications of the course participants. Analyzing content in these forums can uncover useful information for improving student retention and help in initiating instructor intervention. In Chapter 4, I show how to understand content in discussion forums and use that to interpret student course completion abilities. I develop methods using topic models, particularly *seeded topic models* toward this goal. I demonstrate that content analysis of forum posts helps in predicting student survival in MOOCs.

In my analysis of forum posts, I find that a significant number of posts reporting issues go unanswered as they get lost in the mire of thousands of posts in the forums and this often leads to students dropping out of the course. Students often resort to asking fellow students to up-vote their posts to gain instructor's attention. Automatically inferring sentiment and topics of conversation (which I refer to as aspects) in problem-reporting posts would not only help instructors address the problems promptly, but also improve students' learning experience. Labeled aspect-sentiment data for MOOCs are expensive to obtain and may not be transferable between courses, suggesting the need for unsupervised/weakly supervised approaches. In Chapter 5, I present an weakly supervised joint framework for modeling course-related problems (course aspects) and the sentiment associated with them. I demonstrate how to model dependencies between various course aspects and sentiment and show that modeling the dependencies is helpful in detecting fine-grained course aspects.

In order to improve the quality of online courses, instructors need to actively monitor and discern patterns in previous iterations of the course and mould the course better to

suit the ever-changing student population. To enable this, in Chapter 6, I build on the analysis of forums in individual courses and develop models to track evolution of topics across repeated offerings of two long-running online courses. I leverage seeded topic models to perform a detailed analysis of evolution of fine-grained topics in online forums and draw important insights on the nature of students, types of issues, and student satisfaction by modeling the changing topic trends in the course across iterations. I run my models on discussion forums from multiple iterations of two successful long-running MOOCs: i) a business course, and ii) a computer science course. My methods uncover topic trends in both courses including the decline of logistic issues in both courses as the iterations unfold, decline in grading related issues when automatic grading is adopted in the business course, and prevalence and increase of technical issues in the computer science course compared to the business course. My models and analysis are useful for educators and instructors to model the progression of courses and understand how to fine-tune courses to meet student expectations.

In the penultimate chapter of this thesis, Chapter 7, I turn my attention to online professional networks. With professional networks, users have access to tremendous amount of information that influence many aspects of their lives from important life-changing decisions such as job changes to daily activities, and interests. Recently, there has been a growing interest in understanding influence in social networks. Previous work in this area characterize influence as propagation of actions in the social network. However, typically only a single action type is considered in characterizing influence. In online professional networks, users perform a wide variety of actions such as moving jobs, learning a new skill, and pursuing a certain career path, along with other actions commonly observed in

a regular social network such as adding connections, and following content. I present a holistic model to jointly represent different user actions and their respective propagations in online professional networks. My model captures node features such as user seniority in the network, and edge features such as connection strength to characterize influence. My model is capable of representing and combining different kinds of information users assimilate in the network and compute pairwise values of influence taking the different types of actions into account. I evaluate the models on data from online professional network, *LinkedIn* and show the effectiveness of the inferred influence scores in predicting user actions. I further demonstrate that modeling different user actions, node and edge relationships between people leads to around *20%* increase in precision at top k in predicting user actions, when compared to a model based only on general threshold model [Goyal *et al.*, 2010], which is the current state-of-the-art model for inferring influence values in online networks.

1.2 Thesis Organization

The rest of the thesis is organized as follows: Chapter 2 reviews work in the three areas – learning analytics, structured prediction, and topic modeling and describes the tools and methods I use in my work. Chapter 3–6 present work on online courses. In Chapter 3, I present my first work on student engagement in MOOCs, which was published in [Ramesh *et al.*, 2014a; Ramesh *et al.*, 2014b; Ramesh *et al.*, 2013]. In Chapter 4, I present work published in [Ramesh *et al.*, 2014c] on models for using forum content analysis to predict course completion. In Chapter 5, I present models for predicting fine-grained

aspect and sentiment in forums, published in [Ramesh *et al.*, 2015a], and in Chapter 6, I present temporal analysis of forum content over iterations of courses. Chapter 7 presents work published in [Ramesh *et al.*, 2015b], the influence model for online professional networks. I conclude the thesis by presenting future directions in Chapter 8.

Chapter 2: Related Work

My research touches a number of research areas: i) learning analytics, ii) probabilistic graphical models and structured prediction, and iii) topic models. I review work on the related areas below, delving into detail on the models and frameworks I use in my work.

2.1 Learning Analytics

Learning analytics is a rapidly growing area of computer science which uses computer science techniques to improve teaching and learning experience. The tremendous growth and popularity of MOOCs has opened avenues for performing large scale behavioral analysis of students. Various works analyze student dropouts in MOOCs [Kotsiantis *et al.*, 2003; Clow, 2013; Balakrishnan, 2013; Yang *et al.*, 2013]. Guo *et al.* [2014] perform an empirical analysis of how online educational videos affect student engagement. Bruff *et al.* [2013] evaluate the capability of MOOCs as a means to enhance classroom learning experience in a blended learning setting. Student engagement is known to be a significant factor in success of student learning [Kuh, 2003], but there is still limited work studying student engagement in MOOCs. Our work in Chapter 3 is closest to that of Kizilcec *et al.* [2013] and Anderson *et al.* [2014], who attempt to understand student engagement using completely unsupervised techniques (clustering). Qiu *et al.* [2016] analyze student online

behavioral patterns and present models to predict students' learning effectiveness.

There is a growing body of literature on analyzing content of MOOC discussion forums [Cui and Wise, 2015; Ezen-Can *et al.*, 2015; Wong *et al.*, 2015; Stump *et al.*, 2013a; Chaturvedi *et al.*, 2014a]. Stump *et al.* [2013b] propose a framework for taxonomically categorizing forum posts, leveraging manual annotations. Chaturvedi *et al.* [2014a] focus on predicting instructor intervention using lexicon features and thread features. Wong *et al.* [2015] analyze sentiment of forum posts and their relationship with students dropping out of the course. Coetzee *et al.* [2014] evaluate the usefulness of reputation systems in forums. Huang *et al.* [2014] analyze posting behavior in forums and draw correlations to engagement patterns exhibited by students posting in forums.

2.2 Probabilistic Graphical Models and Structured Prediction

Researchers in artificial intelligence and machine learning have long been interested in predicting interdependent unknowns using structural dependencies. Some of the earliest work in this area is inductive logic programming (ILP) [Lavrac and Dzeroski, 1994], in which structural dependencies are described with first-order logic. This enables the construction of intuitive, general-purpose models that are easily applicable or adapted to different domains. Inference then finds the structure(s) that satisfy the given logical constraints. However, ILP is limited by its difficulty in coping with uncertainty. Standard ILP approaches only model dependencies which hold universally, and such dependencies are rare in real-world data.

Another broad area of research, probabilistic models [Pearl, 1988], which provide

mechanisms for directly modeling uncertainty over unknowns. Probabilistic graphical models (PGMs) are a powerful way of modeling uncertainty by enabling compact representations of joint distributions over interdependent unknowns through graphical structures [Koller and Friedman, 2009]. Several approaches that combine the structured representation power of PGMs and rich feature sets of traditional classification techniques have been proposed. A few notable ones include conditional random fields [Lafferty *et al.*, 2001], max-margin Markov networks [Taskar *et al.*, 2003], SEARN [Daumé *et al.*, 2009], and SVM^{struct} [Tsochantaridis *et al.*, 2004].

Statistical relational learning (SRL) [Getoor and Taskar, 2007] builds on probabilistic graphical models and traditional ILP methods by creating effective representations that incorporate in a unified framework two central aspects of modeling in multi-relational domains on the one hand, these representations provide a language for expressing the structural regularities present in a domain, and on the other hand, they provide principled support for probabilistic inference. Several SRL models have been proposed — Markov Logic Networks [Richardson and Domingos, 2006], Relational Dependency Networks [Neville and Jensen, 2007], Sum Product Networks [Poon and Domingos, 2011]. In this thesis, we will explore the use a particular SRL framework—hinge-loss Markov random fields (HL-MRFs) [Bach *et al.*, 2015]. In the following section, I provide an overview of HL-MRFs and a templating language for HL-MRFs—Probabilistic Soft Logic (PSL) [Bach *et al.*, 2015].

2.2.1 Hinge-loss Markov random fields (HL-MRFs) and Probabilistic Soft Logic

Hinge-loss Markov random fields (HL-MRFs) are a scalable class of continuous, conditional graphical models [Bach *et al.*, 2015]. Inference of the most probable explanation in HL-MRFs is a convex optimization problem, which makes working with HL-MRFs very efficient in comparison to many relational modeling tools that use discrete representations. HL-MRFs have achieved state-of-the-art performance in many domains including knowledge graph identification [Pujara *et al.*, 2013], biomedicine and multi-relational link prediction [Fakhraei *et al.*, 2014], and modelling social trust [Huang *et al.*, 2013a]. The probability density function of HL-MRF probabilistic model is given by,

$$P(\mathbf{Y}|\mathbf{X}) \propto \exp\Big(-\sum_{r=1}^{M} \lambda_r \phi_r(\mathbf{Y}, \mathbf{X})\Big)$$
$$\phi_r(\mathbf{Y}, \mathbf{X}) = (\max\{l_r(\mathbf{Y}, \mathbf{X}), 0\})^{\rho_r} \ , \tag{2.1}$$

where $\phi_r(\mathbf{Y}, \mathbf{X})$ is a *hinge-loss potential* corresponding to an instantiation of a rule, and is specified by a linear function l_r and optional exponent $\rho_r \in \{1, 2\}$.

HL-MRFs admit various learning algorithms for fully-supervised training data, and are amenable to point-estimate "hard" expectation maximization for partially-supervised data with latent variables [Bach *et al.*, 2013]. Latent variables can improve the quality of probabilistic models in many ways. Using latent variables to mediate probabilistic interactions can improve generalization by simplifying models. HL-MRFs' capability in representing continuous latent variables is helpful in expressing more nuanced information

when compared to discrete latent variables. HL-MRFs trained with hard EM are accurate and scalable for three reasons: 1) the continuous variables of HL-MRFs can express complex, latent phenomena, such as mixed group memberships, which add flexibility and modeling power to these models, 2) fast, exact MPE inference for HL-MRFs can identify the most probable assignments to variables quickly, and 3) HL-MRFs can easily express dependencies among latent variables creating rich, interpretable models. In Chapter 3, we use this capability to represent student engagement types as a latent variables.

HL-MRF models can be specified using *Probabilistic Soft Logic (PSL)* [Bach *et al.*, 2015]. PSL is a framework for collective, probabilistic reasoning in relational domains, which uses syntax based on first-order logic as a templating language for continuous graphical models over random variables representing soft truth values. Like other statistical relational learning methods, PSL uses weighted rules to model the dependencies in a domain. However, one distinguishing aspect is that PSL uses continuous variables to represent truth values, relaxing Boolean truth values to the interval [0,1]. Triangular norms, which are continuous relaxations of logical connectives AND and OR, are used to combine the atoms in the first-order clauses. As a result of the soft formulation and the triangular norms, the underlying probabilistic model is an HL-MRF[Bach *et al.*, 2015]. An example of a PSL rule is

$$\lambda : P(a) \wedge Q(a, b) \rightarrow R(b),$$

where *P*, *Q*, and *R* are predicates, *a* and *b* are *variables*, and λ is the weight associated with the rule. Inference in HL-MRFs is a convex optimization problem, which makes work-

ing with PSL very efficient in comparison to relational modeling tools that use discrete representations.

2.3 Topic Models

Topic models are statistical models that uncover the hidden abstract semantic structures that occur in document collections. These models are a convenient means to analyze large volumes of text. The earliest known topic model is probabilistic latent semantic analysis (PLSA) [Hofmann, 1999]. PLSA models co-occurrence information under a probabilistic framework to discover the underlying semantic structure of the documents. Blei *et al.* [2003a] developed a generalization of PLSA by using Dirichlet priors to model documents as a mixture of topics. In LDA, each document is modeled as a multinomial distribution over topics, where topics are characterized by multinomial distribution over words. The posterior distribution of latent variables in LDA given the observed documents provides the topic distribution of the collection. Traditional LDA is unsupervised and is applied on the document collection without specifying the nature and type of documents.

The LDA generative process captures the interaction between the observed documents and and the hidden topic structure. Let K be the number of topics, V the vocabulary size, $\vec{\alpha}$ a K-dimensional vector. Let $Dir_V(\vec{\alpha})$ denote a V-dimensional Dirichlet with vector parameter $\vec{\alpha}$ and Dir($\vec{\beta}$) denote a K-dimensional Dirichlet with scalar parameter β. Then, the generative process of LDA is given by,

(1) For each topic,
    (a) Draw a distribution over words $\vec{\phi}_k \sim Dir_V(\vec{\beta})$.
(2) For each document,

(a) Draw a vector of topic proportions $\vec{\theta}_d \sim Dir(\vec{\alpha})$.

(b) For each word,

(i) Draw a topic assignment $Z_{d,n} \sim Mult(\vec{\theta}_d), Z_{d,n} \in 1, ..., K$.

(ii) Draw a word $W_{d,n} \sim Mult(\vec{\phi}_{Z_d,n}), W_{d,n} \in 1, ..., V$.

The latent variable $\vec{\phi}_{1:K}$ represents the topics, $\vec{\theta}_{1:D}$ represents the per-document topic proportions, and $z_{1:D,1:N}$ gives the topic assignments for each word. LDA belongs to a class of mixed-membership models, but are different from classical mixture models where each document is limited to one topic. Document often exhibit multiple topics, LDA can model this additional structure while classical models cannot.

Many extensions to LDA that model the temporal evolution of topics over time have been proposed over the last few years. The topics over time model (ToT) [Wang and McCallum, 2006] assumes that each document chooses its own time stamp based on a topic-specific beta distribution. Each document is a multinomial distribution over topics sampled from Dirichlet and the Beta distribution of each topic generates the document's time stamp. ToT generates a narrow/broad time-distribution topic based on the the duration of a strong word co-occurrence pattern in the documents. The dynamic topic model (DTM) [Blei and Lafferty, 2006] represents documents from each time frame as generated from a normal distribution over topics. The DTM uses a Gaussian prior for the topic parameters instead of Dirichlet prior and can capture the topic evolution over time slices. Multiscale topic tomography [Nallapati *et al.*, 2007] assumes that the topic collection is sorted and grouped into equal-size chunks corresponding to each epoch. This is similar to DTM in formulation but allows more flexibility of studying topic evolution over various time scales. Topic evolution has been applied in many domains, one popular domain

is scientific literature. Jo *et al.* [2011] employ citations to capture relationships between topics and their evolution over time.

The unsupervised nature of LDA often leads to scenarios where the topic distributions do not accurately capture the underlying topic structure of the corpus. Seeded LDA [Jagarlamudi *et al.*, 2012] is another variant of LDA in which user-specified seed words are used to guide topic discovery in documents. The seed words capture specific topics the user is expecting to find in the corpus, thus guiding LDA into finding topics in accordance with the seed words. The Seeded LDA model biases both the topic-word and document-topic distribution using seed words. At the word level, the model uses seed words to bias topics toward producing the given seed words. At the document level, the model influences documents to select topics that contain the seed words.

Acronym	Expansion
HL-MRFs	Hinge-loss Markov Random Fields
PSL	Probabilistic Soft Logic
LDA	Latent Dirichlet Allocation
GTM	General Threshold Model
EM	Expectation Maximization

Table 2.1: Common acronyms

Table 2.1 gives the commonly used acronyms in this thesis for quick reference.

Chapter 3: Latent Variable Models for Student Engagement in MOOCs

3.1 Introduction

The large number of students participating in MOOCs provides the opportunity to perform rich analysis of large-scale online interaction and behavioral data. This analysis is useful in improving student engagement in MOOCs by identifying patterns, suggesting new feedback mechanisms, and guiding instructor interventions. Additionally, insights gained by analyzing online student engagement can also help validate and refine our understanding of engagement in traditional classrooms.

In this chapter, we study the different aspects of online student behavior in MOOCs, develop a large-scale, data-driven approach for modeling student engagement. We use two course *success indicators* for online courses—1) *performance*: whether the student earns a certificate in the course, and 2) *survival*: whether the student follows the course to completion. We demonstrate the construction of a holistic model incorporating content (e.g., language), structure (e.g., social interactions in discussion forums), and outcome data and show that jointly measuring different aspects of student behavior early in the course can provide a strong indication of course success indicators.

Predictive modeling over MOOC data poses a significant technical challenge requiring the ability to combine language analysis of forum posts with graph analysis over very

large networks of entities (students, instructors, assignments, etc.). To address this challenge, we use a recently developed graphical modeling framework—*hinge-loss Markov random fields* (HL-MRFs) [Bach *et al.*, 2015]. This framework provides an easy means to represent and combine behavioral, linguistic, and structural features in a concise manner. Our first contribution is constructing a holistic model to represent and reason about various student activities in the MOOC setting. Our work is a step toward helping educators understand how students interact on MOOCs.

Our second contribution is providing a data-driven formulation that captures student engagement in the MOOC setting. As in the traditional classroom setting, assessing online student engagement requires interpretation of indirect cues. Identifying these cues in an electronic setting is challenging, but the large amounts of available data can offset the loss of in-person communication. We analyze students' online behavior to identify how they engage with course materials and investigate how engagement can be helpful in predicting student performance and successful completion of the course. We extend our HL-MRF model to encode engagement as *latent* variables, which take into account the observed behaviors of online students and their resulting *performance* and *completion* in the class. The latent engagement variables in our model represent three prominent forms of engagement: 1) active engagement, 2) passive engagement, and 3) disengagement. Uncovering these different latent engagement states for students provides a better explanation of students' behavior leading to course completion and resulting grades.

Examining real MOOC data, we observe that there are several indicators useful for gauging students' engagement, such as viewing course content, interacting with other learners or staff on the discussion forums, and the topic and tone of these interactions.

Furthermore, students often engage in different aspects of the course throughout its duration. For example, some students engage in the social aspects of the online community by posting in forums and asking and answering questions, while others only watch lectures and take quizzes without interacting with the community. We take these differences into account and propose models that use the different behavioral aspects to distinguish between forms of engagement: passive, active, and disengagement. We use these engagement types to predict student success, and reason about their behavior over time.

We apply our models to real data collected from seven Coursera* courses and empirically show their ability to capture behavioral patterns of students and predict student success. Our experiments validate the importance of providing a holistic view of students' activities, combining all aspects of online behavior, in order to accurately predict the students' motivation and ability to succeed in the class. We conduct experiments to evaluate two important course success parameters in online courses: course grades (*performance*) and course completion (*survival*). Early detection of changes in student engagement can help educators design interventions and adapt the course presentation to motivate students to continue with the course [Brusilovsky and Millán, 2007]. We show that our models are able to make meaningful predictions using data obtained at an *early* stage in the class. These predictions can help provide the basis for instructor intervention at an early stage in the course, helping to improve student retention rates. Further, we evaluate the importance of each class of feature in predicting student success in MOOCs in different time periods of the course. Our findings strengthen the importance of using a holistic model and uncover important details about student interactions that is helpful for instructors. Fi-

*https://www.coursera.org

nally, we use the latent engagement variables to unearth patterns in student engagement over the course of the class and detect changes in engagement. Our analysis can potentially be used by instructors to understand student movement from one engagement type to another and initiate interventions.

3.2 Related Work

In this section, we discuss prior work related to our work in this chapter. These can be classified into two broad categories: 1) work on classroom and traditional distance education settings, and 2) work on larger settings such as MOOCs. Much of the work before MOOCs concentrate on understanding student engagement using various forms of instructor intervention experiments in classroom settings. Rocca [2010] presents an analysis of student engagement in classroom settings, comparing the effects of different methods of teaching on student participation. These studies primarily analyze the effectiveness of various instructor intervention techniques and teaching methodologies on getting students to participate in classroom discussions. Further, these studies primarily refer to participation in classroom discussions as student engagement. Other forms of student engagement such as attending lectures and giving exams are considered integral part of the class. However, in online settings, the diverse population of the students leads to varied participation levels. This calls for a more nuanced notion of engagement. Drawing analogies from classroom settings and carefully considering student dynamics in online settings, we model three types of student engagement. We refer to participating in discussion forums, which is analogous to participating in classroom discussions as *active engagement*. We refer to

following class materials and tests as *passive engagement* and dropping out of the class as *disengagement*. Kuh [2003] and Carini *et al.* [2006] study the relationship between student engagement and academic performance for traditional classroom courses; they identify several metrics for user engagement (such as student-faculty interaction, level of academic challenge). Carini *et al.* [2006] demonstrate quantitatively that though most engagement metrics are positively correlated to performance, the relationships in many cases can be weak. Our work borrows ideas from Kuh [2003], Carini *et al.* [2006], and from statistical survival models [Richards, 2012] and adapts these to the MOOC setting.

There is also a growing body of work in the area of learning analytics. Various works analyze student dropouts in MOOCs [Kotsiantis *et al.*, 2003; Clow, 2013; Balakrishnan, 2013; Yang *et al.*, 2013]. Our work differs from these in that we analyze a combination of several factors that contribute to student engagement and hence their survival in online courses. We argue that analyzing the ways in which students engage themselves in different phases of online courses can reveal information about factors that lead to their continuous survival. This will pave the way for constructing better quality MOOCs, which will then result in increase in enrollment and student retention. In this work, we analyze the different course-related activities and reason about important factors in determining student survival at different points in the course.

Student engagement is known to be a significant factor in success of student learning [Kuh, 2003], but there is still limited work studying student engagement in MOOCs. Our work is closest to that of Kizilcec *et al.* [2013] and Anderson *et al.* [2014], who attempt to understand student engagement using completely unsupervised techniques (clustering). Our work differs from the above work in that we view types of engagement as

latent variables and learn to differentiate among the engagement types from data. We evaluate two different student success measures in MOOCs—whether the student earns a certificate (*performance*) and whether the student follows the course till the end (*survival*). We use these two student success measures to train the model. We then use this model to predict student success in MOOCs. We model engagement explicitly and demonstrate that it helps in predicting student success.

3.3 Hinge-loss Markov Random Fields

To model the different types of interactions between features and course success, we propose a powerful approach using HL-MRFs. Our HL-MRF model is built on a foundation of observable features from the data and encoded in the templating language, PSL. In Section 3.4.1, we detail the various features we collect from the data. To reason in the first-order logic based syntax of PSL, we encode these features as logical predicates. PSL enables us to encode our observed features and (*latent and target*) variables as logical predicates and design models by writing rules over these predicates. PSL interprets these rules in a parameterized probability model and is able to perform efficient inference and parameter fitting using machine learning algorithms. The expressiveness and flexibility of PSL allows us to easily build different models for MOOC data, and we exploit this by comparing a model that represents multiple forms of latent engagement against a simpler model that directly relates the observable features to student success.

For example, to encode the different behavioral interactions, let U_1 and U_2 be two students interacting in the same thread in the forum, posting posts P_1 and P_2, respectively.

Predicates POST(U_1, P_1) and POST(U_2, P_2) denote student U_1 posting P_1, and U_2 posting P_2. The predicate SAMETHREAD(P_1, P_2) captures if posts P_1 and P_2 are in the same thread. The PSL rule below captures the influence students have on each other when interacting in the forums. Students U_1 and U_2 post in the same threads, hence influencing each other to have similar succeeding abilities.

$$\lambda : \text{POST}(U_1, P_1) \wedge \text{POST}(U_2, P_2) \wedge \text{SAMETHREAD}(P_1, P_2) \wedge \text{SUCCESS}(U_1) \rightarrow \text{SUCCESS}(U_2).$$

We can generate more complex rules connecting the different features and latent variables, which we will demonstrate in Section 3.4.1.4. The HL-MRF model uses these rules to encode domain knowledge about dependencies among the predicates. The continuous value representation further helps in understanding the confidence of predictions.

3.4 Student Success Prediction Models

As students interact on a MOOC, detailed records are generated, including page and video views, forum visits, forum interactions such as voting, posting messages and replies, and graded elements such as quizzes and assignments. In this section, we develop our models for predicting student success in MOOCs. Our models connect performance indicators to complex behavioral, linguistic, temporal, and structural features derived from the raw student interactions. Our first model, referred as the DIRECT model, directly encodes the dependence between student interactions and student success in MOOCs. We then extend the DIRECT model by adding latent variables modeling three types of student engagement:

1) active engagement, 2) passive engagement, and 3) disengagement. We refer to this model as the LATENT model. In the LATENT model, we capture dependencies among student interactions, their different types of engagement, and success measures.

We evaluate the models by employing them to predict student success in MOOCs. We consider two course success indicators in MOOCs: 1) *performance*: whether the student earns a certificate in the course, and 2) *survival*: whether the student follows the course till the end.

3.4.1 Modeling MOOC Student Activity

MOOC students interact with with two main resources on the MOOC website: video lectures and forums. Students can watch lectures multiple times and respond to on-demand quizzes during the lectures[†]. Students can interact by asking and responding to questions in the forums. There are typically multiple forums organized by topics, each consisting of multiple threads, and each thread consisting of multiple posts. Students can respond, vote (up or down) on existing posts and subscribe for updates to forums threads. Each student is given a reputation score based on the votes on posts created by the student. These activities are depicted in Figure 3.1.

We quantify these activities by defining a set of PSL predicates over the raw student data, and capture more complex behaviors by combining these predicates into expressive rules, used as features in our predictive models. We categorize these predicates as either behavioral, linguistic, structural, or temporal, and describe them in the following sections.

[†]These quizzes are generally not used to calculate the final evaluation.

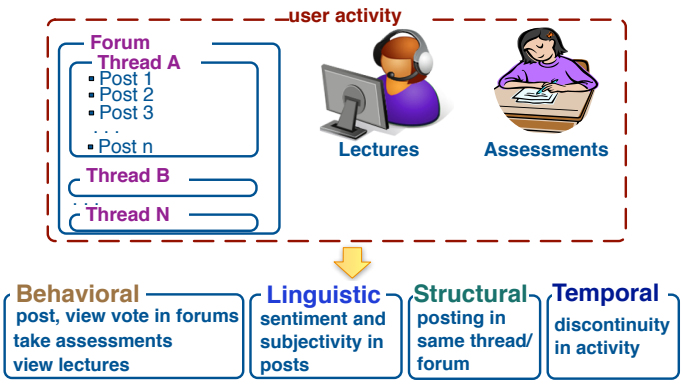


Figure 3.1: Structure of MOOC student activity.

3.4.1.1 Behavioral Features

Behavioral features are derived from various activities that students engage in while interacting on the MOOC website. These features measure the different levels of activity of MOOC participants on the site. We consider three types of student interactions on the discussion forums: posting in the forums, voting on forum posts, and viewing forum posts. We consider two types of behavioral features: aggregate and non-aggregate. Aggregate features are predicates comparing students' activity level to the median. The predicates POST-ACTIVITY(USER), VOTE-ACTIVITY(USER) and VIEW-ACTIVITY(USER) represent aggregate features capturing student activity in the forums. In addition to that, we also measure the reputation of student in the forum taking into account, the total number of upvotes/downvotes gained by the student across all the posts. We refer to this aggregate feature as REPUTATION(USER) in our model. Non-aggregate features directly quantify student's behavior. The predicates POSTS(USER, POST) and VOTES(USER, POST) capture

an instance-level log of users posting and voting on the discussion forums. The predicates POSTS and VOTES are true if the USER posts or votes on POST. Predicate UPVOTE(POST) is true if the post has positive votes and false otherwise, and predicate DOWNVOTE(POST) is true if a post has been down-voted.

The second class of behavioral features capture students' interaction with lectures and quizzes on the MOOC website. We measure the percentage of lectures and accompanying quizzes that were submitted by the student in the course. The features LECTURE-SUBMITTED(USER) captures the fraction of lectures submitted by the student in the course. The feature LECTURE-SUBMITTED-ONTIME(USER) captures the fraction of lectures submitted by the student within the due date. Similarly, for quizzes we derive QUIZ-SUBMITTED and QUIZ-SUBMITTED-ONTIME(USER). These predicates are continuous valued in [0, 1].

3.4.1.2 Forum Content and Interaction Features

MOOC forums are rich with relevant information, indicative of the students' attitudes toward the course and its materials as well as the social interactions between students. We capture this information using two types of features, *linguistic* features capturing the sentiment of the post content, and *structural* features capturing the forum structure, organized topically into threads and forums types.

The attitudes expressed by students on the forums can be captured by estimating sentiment polarity (positive or negative) and identifying subjective posts. Since MOOC forums contain thousands of posts, we use an automated tool, *OpinionFinder* [Wilson *et al.*, 2005a] to avoid manual annotation. The tool segments the forums posts into sentences,

and assigns subjectivity and polarity tags for each sentence. Based on its predictions, we define two predicates, SUBJECTIVE(POST) and POLARITY(POST). Both predicates are calculated by normalizing the number of subjective/objective tags and positive/negative polarity tags marked by OpinionFinder. The normalization keeps these values in the $[0, 1]$ interval.

Forums are structured entities, organized by high-level topics (at the forum level) and specific topics (thread level). Including these structural relationships allows our model to identify structural relations between forum posts and connect them with students participating in the forum discussions. The predicates representing forum structure are SAME-THREAD(POST_1, POST_2) and SAME-FORUM(THREAD_1, THREAD_2), which are true for posts in the same thread and threads in the same forum, respectively. These predicates capture forum interaction among students and propagate *performance*, *survival* and *engagement* values among them.

3.4.1.3 Temporal Features

Student activity levels change over the span of the course. Students are often active at early stages and lose interest as the course progresses. To include signals of how student activity changes over time, we introduce a set of temporal features. We divide the course into three time periods: *start*, *mid*, and *end*. The time period splits are constructed by dividing the course by duration into three equal chunks. The temporal features LAST-QUIZ, LAST-LECTURE, LAST-POST, LAST-VIEW and LAST-VOTE indicate the time-period in which each last interaction of the user occurred. These features measure to what lengths the user

participated in different aspects of the course.

3.4.1.4 Constructing Complex Rules

We use the features above to construct meaningful PSL rules using logical connectives, as demonstrated in Table 3.1. We construct meaningful combinations of predicates to model student engagement and student success. For example, the first rule in Table 3.1 combines the posting activity of user U relative to other students in the class (POST-ACTIVITY) with reputation of the user in the forums to infer student success. This rule captures that students posting high-quality posts (given by reputation) show greater signs of succeeding in the class. This is helpful in discerning between students who post a lot and students who post few highly upvoted posts. Similarly, the third rule combines posting in forums and the polarity of forum posts to capture that students posting positive sentiment posts are more likely to engage and succeed in the course. The PSL models associate these rules with student survival, either directly or indirectly using latent variables. We explain this process in Section 3.5.

• Behavioral Features
POST-ACTIVITY(U) $\wedge$ REPUTATION(U)
LECTURE-SUBMITTED(U) $\wedge$ LECTURE-SUBMITTED-ONTIME(U)
• Forum Content Features
POSTS(U, P) $\wedge$ POLARITY(P)
• Forum Interaction Feature
POSTS(U_1, P_1) $\wedge$ POSTS(U_2, P_2) $\wedge$ SAME-THREAD(P_1, P_2)
• Temporal Features
LAST-QUIZ(U, T_1) $\wedge$ LAST-LECTURE(U, T_1) $\wedge$ LAST-POST(U, T_1)

Table 3.1: Constructing complex rules in PSL

3.4.2 Student Engagement in MOOCs

Student engagement cannot be directly measured from the data. We therefore treat student engagement as latent variables and associate various observed features to one or more forms of engagement. We define three types of engagement variables, denoted ACTIVE-ENGAGEMENT, PASSIVE-ENGAGEMENT and DISENGAGEMENT to capture three types of student engagement in MOOCs. ACTIVE-ENGAGEMENT represents students actively engaged in the course by participating in the forums, PASSIVE-ENGAGEMENT represents students following the class materials but not making an active presence in the forums, and DISENGAGEMENT represents students discontinuing from engaging with the course both actively or passively. We associate different features representing MOOC attributes relevant for each engagement type.

• **Active Engagement** Actively participating in course-related discussions by posting in the forums are signs of active engagement.

• **Passive Engagement** Passively following course material by viewing lectures, viewing/voting/subscribing to posts on discussion forums, and giving quizzes are signs of passive engagement.

• **Disengagement** Temporal features, indicating the last point of user's activity, capture signs of disengagement.

3.5 PSL Models for Student Success Prediction

We construct two different PSL models for predicting student success in a MOOC setting—first, a model (denoted DIRECT) that directly infers student success from observable features, and second, a latent variable model (LATENT) that infers student engagement as a hidden variable to predict student success. By building both models, we are able to evaluate the contribution of the abstraction created by formulating engagement patterns as latent variables.

3.5.1 PSL-DIRECT

In PSL-DIRECT model, we model student success by using the observable behavioral features exhibited by the student, linguistic features corresponding to the content of posts, structural features derived from forum interactions, and temporal features capturing discontinuity in activity. Meaningful combinations of one or more observable behavioral, linguistic, temporal, and structural features are constructed as described in Section 3.4.1 and they are used to predict student SUCCESS. Table 3.2 contains the rules used in the DIRECT model. U and P in tables 3.2, 3.3, and 3.4 refer to USER and POST respectively. The DIRECT model rules allow observable features to directly imply student success. The rules are grouped into four groups based on the features present in them. The first group of rules presents different combinations of student interactions with the three course elements: discussion forums, lectures, and quizzes, to predict student success indicated by SUCCESS. The second group of rules combine the behavioral features with the linguistic features to predict student success. The third set of rules capture the structural interactions

PSL-DIRECT RULES

Rules combining behavioral features

POST-ACTIVITY(U) $\wedge$ REPUTATION(U) $\rightarrow$ SUCCESS(U)

VOTE-ACTIVITY(U) $\wedge$ REPUTATION(U) $\rightarrow$ SUCCESS(U)

VIEW-ACTIVITY(U) $\wedge$ REPUTATION(U) $\rightarrow$ SUCCESS(U)

POST-ACTIVITY(U) $\wedge$ VIEW-ACTIVITIY(U) $\wedge$ VOTE-ACTIVITIY(U) $\rightarrow$ SUCCESS(U)

$\neg$POST-ACTIVITY(U) $\rightarrow$ $\neg$SUCCESS(U)

$\neg$VOTE-ACTIVITY(U) $\rightarrow$ $\neg$SUCCESS(U)

$\neg$VIEW-ACTIVITY(U) $\rightarrow$ $\neg$SUCCESS(U)

POST-ACTIVITY(U) $\wedge$ $\neg$REPUTATION(U) $\rightarrow$ $\neg$SUCCESS(U)

POSTS(U, P) $\wedge$ REPUTATION(U) $\rightarrow$ SUCCESS(U)

SUBMITTED-LECTURE(U) $\rightarrow$ SUCCESS(U)

$\neg$SUBMITTED-LECTURE(U) $\rightarrow$ $\neg$SUCCESS(U)

SUBMITTED-LECTURE(U) $\wedge$ ONTIME(U) $\rightarrow$ SUCCESS(U)

SUBMITTED-LECTURE(U) $\wedge$ $\neg$ONTIME(U) $\rightarrow$ $\neg$SUCCESS(U)

SUBMITTED-QUIZ(U) $\rightarrow$ SUCCESS(U)

$\neg$SUBMITTED-QUIZ(U) $\rightarrow$ $\neg$SUCCESS(U)

SUBMITTED-QUIZ(U) $\wedge$ ONTIME-QUIZ(U) $\rightarrow$ SUCCESS(U)

SUBMITTED-QUIZ(U) $\wedge$ $\neg$ONTIME-QUIZ(U) $\rightarrow$ $\neg$SUCCESS(U)

SUBMITTED-QUIZ(U) $\wedge$ SUBMITTED-QUIZ(U) $\rightarrow$ SUCCESS(U)

Rules combining behavioral and linguistic features

POSTS(U, P) $\wedge$ POLARITY(P) $\rightarrow$ SUCCESS(U)

POSTS(U, P) $\wedge$ $\neg$POLARITY(P) $\rightarrow$ $\neg$SUCCESS(U)

Rules combining behavioral and structural features

POSTS(U_1, P_1) $\wedge$ POSTS(U_2, P_2) $\wedge$ SUCCESS(U_1) $\wedge$ SAME-THREAD(P_1, P_2) $\rightarrow$ SUCCESS(U_2)

POSTS(U_1, P_1) $\wedge$ POSTS(U_2, P_2) $\wedge$ SUCCESS(U_1) $\wedge$ SAME-FORUM(P_1, P_2) $\rightarrow$ SUCCESS(U_2)

Rules combining behavioral and temporal features

LAST-POST(U, *start*) $\rightarrow$ $\neg$SUCCESS(U)

LAST-LECTURE(U, *start*) $\rightarrow$ $\neg$SUCCESS(U)

LAST-QUIZ(U, *start*) $\rightarrow$ $\neg$SUCCESS(U)

LAST-POST(U, *mid*) $\rightarrow$ $\neg$SUCCESS(U)

LAST-LECTURE(U, *mid*) $\rightarrow$ $\neg$SUCCESS(U)

LAST-QUIZ(U, *mid*) $\rightarrow$ $\neg$SUCCESS(U)

LAST-POST(U, *end*) $\rightarrow$ $\neg$SUCCESS(U)

LAST-LECTURE(U, *end*) $\rightarrow$ SUCCESS(U)

LAST-LECTURE(U, *end*) $\rightarrow$ $\neg$SUCCESS(U)

LAST-QUIZ(U, *end*) $\rightarrow$ SUCCESS(U)

LAST-QUIZ(U, *end*) $\rightarrow$ $\neg$SUCCESS(U)

LAST-QUIZ(U, *end*) $\wedge$ LAST-LECTURE(U, *end*) $\wedge$ LAST-POST(U, *end*) $\rightarrow$ SUCCESS(U)

LAST-QUIZ(U, *end*) $\wedge$ LAST-LECTURE(U, *end*) $\wedge$ LAST-POST(U, *end*) $\rightarrow$ $\neg$SUCCESS(U)

Table 3.2: Rules from the PSL-DIRECT model

of students with other fellow students in the forums and how that impacts each other's course succeeding capabilities. The last set of rules capture the interaction between behavioral and temporal features.

3.5.2 PSL-LATENT

In the LATENT model, we enhance reasoning in the DIRECT model by including latent variables semantically based on concepts of student engagement as outlined in Section 3.4.2. We introduce three latent variables ACTIVE-ENGAGEMENT, PASSIVE-ENGAGEMENT, and DISENGAGEMENT to capture the three different types of student engagement. We present the LATENT model in two parts in Tables 3.3 and 3.4. In Table 3.3, we present rules connecting observable features to different forms of engagement. Note that the rules are identical to the rules in the DIRECT model presented in Table 3.2, but in the LATENT model they are changed to imply the latent engagement variables instead of student success.

In this model, some of the observable features (e.g, POST-ACTIVITY, VOTE-ACTIVITY, VIEW-ACTIVITY) are used to classify students into one or more forms of engagement or disengagement. For example, in Table 3.3, conjunction of POST-ACTIVITY and REPUTATION implies ACTIVE-ENGAGEMENT; conjunction of VOTE-ACTIVITY and REPUTATION implies PASSIVE-ENGAGEMENT. Rules that combine observed features that are indicative of more than one form of engagement, such as POST-ACTIVITY and VOTEACTIVITY, are left unchanged from the DIRECT model to directly imply SUCCESS. We then connect the latent engagement variables to student success using the rules in Table 3.4. For exam-

PSL-LATENT RULES (PART 1)

Rules combining behavioral features

POST-ACTIVITY(U) ∧ REPUTATION(U) → ACTIVE-ENGAGEMENT(U)

VOTE-ACTIVITY(U) ∧ REPUTATION(U) → PASSIVE-ENGAGEMENT(U)

VIEW-ACTIVITY(U) ∧ REPUTATION(U) → PASSIVE-ENGAGEMENT(U)

POST-ACTIVITY(U) ∧ VIEW-ACTIVITIY(U) ∧ VOTE-ACTIVITIY(U) → SUCCESS(U)

REPUTATION → SUCCESS(U)

¬POST-ACTIVITY(U) → ¬ACTIVE-ENGAGEMENT(U)

¬VOTE-ACTIVITY(U) → ¬PASSIVE-ENGAGEMENT(U)

¬VIEW-ACTIVITY(U) → ¬PASSIVE-ENGAGEMENT(U)

POST-ACTIVITY(U) ∧ ¬REPUTATION(U) → ¬ACTIVE-ENGAGEMENT(U)

POSTS(U, P) ∧ REPUTATION(U) → ACTIVE-ENGAGEMENT(U)

SUBMITTED-LECTURE(U) → PASSIVE-ENGAGEMENT(U)

¬SUBMITTED-LECTURE(U) → ¬PASSIVE-ENGAGEMENT(U)

SUBMITTED-LECTURE(U) ∧ ONTIME(U) → PASSIVE-ENGAGEMENT(U)

SUBMITTED-LECTURE(U) ∧ ¬ONTIME(U) → ¬PASSIVE-ENGAGEMENT(U)

SUBMITTED-LECTURE(U) ∧ POST-ACTIVITY(U) → PASSIVE-ENGAGEMENT(U)

SUBMITTED-QUIZ(U) → PASSIVE-ENGAGEMENT(U)

SUBMITTED-QUIZ(U) → ¬PASSIVE-ENGAGEMENT(U)

SUBMITTED-QUIZ(U) ∧ ONTIME-QUIZ(U) → PASSIVE-ENGAGEMENT(U)

Rules combining behavioral and linguistic features

POSTS(U, P) ∧ POLARITY(P) → ACTIVE-ENGAGEMENT(U)

POSTS(U, P) ∧ ¬POLARITY(P) → ¬ACTIVE-ENGAGEMENT(U)

Rules combining behavioral and structural features

POSTS(U_1, P_1) ∧ POSTS(U_2, P_2) ∧ ACTIVE-ENGAGEMENT(U_1) ∧ SAME-THREAD(P_1, P_2) → ACTIVE-ENGAGEMENT(U_2)

POSTS(U_1, P_1) ∧ POSTS(U_2, P_2) ∧ ACTIVE-ENGAGEMENT(U_1) ∧ SAME-FORUM(P_1, P_2) → ACTIVE-ENGAGEMENT(U_2)

Rules combining behavioral and temporal features

LAST-POST(U, *start*) → DISENGAGEMENT(U)

LAST-LECTURE(U, *start*) → DISENGAGEMENT(U)

LAST-QUIZ(U, *start*) → DISENGAGEMENT(U)

LAST-POST(U, *mid*) → DISENGAGEMENT(U)

LAST-LECTURE(U, *mid*) → DISENGAGEMENT(U)

LAST-QUIZ(U, *mid*) → DISENGAGEMENT(U)

LAST-POST(U, *end*) → DISENGAGEMENT(U)

LAST-POST(U, *end*) → ACTIVE-ENGAGEMENT(U)

LAST-LECTURE(U, *end*) → DISENGAGEMENT(U)

LAST-LECTURE(U, *end*) → PASSIVE-ENGAGEMENT(U)

LAST-QUIZ(U, *end*) → DISENGAGEMENT(U)

LAST-QUIZ(U, *end*) → PASSIVE-ENGAGEMENT(U)

LAST-QUIZ(U, *end*) ∧ LAST-LECTURE(U, *end*) ∧ LAST-POST(U, *end*) → SUCCESS(U)

LAST-QUIZ(U, *end*) ∧ LAST-LECTURE(U, *end*) ∧ LAST-POST(U, *end*) → ¬SUCCESS(U)

Table 3.3: Rules from the PSL-LATENT model capturing dependencies between observed features and latent engagement variables

PSL-LATENT RULES (PART 2)
Rules combining latent engagement variables
PASSIVE-ENGAGEMENT(U) $\rightarrow$ SUCCESS(U)
$\neg$PASSIVE-ENGAGEMENT(U) $\rightarrow$ $\neg$SUCCESS(U)
ACTIVE-ENGAGEMENT $\rightarrow$ SUCCESS(U)
$\neg$ACTIVE-ENGAGEMENT $\rightarrow$ $\neg$SUCCESS(U)
PASSIVE-ENGAGEMENT(U) $\wedge$ ACTIVE-ENGAGEMENT $\rightarrow$ SUCCESS(U)
PASSIVE-ENGAGEMENT(U) $\wedge$ $\neg$ACTIVE-ENGAGEMENT $\rightarrow$ SUCCESS(U)
PASSIVE-ENGAGEMENT(U) $\wedge$ $\neg$ACTIVE-ENGAGEMENT $\rightarrow$ $\neg$SUCCESS(U)
$\neg$PASSIVE-ENGAGEMENT(U) $\wedge$ ACTIVE-ENGAGEMENT $\rightarrow$ SUCCESS(U)
$\neg$PASSIVE-ENGAGEMENT(U) $\wedge$ ACTIVE-ENGAGEMENT $\rightarrow$ $\neg$SUCCESS(U)
$\neg$PASSIVE-ENGAGEMENT(U) $\wedge$ $\neg$ACTIVE-ENGAGEMENT $\rightarrow$ $\neg$SUCCESS(U)
DISENGAGEMENT $\rightarrow$ $\neg$SUCCESS(U)

Table 3.4: Rules from the PSL-LATENT model capturing dependencies between latent engagement variables and student success

ple, ACTIVE-ENGAGEMENT and PASSIVE-ENGAGEMENT implies SUCCESS. We consider various combinations of engagement and their relationship to SUCCESS. For example, exhibiting both passive and active forms of engagement implies SUCCESS. Also, exhibiting only one form of engagement, either active or passive, implies SUCCESS. We train the weights for the model by performing expectation maximization with SUCCESS as the target variable. The weighted combinations of different engagement types encodes variations in student engagement types and their relationship to student success. We consider two measures of success—1) performance, and 2) survival. In Section 4.4, we present results from training and testing our models on these two success measures. The resulting model with latent engagement suggests which forms of engagement are good indicators of student success. We demonstrate that the LATENT model not only produces better predictive performance, but also provides more insight into MOOC user behavior when compared to the DIRECT model.

3.6 Empirical Evaluation

We conduct experiments to answer the following questions. First, how effective are our models at predicting student performance and student survival in online courses? Second, how effective are our models at predicting student survival considering student interactions only from early part of the course? Third, we evaluate the importance of the different classes of features we use in our models in predicting student success in different time periods in the course. Then, we qualitatively analyze student engagement values at different points in the course to uncover engagement patterns. Lastly, we analyze forum posts by students with different kinds of engagement and discuss the sentiment in the posts as it pertains to their engagement.

3.6.1 Datasets and Experimental Setup

We evaluate our models on seven Coursera MOOCs at University of Maryland: *Surviving Disruptive Technologies*, *Women and the Civil Rights Movement*, two iterations of *Gene and the Human Condition*, and three iterations of *Developing Innovative Ideas for New Companies*. These courses cover a broad spectrum of topics spanning across humanities, business, and sciences. We refer to these courses as DISR, WOMEN, GENE-1, GENE-2, INNO-1, INNO-2 and INNO-3, respectively. Our data consists of anonymized student records, grades, and online behavior recorded during each course duration.

Figure 3.2 shows the number of participants in different course-related activities. Of the total number of students registered, around 5% of the students in DISR-TECH and WOMEN, 14% in GENE-1, 21% in GENE-2, 7% in INNO-1, 15% in INNO-2, and 5% in

INNO-3 complete the course. In all the courses, the most prominent activity exhibited by students while on the site is viewing lectures. Hence, we rank students based on number of lectures viewed, as a baseline (denoted LECTURE-RANK in our tables) for comparison. The other prevalent activities include submitting quizxes and viewing forum content. Observing the statistics, DISR and WOMEN have a higher percentage of total registered students participating in forums compared to GENE and INNO courses.

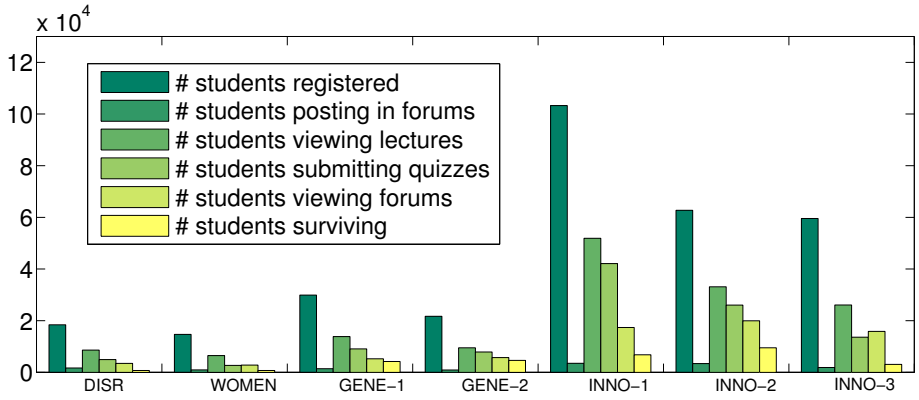


Figure 3.2: Comparison of number of students participating in course-related activities in seven courses.

We evaluate the model on the following metrics: area under the precision-recall curve for positive and negative labels and area under the ROC curve. We use ten-fold cross-validation, leaving out 10% of the data for testing and revealing the rest for training the model weights.

3.6.2 Student Performance Analysis

We conduct experiments to assess how effective our models are in predicting student performance, as measured both by their official grade and whether they complete the

course requirements. We also look at the key factors influencing student performance in the online setting as determined by our model. We filter the dataset to include only students that participated in at least one of the possible course related activities. For these students, we label the ones who earn a certificate from the course as positive instances (PERFORMANCE = 1.0) and students that did not as negative instances (PERFORMANCE = 0.0). These labels are used as ground truth to train and test the models. Our experimental results are summarized in Tables 3.5, and show performance values for the DIRECT and LATENT PSL models compared to the LECTURE-RANK baseline. We observe that the LATENT PSL model performs better at predicting students performance, outperforming both the DIRECT and LECTURE-RANK models.

To better understand which behavioral factors provide more predictive information, we examine the weights our models learned at training time. The rules involving viewing lectures and viewing forum posts have highest weights in the DIRECT learned model, indicating the importance of these features in predicting performance. The other prominent features which get high weights in the learned model are posting in forums, and reputation of student in the forums. In the LATENT model, rules corresponding to passive engagement have highest weights in the learned model for predicting performance. This emphasizes the importance of passive forms of engagement in online settings. This is followed by rules corresponding to active engagement, indicating that active forms of engagement are also predictive of student success in online courses, but fall second to active forms of engagement. Rules corresponding to disengagement gain high weights for predicting student drop out.

COURSE	MODEL	AUC-PR Pos.	AUC-PR Neg.	AUC-ROC
DISR	LECTURE-RANK	0.630	0.421	0.512
	DIRECT	0.739	0.546	0.667
	LATENT	0.749	0.575	0.692
WOMEN	LECTURE-RANK	0.263	0.761	0.503
	DIRECT	0.557	0.881	0.767
	LATENT	0.732	0.959	0.909
GENE-1	LECTURE-RANK	0.503	0.482	0.476
	DIRECT	0.814	0.755	0.817
	LATENT	0.943	0.879	0.931
GENE-2	LECTURE-RANK	0.466	0.522	0.482
	DIRECT	0.806	0.783	0.831
	LATENT	0.923	0.941	0.932
INNO-1	LECTURE-RANK	0.376	0.651	0.507
	DIRECT	0.714	0.858	0.815
	LATENT	0.850	0.920	0.899
INNO-2	LECTURE-RANK	0.536	0.984	0.938
	DIRECT	0.785	0.790	0.811
	LATENT	0.892	0.876	0.881
INNO-3	LECTURE-RANK	0.239	0.813	0.543
	DIRECT	0.586	0.930	0.835
	LATENT	0.833	0.983	0.945

Table 3.5: Performance of LECTURE-RANK, DIRECT and LATENT models in predicting student performance

3.6.3 Student Survival Analysis

Our experiments in the student survival models are aimed at measuring student survival by understanding factors influencing students' survival in the course, engagement types and changes in engagement, and the effectiveness of prediction at different time periods of the course.

3.6.3.1 Student Survival Results

In our first set of experiments, we consider all student activity during the entire course to predict whether each student takes the final quiz. We consider all registered students in the course. The scores for our DIRECT and LATENT survival models and LECTURE-RANK baseline are listed in Table 3.6. As can be observed from Figure 3.2, a high proportion

of students drop out from MOOCs, leading to a huge class imbalance in the data. Hence, models that can identify students who will complete the course are more valuable in this setting. The LECTURE-RANK baseline can predict dropouts reasonably well, but its comparatively low precision and recall for positive survival (AUC-PR pos.) indicates that using this feature alone is suboptimal. The strength of our models comes from combining behavioral, linguistic, temporal, and structural features for predicting student survival. Our models DIRECT and LATENT significantly improve on the baseline, and the LATENT model outperforms the DIRECT model.

COURSE	MODEL	AUC-PR Pos.	AUC-PR Neg.	AUC-ROC
DISR	LECTURE-RANK	0.333	0.998	0.957
	DIRECT	0.393	0.997	0.936
	LATENT	0.546	0.998	0.969
WOMEN	LECTURE-RANK	0.508	0.995	0.946
	DIRECT	0.565	0.995	0.940
	LATENT	0.816	0.998	0.983
GENE-1	LECTURE-RANK	0.688	0.984	0.938
	DIRECT	0.793	0.997	0.976
	LATENT	0.818	0.985	0.944
GENE-2	LECTURE-RANK	0.610	0.983	0.916
	DIRECT	0.793	0.985	0.939
	LATENT	0.848	0.997	0.980
INNO-1	LECTURE-RANK	0.473	0.992	0.930
	DIRECT	0.597	0.995	0.950
	LATENT	0.694	0.997	0.968
INNO-2	LECTURE-RANK	0.653	0.984	0.928
	DIRECT	0.680	0.985	0.930
	LATENT	0.753	0.988	0.936
INNO-3	LECTURE-RANK	0.353	0.994	0.922
	DIRECT	0.492	0.995	0.937
	LATENT	0.822	0.999	0.984

Table 3.6: Performance of LECTURE-RANK, DIRECT and LATENT models in predicting student survival

3.6.3.2 Early Prediction

Predicting student survival can provide instructors with a powerful tool if these predictions can be made reliably before the students disengage and drop out. We simulate this scenario by training our model over data collected early in the course. The student survival labels are the same as for the complete dataset (i.e., whether the student submitted the final quizzes/assignments at the end of the course), but our models are only given access to data from the early parts of the course. We divide the course into three equal parts according to the duration of the course: *start*, *mid*, and *end*. We combine *start* and *mid* time periods to get data till *mid* part of the course, which we refer to as *start-mid*. *start-end* refers to data collected over the entire course.

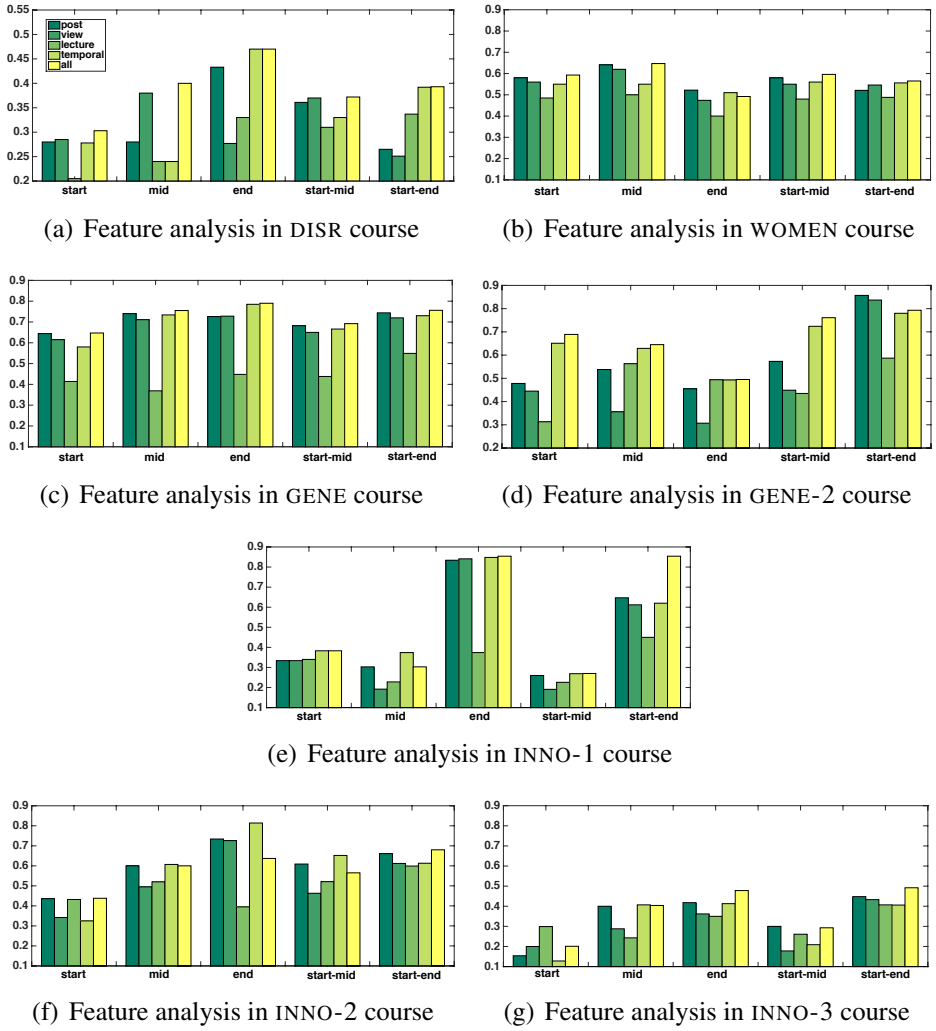
Table 3.7 lists the performance metrics for our two models using different splits in the data. Similar to the results in Table 3.6, the change in the AUC-PR (Neg.) scores are negligible and close to optimal for all models because of class imbalance. To highlight the strength our models, we only report the AUC-PR (Pos.) scores of the models. Early prediction scores under *start*, *mid*, and *start-mid* indicate that our model can indeed make early survival predictions reliably. As the data available is closer to the end of the course, models make better predictions. Similar to the previous experimental setting, the LATENT model achieves the highest prediction quality. We observe that the LATENT model consistently outperforms the DIRECT model on all time periods across seven courses. The LATENT model also significantly outperforms the DIRECT model in the *start* time period, making it a very useful tool for instructors to predict student survival early on in the course.

COURSE	MODEL	*start*	*mid*	*end*	*start-mid*
DISR	LECTURE-RANK	0.204	0.280	0.324	0.269
	DIRECT	0.304	0.400	0.470	0.372
	LATENT	0.417	0.454	0.629	0.451
WOMEN	LECTURE-RANK	0.538	0.518	0.415	0.533
	DIRECT	0.593	0.647	0.492	0.596
	LATENT	0.674	0.722	0.733	0.699
GENE-1	LECTURE-RANK	0.552	0.648	0.677	0.650
	DIRECT	0.647	0.755	0.784	0.692
	LATENT	0.705	0.755	0.789	0.778
GENE-2	LECTURE-RANK	0.449	0.431	0.232	0.699
	DIRECT	0.689	0.645	0.494	0.761
	LATENT	0.754	0.755	0.809	0.820
INNO-1	LECTURE-RANK	0.221	0.118	0.403	0.378
	DIRECT	0.383	0.304	0.846	0.692
	LATENT	0.571	0.460	0.854	0.778
INNO-2	LECTURE-RANK	0.232	0.464	0.456	0.301
	DIRECT	0.438	0.600	0.637	0.565
	LATENT	0.605	0.676	0.794	0.648
INNO-3	LECTURE-RANK	0.104	0.188	0.203	0.113
	DIRECT	0.202	0.405	0.478	0.293
	LATENT	0.309	0.574	0.803	0.428

Table 3.7: Early prediction performance of LECTURE-RANK, DIRECT and LATENT models in time-periods *start*, *mid*, *end*, and *start-mid*

From the results, it appears that the middle phase (*mid*) is the most important phase to monitor student activity for predicting whether the student will survive the length of the course. Our model produces higher AUC-PR values when using data from the *mid* phase, compared to the settings where we use data from the *start* phase, and an almost equal value when compared to *start-mid*. We hypothesize that this is due to the presence of a larger student population in the *start* phase that fails to remain engaged until the end. This phenomenon is typical in both traditional and online classrooms where students familiarize themselves with the course and then decide whether to stay or drop out. Eliminating data collected from this population helps improve our prediction of student survival, as indicated by an increase in performance values for *mid*.

3.6.4 Feature Analysis



(a) Feature analysis in DISR course

(b) Feature analysis in WOMEN course

(c) Feature analysis in GENE course

(d) Feature analysis in GENE-2 course

(e) Feature analysis in INNO-1 course

(f) Feature analysis in INNO-2 course

(g) Feature analysis in INNO-3 course

Figure 3.3: Bar graph showing AUC-PR (Pos.) value upon removal of each feature from the DIRECT model across time periods

We evaluate the contribution of each feature by leaving each feature out and observing the resulting change in prediction performance values. The features considered are: posting in forums including linguistic and structural forum features (*post*), viewing forum content (*view*), viewing lectures and submitting quizzes part of the lecture (*lecture*), and temporal features (*temporal*). The model with all the features included is given by *all*. For

each of the five features above, we construct a PSL model by omitting the relevant feature from all PSL rules. Figure 3.3 plots the results from these tests for phases—*start*, *mid*, *end*, *start-mid*, and *start-end*. The decrease in value from *all* corresponds to the importance of each class of features in the model. The *lecture* feature is consistently important for predicting student survival, indicating that it is the most prevalent form of interaction of MOOC participants on the MOOC website. This is especially evident in the *mid* and *end* phases, where *lecture* is a very important feature. In some courses, it is a very strong feature from the *start* phase (DISR, WOMEN, GENE-1, and GENE-2), while in the INNO courses, it only becomes relevant in the *mid* and *end* phases. Discussion forums serve as a platform connecting students worldwide enrolled in the course, hence activity in the discussion forums also turns out to be a strongly contributing feature. Since, the concentration of forum posts in the courses analyzed is more in the *mid* and *end* phases, posting in forums is accordingly more important during the *mid* and *end* phases. Also, in the *start* phase of the course, most posts are about students introducing themselves and getting to know other people enrolled in the course. These posts are not very predictive of student engagement and their subsequent performance or survival in the course. Simply viewing content on the forums is also a strong feature, contributing consistently in all phases across all courses. In fact, from Figure 3.3, we can see that the feature strength of forum views is second only to lecture views. This further ascertains the importance of passive engagement in online courses. *Temporal* features are a strong feature in the early part of the course, particularly in the *start* phase across all seven courses. But, they decline as a predictive feature in the *mid* and *end* phases. The data suggests that this is due to the larger volume of students dropping out in the early part of the course, making

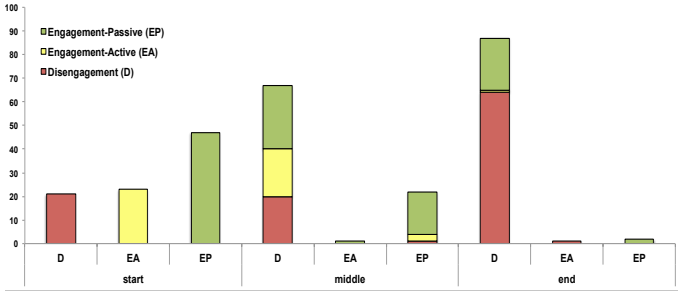
it an excellent predictor for student survival in the *start* phase. As the student population grows steady, *temporal* features start to decline as a predictive feature.

3.6.5 Gaining Insight from Latent Engagement Assignments

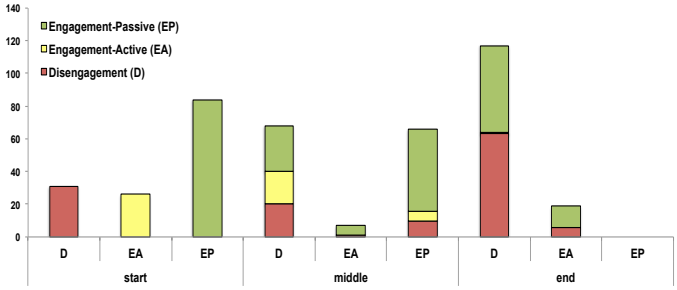
Going beyond measuring the impact of engagement on performance prediction, we are interested in understanding the value of the engagement information our model uncovers. We look into two possible applications: the first uses this information to analyze temporal engagement patterns of students. The second application provides a qualitative analysis of forum activity by observing representative forum content posted by students with different engagement assignments.

3.6.5.1 Analyzing Engagement Pattern Dynamics

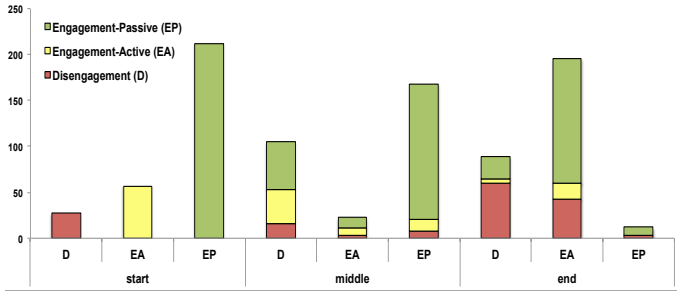
In this section, we take a first step toward understanding how student engagement changes as the course progresses. We track the changes in engagement assignments patterns for several interesting student populations and discuss potential explanations for these changes. We categorize students that drop out of the course according to the time period in which they dropped out. We analyze the student engagement values predicted by the model for three groups of students—(1) students dropping out in the *mid* phase, (2) students dropping out in the *end* phase, and (3) students continuing until course completion. Students dropping out in the *mid* phase stop participating in course activities sometime during middle phase. Similarly, students dropping out in the *end* phase stop participating in the course sometime during the *end* phase. We infer engagement assignments for these groups



(a) Engagement patterns of students that dropped out of the class in the middle phase



(b) Engagement patterns of students that dropped out of the class in the end phase



(c) Engagement patterns of students that survived the complete class

Figure 3.4: Bar-graph showing the distribution of engagement label assignments at three time points throughout the class. We capture engagement transition patterns by coloring the bars according to the engagement assignments of students at the previous time point.

by training on data from the start and middle respectively. The students are classified into one of the engagement types by considering the dominant value of engagement as predicted by the model. This helps distinguish between the different engagement types for different populations of students, uncovering their movement from one engagement type to another and how engagement-mobility patterns relate to student survival.

Figure 3.4 describes the student engagement values predicted by the model for the three classes of students. The labels D, EA and EP refer to values for latent variables DISENGAGEMENT, ACTIVE-ENGAGEMENT and PASSIVE-ENGAGEMENT, respectively. For each student group, we provide a bar graph, showing the different engagement assignment levels at each time span (*start, middle, end*). In order to track student engagement patterns, we color code the bars according to the previous engagement assignments of the students. Each bar therefore consists of the combination of three smaller bars, colored differently, capturing the previous engagement values. In Figure 3.4(a), EA students start to move toward disengagement in the middle phase. While some EP students, who are not taking quizzes in middle phase, still follow the course passively, placing them in EP rather than D. *We hypothesize that these students may be more likely to respond to intervention than the already disengaged students.* In Figure 3.4(b), it can be seen that, out of the students that drop out eventually in the *end* phase, about half of them are in *EP*. Finally, Figure 3.4(c) suggests that most engaged students only exhibit passive forms of engagement in the *start* and *mid* phases of the course. While in the *end* phase, students tend to become more actively engaged in the course. All these results corroborate the importance of taking into account passive engagement. In all these classes of students, passive engagement is a more prevalent type of engagement than active, stressing the fact that careful observation of passive engagement (which includes subtle activities such as viewing forum posts) can help MOOC instructors assess student health.

Engagement	Sentiment	Values	Post
Engaged	positive	performance = 0.75 disengagement = 0.0	Thank you for a great course! And thank you Coursera!
Engaged	negative	performance = 0.8 disengagement = 0.0	I have also received a 9, the most disappointing thing is that I have only received good or passing comments from my peers, 3 of 5 did not post any comment about my work.
Disengaged	negative	performance = 0.5 disengagement = 0.7	I agree completely. I used a lot of time on my assignment and got 7.5. think the evaluation criteria were wrong.
Disengaged	negative	performance = 0.3 disengagement = 1.0	The grades I received are ridiculous! I've re-read my assignment and I still can't believe in my grade. Is it really fair?.
Auditor	positive	performance = 0.0 survival = 1.0 disengagement = 0.3	This has been an otherwise fantastic course. Too bad the potential for success is so heavily weighted on two assignments.
Engaged	positive	survival = 0.9 disengagement = 0.2	I didn't have a problem with the video lectures. Combining the lectures with the written transcripts will really drive home the points. Hope everyone is enjoying the course as much as I am.
Disengaged	negative	survival = 0.1 disengagement = 0.8	I've tried the first quiz but have not submitted yet. So in this course you submit and not know if you made a mistake or not until the deadline? That is strange all the rest of the course I took so far had immediate feedback once you submit your answers.
Engaged	negative	survival = 0.7 disengagement = 0.2	I'm a native English speaker and I lecture myself and I agree with the others who say that the instructor would be a more effective lecturer if she'd slow down a bit. Her presentation is otherwise clear and well-organized; but I find myself frequently having to go back and re-listen to parts of her lectures simply because she raced through them so quickly.

Table 3.8: Relevant forum content by students assigned different engagement labels by our model.

3.6.5.2 Using Engagement for Qualitative Language Analysis

In addition to predicting student success, the engagement variables are helpful in interpreting the different facets of student participation in the course. Of particular interest is the content of the posts made by students and how it corresponds to the values predicted by the our model. Table 3.8 shows some examples of posts made by students and the engagement and performance/survival scores predicted by our models. In general, we observe that positive sentiment posts are a sign of positive student engagement. However, it is interesting to note that both engaged and disengaged students post content with negative sentiment on the course. For example, the second and eighth student in Table 3.8 both post negative sentiment posts, but are engaged students completing the course. While, the third and seventh students also post negative sentiment posts, they are disengaged students who do not complete the course. Considering other forms of interaction and changes in activity helps discern them from the engaged ones. Another very interesting example is the fifth student, who is an auditor, only viewing lectures but not completing the assignments. The performance prediction for this student is 0.0, as she does not complete the assignments, but the survival prediction is 1.0. Hence, considering different methods of evaluating MOOC participants (such as, performance and survival) is essential in understanding their varied needs and engagement in the course.

3.7 Discussion

In this work, we take a step toward helping MOOC instructors and optimizing experience for MOOC participants by modeling latent student engagement using data-driven

methods. We formalize, using HL-MRFs, that student engagement can be modeled as a complex interaction of behavioral, linguistic and social cues, and we model student engagement types as latent variables over these cues. Our models construct interpretations for latent engagement variables from data and predict student course success indicators reliably, even at early stages in the course. These results are a first step toward facilitating instructors' intervention at critical points, thus helping improve course retention rates. The latent formulation we present can be extended to more sophisticated modeling by including additional latent factors that affect academic performance such as motivation, self-regulation and tenacity. These compelling directions for future interdisciplinary investigation can provide a better understanding of MOOC students.

Chapter 4: Seeded Topic Models for MOOC Discussion Forums

4.1 Introduction

MOOC discussion forums provide a platform for exchange of ideas, course administration and logistics questions, reporting errors in lectures, and discussions about course material. Unlike classroom settings, where there is face-to-face interaction between the instructor and the students and among the students, MOOC forums are the primary means of interaction in MOOCs. Due to the open nature of MOOCs, they attract people from all over the world leading to large numbers of participants and hence, large numbers of posts in the discussion forums. In the courses we worked with, we found that over the course of the class there were typically over 10,000 posts.

However, due to the large number of students and the large volume of posts generated by them, MOOC forums are not monitored completely. Forums can include student posts expressing difficulties in course-work, grading errors, dissatisfaction in the course, which are possible precursors to students dropping out. In this chapter, I will explore the importance of mining content in MOOC discussion forums. I present analysis of MOOC discussion content and demonstrate that analyzing discussion forum content is helpful in predicting student course completion. In this analysis, we observe that posts discussing course logistics correlate well with student course completion.

Our main contributions in this chapter are as follows:

- We employ seeded topic models to map discussion forum posts to topics in an unsupervised manner. We employ background knowledge from the course syllabus and manual inspection of discussion forum posts to seed topic models to get better separated topics.

- We incorporate the topic distributions so obtained as features in our course success prediction models and encode meaningful combinations with other behavioral and linguistic features. We demonstrate that inclusion of topic features is helpful in predicting student survival in online courses across three Coursera MOOCS: *Surviving Disruptive Technologies* (DISR), *Women and the Civil Rights Movement* (WOMEN), and *Gene and the Human Condition* (GENE).

4.2 MOOC Forum content analysis for Modeling Student Survival

Previous work analyzing discussion forum content manually label posts with categories of interest [Stump *et al.*, 2013b]. Unfortunately, the effort involved in manually annotating the large amounts of posts prevents using such solutions on a large scale. Instead, we suggest using natural language processing tools for identifying relevant aspects of forum content automatically. Specifically, we explore *SeededLDA* [Jagarlamudi *et al.*, 2012], a recent extension of topic models which can utilize a lexical seed set to bias the topics according to relevant domain knowledge.

Exploring data from three MOOCs, we find that forum posts usually belong to these three categories—a) course content, which include academic discussions about course

material (ACADEMIC), b) meta-level discussions about the course, including feedback and course logistics (LOGISTICS), and c) social posts, which include student introductions and formation of study groups (SOCIAL). In order to capture these categories automatically we provide seed words for each category. For example, we extract seed words for the ACADEMIC topic from each course's syllabus.

In addition to the automatic topic assignment, we capture the sentiment polarity using *Opinionfinder* [Wilson *et al.*, 2005a]. We use features derived from topic assignments and sentiment to predict student course completion(which we refer to as *student survival*). We measure course completion by examining if the student attempted the final exam/ last few assignments in the course. We follow the observation that LOGISTICS posts contain feedback about the course. Finding high-confidence LOGISTICS posts can give a better understanding of student opinion about the course. Similarly, posting in ACADEMIC topic and receiving good feedback (i.e., votes) is an indicator of student success. We show that modeling these intuitions using topic assignments together with sentiment scores, helps in predicting student survival. In addition, we examine the topic assignment and sentiment patterns of some users and show that topic assignments help in understanding student concerns better.

This chapter builds on our work in Chapter 3 on modeling student success using PSL. In Chapter 3, we model sentiment without modeling the context in which the sentiment was expressed: associating positive sentiment with course success and negative sentiment with failure. In this work, we introduce context by adding topics and enable reasoning about sentiment in specific types of posts. While sentiment of posts can indicate general dissatisfaction, we expect this to be more pronounced in LOGISTICS posts

as posts in this category correspond to issues and feedback about the course. In contrast, sentiment in posts about course material may signal a particular topic of discussion in a course and may not indicate attitude of the student toward the course. In Section 4.4.3, we show some examples of course-related posts and their sentiment, and we illustrate that they are not suggestive of student survival. For example, in WOMEN course, the post—*"I think our values are shaped by past generations in our family as well, sometimes negatively."*—indicates an attitude towards an issue discussed as part of the course. Hence, identifying posts that fall under LOGISTICS can improve the value of sentiment in posts. In Section 4.3, we show how these are translated into rules in our model.

4.3 Enhancing Student Survival Models with Topic Modeling

Discussion forums in online courses are organized into threads to facilitate grouping of posts into topics. For example, a thread titled *errata, grading issues* is likely a place for discussing course logistics and a thread titled *week 1, lecture 1* is likely a place for discussing course content. But a more precise examination of such threads reveals that these heuristics do not always hold. We have observed that *course content* threads often house *logistic content* and vice-versa. This demands the need to use computational methods to classify the content in discussion forums.

4.3.1 Latent Dirichlet Allocation

Table 4.1 gives the topics given by *latent Dirichlet allocation* (LDA) on discussion forum posts. The words that are likely to fall under LOGISTICS are underlined in the table. It

can be observed that these words are spread across more than one topic. Since we are especially interested in posts that are on LOGISTICS, we use *Seeded LDA* [Jagarlamudi *et al.*, 2012], which allows one to specify *seed* words that can influence the discovered topics toward our desired three categories.

topic 1: kodak, management, great, innovation, problem, film, businesses, changes, needs
topic 2: good, change, publishing, brand, companies, publishers, history, marketing, traditional, authors
topic 3: think, work, technologies, newspaper, content, paper, disruptive, print, media, *course*, *assignment*
topic 4: digital, kodak, company, camera, market, quality, phone, development, future, failed, high
topic 5: amazon, books, netflix, blockbuster, stores, online, experience, products, apple, strategy
topic 6: time, *grading*, different, *class*, *course*, *major*, *focus*, product, like, years
topic 7: companies, *interesting*, *class*, *thanks*, going, printing, far, wonder, article, sure

Table 4.1: Topics identified by LDA

4.3.1.1 Seeded LDA

We experiment by providing seed words for topics that fall into the three categories. The seed words for the three courses are listed in tables 4.2 and 4.3. The seed words for LOGISTICS and SOCIAL topics are common across all the three courses. The seed words for the ACADEMIC topic are chosen from the course-syllabus of the courses. This construction of seed words enables the model to be applied to new courses easily. Topics in Table 4.3 denote the course specific seed words for DISR, WOMEN, and GENE courses respectively. Since the syllabus is only an outline of the class, it does not contain all the terms that will be used in class discussions. To capture other finer course content discussions as separate topics, we include k more topics when we run the SeededLDA. We notice that not including more topics here and only including the seeded topics (i.e., running SeededLDA with exactly three topics) results in some words from course content discussions,

which were not specified in the course-seed words, appearing in the LOGISTICS or SOCIAL topics. Thus, the k extra topics help isolate ACADEMIC topics that are not captured by academic seed words in Table 4.3. Note that these *extra* topics are not seeded. We experimented with different values of k on our experiments and found by manual inspection that the topic-terms produced by our model are well separated for $k = 3$. Thus, we run *SeededLDA* with 7 total topics. Tables 4.4, 4.5, and 4.6 give the topics identified for DISR, WOMEN and GENE by *SeededLDA*. The topic assignments so obtained are used as input features to the PSL model—the predicate for the first topic is LOGISTICS, the second one is SOCIAL and the rest are summed up to get the topic assignment for the ACADEMIC topic.

LOGISTICS: thank, professor, lectures, assignments, concept, love, thanks, learned, enjoyed, forums subject, question, hard, time, grading, peer, lower, low
SOCIAL: introduction, study, moocs, courses, students, online, group, coursera

Table 4.2: Seed words in LOGISTICS and SOCIAL for DISR-TECH, WOMEN and GENE courses

ACADEMIC DISR: disruptive, technology, innovation, survival, digital, disruption, survivor
ACADEMIC WOMEN: women, civil, rights, movement, american, black, struggle, protests, african, status
ACADEMIC GENE: genomics, genome, egg, living, ancestors, genes, behavior, genetic, biotechnology

Table 4.3: Seed words for COURSE topic for DISR-TECH, WOMEN and GENE courses

4.3.2 Using Topic Assignments in PSL

We build on the DIRECT model discussed in Chapter 3 and include topic assignments as features in the model to construct the DIRECT+TOPIC model. We compare the DIRECT+TOPIC model to the DIRECT model in our experiments.

Table 4.7 contains examples of rules in the DIRECT model and the corresponding

topic 1: time, thanks, one, low, hard, question, course, love, professor, lectures, lower, concept, peer, point
topic 2: online, education, coursera, students, university, courses, classroom, moocs, teaching, video
topic 3: digital, survival, management, disruption, technology, development, market, business, innovation
topic 4: publishing, publisher, traditional, companies, money, history, brand
topic 5: companies, social, internet, work, example
topic 6: business, company, products, services, post, consumer, market, phone, changes, apple
topic 7: amazon, book, nook, readers, strategy, print, noble, barnes

Table 4.4: Topics identified by SeededLDA for DISR

topic 1: time, thanks, one, hard, question, course, love, professor, lectures, forums, help, essays, problem, thread, concept, subject
topic 2: online, education, coursera, students, university, courses, classroom, moocs, teaching, video, work, english, interested, everyone
topic 3: women, rights, black, civil, movement, african, struggle, social, citizenship, community, lynching, class, freedom, racial, segregation
topic 4: violence, public, people, one, justice, school,s state, vote, make, system, laws
topic 5: idea, believe, women, world, today, family, group, rights
topic 6: one, years, family, school, history, person, men, children, king, church, mother, story, young
topic 7: lynching, books, mississippi, march, media, youtube, death, google, woman, watch, south, film

Table 4.5: Topics identified by SeededLDA for WOMEN

topic 1: time, thanks, one, answer, hard, question, course, love, professor, lectures, brian, lever, concept, agree, peer, material, interesting
topic 2: online, education, coursera, students, university, courses, classroom, moocs, teaching, video
topic 3: genes, genome, nature, dna, gene, living, behavior, chromosomes, mutation, processes
topic 4: genetic, biotechnology, engineering, cancer, science, research, function, rna
topic 5: reproduce, animals, vitamin, correct, term, summary, read, steps
topic 6: food, body, cells, alleles blood, less, area, present, gmo, crops, population, stop
topic 7: something, group, dna, certain, type, early, large, cause, less, cells

Table 4.6: Topics identified by SeededLDA for GENE

DIRECT	DIRECT+TOPIC
$\text{POSTS}(U, P) \wedge \text{POLARITY}(P) \rightarrow \text{SURVIVAL}(U)$	$\text{POSTS}(U, P) \wedge \text{TOPIC(P, LOGISTICS)} \wedge \neg\text{POLARITY}(P) \rightarrow \text{SURVIVAL}(U)$
$\text{POSTS}(U, P) \wedge \neg\text{POLARITY}(P) \rightarrow \neg\text{SURVIVAL}(U)$	$\text{POSTS}(U, P) \wedge \text{TOPIC(P, LOGISTICS)} \wedge \neg\text{POLARITY}(P) \rightarrow \text{SURVIVAL}(U)$
$\text{POSTS}(U, P) \rightarrow \text{SURVIVAL}(U)$	$\text{POSTS}(U, P) \wedge \text{TOPIC(P, SOCIAL)} \rightarrow \neg\text{SURVIVAL}(U)$
$\text{POSTS}(U, P) \wedge \text{UPVOTE}(P) \rightarrow \text{SURVIVAL}(U)$	$\text{POSTS}(U, P) \wedge \text{TOPIC(P, COURSE)} \wedge \text{UPVOTE}(P) \rightarrow \text{SURVIVAL}(U)$
	$\text{POSTS}(U_1, P_1) \wedge \text{POSTS}(U_2, P_2) \wedge \text{TOPIC}(P_1, \text{COURSE}) \wedge$
	$\text{TOPIC}(P_2, \text{COURSE}) \wedge \text{SURVIVAL}(U_1) \rightarrow \text{SURVIVAL}(U_2)$

Table 4.7: Rules modified to include topic features

rules including topic assignments in DIRECT+TOPIC model. Section 3.5.1 in Chapter 3 contains a detailed discussion of rules in the DIRECT model. Only rules that are enhanced to include the topic features are presented here. The first and second rules in Table 4.7 containing polarity are changed to include LOGISTICS topic feature, following our observation that polarity matters in *meta-course* posts. While the DIRECT model regards posting in forums as an indication of survival, in the DIRECT+TOPIC model, this rule is enhanced to capture that students that predominantly post *social* posts on the forums do not necessarily participate in course-related discussions. The fourth rule containing *upvote* predicate, which signifies posts that received positive feedback in the form of votes, is changed to include the topic-feature ACADEMIC. This captures the significance of posting *academic* content that gets positive feedback as opposed to *logistics* or *social* content in the forums. This rule helps us discern posts in social/logistic category that can get a significant number of positive votes (*upvote*), but do not necessarily indicate student survival. For example, some introduction posts receive many positive votes, but do not necessarily signify student survival.

4.4 Empirical Evaluation

We conduct experiments to answer the following question—how much do the topic assignments from *SeededLDA* help in predicting student survival? We also perform a qualitative analysis of topic assignments, the sentiment of posts, and their correspondence with student survival.

4.4.1 Datasets and Experimental Setup

We evaluate our models on three Coursera MOOCs: DISR, WOMEN, and GENE. Our data consists of anonymized student records, grades, and online behavior recorded during the seven week duration of each course. We label students as *survival = 1.0* if they take the final exam/quiz and *survival = 0.0* otherwise. In our experiments, we only consider students that completed at least *one* quiz/assignment in the course. We evaluate our models using area under precision-recall curve for positive and negative survival labels and area under ROC curve. We use ten-fold cross-validation on each of the courses, leaving out 10% of users for testing and revealing the rest of the users for training the model weights. We evaluate statistical significance using a paired t-test with a rejection threshold of *0.05*.

4.4.2 Survival Prediction using topic features

Table 4.8 shows the prediction performance of the DIRECT and DIRECT+TOPIC model. The inclusion of the topic-features improves student survival prediction in all the three courses.

COURSE	MODEL	AUC-PR POS.	AUC-PR NEG.	AUC-ROC
DISR	DIRECT	0.764	0.628	0.688
	DIRECT+TOPIC	**0.794**	0.638	**0.708**
WOMEN	DIRECT	0.654	0.899	0.820
	DIRECT+TOPIC	**0.674**	0.900	**0.834**
GENE	DIRECT	0.874	0.780	0.860
	DIRECT+TOPIC	**0.894**	**0.791**	**0.873**

Table 4.8: Performance of DIRECT and DIRECT+TOPIC models in predicting student survival. Statistically significant scores typed in bold.

4.4.3 Discussion topic analysis using topic features

Table 4.9 shows some posts by users that did not survive the class. All these posts have negative sentiment scores by *Opinionfinder* and belong to LOGISTICS. Also, in the forum, all these posts were not answered. This suggests that students display a tendency to drop out of the course if their logistics questions are not answered. Table 4.10 gives examples of student posts that also have a negative sentiment. But the sentiment of the thread changes when the issue is resolved (last row in the table). We observe that these two students survive the course and a timely answer to their posts might have been a reason influencing these students to complete the course.

Survival	Sentiment	Topic	Post
survival = 0.0	polarity = 0.25	logistics = 0.657 general = 0.028 course = 0.314	JSTOR allowed 3 items (texts/writings) on my 'shelf' for 14 days. But, I read the items and wish to return them, but cannot, until 14 days has expired. It is difficult then, to do the extra readings in the "Exploring Further" section of Week 1 reading list in a timely manner. Does anyone have any ideas for surmounting this issue?
survival = 0.0	polarity = 0.0	logistics = 0.643 general = 0.071 course = 0.285	There are some mistakes on quiz 2. Questions 3, 5, and 15 mark you wrong for answers that are correct.
survival = 0.0	polarity = 0.25	logistics = 0.652 general = 0.043 course = 0.304	I see week 5 quiz is due April 1(by midnight 3/31/13).I am concerned about this due date being on Easter, some of us will be traveling, such as myself. Can the due date be later in the week? Thank you

Table 4.9: Logistics posts containing negative sentiment for dropped-out students

Tables 4.9 and 4.10 show how student survival may depend on forum interaction and responses they receive. Our approach can help discover potential points of contention in the forums, identifying potential drop outs that can be avoided by intervention. Table 4.11 shows posts flagged as ACADEMIC by the *SeededLDA*. The polarity scores in the ACADEMIC posts indicate opinions and attitude toward course specific material. For ex-

Survival	Sentiment	Topic	Post
survival = 1.0	polarity = 0.0	logistics = 0.67 general = 0.067 course = 0.267	I was just looking at the topics for the second essay assignments. The thing is I dont see what the question choices are. I have the option of Weeks and I have no idea what that even means. Can someone help me out here and tell me what the questions for the second essay assignment are I think my computer isnt allowing me to see the whole assignment! Someone please help me out and let me know that the options are.
survival = 1.0	polarity = 0.25	logistics = 0.769 general = 0.051 course = 0.179	I'd appreciate someone looks into the following: Lecture slides for the videos (week 5) don't open (at all) (irrespective of the used browser). Some required reading material for week 5 won't open either (error message). I also have a sense that there should be more material posted for the week (optional readings, more videos, etc). Thanks. — I am not seeing a quiz posted for Week 5.
survival = 1.0	polarity = 0.78	logistics = 0.67 general = 0.067 course = 0.267	Hopefully the Terrell reading and the Lecture PowerPoints now open for you. Thanks for reporting this.

Table 4.10: Example of change in sentiment in a course logistic thread

Survival	Sentiment	Topic	Post
survival = 1.0	polarity = 0.25	logistics = 0.372 social = 0.163 academic = 0.465	I've got very interested in the dynamic of segregation in terms of space and body pointed by Professor Brown and found a document written by GerShun Avilez called "Housing the Black Body: Value, Domestic Space,and Segregation Narratives".
survival = 1.0	polarity = 0.9	logistics = 0.202 social = 0.025 academic = 0.772	I think that you hit it on the head, the whole idea of Emancipation came as a result not so much of rights but of the need to get the Transcontinental Railroad through the mid-west and the north did not want the wealth of the southern slave owners to overshadow the available shares. There are many brilliant people "good will hunting", and their brilliance either dies with them or dies while they are alive due to intolerance. Many things have happened in my life to cause me to be tolerant to others and see what their debate is, Many very evil social ills and stereotypes are a result of ignorance. It would be awesome if the brilliant minds could all come together for reform and change.
survival = 1.0	polarity = 0.167	logistics = 0.052 social = 0.104 academic = 0.844	I think our values are shaped by past generations in our family as well – sometimes negatively. In Bliss, Michigan where I come from, 5 families settled when the government kicked out the residents – Ottowa Tribe Native Americans. I am descended from the 5 families. All of the cultural influences in Bliss were white Christian – the Native American population had never been welcomed back or invited to stay as they had in Cross Village just down the beach. My family moved to the city for 4 years during my childhood, and I had African American, Asian, and Hispanic classmates and friends. When we moved back to the country I was confronted with the racism and generational wrong-doings of my ancestors. At the tender age of 10 my awareness had been raised! Was I ever pissed off when the full awareness of the situation hit me! I still am.

Table 4.11: Posts talking about ACADEMIC content

ample, post #3 in Table 4.11 indicates opinion towards human rights. While the post's polarity is negative, it is clear that this polarity value is not directed at the course and should not be used to predict student survival. In fact, all these users survive the course. We find that participation in course related discussion is a sign of survival. These examples demonstrate that analysis on ACADEMIC posts can mislead survival and justify our using topic predictions to focus sentiment analysis on LOGISTICS posts.

4.5 Discussion

In this chapter, we have taken a step toward understanding discussion forum content in massive open online courses. Our topic analysis is coarse-grained, grouping posts into three categories. In our analysis, all the meta-content—course logistics and course feedback—were grouped under the same topic category. Instead, a finer-grained topic model could be seeded with different components of meta-content as separate topics. The same applies for course-related posts too, where a finer-grained analysis could help identify difficult topics that may cause student frustration and dropout. In the following chapter, we delve deeper into finer-grained topics of conversation in online course discussion forums.

Chapter 5: Weakly Supervised Aspect-Sentiment Models for MOOC Discussion Forums

5.1 Introduction

Discussion forums are the primary means of communication between MOOC participants (students, TAs, and instructors). Due to the open nature of these courses, they attract people from all over the world leading to large numbers of participants and hence, large numbers of posts in the discussion forums. In the courses we worked with, we found that over the course of the class there were typically over 10,000 posts. Within this slew of posts, there are valuable *problem-reporting* posts that identify issues such as broken links, audio-visual glitches, and inaccuracies in the course materials. Automatically identifying these reported problems is important for several reasons: *i*) it is time-consuming for instructors to manually screen through all of the posts due to the highly skewed instructor-to-student ratio in MOOCs, *ii*) promptly addressing issues could help improve student retention, and *iii*) future iterations of the course could benefit from identifying technical and logistical issues currently faced by students. In this chapter, we investigate the problem of determining the fine-grained topics of posts (which we refer to as "MOOC aspects") and the sentiment toward them, which can potentially be used to improve the course.

While aspect-sentiment has been widely studied, the MOOC discussion forum scenario presents a unique set of challenges. Labeled data are expensive to obtain, and posts containing fine-grained aspects occur infrequently in courses and differ across courses, thereby making it expensive to get sufficient coverage of all labels. Few distinct aspects occur per course, and only $5 - 10\%$ of posts in a course are relevant. Hence, getting labels for fine-grained labels involves mining and annotating posts from a large number of courses. Further, creating and sharing labeled data is difficult as data from online courses is governed by IRB regulations. Privacy restrictions are another reason why unsupervised/weakly-supervised methods can be helpful. Lastly, to design a system capable of identifying all possible MOOC aspects across courses, we need to develop a system that is not fine-tuned to any particular course, but can adapt seamlessly across courses.

To this end, we develop a weakly supervised system for detecting aspect and sentiment in MOOC forum posts and validate its effectiveness on posts sampled from twelve MOOC courses. Our system can be applied to any MOOC discussion forum with no or minimal modifications.

Our contributions in this chapter are as follows:

- We show how to encode weak supervision in the form of seed words to extract extract course-specific features in MOOCs using SeededLDA, a seeded variation of topic modeling [Jagarlamudi *et al.*, 2012].

- Building upon our SeededLDA approach, we develop a joint model for aspects and sentiment using the *hinge-loss Markov random field (HL-MRF)* probabilistic modeling framework. This framework is especially well-suited for this problem

because of its ability to combine information from multiple features and jointly reason about aspect and sentiment.

- To validate the effectiveness of our system, we construct a labeled evaluation dataset by sampling posts from twelve MOOC courses, and annotating these posts with fine-grained MOOC aspects and sentiment via *crowdsourcing*. The annotation captures fine-grained aspects of the course such as content, grading, deadlines, audio and video of lectures and sentiment (i.e., *positive, negative, and neutral*) toward the aspect in the post.

- We demonstrate that the proposed HL-MRF model predicts fine-grained aspects and sentiment and outperforms the model based only on SeededLDA.

5.2 Related Work

To the best of our knowledge, the problem of predicting aspect and sentiment in MOOC forums has not yet been addressed in the literature. We review prior work in related areas here.

5.2.1 Aspect-Sentiment in Online Reviews

It is valuable to identify the sentiment of online reviews towards aspects such as hotel cleanliness and cellphone screen brightness, and sentiment analysis at the aspect-level has been studied extensively in this context [Liu and Zhang, 2012]. Several of these methods use latent Dirichlet allocation topic models [Blei *et al.*, 2003b] and variants of it for detecting aspect and sentiment [Lu *et al.*, 2011; Lin and He, 2009]. [Liu and Zhang, 2012]

provide a comprehensive survey of techniques for aspect and sentiment analysis. Here, we discuss works that are closely related to ours.

Titov and McDonald [2008] emphasize the importance of an unsupervised approach for aspect detection. However, the authors also indicate that standard LDA [Blei *et al.*, 2003b] methods capture global topics and not necessarily pertinent aspects — a challenge that we address in this work. Brody and Elhadad [2010], Titov and McDonald Titov and McDonald [2008], and Jo and Oh [2011] apply variations of LDA at the sentence level for online reviews. We find that around 90% of MOOC posts have only one aspect, which makes sentence-level aspect modeling inappropriate for our domain.

Most previous approaches for sentiment rely on manually constructed lexicons of strongly positive and negative words [Fahrni and Klenner, 2008; Brody and Elhadad, 2010]. These methods are effective in an online review context, however sentiment in MOOC forum posts is often implicit, and not necessarily indicated by standard lexicons. For example, the post "Where is my certificate? Waiting for it for over a month." expresses negative sentiment toward the certificate aspect, but does not include any typical negative sentiment words. In our work, we use a data-driven model-based approach to discover domain-specific lexicon information guided by small sets of seed words.

There has also been substantial work on joint models for aspect and sentiment [Kim *et al.*, 2013; Diao *et al.*, 2014; Zhao *et al.*, 2010; Lin *et al.*, 2012], and we adopt such an approach in this work. Kim *et al.* [2013] use a hierarchical aspect-sentiment model and evaluate it for online reviews. Mukherjee and Liu [2012] use seed words for discovering aspect-based sentiment topics. Drawing on the ideas of Mukherjee and Liu [2012] and Kim *et al.* [2013], we propose a statistical relational learning approach that combines

the advantages of seed words, aspect hierarchy, and flat aspect-sentiment relationships. It is important to note that a broad majority of the previous work on aspect sentiment focuses on the specific challenges of online review data. As discussed in detail above, MOOC forum data have substantially different properties, and our approach is the first to be designed particularly for this domain.

5.2.2 Learning Analytics

In another line of research, there is a growing body of work on the analysis of online courses. Regarding MOOC forum data, Stump *et al.* [2013b] propose a framework for taxonomically categorizing forum posts, leveraging manual annotations. We differ from their approach in that we develop an automatic system to predict MOOC forum categories without using labeled training data. Chaturvedi *et al.* [2014b] focus on predicting instructor intervention using lexicon features and thread features. In the previous chapter, we categorize forum posts into three broad topic categories in order to predict student success. In this chapter, we expand our work to develop a system capable of fine-grained categorization of aspects in MOOCS. Our system is capable of predicting fine MOOC aspects and sentiment in forum posts and thus provides a more informed analysis of MOOC posts.

5.3 Problem Setting and Data

MOOC participants primarily communicate through discussion forums, consisting of posts, which are short pieces of text. Table 5.1 provides examples of posts in MOOC

forums. Posts *1* and *2* report issues and feedback for the course, while post *3* is a social interaction message. Our goal is to distinguish *problem-reporting* posts such as *1* and *2* from *social* posts such as *3*, and to *identify the issues* that are being discussed.

Post 1: I have not received the **midterm**.
Post 2: No **lecture subtitles** week, will they be uploaded?
Post 3: I am ... and I am looking forward to learn more ...

Table 5.1: Example posts from MOOC forums. Aspect words are highlighted in **bold**.

We formalize this task as an *aspect-sentiment* prediction problem [Liu and Zhang, 2012]. The issues reported in MOOC forums can be related to the different elements of the course such as *lectures* and *quizzes*, which are referred to as *aspects*. The aspects were selected based on MOOC domain expertise and inspiration from Stump et al. [2013b], aiming to cover common concerns that could benefit from intervention. The task is to predict these aspects for each post, along with the *sentiment* polarity toward the aspect, which we code as *positive*, *negative*, or *neutral*. The negative-sentiment posts, along with their aspects, allow us to identify potentially correctable issues in the course. As labels are expensive in this scenario, we formulate the task as *weakly supervised* prediction problem. In our work, we assume that a post has at most one fine-grained aspect, as we found that this was true for 90% of the posts in our data. This property is due in part to the brevity of forum posts, which are much shorter documents than those considered in other aspect-sentiment scenarios such as product reviews.

5.3.1 Aspect Hierarchy

While we do not require labeled data, our approaches allow the analyst to instead relatively easily encode a small amount of domain knowledge by seeding the models with a few words relating to each aspect of interest. Hence, we refer to our approach as *weakly supervised*. Our models can further make use of hierarchical structure between the aspects. The proposed approach is flexible, allowing the aspect seeds and hierarchy to be selected for a given MOOC domain.

For the purposes of this study, we represent the MOOC aspects with a two-level hierarchy. We identify a list of *nine* fine-grained aspects, which are grouped into *four* coarse topics. The *coarse* aspects consist of LECTURE, QUIZ, CERTIFICATE, and SOCIAL topics. Table 5.2 provides a description of each of the aspects and also gives the number of annotated posts in each aspect category.

COARSE-TOPIC	FINE-TOPIC	# of posts
LECTURE	LECTURE-CONTENT	559
	LECTURE-VIDEO	215
	LECTURE-SUBTITLES	149
	LECTURE-AUDIO	136
	LECTURE-LECTURER	69
QUIZ	QUIZ-CONTENT	439
	QUIZ-GRADING	360
	QUIZ-SUBMISSION	329
	QUIZ-DEADLINE	142
CERTIFICATE		194
SOCIAL		1187

Table 5.2: Breakdown of number of posts per label category

As both LECTURE and QUIZ are key coarse-level aspects in online courses, and

more nuanced aspect information for these is important to facilitate instructor interventions, we identify fine-grained aspects for these topics. For LECTURE we identify LECTURE-CONTENT, LECTURE-VIDEO, LECTURE-AUDIO, LECTURE-SUBTITLES, and LECTURE-LECTURER as fine aspects. For QUIZ, we identify the fine aspects QUIZ-CONTENT, QUIZ-GRADING, QUIZ-DEADLINES, and QUIZ-SUBMISSION. We use the label SOCIAL to refer to social interaction posts that do not mention a problem-related aspect.

5.3.2 Dataset

We construct a dataset by sampling posts from MOOC courses to capture the variety of aspects discussed in online courses. We include courses from different disciplines (business, technology, history, and the sciences) to ensure broad coverage of aspects. Although we adopt an approach that does not require labeled data for training, which is important for most practical MOOC scenarios, in order to validate our methods we obtain labels for the sampled posts using *Crowdflower*,* an online crowd-sourcing annotation platform. Each post was annotated by at least 3 annotators. Crowdflower calculates *confidence* in labels by computing trust scores for annotators using test questions. Kolhatkar *et al.* [2013] provide a detailed analysis of Crowdflower trust calculations and the relationship to inter-annotator agreement. We follow their recommendations and retain only labels with *confidence* > 0.5.

*http://www.crowdflower.com/

5.4 Aspect-Sentiment Prediction Models

In this section, we develop models and feature-extraction techniques to address the challenges of aspect-sentiment prediction for MOOC forums. We present two weakly-supervised methods—first, using a seeded topic modeling approach [Jagarlamudi *et al.*, 2012] to identify aspects and sentiment. Second, building upon this method, we then introduce a more powerful statistical relational model which reasons over the seeded LDA predictions as well as sentiment side-information to encode hierarchy information and correlations between sentiment and aspect.

5.4.1 Seeded LDA Model

Topic models [Blei *et al.*, 2003b], which identify latent semantic themes from text corpora, have previously been successfully used to discover aspects for sentiment analysis [Diao *et al.*, 2014]. By equating the topics, i.e. discrete distributions over words, with aspects and/or sentiment polarities, topic models can recover aspect-sentiment predictions. In the MOOC context we are specifically interested in problems with the courses, rather than general topics which may be identified by a topic model, such as the topics of the course material. To guide the topic model to identify aspects of interest, we use *SeededLDA* [Jagarlamudi *et al.*, 2012], a variant of LDA which allows an analyst to "seed" topics by providing key words that should belong to the topics.

We construct SeededLDA models by providing a set of seed words for each of the coarse and fine aspects in the aspect hierarchy of Table 5.2. We also seed topics for *positive*, *negative* and *neutral* sentiment polarities. The seed words for coarse topics are

LECTURE: lectur, video, download, volum, low, headphon, sound, audio, transcript, subtitl, slide, note
QUIZ: quiz, assignment, question, midterm,exam, submiss, answer, grade, score, midterm, due, deadlin
CERTIFICATE: certif, score, signatur, statement, final, course, pass, receiv, coursera, accomplish, fail
SOCIAL: name, course, introduction, stud, group, everyon, student

Table 5.3: Seed words for *coarse* aspects

LECTURE-VIDEO: video, problem, download, play, player, watch, speed, length, fast, slow, render, qualiti
LECTURE-AUDIO: volum, low, headphon, sound, audio, hear, maximum, troubl, qualiti, high, loud, heard
LECTURE-LECTURER: professor, fast, speak, pace, follow, speed, slow, accent, absorb, quick, slowli
LECTURE-SUBTITLES: transcript, subtitl, slide, note, lectur, difficult, pdf
LECTURE-CONTENT: typo, error, mistak, wrong, right, incorrect, mistaken
QUIZ-CONTENT: question, challeng, difficult, understand, typo, error, mistak, quiz, assignment
QUIZ-SUBMISSION: submiss, submit, quiz, error, unabl, resubmit
QUIZ-GRADING: answer, question, answer, grade, assignment, quiz, respons ,mark, wrong, score
QUIZ-DEADLINE: due, deadlin, miss, extend, late

Table 5.4: Seed words for *fine* aspects

POSITIVE: interest, excit, thank, great, happi, glad, enjoy, forward, insight, opportun, clear, fantast, fascin
NEGATIVE: problem, difficult, error, issu, unabl, misunderstand, bother, hate, wrong, mistak, fear, troubl
NEUTRAL: coursera, class, hello, everyon, greet, nam, meet, group, studi, join, introduct, question

Table 5.5: Seed words for *sentiment*

DIFFICULTY: difficult, understand, ambigu, disappoint, hard, follow, mislead, difficulti, challeng, clear
CONTENT: typo, error, mistak, wrong, right, incorrect, mistaken, score
AVAILABILITY: avail, nowher, find, access, miss, view, download, broken, link, bad, access, deni, miss
COURSE-1: develop, eclips, sdk, softwar, hardware, accuser, html, platform, environ, lab, ide, java,
COURSE-2: protein, food, gene, vitamin, evolut, sequenc, chromosom, evolv, mutat, ancestri
COURSE-3: compani, product, industri, strategi, decision, disrupt, technolog, market

Table 5.6: Seed words for sentiment specific to online courses

provided in Table 5.3, and fine aspects in Table 5.4. For the sentiment topics (Table 5.5), the seed words for the topic *positive* are positive words often found in online courses such as *thank*, *congratulations*, *learn*, and *interest*. Similarly, the seed words for the *negative* topic are negative in the context of online courses, such as *difficult*, *error*, *issue*, *problem*, and *misunderstand*.

Additionally, we also use SeededLDA for isolating some common problems in online courses that are associated with sentiment, such as *difficulty*, *availability*, *correctness*,

and course-specific seed words from the syllabus as described in Table 5.6. Finally, having inferred the SeededLDA model from the data set, for each post p we predict the most likely aspect and the most likely sentiment polarity according to the post's inferred distribution over topics $\theta^{(p)}$.

In our experiments, we tokenize and stem the posts using NLTK toolkit [Loper and Bird, 2002], and use a stop word list tuned to online course discussion forums. The topic model Dirichlet hyperparameters are set to α = 0.01, β = 0.01 in our experiments. For SeededLDA models corresponding to the seed sets in Tables 5.3, 5.4, and 5.5, the number of topics is equal to the number of seeded topics. For SeededLDA models corresponding to the seed words in Tables 5.6 and 5.3, we use 10 topics, allowing for some *unseeded* topics that are not captured by the seed words.

5.4.2 Hinge-loss Markov Random Fields

The approach described in the previous section automatically identifies user-seeded aspects and sentiment, but it does not make further use of structure or dependencies between these values, or any additional side-information. To address this, we propose a more powerful approach using hinge-loss Markov random fields (HL-MRFs) [Bach *et al.*, 2015]. Section 2.2.1 provides background on HL-MRFs. In our MOOC aspect-sentiment model, if *P* and *F* denote *post P* and *fine aspect F*, then we have predicates SEEDLDA-FINE(P, F) to denote the value corresponding to topic F in SeededLDA, and FINE-ASPECT(P, F) is the target variable denoting the fine aspect of the post *P*. A PSL rule to encode that the

SeededLDA topic *F* suggests that aspect *F* is present is

$$\lambda : \text{SEEDLDA-FINE}(P, F) \rightarrow \text{FINE-ASPECT}(P, F).$$

We can generate more complex rules connecting the different features and target variables, e.g.,

$$\lambda : \text{SEEDLDA-FINE}(P, F) \wedge \text{SENTIMENT}(P, S) \rightarrow \text{FINE-ASPECT}(P, F).$$

This rule encodes a dependency between SENTIMENT and FINE-ASPECT, namely that the SeededLDA topic and a strong sentiment score increase the probability of the fine aspect. The HL-MRF model uses these rules to encode domain knowledge about dependencies among the predicates. The continuous value representation further helps in understanding the confidence of predictions.

5.4.3 Joint Aspect-Sentiment Prediction using Probabilistic Soft Logic (PSL-Joint)

In this section, we describe our joint approach to predicting aspect and sentiment in online discussion forums, leveraging the strong dependence between aspect and sentiment. We present a system designed using HL-MRFs which combines different features, accounting for their respective uncertainty, and encodes the dependencies between aspect and sentiment in the MOOC context.

Table 7.1 provides some representative rules from our model. The rules are repeated

PSL-JOINT RULES
Rules combining features
SEEDLDA-FINE(POST, LECTURE-LECTURER) ∧ SEEDLDA-COARSE(POST, LECTURE) → FINE-ASPECT(POST, LECTURE-LECTURER)
SEEDLDA-SENTIMENT-COURSE(POST, NEGATIVE) ∧ SEEDLDA-SENTIMENT(POST, NEGATIVE) → SENTIMENT(POST, NEGATIVE)
SEEDLDA-SENTIMENT-COURSE(POST, NEGATIVE)∧SEEDLDA-FINE(POST, QUIZ-GRADING) → FINE-ASPECT(POST, QUIZ-GRADING)
Encoding dependencies between aspect and sentiment
SEEDLDA-FINE(POST, QUIZ-DEADLINES) ∧ SENTIMENT(POST, NEGATIVE) → FINE-ASPECT(POST, QUIZ-DEADLINES)
SEEDLDA-FINE(POST, QUIZ-SUBMISSION) ∧ FINE-ASPECT(POST, QUIZ-SUBMISSION) → COARSE-ASPECT(POST, QUIZ)

Table 5.7: Representative rules from PSL-Joint model

for all sentiment, coarse and fine aspect values.[†] The rules can be classified into two broad categories—1) rules that combine multiple features, and 2) rules that encode the dependencies between aspect and sentiment.

5.4.3.1 Combining Features

The first set of rules in Table 7.1 combine different features extracted from the post. SEEDLDA-FINE, SEEDLDA-COARSE and SEEDLDA-SENTIMENT-COURSE predicates in rules refer to SeededLDA posterior distributions using *coarse*, *fine*, and course-specific *sentiment* seed words respectively. The strength of our model comes from its ability to encode different combinations of features and weight them according to their importance. The first rule in Table 7.1 combines the SeededLDA features from both SEEDLDA-FINE and SEEDLDA-COARSE to predict the fine aspect. Interpreting the rule, the fine aspect of the post is more likely to be LECTURE-LECTURER if the coarse SeededLDA score for the post is LECTURE, *and* the fine SeededLDA score for the post is LECTURE-LECTURER. Similarly, the second rule provides combinations of some of the other features used by the model—two different SeededLDA scores for sentiment, as indicated by seed words in Tables 5.5 and 5.6. The third rule states that certain fine aspects occur together with

[†]Full model available at https://github.com/artir/ramesh-acl15

certain values of sentiment more than others. In online courses, posts that discuss grading usually talk about grievances and issues. The rule captures that QUIZ-GRADING occurs with negative sentiment in most cases.

5.4.3.2 Encoding Dependencies Between Aspect and Sentiment

In addition to combining features, we also encode rules to capture the taxonomic dependence between coarse and fine aspects, and the dependence between aspect and sentiment (Table 7.1, bottom). Rules 4 and 5 encode pair-wise dependency between FINE-ASPECT and SENTIMENT, and COARSE-ASPECT and FINE-ASPECT respectively. Rule 4 uses the SeededLDA value for QUIZ-DEADLINES to predict both SENTIMENT, and FINE-ASPECT jointly. This together with other rules for predicting SENTIMENT and FINE-ASPECT individually creates a constrained satisfaction problem, forcing aspect and sentiment to agree with each other. Rule 5 is similar to rule 4, capturing the taxonomic relationship between target variables COARSE-ASPECT and FINE-ASPECT.

Thus, by using conjunctions to combine features and appropriately weighting these rules, we account for the uncertainties in the underlying features and make them more robust. The combination of these two different types of weighted rules, referred to below as *PSL-Joint*, is able to reason collectively about aspect and sentiment.

5.5 Empirical Evaluation

In this section, we present the quantitative and qualitative results of our models on the annotated MOOC dataset. Our models do not require labeled data for training; we use the la-

bel annotations only for evaluation. Tables 5.8 – 5.11 show the results for the SeededLDA and PSL-Joint models. Statistically significant differences, evaluated using a paired t-test with a rejection threshold of *0.01*, are typed in bold.

Model	LECTURE-CONTENT			LECTURE-VIDEO			LECTURE-AUDIO			LECTURE-LECTURER			LECTURE-SUBTITLES		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
SEEDEDLDA	0.137	0.057	0.08	0.156	0.256	0.240	**0.684**	**0.684**	**0.684**	0.037	0.159	0.06	0.289	**0.631**	0.397
PSL-JOINT	**0.407**	**0.413**	**0.410**	**0.411**	**0.591**	**0.485**	0.635	0.537	0.582	**0.218**	**0.623**	**0.323**	**0.407**	0.53	**0.461**

Table 5.8: Precision, recall and F1 scores for LECTURE fine aspects

Model	QUIZ-CONTENT			QUIZ-SUBMISSION			QUIZ-DEADLINES			QUIZ-GRADING		
	Prec	Rec.	F1	Prec	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
SEEDEDLDA	0.042	0.006	0.011	0.485	**0.398**	**0.437**	0.444	0.141	0.214	0.524	0.508	0.514
PSL-JOINT	**0.324**	**0.405**	**0.36**	**0.521**	0.347	0.416	**0.667**	**0.563**	**0.611**	**0.572**	**0.531**	**0.550**

Table 5.9: Precision, recall and F1 scores for QUIZ fine aspects

Model	LECTURE			QUIZ			CERTIFICATE			SOCIAL		
	Prec	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
SEEDEDLDA	0.597	0.673	0.632	**0.752**	0.583	0.657	0.315	**0.845**	0.459	**0.902**	0.513	0.654
PSL-JOINT	0.563	**0.715**	0.630	0.724	**0.688**	**0.706**	**0.552**	0.711	**0.621**	0.871	**0.530**	**0.659**

Table 5.10: Precision, recall and F1 scores for coarse aspects

Model	POSITIVE			NEGATIVE			NEUTRAL		
	Prec	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
SEEDEDLDA	0.104	**0.721**	0.182	0.650	0.429	0.517	0.483	0.282	0.356
PSL-JOINT	**0.114**	0.544	**0.189**	**0.571**	**0.666**	**0.615**	**0.664**	**0.322**	**0.434**

Table 5.11: Precision, recall and F1 scores for sentiment

5.5.1 SeededLDA for Aspect-Sentiment

For SeededLDA, we use the seed words for *coarse*, *fine*, and *sentiment* given in Tables 5.3 – 5.5. After training the model, we use the SeededLDA multinomial posterior distribution

to predict the target variables. We use the maximum value in the posterior for the distribution over topics for each post to obtain predictions for coarse aspect, fine aspect, and sentiment. We then calculate precision, recall and F1 values comparing with our ground truth labels.

5.5.2 PSL for Joint Aspect-Sentiment (PSL-Joint)

Tables 5.8 and 5.9 give the results for the fine aspects under LECTURE and QUIZ. PSL-JOINT performs better than SeededLDA in most cases, without suffering any statistically significant losses. Notable cases include the increase in scores for LECTURE-LECTURER, LECTURE-SUBTITLES, LECTURE-CONTENT, QUIZ-CONTENT, QUIZ-GRADING, and QUIZ-DEADLINES, for which the scores increase by a large margin over SeededLDA. We observe that for LECTURE-CONTENT and QUIZ-CONTENT, the increase in scores is more significant than others with SeededLDA performing very poorly. Since both lecture and quiz content have the same kind of words related to the course material, SeededLDA is not able to distinguish between these two aspects. We found that in 63% of these missed predictions, SeededLDA predicts LECTURE-CONTENT, instead of QUIZ-CONTENT, and vice versa. In contrast, PSL-Joint uses both coarse and fine SeededLDA scores and captures the dependency between a coarse aspect and its corresponding fine aspect. Therefore, PSL-Joint is able to distinguish between LECTURE-CONTENT and QUIZ-CONTENT. In the next section, we present some examples of posts that SEEDEDLDA misclassified but were predicted correctly by PSL-Joint.

Table 5.10 presents results for the coarse aspects. We observe that PSL-Joint per-

forms better than SeededLDA for all classes. In particular for CERTIFICATE and QUIZ, PSL-Joint exhibits a marked increase in scores when compared to SeededLDA. This is also true for sentiment, for which the scores for NEUTRAL and NEGATIVE sentiment show significant improvement (Table 5.11).

Correct Label	PSL	SeededLDA	Post
QUIZ-CONTENT	QUIZ-CONTENT	LECTURE-CONTENT	There is a typo or other mistake in the assignment instructions (e.g. essential information omitted) Type ID: programming-content Problem ID: programming-mistake Browser: Chrome 32 OS: Windows 7
QUIZ-CONTENT	QUIZ-CONTENT	LECTURE-CONTENT	There is a typo or other mistake on the page (e.g. factual error information omitted) Week 4 Quiz Question 6: Question 6 When a user clicks on a View that has registered to show a Context Menu which one of the following methods will be called?
LECTURE-AUDIO	LECTURE-AUDIO	LECTURE-SUBTITLES	Thanks for the suggestion about downloading the video and referring to the subtitles. I will give that a try but I would also like to point out that what the others are saying is true for me too: The audio is just barely audible even when the volume on my computer is set to 100%.
SOCIAL	SOCIAL	LECTURE-VIDEO	Let's start a group for discussing the lecture videos.

Table 5.12: Example posts that PSL-Joint predicted correctly, but were misclassified by SeededLDA

Correct Label	Predicted Label	Second Prediction	Post
LECTURE-CONTENT	QUIZ-CONTENT	LECTURE-CONTENT	I have a difference of opinion to the answer for Question 6 too. It differs from what is presented in lecture 1.
SOCIAL	LECTURE-SUBTITLES	SOCIAL	Hello guys!!! I am ... The course materials are extraordinary. The subtitles are really helpful! Thanks to instructors for giving us all a wonderful opportunity.
LECTURE-CONTENT	QUIZ-CONTENT	LECTURE-CONTENT	As the second lecture video told me I started windows telnet and connected to the virtual device. Then I typed the same command for sending an sms that the lecture video told me to. The phone received a message all right and I was able to open it but the message itself seems to be written with some strange characters.

Table 5.13: Example posts whose second-best prediction is correct

5.5.3 Interpreting PSL-Joint Predictions

Table 5.12 presents some examples of posts that PSL-Joint predicted correctly, and which SeededLDA misclassified. The first two examples illustrate that PSL can predict the subtle difference between LECTURE-CONTENT and QUIZ-CONTENT. Particularly notable is

the third example, which contains mention of both *subtitles* and *audio*, but the negative sentiment is associated with *audio* rather than *subtitles*. PSL-Joint predicts the fine aspect as LECTURE-AUDIO, even though the underlying SeededLDA feature has a high score for LECTURE-SUBTITLES. This example illustrates the strength of the joint reasoning approach in PSL-Joint. Finally, in the last example, the post mentions starting a *group* to discuss videos. This is an ambiguous post containing the keyword *video*, while it is in reality a social post about starting a group. PSL-Joint is able to predict this because it uses both the sentiment scores associated with the post and the SeededLDA scores for fine aspect, and infers that social posts are generally positive. So, combining the feature values for social aspect and positive sentiment, it is able to predict the fine aspect as SOCIAL correctly.

The continuous valued output predictions produced by PSL-Joint allow us to rank the predicted variables by output prediction value. Analyzing the predictions for posts that PSL-Joint misclassified, we observe that for *four* out of *nine* fine aspects, more than 70% of the time the correct label is in the top three predictions. And, for all fine aspects, the correct label is found in the top 3 predictions around 40% of the time. Thus, using the top three predictions made by PSL-Joint, we can understand the fine aspect of the post to a great extent. Table 5.13 gives some examples of posts for which the second best prediction by PSL-Joint is the correct label. For these examples, we found that PSL-Joint misses the correct prediction by a small margin(< 0.2). Since our evaluation scheme only considers the maximum value to determine the scores, these examples were treated as misclassified.

5.5.4 Understanding Instructor Intervention using PSL-Joint Predictions

In our 3275 annotated posts, the instructor replied to 787 posts. Of these, 699 posts contain a mention of some MOOC aspect. PSL-Joint predicts 97.8% from those as having an aspect and 46.9% as the correct aspect. This indicates that PSL-Joint is capable of identifying the most important posts, i.e. those that the instructor replied to, with high accuracy. PSL-Joint's MOOC aspect predictions can potentially be used by the instructor to select a subset of posts to address in order to cover the main reported issues. We found in our data that some fine aspects, such as CERTIFICATE, have a higher percentage of instructor replies than others, such as QUIZ-GRADING. Using our system, instructors can sample from multiple aspect categories, thereby making sure that all categories of problems receive attention.

5.6 Discussion

In this chapter, we developed a weakly supervised joint probabilistic model (PSL-Joint) for predicting aspect-sentiment in online courses. Our model provides the ability to conveniently encode domain information in the form of seed words, and weighted logical rules capturing the dependencies between aspects and sentiment. We validated our approach on an annotated dataset of MOOC posts sampled from twelve courses. We compared our PSL-Joint probabilistic model to a simpler SeededLDA approach, and demonstrated that PSL-Joint produced statistically significantly better results, exhibiting a 3–5 times improvement in F1 score in most cases over a system using only SeededLDA. As further shown by our qualitative results and instructor reply information, our system can poten-

tially be used for understanding student requirements and issues, identifying posts for instructor intervention, increasing student retention, and improving future iterations of the course. An interesting future direction is analyzing the evolution of topics over iterations to examine how the emphasis on them are changing with time. In the following chapter, we develop models to understand the progression of topics in discussion forums as iterations unfold.

Chapter 6: Topics Evolution Models for Long-running MOOCs

6.1 Introduction

As MOOCs continue to grow, instructors are faced with the problem of understanding the needs and expectations of the ever-changing student population, molding the course to better suit their interests, identifying issues in past iterations, and addressing them in future iterations. This endeavor ensures a smoother delivery of the course and helps in fostering a superior learning experience. With MOOCs, there is tremendous opportunity to develop methods to automatically gauge feedback by interpreting textual content in the discussion forums. The textual content in the forums reflects many important aspects of the course such as the student population and their changing interests, parts of the course that were well received and parts needing attention, and common misconceptions faced by students. While most previous work in this space interpret text in the discussion forums of individual courses [Cui and Wise, 2015; Ezen-Can *et al.*, 2015; Wong *et al.*, 2015; Stump *et al.*, 2013a; Chaturvedi *et al.*, 2014a], it is important to develop models that analyze text in the forums across repeated offerings of a course to provide a panoramic view of course progression. These models can potentially help instructors discern topic patterns corresponding to relevant topics such as course materials and issues, and focus limited instructor resources on addressing the most prevalent and important problems.

In this chapter, we develop models to analyze textual content in discussion forums and draw insights on topic patterns across repeated sequential offerings of two successful long-running MOOCs: i) *thirty four* iterations of a business course (BUSINESS), and 2) *fifteen* iterations of a computer science course (CS). We leverage seeded topic modeling to induce and track evolution of specific topic clusters relevant to online courses across iterations. To the best of our knowledge, ours is the first work in this direction of modeling topic evolution across repeated offerings of a course.

Our main contributions in this chapter are as follows:

- We first categorize the posts according to the following topics: i) social, ii) issue, and iii) technical, to identify and model the important topic themes in the course. The *social* topic captures the social interactions in the forum, the *issue* topic captures posts that report problems in the course, and the *technical* posts refer to course content related discussions in the course. In the BUSINESS course, we observe that issue posts decrease steadily over time, reducing to negligible numbers after 30^{th} iteration. This confirms with our intuition that as iterations unfold, issues in previous iterations are addressed and hence lesser issues are reported. However, in the CS course, we observe an increase in issue posts after the fourth iteration, warranting a finer grained analysis of issues. The course splits into two parts starting from the fourth iteration and our temporal analysis is helpful in understanding how big changes such as splitting the course are received by the students.

- Secondly, we categorize the posts referring to the three most important *course elements* in online courses: i) lectures, ii) quizzes, and iii) certificate, to understand

the emphasis on each of them, respectively. This analysis provides insight on which course elements get more attention in the course and how that is changing with iterations. While lectures are the most popular course element in the BUSINESS course, quizzes surpass lectures in the CS course. Also, we observe that certificate receives more attention in BUSINESS course when compared to CS course. This analysis provides insight on the nature of students in both the courses and how their interests are changing with time, helping instructors to effectively mold their courses to the changing student population.

- We then analyze the distribution of issue posts across the three course elements over the iterations. We find that though lectures are the most dominant course element in the BUSINESS course, most issues are reported on quizzes. We further perform a detailed fine-grained analysis of issues for lecture and quiz course elements, and study the distribution of issue posts across fine-grained lecture and quiz sub-topics. While grading issues tend to occur across both the courses, submission issues predominate the forums in the CS course. Interestingly, we notice that grading issues decline in the BUSINESS course after peer grading is replaced by automatic grading, indicating a general preference for the latter. Our fine-grained analysis throws light on issues faced by students across iterations that could negatively impact student satisfaction and enrollment in future iterations.

- In the CS course, we observe another important dimension, *technical* posts, dominating the issue posts. Technical issues such as software installation are unique to computer science courses and this finding explains the increase in issue posts in the

CS course as iterations progress. We use a combination of seeded LDA models to separate logistic issues from technical issues and find that logistic issues follow a similar trend to the BUSINESS course, declining with time, indicating the need to focus on technical issues.

6.2 Related Work

Recently, there has been a growing interest in understanding text in discussion forums to improve experience of MOOC participants [Cui and Wise, 2015; Ezen-Can *et al.*, 2015; Wong *et al.*, 2015; Stump *et al.*, 2013a; Chaturvedi *et al.*, 2014a]. In this chapter, we build on our previous work on predicting fine-grained aspect-sentiment in online course discussion forums in Chapter 5. We present detailed analysis on how topics of conversation evolve as courses mature. Most previous work in this space focus on understanding forum content in individual courses, but do not model the evolution of content over time. To the best of our knowledge, ours is the first work addressing evolution of courses over time.

Topic evolution has been studied previously on scientific research papers, analyzing their evolution over years. Hall *et al.* [2008] analyze the history of ideas in NLP conferences using topic modeling. They use LDA to detect topic clusters and model the strength of topics over time. Temporal variants of LDA such as Dynamic Topic Model (DTM) [Blei and Lafferty, 2006] and Topics over Time [Wang and McCallum, 2006] model LDA topic and word distributions over time. As Hall *et al.* note, both of these models impose constraints on the time periods making them inflexible and unsuitable for modeling

documents which can change dramatically from one time period to another. The DTM penalizes large changes between subsequent time periods and the beta distributions in Topics over Time are not flexible to large topic changes.

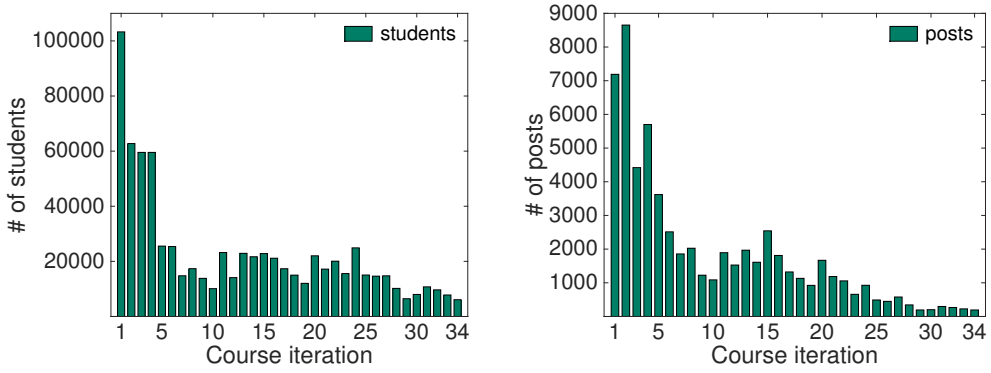
In online courses, each iteration of a course attracts an entirely new set of people and the content in the forums varies significantly according to their interests and backgrounds, rendering the temporal variants of LDA unsuitable for our problem. LDA alone is insufficient as we want to track how specific topics are changing across iterations. Seeding topic models with words corresponding to topics of interest is a simple, yet effective means to track evolution of topics. In our work, we employ a seeded variant of LDA, Seeded LDA [Jagarlamudi *et al.*, 2012] to track specific topics of conversation in the forums through iterations of online courses. We use LDA to identify the possible different topics of interest in online courses and then use these words as seed words to guide discovery of these topics in forum posts.

6.3 Data

We use data from two popular long running Coursera MOOCs: a) a business course, and b) a computer science course, in our analysis. We will refer to these courses as BUSINESS course and CS course, respectively. Both courses are active courses attracting thousands of students every iteration.

6.3.1 Business Course

We analyze 34 iterations of BUSINESS course, each iteration spanning 6 weeks. Figure 6.1 gives course statistics across iterations. The number of students enrolled in the course, given by Figure 6.1(a), shows a steady decline as the iterations unfold.



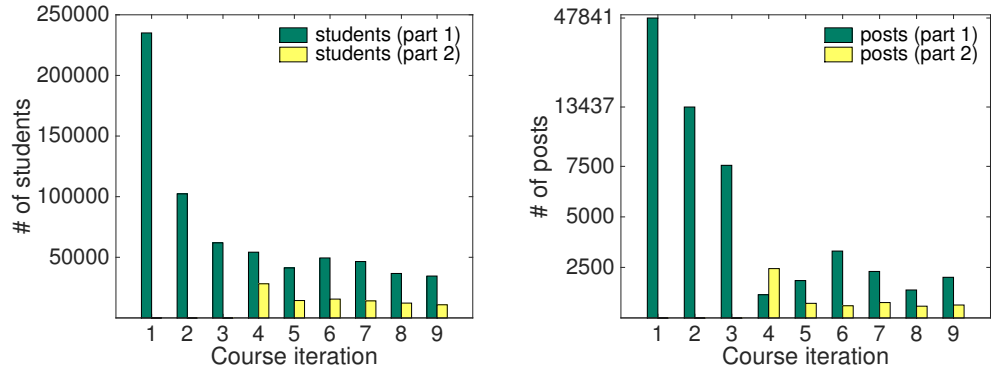
(a) Number of students (b) Number of posts

Figure 6.1: Business course: statistics across iterations

The number of students posting in forums remains consistently 3–5% of the number of students registered in the course across all the iterations. While only a small percentage of registered students post in the forums, a larger percentage of students ($\sim$40%) view forum posts, making forums a very integral part of the course. Figure 6.1(b) gives the number of posts in the forums across iterations. Note that the number of posts follows a similar trend as number of students, declining with time.

6.3.2 Computer Science (CS) Course

The CS course spans 8 weeks for the first three iterations. From the fourth iteration, the course splits into two parts spanning 6 weeks each, which we refer to as CS-1 and CS-2.



(a) Number of students

(b) Number of posts

Figure 6.2: CS course: statistics across iterations

We analyze 6 iterations each of CS-1 and CS-2. In all, we analyze data from 15 course offerings. Figure 6.2 gives the statistics of number of registered students and students posting in the forums in the CS course. Similar to the BUSINESS course, we observe a decline in the number of registered students (6.2(a)) and number of posts (6.2(b)) as the iterations progress. The CS course has a slightly higher percentage of students posting in the forums, around $4 - 7\%$. The highest percentage of posting students is in the first iteration and then it declines slowly, dropping to 4% in the last iteration.

6.4 Topic Discovery in Online Courses

In this section, we build models to discover topics in online courses. We explore LDA to understand discussion forum posts and identify seed words that are relevant to specific topics of interest in online courses.

6.4.1 Topics in Online Courses

We run LDA on posts from both courses to understand the nature of topics and identify topics that are interesting to model over time. Table 6.1 gives the topics identified by LDA. Words that occur in more than one topic are indicated in italics. We hypothesize that this is because students often talk about a variety of topics and some words such as *course*, *learn* tend to occur across posts, making these words part of multiple topics. Seeding topics with words related to specific topics will help us track the evolution of these topics in the course. For this, we turn to a guided LDA variant: seeded LDA [Jagarlamudi *et al.*, 2012].

topic 1: *cours*, *everyon*, lectur, video, assign, problem, grade
topic 2: *cours*, *busi*, *everyon*, *great*, *hello*, *learn*
topic 3: *answer*, *assign*, *grade*, evalu, *cours*, student
topic 4: *compani*, *busi*, everyon, great, interest, hope, *hello*
topic 5: develop, market, product, *compani*, *interest*, *learn*

Table 6.1: Topics identified by LDA

6.4.2 Seeded LDA for Online Courses

In this section, we present the seeded LDA topic models for discovering topics in forum posts. We present models for slicing the data in many different ways to discover different possible topic themes. Tables 6.2, 6.3, and 6.4 give the seed words for our seeded LDA models. We encode seed words for discovering the following different topic categories in online courses.

In the first categorization, we classify the posts into three categories: i) social, ii)

social: hello, everyon, greet, name, meet, group, studi, join, introduct, linkedin
issue: problem, error, issu, unabl, misunderstand, bother, hate, wrong, mistak, fear, troubl

Table 6.2: Seed words for identifying negative sentiment/issues

lecture: lectur, video, download, subtitl, slide
quiz: quiz, assign, exam, assess, score, submit, submiss
certificate: certif, signatur, accomplish

Table 6.3: Seed words for identifying course elements

Fine grained lecture seed words
video/audio: video, play, player, watch, lectur, volum, headphon, audio
subtitles: subtitl, transcript, slide, note, pdf, book
Fine grained quiz seed words
submission: submiss, submit, quiz, resubmit
grading: answer, grade, assignment, quiz, respons, mark, score
deadline: due, deadlin, miss, extend, late

Table 6.4: Seed words for identifying fine-grained issues

technical: compani, product, industri, strategi, entrepreneur, innov, entrepreneuri, busi

Table 6.5: Business course: Seed words for isolating technical issues

technical 1: android, java, eclips, studio, adt, jdk
technical 2: instal, emul, app, run, devic, sdk

Table 6.6: CS course: Seed words for isolating technical issues

issues, and iii) course content topics. These three categories reflect the three primary purposes of the forums in online courses. Social posts are posts that bring out the social element in discussion forums, where students meet and e-socialize with their fellow classmates. These posts usually fall into one of the following subcategories: a) introductions, and b) study groups, as captured by the seed words in *social* topic in Table 6.2. Issue posts are posts that intend to bring issues in the course to the attention of the instructor and fellow classmates or ask for their help in solving them. Topic 2 in Table 6.2 gives the seed words for identifying issue posts in the course. Course content posts discuss course

related material. Tables 6.5 and 6.6 give the course content related seed words pertaining to BUSINESS and CS course, respectively. We will be referring to these posts as *technical* posts.

Unlike most classroom courses, online courses attract diverse set of students with varied interests and expectations from the course. Anderson *et al.* [2014] classify students according to their interaction on the MOOC. Of these, three most common types of students include: 1) students interested in the video lectures, 2) students interested in the assignments, and 3) students taking the course for the certificate. These three types of students map to the three corresponding course elements: i) lectures, ii) quizzes/assignments, and iii) certificate. In the second categorization, we identify posts talking about important course related elements. Analyzing references to these elements in posts help us understand the different types of students in the course and which course elements to focus on improving for future iterations. The *lecture*, *quiz*, and *certificate* topics in Table 6.3 gives the seed words for the three course elements, respectively.

We further drill down on issue posts to identify how they are distributed across fine-grained topics related to course elements. We identify fine grained logistic issues corresponding to course elements *lecture* and *quiz* and categorize the logistic issues in the course into these fine-grained logistic issues. Topics *video* and *subtitles* in Table 6.4 give the fine-grained lecture topics. Topics *submission*, *grading*, and *deadline* are fine-grained quiz topics corresponding to quiz submission, grading, and quiz deadlines, respectively.

6.5 Topic Trends in Online Courses

In this section, we present in-depth analysis of how topics are evolving across iterations of BUSINESS course, and CS course. For each course, we conduct experiments to answer the following questions:

1. How are posts distributed across the three topics constituting the three primary purposes of forums: a) social, b) issues, and c) technical topics, and how is that evolving with time?

2. Next, we answer the question of which course elements are most popular in the course and how is the emphasis on them changing with time?

3. Finally, we drill down deeper on issue posts and analyze what topics constitute the focus of issue posts and how are they changing as iterations unfold?

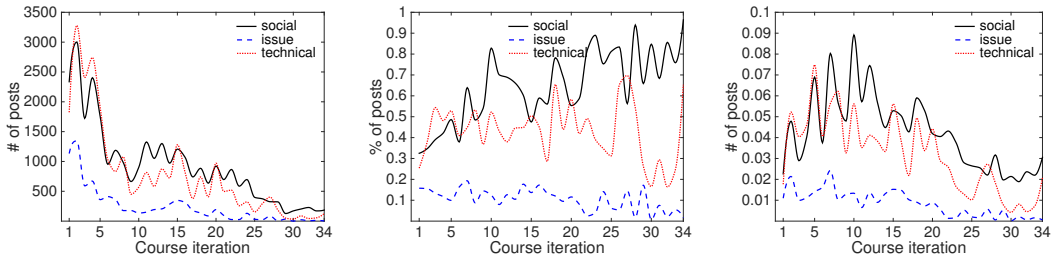
6.5.1 BUSINESS course

In this section, we present topic evolution analysis of posts in BUSINESS course.

6.5.1.1 Primary Purpose of Forums

In our first set of experiments, we run seeded LDA corresponding to the three central topic themes in online courses: 1) social, 2) issues, and 3) technical posts. We run seeded LDA corresponding to seed words in Table 6.2 and 6.5. We include a total of seven topics, including three un-seeded topics to capture posts that fall into other categories. For all our seeded LDA models, we use α = 0.0001 and β = 0.0001. We add the multinomial

distribution given by seeded LDA for each iteration of the course and plot the number of posts in each topic across iterations. Figure 6.3(a) gives the number of *social*, *issue*, and *technical* posts across iterations. In the BUSINESS course, we observe that social posts contribute to a significant number of posts in the forum, stressing the importance of forums as a socializing platform. This is closely followed by technical posts. Issue posts are fewer in number when compared to social and technical posts and decline to negligible numbers in the later iterations. Social and issue posts also decline over time, but always remain higher than issue posts.

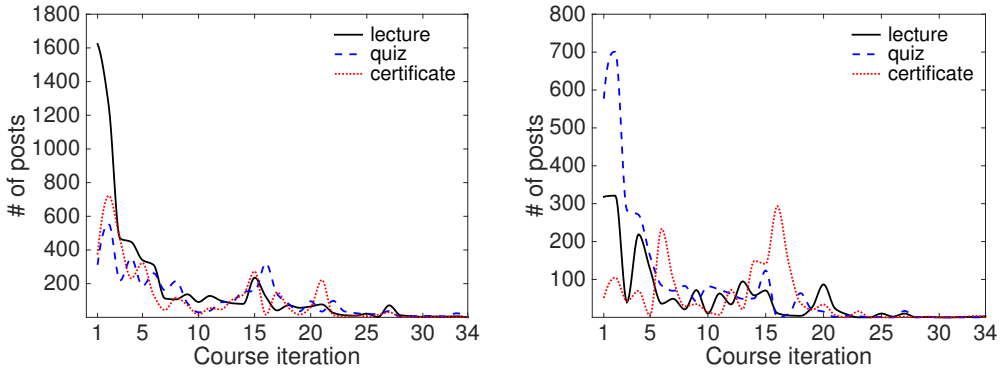


(a) # social, issue, and technical posts (b) % of social, issue, and technical posts (c) # social, issue, and technical posts/student

Figure 6.3: BUSINESS course: evolution of social, issue, and technical posts across iterations

Analyzing the percentage of social, issue, and technical posts in the total number of posts in each iteration, we observe that social and technical topics together constitute a significant percentage ($\sim$80%) of posts. Issues contribute to less than 20% of posts in the early iterations, declining steadily, dropping to less than 10% after 30 iterations. Analyzing the number of social and issue posts per student, we observe that all three categories increase steadily for the initial iterations and then decline, indicating that fewer students participate in the forums as the course stabilizes.

6.5.1.2 Course Elements



(a) # lecture, quiz, certificate posts

(b) # lecture, quiz, certificate *issue* posts

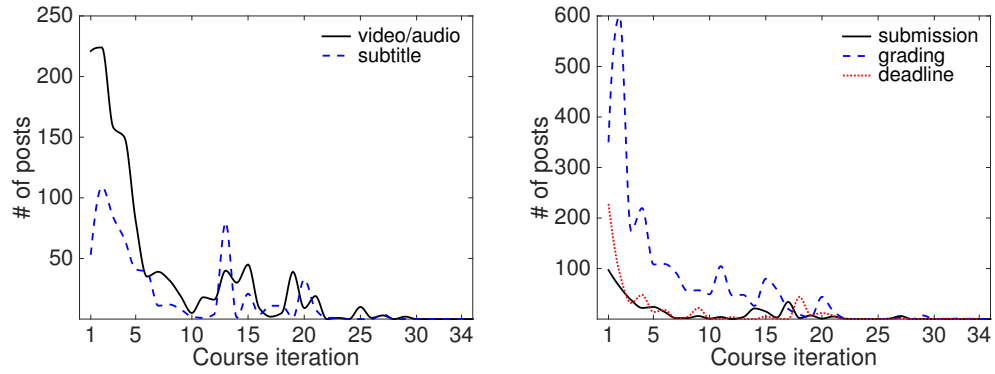
Figure 6.4: BUSINESS course: distribution of posts and issue posts across three course elements: lecture, quiz, and certificate

We run seeded LDA corresponding to the seed words in Table 6.3 with 8 topics, 3 seeded topics and 5 un-seeded topics, we get another categorization of the posts corresponding to emphasis on course elements. Figure 6.4(a) gives the number of posts in the three course elements across iterations. We observe that lectures are the most dominant course element in BUSINESS course across all iterations, followed by quiz, and then certificate.

6.5.1.3 Fine-grained Analysis of Issue Posts

In our third set of experiments, we further drill down on issues and analyze how they are distributed across the course elements. First, we combine the two seeded LDA distributions given by Table 6.2 and 6.3, to categorize issue posts across the three course elements. Figure 6.4(b) gives the distribution of issues across the course elements. It is interesting

to note that while lectures are the most talked about course element, most issues are reported on quizzes in the initial iterations. Throughout all the iterations, we observe that certificates are a popular course element consistently attracting posts in the category. In the middle part of the course around 15^{th} iteration, we observe an increase in interest in certificate in Figure 6.4(a) and a corresponding increase in issues reported on certificate.



(a) # lecture related issue posts (b) # quiz related issue posts

Figure 6.5: Business course: distribution of posts across fine-grained topics

Next, we perform a finer-grained analysis on issues in lecture and quiz course elements and analyze how issues are distributed across finer-grained lecture and quiz sub-topics, given by seed words in Table 6.4. Figure 6.5(a) gives the distribution of issues across lecture sub-topics: video/audio and subtitles. We notice that video/audio issues are more prominent in the earlier iterations. Video/audio and subtitle issues both decline and contribute almost equally to lecture issues in the middle iterations before declining to negligible number of posts in the later iterations. Figure 6.5(b) gives the distribution of issues across quiz sub-topics. We observe that a major proportion of quiz issues fall under grading, with submission and deadline hardly contributing to the issues. Often instructors make modifications to the course responding to feedback from students. Our analysis not

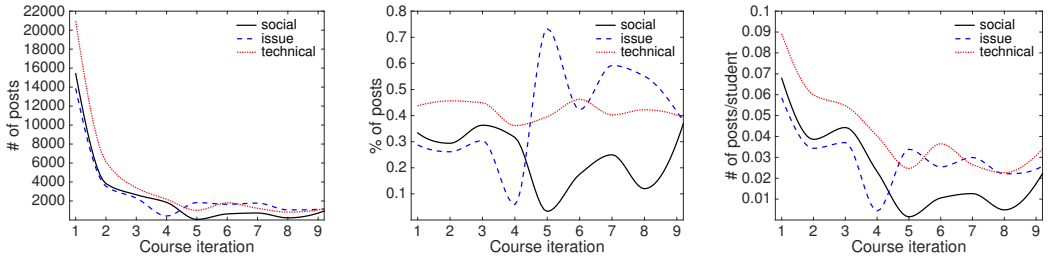
only helps them identify the issues but also provides them with a simple and effective tool to evaluate the success of their fixes. One such example is replacing peer grading with automatic grading from the 3^{rd} iteration. Grading issues follow a steep decline from the third iteration with the introduction of automatic grading in the course, indicating a preference for this grading methodology.

6.5.2 CS course

In this section, we present topic evolution results for nine iterations of CS course.

6.5.2.1 Primary Purpose of Forums

Figure 6.6(a) gives the number of posts in the social, issue, and technical topics across iterations of the course. We notice a very different trend in the CS course when compared to the BUSINESS course. While social posts dominate the forums in the initial iterations, they slowly decline from the fourth iteration. Technical and issue posts primarily dominate the forums from the fifth iteration, with issue posts being slightly more predominant when compared to technical posts.



(a) # social and issue posts (b) % of social and issue posts (c) # social and issue posts per student

Figure 6.6: CS course: change in social and issue posts across iterations

Figure 6.6(b) gives the percentage of social and issue posts in the total number of posts in the forum. Again, we clearly see that as the iterations unfold, the percentage of technical posts remains the same, while there is a marked increase in issue posts from the fifth iteration. Percentage of issue posts reaches the highest value at the fifth iteration and declines thereafter, but still remains higher than social and technical posts. Intuitively, as courses stabilize, it is expected that issues reported in the previous iterations are fixed causing issue posts to decline with time. But in the CS course, we observe the opposite, which calls for a detailed analysis on why issue posts exhibit an increasing trend and what kind of issues are being reported by students. Another interesting trend to note is in Figure 6.6(c), we observe that higher percentage of issue posts come from a fewer number of students when compared to the technical posts.

Comparing issue posts in the CS course to BUSINESS course, we find that the issues reported vary significantly across both courses. Computer science courses often have software installation prerequisites that could potentially trigger a large number of posts around errors in installing/compiling software. Unlike logistic issues, these issues are inherently different in nature and in most cases cannot be easily fixed by the instructor, especially in an online setting. Upon careful analysis of issues in both the courses, we notice that issue posts in CS course mostly revolve around technical category, while BUSINESS course issue posts are mostly around logistic issues. Table 6.7 gives some example posts that are categorized as issues in CS and BUSINESS course.

For the CS course, isolating logistic issues from technical issues is challenging as these posts contain words similar to logistic issues along with specific technical terms. To better understand technical issue posts in CS course, we include extra seeded topics

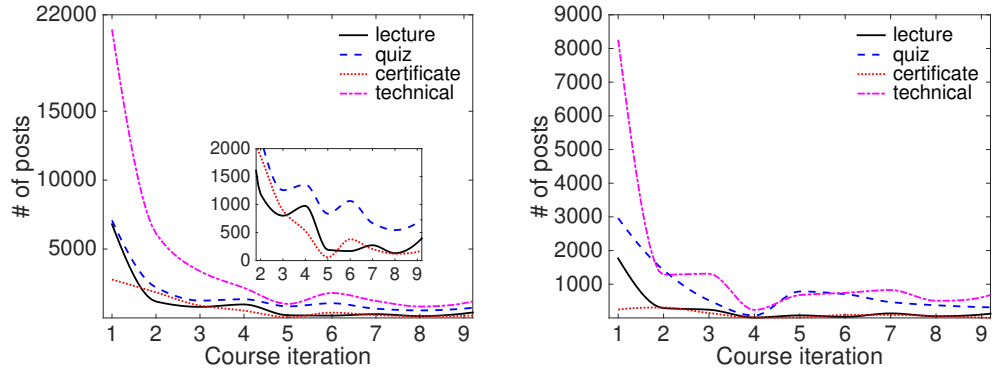
Course	Topic	Post
CS	Technical	The build process constructs this file automatically. If you get errors about not having one, it's almost always because you have errors in your XML files.
	Technical	I tried several times but get NullpointerException when i run my app.
BUSINESS	Logistic	There's no voice in lecture 1.2!
	Logistic	It's troubling that the grading servers do not come to consistent answers when evaluating our code.

Table 6.7: Examples of issue posts from BUSINES and CS courses. CS course has a significant number of technical issue posts, while the issue posts in BUSINESS course are primarily logistic issues.

corresponding to the technical issues as described in Table 6.6. We delve deeper on the issue posts in Section 6.5.2.3.

6.5.2.2 Course Elements

In our next set of experiments, we analyze the evolution of topics corresponding to the three important course elements. We add a *technical* topic in this classification for readability, as we will be drilling deeper into the technical issues along with issues reported in course elements in Section 6.5.2.3. For the technical topic, we add the topic distribution values across all technical topics and present their evolution over time. Figure 6.7(a) gives the evolution of course elements. Notice that quizzes are the dominating course element in the CS course, followed by lectures and certificate in that order. This analysis helps instructors focus on course elements that students care about the most.



(a) # lecture, quiz, certificate, and technical posts

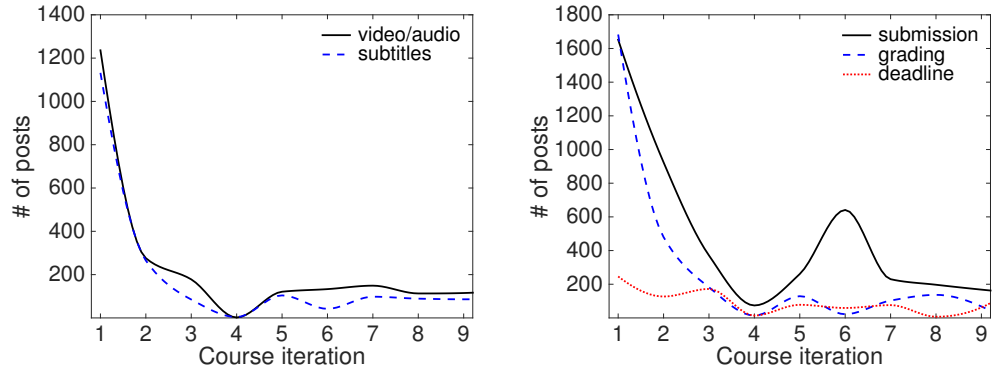
(b) # lecture, quiz, certificate, and technical issue posts

Figure 6.7: CS Technical course: distribution of posts across lecture, quiz, certificate, and technical topics

6.5.2.3 Fine-grained Analysis of Issue Posts

Next, we analyze the course elements and issues to investigate how issues are distributed across the course elements. As we observe in Table 6.7, issue posts in the CS course also fall under the technical topic category. Hence, we add a *technical* category to the list of course elements to understand evolution of technical issue posts. Figure 6.7(b) gives the evolution course element and technical issue topics. We find that the technical issues dominate the issue posts across all the iterations, followed by quiz issues. At iteration 5, where there is an overall increase in issues as indicated by Figure 6.6, we observe a similar spike in the quiz and technical issue posts as well. Lecture and certificate topics hardly contribute to the issue posts and decline to small numbers as iterations progress.

A finer analysis of the distribution of issue posts across lecture and quiz sub-topics are given in Figure 6.8. Figure 6.8(a) gives the distribution of issue posts across lecture sub-topics: video/audio and subtitles. As we observed in Figure 6.7(b), there are only a



(a) # lecture related issue posts (b) # quiz related issue posts

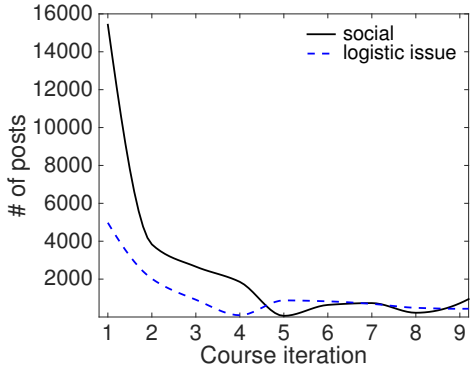
Figure 6.8: CS course: distribution of posts across fine-grained topics

few lecture issue posts in each iteration and this reflects in the finer analysis as well. Between the lecture subtopics, video/audio is the most contributing sub-category. Performing a similar analysis on quiz sub-topics, we find that most of the quiz issue posts fall under the submission category, followed by grading, which is then followed by deadlines.

The submission category refers to assignment submissions, which includes programming assignments. Some submission issue posts can also be perceived as technical issue posts as technical issues sometimes can prevent student from successfully submitting their assignments. While grading consistently remains a contributing issue category across both the courses, we note that the structure of CS course requires submitting computer programs in an online platform which incites a significant number of issue posts in the submission category.

6.5.3 Isolating Logistic Issues in CS Course

The BUSINESS course and the CS course primarily differ in the issue posts. Figure 6.9 gives the distribution of logistic issues and social posts in the CS course. Comparing this to Figure 6.3(a), we find that when we isolate the logistic issues in the CS course, it follows a similar pattern to BUSINESS course, declining over time. This follows our hypothesis that as courses stabilize lesser logistic issues surface and hence they are reported less in the forums.



(a) # social and logistic issues

Figure 6.9: CS course: distribution of social and logistic issues

6.5.4 Analyzing Sentiment in BUSINESS Course

The 34 iterations were divided into 3 phases: early phase, middle phase and the late phase and we analyze the issue posts across these three phases for expression of negative sentiment using OpinionFinder [Wilson *et al.*, 2005b]. Table 6.8 shows posts in the issue category across iterations with highest negative sentiment. Words which are indicative of negative sentiment are italicized. We find that in the early iterations, there is increased

expression of frustration in the posts. The nature of issues reported also vary significantly. The surface level issues such as grading, submission which can be easily fixed by the instructor are more common in the early iterations, while in the later iterations, issues are more focused around understanding course content and availability of certificates. After the 23^rd iteration, issue posts decline and there aren't any issue posts which have negative sentiment. Iteration 23 is the latest iteration which has explicit negative sentiment words in the posts.

Iteration	Post
early	I haven't received my grades too, quite disappointed actually! :(
early	What was wrong with my answers? I answered carefully and everything needed! Why do you grade like that? I graded with 2?s at least they would give me nothing as an answer!! This is unfairr!!! The worst part is that I don't get feedback at all!!!
early	I am also disappointed with the peer evaluation, and not because of the grade I got, I was expecting to a higher score, but it is ok. What I find really annoying is that there is no feedback. What is wrong with my answers? I spent a lot of time reading and evaluating the BP and on return I just ok cold numbers. Not a good learning experience.
middle	I have attempted to submit Week 5 assignment twice now, once inside the deadline and the other a day or so after the soft deadline. In both cases I was using the iPad Coursera app and after submitting all answers, the coursera app logged me out and advised me I was not signed in. This is very frustrating, in terms of the inconvenience and loss of time. Does anyone know of a solution - I cannot use the web platform. thanks
middle	I too have a same problem.am doing assignments and also submitting in time.I would like to request coursera if possible to extend time for signature track.or let us know any others method how our efforts wouldn't go in vain
end	I have the same issue, which was quite confusing as the 2nd edition only goes up to Chapter 11.
end	I didn't receive the Certificate yet plz what is wrong? I checked my accomplishment page every day

Table 6.8: Example of issue posts in business course across iterations

6.5.5 CS Technical and Logistic Posts

In the CS course, we observe that a considerable percentage of posts express problems of both technical and logistic nature in a single post. The seeded LDA multinomial distribution for these posts assign high values for both the respective logistic and technical topics. Table 6.9 gives examples of these posts along with the topic distribution values. Notice that all these posts have high values for both logistic and technical topic categories. Though these posts use words related to technical content, the content is related to logistic

issues. The technical words in these posts are italicized. For example, the first post mentions a *submission* issue, but has the word *compile*, which makes it ambiguous for seeded LDA to classify it.

Logistic	Technical	Post
0.61	0.38	All my assignments run on my system. But my submission says File not found and cannot *compile*. Also my submission has all the required files. Do I have to submit my files again? Please advice.
0.37	0.31	I am also getting the same error - have tried submitting 3 times now. The files compile and tests pass in *Android Studio* on my machine. I even tried creating the submission package from scratch by copying the files directly from the lab skeletons and then only making the required changes.
0.59	0.41	I found my problem. I was calling the *enumerator STATUS* without doing it through the *ToDoItem class*. Apparently, "Status" also have the DONE and NOT DONE keywords.. thanks for your help!

Table 6.9: Posts presenting both logistic and technical issues in CS course

6.6 Discussion

In this chapter, we presented a detailed temporal analysis of discussion forum posts in online courses across different topics relevant in online courses. Our analysis revealed insights on how forums are utilized in the different iterations and which course elements receive more attention and how that varies with time. Our analysis also revealed trends across topics relevant to online courses such as social, and issue-reporting posts. We compared the evolution of these topics across two long-running MOOCs from different disciplines and identified the similarities and differences between them. We also presented a in-depth analysis of issues across the different course elements. Our methodology and analysis is useful for instructors and educators to evaluate the progress of their courses and how big changes to the course such as changing the grading methodology and splitting the course affect the student population. Our analysis is helpful in determining the

stability of these courses and identifying opportunities for improvement. There are several exciting directions to go from here. The temporal analysis can potentially be integrated with a automatic feedback mechanism to help instructors get notified of abrupt changes in the forums and allow them to address these promptly.

Chapter 7: Multi-relational Influence Models for Online Professional Networks

7.1 Introduction

The last decade has witnessed the rise of social networks and their prevalence in our everyday lives. Users perform several actions (e.g., browsing content, adding connections, joining groups) and interactions (e.g., sharing/commenting on content, following people) in a social network. Multiple factors affect user actions and interactions in social networks: personal interests, popularity of an action, or social contacts performing the action *influencing* them to perform the same action. Several works in the past have studied the effect of users' actions on their connections in the social network, which they refer to as *influence* [Goyal *et al.*, 2010; Bakshy *et al.*, 2011]. For example, a user witnessing her friends perform a certain action on a social networking site might be influenced into performing the same action herself. Detecting and quantifying influence is a hard but a very useful problem having a number of applications, which include personalized recommendations [Song *et al.*, 2006; Song *et al.*, 2007], trust modeling [Guha *et al.*, 2004; Ziegler and Lausen, 2005; Golbeck and Hendler, 2006; Taherian *et al.*, 2008], feed ranking [Agarwal *et al.*, 2014], and viral marketing [Domingos and Richardson, 2001;

Richardson and Domingos, 2002; Kempe *et al.*, 2003].

Our work is closest to Goyal *et al.* [2010], who use the action log and the connection graph to learn pairwise influence probabilities between users. Their model is an instance of the General Threshold Model (GTM) [Kempe *et al.*, 2003] for modeling influence propagation in networks. However, their model for calculating influence probabilities only takes a single action type into account. For example, in their evaluation on Flickr social network, they consider only the action of users joining groups. They also do not consider other edge relationships such as organization hierarchy, relationship strength, and individual's seniority in the network that could affect the presence and amount of influence between individuals. Therefore, in this work, we build on Goyal et al.'s approach to design a holistic model that takes into account various action propagations, and other edge relationships between individuals to compute pairwise influence scores.

Our framework based on *hinge-loss Markov Random Fields (HL-MRFs)* combines different heterogenous relationships between individuals to learn influence probabilities. We demonstrate how to encode multiple action propagations, edge relationships, and node features in social networks and use that to learn a combined value of influence that integrates many different interactions between users. We show that influence probabilities between users is a measure of social influence a person exerts on another person in the network and calculating them involves meticulously taking into account all user actions and interactions. Our framework can easily be extended to add other node and edge relationships.

Our main contributions are as follows:

1. We construct a holistic framework capable of encoding multiple pairwise interactions between individuals using a recently developed statistical relational learning method, Hinge-loss Markov random fields (HL-MRFs). We demonstrate how to encode different edge and node relationships that exist in graphs and combine them efficiently to infer influence.

2. We test our models on data from the professional social network, LinkedIn. We generate features that take into account the richness of the dataset and capture different kinds of user interactions. We show that our framework is capable of encoding the rich features in this domain as opposed to previous efforts that can only encode a single action type. Our dataset consists of millions of users and millions of actions comprising of four different types of actions: joining groups, following content, moving jobs, and adding skills to LinkedIn profile.

3. We compare our approach to the state-of-the-art for inferring influence values that extend GTM, using a predictive modeling setup for predicting user actions. We evaluate precision at top *k* for predicting user actions and demonstrate that our models are capable of predicting user actions better than the existing approaches for inferring influence values.

7.2 Problem Definition

Consider a graph G, of the form $G = (V, E, T)$, where nodes V are users, with time-stamped edges E between pairs of users. $E(u, v, t) \in$ E between users u and v represents

the presence of a social network link between u and v, time-stamped with time t when the connection was made. In addition to the social network, we construct an action log by observing the various actions performed by users. Each entry in the action log identifies a single action by the user. We classify user actions into four broad types—1) joining groups, 2) following content, 3) moving jobs, and 4) adding a new skill. The action log is a relation *Actions(User, Action-Type, Action, τ)*, each tuple in the relation representing a user action in the four categories mentioned above. For instance, *(u, group, group-id, τ)* captures that user u joined group group-id at time τ.

Using the action log and the connection graph, we construct an action propagation graph, to capture propagation of actions in the network. The action propagation graphs capture how users' react to actions performed by their connections. Our definition of action propagation is very similar to Goyal *et al.* [2010], except that we add an additional term a_t to identify the action-type.

DEFINITION 1. *An action $a \in A$ of type $a_t \in A_t$ propagates from user v_i to v_j, iff: (i) (v_i, v_j) $\in$ E; (ii) $\exists$(v_i, a_t, a, τ_i), (v_j, a_t, a, τ_j) $\in$ Actions with $\tau_i < \tau_j$; and (iii) $T_{v_i,v_j} \leq \tau_i$. We refer to the action propagation as prop(a, a_t, v_i, v_j, $\triangle\tau$).*

Note that users v_i and v_j should be connected in the social network before either of them perform the action, for it to be considered an action propagation. Using the action propagations, a propagation graph can be constructed for each of the action types mentioned above.

DEFINITION 2. *For each action a of type a_t we define an action propagation graph PG(a, a_t) = (V, E) with unidirectional edges. V= $\{v \mid \exists\tau$: (v, a_t, a, τ) $\in$ Actions$\}$; there is a*

directed edge between $v_i \rightarrow v_j$ in E, whenever prop(a, a_t, v_i, v_j, $\triangle\tau$).

Note that we generate four propagation graphs, for the four types of actions. We refer to our propagation graphs as GROUP-PROP(v_i, v_j), CONTENT-PROP(v_i, v_j), JOB-PROP(v_i, v_j), and SKILL-PROP(v_i, v_j), respectively. We utilize the propagation graphs as features in our model. Section 7.3 gives more details about the action propagation features.

The problem we address in this work is—how to construct rich models of influence that combine information from the social connection graph, action propagation graphs and other node and edge relationships in social graphs such as user seniority in the network, and strength of social connection. For achieving this, we explore HL-MRFs. Section 7.3 gives more details about our framework and features we use in our models.

7.3 Influence Prediction Models

In this section, we first present an overview of GTM and then develop our HL-MRF influence models by incorporating various node features and edge relationships, including the influence values predicted by GTM.

7.3.1 General Threshold Model (GTM)

The GTM formulates any user *u* as either active (already an adopter, in the case of actions, already has performed the action), or inactive. The user *u* is more likely to perform an action when more connections become active, given by the monotonic nature of the activation function. Time unfolds in discrete steps and when user *u* activates, *u* further

can activate other connections of *u* that are not active yet. Equation 7.1 gives probability of user u performing an action ($P_u(S)$), using influence values $P_{v,u}$, where $v \in$ set S of users, who have already performed the action.

$$P_u(S) = 1 - \prod_{v \in S}(1 - P_{v,u}) \tag{7.1}$$

Goyal *et al.*'s model is an instance of GTM. They compute $P_{v,u}$ via the following three approaches: 1) using maximum likelihood estimation, 2) using Jaccard index, and 3) using a discrete time variation model. The discrete time variation model assumes that influence of an active user v on its neighbor remains constant at $P_{v,u}$ for time window of $\tau_{v,u}$ after the v performs the action, and drop to 0 after $\tau_{v,u}$. More details are available in [Goyal *et al.*, 2010].

7.3.2 Hinge-loss Markov Random Fields (HL-MRFs)

The GTM model proposed by Goyal *et al.* is capable of only examining the effect of a single action type on users. To represent and combine different heterogenous relationships between users, we propose a more powerful approach using HL-MRFs.

In our influence model, if *U*, and *V* denote users, then we have predicates JOB-PROP(U, V) to denote the propagation of job from user *U* to user *V* in the action propagation graph, and INFLUENCE (U, V) is the target variable denoting the probability of influence of *U* on *V*. A PSL rule to encode that job propagation from U to V suggest that

U influences V is

$$\lambda : \text{JOB-PROP}(U, V) \rightarrow \text{INFLUENCE}(U, V).$$

We can generate more complex rules connecting the different features and target variables, e.g.,

$$\lambda : \text{JOB-PROP}(U, V) \wedge \text{MANAGES}(U, V) \rightarrow \text{INFLUENCE}(U, V).$$

This rule encodes that if user U propagates job to user V and user U manages user V, then user U influences user V. These rules can be weighted according to their importance using domain knowledge expertise. The HL-MRF model uses these rules to encode domain knowledge about dependencies among the predicates.

7.3.3 Feature Engineering

In this section, we develop the features in our influence models that capture user pairwise interactions and relationships between individuals in a network.

7.3.3.1 Action Propagations

We derive action propagation graphs according to the definition in Section 7.2 for four types of user actions on the site: 1) joining groups, 2) following content, 3) moving jobs/companies, and 4) editing profile, particularly updating skills in the profile. We refer to them as *group propagation*, *content propagation*, *job propagation*, and *skill propa-*

gation, respectively. These features are computed using Definitions 1 and 2. We extract features from the action propagation graphs for these four actions as follows.

If there exists an edge in the action propagation graph for users U and V, then, value of PROPAGATION = 1, else 0. Following this, we generate the features: JOB-PROP, GROUP-PROP, CONTENT-PROP, and SKILL-PROP from the action propagation graphs. For content propagation, we only capture if two people act on the same article, and do not differentiate between different kinds of sub-actions such as liking, sharing, commenting on content.

We determine the sequential nature of the actions, by looking at the time difference between the users making the same action. For jobs, we use the date in users' profile associated with the job rather than using the timestamp when the update was made as users sometimes do not update their positions exactly when they start. To eliminate any uncertainty around propagations, we measure the time difference in months for job propagation. For groups and skills, we measure the time difference in days/minutes, and for content, the time difference is measured in minutes/seconds.

7.3.3.2 Relationship Strength (People You May Know score)

We capture the strength of relationship between two users using the *People You May Know* score [Huang *et al.*, 2013b; Lee *et al.*, 2014]. The score is part of the people recommendation framework at LinkedIn. This score is a unidirectional score in [0, 1]. In our models, we refer to this score by STRENGTH(U, V).

7.3.3.3 Manager-managee Relationship

For employees within LinkedIn, we have the manager-managee relationships available via an internal portal. The predicate MANAGES(U, V) captures the manager-managee relationship in the model, where user U is the manager of user V.

7.3.3.4 Member Seniority score

We use member seniority scores indicating the popularity and reputation of the member in LinkedIn. The predicate SENIORITY(U) captures the seniority of user U within the social network. This is a continuous score in [0, 1].

7.3.3.5 Content Follower-Followee Score

Similar to the relationship score, we can also generate a score for a user following another user's content. This is done by weighting all interactions involving content between two users. Each action has a score according to its importance. For example, *likes* are weighted less than *comments*, which are in turn weighted less than *shares*. This score also is a continuous score in [0, 1]. The People You May Know score, Seniority score and the Content Follower-Followee score are scores part of existing prediction models at LinkedIn.

7.3.3.6 User Influenceability Score

Following Goyal *et al.* [2010], we construct user influenceability score INFL(USER) for users based on how easily they can be influenced by their connections. This is calculated by taking the ratio of number of actions that were propagated to the user and total number

of actions performed by the user.

7.3.3.7 GTM Features

We use the influence values computed by Goyal *et al.* [2010] in their GTM framework as features in our model. We refer to influence scores obtained using maximum likelihood estimation as GTM_{mle}, using Jaccard index as $GTM_{jaccard}$, and the discrete time variation of maximum likelihood estimation as GTM_{DT}.

7.3.4 PSL Influence Models

7.3.4.1 PSL-Influence

We construct weighted logical rules to encode dependencies between the features described in Section 7.3.3 to infer influence. INFLUENCE(U, V) gives the value of influence for pairs of users. The weights in our models are manually specified, taking into account the importance of the feature or combination of features. Table 7.1 gives some representative rules from our PSL-Influence model. The table gives six different combinations of predicates from our PSL-Influence model. The rules combine various edge and node features together to reason about influence. For example, the first rule specifies that if USER-A propagates job to USER-B, then USER-A influences USER-B. The second rule builds on the first rule by adding group propagation to job propagation. It specifies that if USER-A propagated both job and group to USER-B, then, USER-A influences USER-B. By weighting these rules appropriately, we combine the effects of propagation on influence.

Similarly, we use seniority of user along with the propagation graphs to encode

that a user who is senior is more likely to influence other users. For employees within LinkedIn, we also have the MANAGES relationship and we use that along with the propagation graphs to encode that managers usually have an influence on their reports. Since the rule is weighted, it does not mandate that influence relationships should follow manager-managee relationship, but helps to also identify influence that flows from employees to their managers.

Combining user influenceability score and action propagations, we can model that influenceable users are more susceptible to action propagations from their connections. We also incorporate the influence scores from Goyal *et al.*'s model (GTM features), and combine them with seniority scores to infer influence. Also, our framework combines together different inferred influence values from $GTM_{group\text{-}mle}$ and $GTM_{group\text{-}jaccard}$, to eliminate uncertainty and strengthen the scores. The last two rules in our model capture propagation of influence—if USER-A propagates an action to USER-B and USER-B influences USER-C, then USER-A influences USER-C.

7.3.4.2 PSL-Influential

The PSL-Influential model summarizes the edge scores for influencer nodes to measure how influential a person is in the network. This is particularly useful in determining the top influencers in the social network, which has many uses in viral marketing and information diffusion. The predicate to determine if a user is influential is given by *Influential(user)*.

Table 7.2 gives the rules in the model for inferring *infuential* users. If a user propagated multiple actions to other users, then the user is more influential. Also, it is important

PSL-INFLUENCE RULES
Rules combining action propagations
JOB-PROP(USER-A, USER-B) $\rightarrow$ INFLUENCE(USER-A, USER-B)
JOB-PROP(USER-A, USER-B) $\wedge$ GROUP-PROP(USER-A, USER-B) $\rightarrow$ INFLUENCE(USER-A, USER-B)
GROUP-PROP(USER-A, USER-B) $\wedge$ SENIORITY(USER-A) $\rightarrow$ INFLUENCE(USER-A, USER-B)
Rules combining user influenceability and action propagation
GTM_{group}(USER-A, USER-B) $\wedge$ INFL(USER-B) $\rightarrow$ INFLUENCE(USER-A, USER-B)
Rules combining GTM influence values
GTM_{group}(USER-A, USER-B) $\wedge$ SENIORITY(USER-A) $\rightarrow$ INFLUENCE(USER-A, USER-B)
$\text{GTM}_{group\text{-}mle}$(USER-A, USER-B) $\wedge$ $\text{GTM}_{group\text{-}jaccard}$(USER-A) $\rightarrow$ INFLUENCE(USER-A, USER-B)
GTM_{group}(USER-A, USER-B) $\wedge$ $\text{GTM}_{content}$(USER-A, USER-B) $\rightarrow$ INFLUENCE(USER-A, USER-B)
Rules combining seniority and relationship strength
RELATIONSHIP-STRENGTH(USER-A, USER-B) $\wedge$ SENIORITY(USER-A) $\rightarrow$ INFLUENCE(USER-A, USER-B)
Rules combining propagation and manager-managee relationship
GROUP-PROP(USER-A, USER-B) $\wedge$ MANAGES(USER-A, USER-B) $\rightarrow$ INFLUENCE(USER-A, USER-B)
Transitive Rules
GROUP-PROP(USER-A, USER-B) $\wedge$ INFLUENCE(USER-B, USER-C) $\rightarrow$ INFLUENCE(USER-A, USER-C)
CONTENT-PROP(USER-A, USER-B) $\wedge$ INFLUENCE(USER-B, USER-C) $\rightarrow$ INFLUENCE(USER-A, USER-C)

Table 7.1: Representative rules from PSL-Influence model

to notice that apart from action propagations, features such as hierarchical relationship between users inside organization, their connection strength and seniority play an important role in determining influential users. In Section 7.4, we show how we use the influential scores to filter users and improve the influence scores to make more informed predictions. Influential scores, together with the influenceability scores create possibilities for modeling characteristics of both influencer and the person influenced to create more meaningful influence models.

7.4 Experimental Results

In this section, we conduct experiments to: 1) evaluate the the effectiveness of the computed influence values, and 2) interpret influence values and use them to understand social interactions in the social network.

PSL-INFLUENTIAL RULES
Rules combining action propagations
JOB-PROP(USER-A, USER-B) $\rightarrow$ INFLUENTIAL(USER-A)
JOB-PROP(USER-A, USER-B) $\wedge$ GROUP-PROP(USER-A, USER-B) $\rightarrow$ INFLUENTIAL(USER-A)
GROUP-PROP(USER-A, USER-B) $\wedge$ SENIORITY(USER-A) $\rightarrow$ INFLUENTIAL(USER-A)
Rules combining GTM influence values
GTM_{group}(USER-A, USER-B) $\wedge$ SENIORITY(USER-A) $\rightarrow$ INFLUENTIAL(USER-A)
$\text{GTM}_{group\text{-}mle}$(USER-A, USER-B) $\wedge$ $\text{GTM}_{group\text{-}jaccard}$(USER-A, USER-B) $\rightarrow$ INFLUENTIAL(USER-A)
Rules combining propagation and manager-managee relationship
JOB-PROPAGATION(USER-A, USER-B) $\wedge$ MANAGES(USER-A, USER-B) $\rightarrow$ INFLUENTIAL(USER-A)

Table 7.2: Representative rules from PSL-Influential model

7.4.1 Dataset

We test our models on data from the professional social networking site LinkedIn. LinkedIn is the world's largest professional networking site, which enables users to make professional connections, and search jobs. LinkedIn users have a profile page, where they can enlist their education, professional experiences and skills. The presence of an enhanced professional profile contributes to interesting action propagations such as job propagation and skill propagation and makes professional networks very unique and interesting. In addition to that, LinkedIn also has other actions similar to social networks such as a feed customized for each user, which captures the highlights of their connections' activities, and opportunity to create and join groups.

7.4.2 Predicting Actions using Influence scores

First, we run experiments to evaluate the effectiveness of the influence scores. We consider the subset of users comprising of employees at LinkedIn and their social connections. Since there are no true labels for evaluating the influence values, we use these values

to predict user actions of joining groups and following content. For groups, we consider actions in the last five years. For content, we consider actions from last 100 days. We split the data into training and test based on actions and use 90% of data for training and 10% for testing. For the groups data, our test dataset has user-action pairs in the order of millions, around hundreds of thousands of users and tens of thousands of actions. We compare PSL-Influence models to a model based only on GTM. For the GTM models, the parameters $P_{v,u}$, and $\tau_{v,u}$ are calculated at training time. At test time, $P_u(S)$ is calculated using Equation 7.1. For the PSL models, we substitute values from INFLUENCE(V,U) predicate in place of $P_{v,u}$ in Equation 7.1 to predict user actions. We evaluate the models by measuring if the user performs an action in the *top k* predictions generated by the model. We consider *k = 15, 10, 5, and 3* respectively. Tables 7.3 and 7.4 give the precision at top k for GTM and PSL models. The PSL-Influence model performs better than the GTM models in predicting if the users will perform the action.

Further, we use the *influential* scores given by our models to filter influential users and only consider their influence on users. We rank users' connections using the influential scores and retain influencers with influential score greater than 0.5. This is given by *PSL-Influence (Influential Users)* in Tables 7.3 and 7.4. Retaining only influential users in the prediction further improves action prediction scores. Statistically significant differences, evaluated using a paired t-test with a rejection threshold of 0.01, are typed in bold.

Models	*top 15*	*top 10*	*top 5*	*top 3*
GTM-MLE	14.60	14.60	14.53	14.22
GTM-Jaccard	15.30	15.1	14.49	14.10
GTM-DT	15.68	13.56	13.21	13.09
PSL-Influence	16.76	16.67	14.96	13.32
PSL-Influence (Influential users)	**19.01**	**18.89**	**15.83**	**13.33**

Table 7.3: Precision at top k for GTM models, PSL-Influence, and PSL-Influential for predicting users joining groups

Models	*top 15*	*top 10*	*top 5*	*top 3*
GTM-MLE	13.45	13.30	12.53	10.90
GTM-Jaccard	15.48	15.09	13.46	13.01
GTM-DT	16.78	15.66	13.45	12.22
PSL-Influence	18.01	17.86	16.65	16.04
PSL-Influence (Influential users)	**20.22**	**20.12**	**17.66**	**17.01**

Table 7.4: Precision at top k for GTM models, PSL-Influence, and PSL-Influential for predicting users following content

7.4.3 Interpreting Influence scores

The influence scores given by our models capture the pair-wise influence among users in the network. Our experiments in Section 7.4.2 demonstrate that the influence scores can be very useful in predicting user actions. However, the scores themselves carry weight, as they bring out the strength of connections in the social network and also can potentially be helpful in a number of applications such as personalization, recommendations, and ranking relevant content. In this section, we present qualitative results of understanding the scores and comparing them to other edge relationships that can exist in the network.

Two other edge relationship scores that are worth comparing with the influence scores are *relationship-strength* scores, and *organization hierarchy*. We compare the *influence* scores to both these scores to see how the influence scores between the same pair of individuals are different. Around 12% of times, the influence flows in the reverse direction when compared to the manages relationship, i.e., if User A is User B's manager, then the influence is in the opposite direction User B to User A. In such cases, we find that

the employee is often more active in the network, contributing to more actions, which are reciprocated by managers. In around 20% cases, influence between individuals in the same organization is characterized by peers. This verifies how influence relationships do not always flow from top-down in an organization.

Comparing *influence* scores to People You May Know scores, we find that in about 10% of cases, the influence flows in opposite direction to relationship strength. For example, if User *A* and User *B* are connected in a network and $\text{STRENGTH}(A, B) > \text{STRENGTH}(B, A)$, in 10% of cases, $\text{INFLUENCE}(A, B) < \text{INFLUENCE}(B, A)$, and vice-versa.

7.5 Discussion

In this chapter, we presented work on understanding influence in rich behavioral settings, such as online professional networks, by examining multiple edge and node relationships. Our system can be easily extended to more edge relationships, node features and more action types or contexts. There are many exciting directions to go: can we use influence scores in one context to predict influence in other types of actions? Our influence scores can also be potentially used to recommend feed content for users, which primarily consists of the four action types that we consider. Using our influence and influential models, we can generate more meaningful ranking of feed content, by taking into account the top influencers for each person. Our system can also be extended to combine coarse and fine grained interactions between users and to infer action-specific top influencers in the network to make more personalized recommendations.

Chapter 8: Conclusion and Future Work

8.1 Thesis Summary

This thesis focused on developing models for representing and modeling rich socio-behavioral interactions present in online networks, particularly focusing on two prominent types of networks: online courses and online professional networks.

In Chapter 3, I demonstrated how to construct a data-driven model of student engagement for online courses as a complex interaction of different linguistic, structural, and behavioral attributes of student interaction with the course using latent hinge-loss Markov random fields. I then showed how to use the model to predict two different measures of student success: course performance and completion in MOOCs. I then utilized this model to predict student engagement early on in the course, allowing instructors to assess and address student disengagement before students drop out. Further, I performed a detailed quantitative analysis of the different features and their respective importance in predicting student success in MOOCs. This analysis helped understand the various characteristics of online students and how to use that to improve their learning experience.

Next, in Chapter 4, I delved deeper into discussion forums in online courses and developed models to understand and interpret communications of online course participants. I enhanced the student engagement model in Chapter 3 by adding topic and sentiment

features derived from linguistic analysis of discussion forums and employed this model to predict student course completion. I demonstrated that content analysis of discussion forums is helpful in predicting course completion in online courses.

In the following chapters, Chapter 5 and Chapter 6, I further drill down on the MOOC discussion forums to better understand student concerns. In Chapter 5, I developed a weakly-supervised model for modeling fine-grained topics of conversation in MOOC discussion forums. I combined lexical weak supervision in the form of seed words and structural weak supervision in the form of logical PSL rules capturing relationships between the various topics, to predict fine-grained logistic issues and sentiment. My model is capable of encoding both hierarchical and flat structural relationships between various aspects and sentiment in online courses. I demonstrated that my model is capable of predicting fine-grained logistic issues in forums by evaluating on crowdsourced posts sample from twelve online courses.

Building on the discussion forum analysis, in Chapter 6, I performed a temporal analysis of topics in MOOC discussion forums, across iterations of two long running courses. I employed seeding to track the evolution of specific topics in discussion forums. My analysis provided important insights on the nature of students, issues in the course, and their evolution as iterations unfold. I performed a detailed analysis comparing two different courses and highlight similarities and differences in the progression of the courses.

In the penultimate chapter of this thesis, Chapter 7, I shifted my attention to large real-world social networks and developed multi-relational models for understanding influence. I developed an easily extensible system that can represent multiple edge rela-

tionships, node features, characteristic to real-world social networks, to learn values of influence between pairs of people. Then, I used the influence scores to predict actions in social networks and showed that my models are able to predict user actions effectively among a large set of possible user actions.

8.2 Future Work

This thesis brings together several areas in computer science including learning analytics, structured prediction, and topic modeling. I outline some possible extensions along these different sub-areas here.

8.2.1 Learning Analytics

Consider the work on student engagement in Chapter 3. My student engagement models can be extended to model various latent variables influencing student success in online courses, such as motivation and learning. My latent engagement models can be extended to represent these different latent variables and relationships between them. Another interesting improvement to my models is extending them to a continuous setting, which is helpful in offering real-time feedback to instructors regarding student progress. Especially, my early success prediction models when extended to a continuous setting, will be very helpful in alerting instructors as and when students get disengaged. Modeling transition between different engagement and learning states in a continuous manner can help accurately assess student's learning patterns and interests.

My work on discussion forums open up many avenues for future research. In Chap-

ter 5, I model fine-grained logistic aspects that students talk about in forums and in Chapter 6, I track the evolution of both logistic and course related topics across iterations. Using both these techniques to perform a similar in-depth analysis of course content is helpful in understanding common misconceptions students develop in the course. This analysis will shed light on which sub-topics in the course students found interesting, easy to understand versus topics needing more instructor attention. This model combined with a latent student engagement and learning model in Chapter 3 can function as a complete framework to accurately assess student engagement, reasons behind student dropout, identify posts and points in the course for instructor intervention.

My models in Chapters 3, 4, 5, and 6 focus on online courses, but the problems addressed are present across the domain of online networks. My models can be extended to model user engagement in various other online settings, such as site engagement, and job satisfaction. Similarly, my aspect-sentiment model in Chapter 5 can be extended to other natural text settings to model fine-grained topics of discussion in online networks.

8.2.2 Generative Models with Structured Priors

My work on combining structured prediction and generative methods paves way for interesting research directions. In my work, I combine structured prediction and generative methods, by using distributions given by generative models such as LDA as features in the structured prediction framework. The structured prediction frameworks can also be used as priors in the generative models to better model the intricate structural relationships in the data. Foulds *et al.* [2015] propose a variant of LDA using HL-MRF priors,

called Latent Topic Networks (LTN). They incorporate structured priors using hinge-loss terms to constrain the LDA topic and word distributions and derive a EM-based inference algorithm for inferring the hinge-loss terms and LDA topic and word distributions. Their LDA distribution however does not include lexical weak supervision in the form of seeding. Seeding LDA distributions is essential to guide LDA to discover topics of interest and also helps in constraining topic and word distributions better. A natural extension to their model is a seeded version of LTN, that can incorporate seeding as well as structured priors to guide better topic discovery. To accomplish this, seeding must be performed at both topic and word distributions and use seeding to identify relationships between topics and words.

Similarly, mixed membership stochastic blockmodels (MMSB) are a class of generative models for modeling latent memberships in networks [Airoldi *et al.*, 2008]. These models also have similar disadvantages with flat Dirichlet priors and can benefit from structured priors, such as HL-MRFs. Structured priors can be employed to constrain network relationships between groups by taking into account other node and edge attributes, which is not possible with the current formulation.

Inference in latent variable HL-MRFs employs EM point estimates to infer values for the latent variables. Approximating the full distribution with point estimates does not bring forth the full advantages of using continuous variables in HL-MRFs. Developing variational inference methods that preserve the full distribution is very helpful in addressing this problem. Variational inference methods can benefit other models discussed here as well, including LTNs.

Bibliography

[Agarwal *et al.*, 2014] Deepak Agarwal, Bee-Chung Chen, Rupesh Gupta, Joshua Hartman, Qi He, Anand Iyer, Sumanth Kolar, Yiming Ma, Pannagadatta Shivaswamy, Ajit Singh, and Liang Zhang. Activity ranking in linkedin feed. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, 2014.

[Airoldi *et al.*, 2008] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, June 2008.

[Anderson *et al.*, 2014] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Engaging with massive online courses. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2014.

[Bach *et al.*, 2013] Stephen H. Bach, Bert Huang, and Lise Getoor. Learning latent groups with hinge-loss Markov random fields. In *ICML Workshop on Inferning: Interactions between Inference and Learning*, 2013.

[Bach *et al.*, 2015] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. arXiv:1505.04406 [cs.LG], 2015.

[Bakshy *et al.*, 2011] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, WSDM, 2011.

[Balakrishnan, 2013] Girish Balakrishnan. Predicting student retention in massive open online courses using hidden Markov models. Master's thesis, EECS Department, University of California, Berkeley, 2013.

[Blei and Lafferty, 2006] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2006.

[Blei *et al.*, 2003a] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.

[Blei *et al.*, 2003b] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[Brody and Elhadad, 2010] Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The*

2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 804–812. Association for Computational Linguistics, 2010.

[Bruff *et al.*, 2013] Derek O Bruff, Douglas H Fisher, Kathryn E McEwen, and Blaine E Smith. Wrapping a mooc: Student perceptions of an experiment in blended learning. *Journal of Online Learning and Teaching*, 9(2):187, 2013.

[Brusilovsky and Millán, 2007] Peter Brusilovsky and Eva Millán. User models for adaptive hypermedia and adaptive educational systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *Lecture Notes In Computer Science, Vol. 4321*, pages 3–53. 2007.

[Carini *et al.*, 2006] RobertM. Carini, GeorgeD. Kuh, and StephenP. Klein. Student engagement and student learning: Testing the linkages. *Research in Higher Education*, 47(1):1–32, 2006.

[Chaturvedi *et al.*, 2014a] Snigdha Chaturvedi, Dan Goldwasser, and Hal Daumé III. Predicting instructor's intervention in mooc forums. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014.

[Chaturvedi *et al.*, 2014b] Snigdha Chaturvedi, Dan Goldwasser, and Hal Daumé III. Predicting instructor's intervention in mooc forums. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1501–1511. Association for Computational Linguistics, 2014.

[Clow, 2013] Doug Clow. MOOCs and the funnel of participation. In *Proceedings of the International Conference on Learning Analytics and Knowledge (LAK)*, 2013.

[Coetzee *et al.*, 2014] Derrick Coetzee, Armando Fox, Marti A. Hearst, and Björn Hartmann. Should your mooc forum use a reputation system? In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*, 2014.

[Cui and Wise, 2015] Yi Cui and Alyssa Friend Wise. Identifying content-related threads in mooc discussion forums. In *Proceedings of the ACM Conference on Learning @ Scale (L@S)*, 2015.

[Daumé *et al.*, 2009] Hal Daumé, Iii, John Langford, and Daniel Marcu. Search-based structured prediction. *Journal of Machine Learning (JMLR)*, 75(3):297–325, June 2009.

[Diao *et al.*, 2014] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J. Smola, Jing Jiang, and Chong Wang. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 193–202, New York, NY, USA, 2014. ACM.

[Domingos and Richardson, 2001] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, 2001.

[Ezen-Can *et al.*, 2015] Aysu Ezen-Can, Kristy Elizabeth Boyer, Shaun Kellogg, and Sherry Booth. Unsupervised modeling for understanding mooc discussion forums:

A learning analytics approach. In *Proceedings of the International Conference on Learning Analytics And Knowledge (LAK)*, 2015.

[Fahrni and Klenner, 2008] Angela Fahrni and Manfred Klenner. Old wine or warm beer: Target-specific sentiment analysis of adjectives. In *Proc. of the Symposium on Affective Language in Human and Machine, AISB*, pages 60–63, 2008.

[Fakhraei *et al.*, 2014] Shobeir Fakhraei, Bert Huang, Louiqa Raschid, and Lise Getoor. Network-based drug-target interaction prediction with probabilistic soft logic. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(5):775–787, Sept 2014.

[Foulds *et al.*, 2015] James Foulds, Shachi Kumar, and Lise Getoor. Latent topic networks: A versatile probabilistic programming framework for topic models. In *International Conference on Machine Learning (ICML)*, 2015.

[Getoor and Taskar, 2007] Lise Getoor and Ben Taskar. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[Golbeck and Hendler, 2006] Jennifer Golbeck and James Hendler. Inferring binary trust relationships in web-based social networks. *ACM Trans. Internet Technol.*, 6(4):497–529, November 2006.

[Goyal *et al.*, 2010] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the International Conference on Web Search and Data Mining*, WSDM, 2010.

[Guha *et al.*, 2004] R. Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of trust and distrust. In *Proceedings of the International Conference on World Wide Web*, WWW, 2004.

[Guo *et al.*, 2014] Philip J. Guo, Juho Kim, and Rob Rubin. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the ACM Conference on Learning @ Scale Conference (L@S)*, 2014.

[Hall *et al.*, 2008] David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1999.

[Huang *et al.*, 2013a] Bert Huang, Angelika Kimmig, Lise Getoor, and Jennifer Golbeck. A flexible framework for probabilistic models of social trust. In *The 2013 International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction (SBP 2013)*, 2013.

[Huang *et al.*, 2013b] Xinyi (Lisa) Huang, Mitul Tiwari, and Sam Shah. Structural diversity in social recommender systems. In *Proceedings of the RecSys Workshop on Recommender Systems and the Social Web*, 2013.

[Huang *et al.*, 2014] Jonathan Huang, Anirban Dasgupta, Arpita Ghosh, Jane Manning, and Marc Sanders. Superposter behavior in mooc forums. In *Proceedings of the First ACM Conference on Learning @ Scale Conference (L@S)*, 2014.

[Jagarlamudi *et al.*, 2012] Jagadeesh Jagarlamudi, Hal Daumé, III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 204–213, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[Jo and Oh, 2011] Y. Jo and A.H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824. ACM, 2011.

[Jo *et al.*, 2011] Yookyung Jo, John E. Hopcroft, and Carl Lagoze. The web of topics: Discovering the topology of topic evolution in a corpus. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2011.

[Kempe *et al.*, 2003] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, 2003.

[Kim *et al.*, 2013] Suin Kim, Jianwen Zhang, Zheng Chen, Alice Oh, and Shixia Liu. A hierarchical aspect-sentiment model for online reviews. In *Proceedings of The Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13)*. AAAI, July 2013.

[Kizilcec *et al.*, 2013] René F. Kizilcec, Chris Piech, and Emily Schneider. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the International Conference on Learning Analytics and Knowledge (LAK)*, 2013.

[Kolhatkar *et al.*, 2013] Varada Kolhatkar, Heike Zinsmeister, and Graeme Hirst. Annotating anaphoric shell nouns with their antecedents. In *7th Linguistic Annotation Workshop and Interoperability with Discourse*. ACL, 2013.

[Koller and Friedman, 2009] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[Kotsiantis *et al.*, 2003] S.B. Kotsiantis, C.J. Pierrakeas, and P.E. Pintelas. Preventing student dropout in distance learning using machine learning techniques. In Vasile Palade, RobertJ. Howlett, and Lakhmi Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, volume 2774 of *Lecture Notes in Computer Science*, pages 267–274. 2003.

[Kuh, 2003] G. D. Kuh. What were learning about student engagement from nsse: Benchmarks for effective educational practices. *Change: The Magazine of Higher Learning*, 2003.

[Lafferty *et al.*, 2001] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence

data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2001.

[Lavrac and Dzeroski, 1994] Nada Lavrac and Saso Dzeroski. Inductive logic programming. In *WLP*, pages 146–160. Springer, 1994.

[Lee *et al.*, 2014] Pei Lee, Laks V. S. Lakshmanan, Mitul Tiwari, and Sam Shah. Modeling impression discounting in large-scale recommender systems. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

[Lin and He, 2009] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 375–384, New York, NY, USA, 2009. ACM.

[Lin *et al.*, 2012] Chenghua Lin, Yulan He, R. Everson, and S. Ruger. Weakly supervised joint sentiment-topic detection from text. *Knowledge and Data Engineering, IEEE Transactions on*, 24(6):1134–1145, June 2012.

[Liu and Zhang, 2012] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 415–463. Springer, 2012.

[Loper and Bird, 2002] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[Lu *et al.*, 2011] Bin Lu, Myle Ott, Claire Cardie, and Benjamin K. Tsou. Multi-aspect sentiment analysis with topic models. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11, pages 81–88, Washington, DC, USA, 2011. IEEE Computer Society.

[Mukherjee and Liu, 2012] Arjun Mukherjee and Bing Liu. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 339–348, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[Nallapati *et al.*, 2007] Ramesh M. Nallapati, Susan Ditmore, John D. Lafferty, and Kin Ung. Multiscale topic tomography. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

[Neville and Jensen, 2007] Jennifer Neville and David Jensen. Relational dependency networks. *The Journal of Machine Learning Research*, 8:653–692, 2007.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

[Poon and Domingos, 2011] Hoifung Poon and Pedro Domingos. Sum-product networks: A new deep architecture. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 689–690. IEEE, 2011.

[Pujara *et al.*, 2013] Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. Knowledge graph identification. In *International Semantic Web Conference (ISWC)*, 2013. Winner of Best Student Paper award.

[Qiu *et al.*, 2016] Jiezhong Qiu, Jie Tang, Tracy Xiao Liu, Jie Gong, Chenhui Zhang, Qian Zhang, and Yufei Xue. Modeling and predicting learning behavior in moocs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM)*, 2016.

[Ramesh *et al.*, 2013] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. Modeling learner engagement in moocs using probabilistic soft logic. In *NIPS Workshop on Data Driven Education*, 2013.

[Ramesh *et al.*, 2014a] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. Learning latent engagement patterns of students in online courses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014.

[Ramesh *et al.*, 2014b] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. Uncovering hidden engagement patterns for predicting learner performance in moocs. In *ACM Conference on Learning at Scale (L@S)*, 2014.

[Ramesh *et al.*, 2014c] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daumé Understanding mooc discussion forums using seeded lda. In *9th ACL Workshop on Innovative Use of NLP for Building Educational Applications*. ACL, 2014.

[Ramesh *et al.*, 2015a] Arti Ramesh, Shachi Kumar, James Foulds, and Lise Getoor. Weakly supervised models of aspect-sentiment for online course discussion forums. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015.

[Ramesh *et al.*, 2015b] Arti Ramesh, Mario Rodriguez, and Lise Getoor. Understanding influence in online professional networks. In *NIPS Workshop on Networks in Social and Information Sciences*, 2015.

[Richards, 2012] S. J. Richards. A handbook of parametric survival models for actuarial use. *Scandinavian Actuarial Journal*, 2012(4):233–257, 2012.

[Richardson and Domingos, 2002] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, 2002.

[Richardson and Domingos, 2006] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1-2):107–136, 2006.

[Rocca, 2010] Kelly A Rocca. Student participation in the college classroom: An extended multidisciplinary literature review. *Communication Education*, 59(2):185–213, 2010.

[Song *et al.*, 2006] Xiaodan Song, Belle L. Tseng, Ching-Yung Lin, and Ming-Ting Sun. Personalized recommendation driven by information flow. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, 2006.

[Song *et al.*, 2007] Xiaodan Song, Yun Chi, Koji Hino, and Belle L. Tseng. Information flow modeling based on diffusion rate for prediction and ranking. In *Proceedings of the International Conference on World Wide Web*, WWW, 2007.

[Stump *et al.*, 2013a] Glenda S. Stump, Jennifer DeBoer, Jonathan Whittinghill, and Lori Breslow. Development of a framework to classify mooc discussion forum posts: Methodology and challenges. In *NIPS Workshop on Data Driven Education*, 2013.

[Stump *et al.*, 2013b] Glenda S. Stump, Jennifer DeBoer, Jonathan Whittinghill, and Lori Breslow. Development of a framework to classify mooc discussion forum posts: Methodology and challenges. In *NIPS Workshop on Data Driven Education*, 2013.

[Taherian *et al.*, 2008] Mohsen Taherian, Morteza Amini, and Rasool Jalili. Trust inference in web-based social networks using resistive networks. In *Proceedings of the International Conference on Internet and Web Applications and Services*, ICIW, 2008.

[Taskar *et al.*, 2003] Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin markov networks. In *Proceedings of the advances in Neural Information Processing Systems (NIPS)*, 2003.

[Titov and McDonald, 2008] Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 111–120, New York, NY, USA, 2008. ACM.

[Tsochantaridis *et al.*, 2004] Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

[Wang and McCallum, 2006] Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2006.

[Wilson *et al.*, 2005a] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, 2005.

[Wilson *et al.*, 2005b] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, 2005.

[Wong *et al.*, 2015] Jian-Syuan Wong, Bart Pursel, Anna Divinsky, and Bernard J. Jansen. *An Analysis of MOOC Discussion Forum Interactions from the Most Active Users*, pages 452–457. Springer International Publishing, 2015.

[Yang *et al.*, 2013] Diyi Yang, Tanmay Sinha, David Adamson, and Rose Penstein. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *NIPS Workshop on Data Driven Education*, 2013.

[Zhao *et al.*, 2010] Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 56–65, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Ziegler and Lausen, 2005] Cai-Nicolas Ziegler and Georg Lausen. Propagation models for trust and distrust in social networks. *Information Systems Frontiers*, 7(4-5):337–358, December 2005.