A Probabilistic Approach to Modeling Socio-Behavioral Interactions

Arti Ramesh

Dissertation proposal submitted to the
Department of Computer Science
University of Maryland, College Park, MD 20742

April 30, 2015

Advisor: Dr. Lise Getoor (U. of Maryland)
Committee: Dr. Hal Daume III (U. of Maryland)
            Dr. Amol Deshpande (U. of Maryland)
Abstract

The vast growth and reach of internet and social media have led to a tremendous increase in socio-behavioral interaction content on the web. The ever-increasing number of online interactions have led to a growing interest to understand and interpret online communications to enhance user experience. This includes personalization, user retention, predicting user interests, and product recommendations. In this thesis, I address how to use machine learning methods to model socio-behavioral interactions and predict user behavior patterns in online networks. In the first part of this proposal, I focus on one such emerging online interaction platform—online courses (MOOCs). Structured data from these courses contain behavioral, and interaction data and provide opportunity to design machine learning methods for understanding user interaction. The data also contains unstructured data, such as natural language text from forum posts and other online discussions. I present a family of probabilistic models that I have developed for: 1) modeling student engagement, 2) predicting student completion and dropouts, 3) modeling student sentiment toward various course aspects (e.g., content vs. logistics), and 4) detecting coarse and fine-grained course aspects (e.g., grading, video, content) in 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 second part of the proposal, I describe how I plan to extend the above-mentioned models to model socio-behavioral interactions at multiple scales in networks. I plan to test the effectiveness of this model via experimentation on different types of platforms such as MOOCs and professional networks (e.g., LinkedIn).
Contents

1 Introduction 4

2 Background 8
   2.1 Related Work 8
      2.1.1 Learning Analytics 8
      2.1.2 Structured Prediction 9
      2.1.3 Sentiment Analysis and Topic Modeling 11

3 Latent Variable Models for Student Engagement in MOOCs 12
   3.1 Introduction 12
   3.2 Modeling Student Course Success Indicators:
      Course Performance and Survival 14
      3.2.1 Modeling MOOC Student Activity 14
      3.2.2 PSL Student Survival Models 16
   3.3 Empirical Evaluation 19
      3.3.1 Datasets and Experimental Setup 19
      3.3.2 Student Performance Analysis 19
      3.3.3 Student Survival Analysis 21
      3.3.4 Feature Analysis 22
      3.3.5 Gaining Insight from Latent Engagement Assignments 24
   3.4 Conclusion 27

4 Content Analysis of MOOC Forums 29
   4.1 Introduction 29
   4.2 MOOC Forum content analysis for Modeling Student Survival 30
      4.2.1 Background 31
      4.2.2 Enhancing Student Survival Models with Topic Modeling 32
      4.2.3 Empirical Evaluation 35
      4.2.4 Datasets and Experimental Setup 35
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 on the web. 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 had lead to a growing interest to understand and interpret online interactions to enhance user experience. This includes personalization, user retention, predicting user interests, and product and friend recommendations.

In the first part of this thesis proposal, I will focus on an important such online interaction platform—online courses. 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 free of cost to people around the world. 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 use machine learning for studying 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 sta-
tical relational learning based methods to improve the teaching and learning experience for MOOC participants. Chapter 3 of this proposal focusses on understanding and defining engagement in the context of online courses. 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. In Chapter 4, I delve more in detail into MOOC discussion forums, which serve as a bridge connecting MOOC participants across the world, and present methods to understand issues and feedback about the course. I develop methods for mining MOOC forums for cues about student interests and problems and investigate how that relates to their course completion. I present methods to automatically identify problem-reporting posts and the fine-grained problems reported that will be helpful for instructors. In Chapter 5, I identify other domains where there are similar socio-behavioral interactions and extensions to the framework toward understanding user interactions at multiple scales in socio-behavioral platforms, toward a thesis dissertation.

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. In Chapter 3, I study factors that affect student engagement in MOOCs. I develop a framework for modeling and understanding student engagement in online courses based on student behavioral cues.

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 types of 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 for student success at different points in the course.

Discussion forums serve as a platform for student discussions in massive open online courses (MOOCs). With the increase in popularity of MOOCs, there is a corresponding increase in the 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 my analysis of forum posts, I find that a lot of forum posts reporting issues go unanswered as they get lost in the mire of thousands of posts and often leads to students dropping out of the course. Students often resort to asking to up-vote posts to gain instructor attention. Automatically detecting posts that report problems would not only help instructors address the problems, but also improve the learning experience for the students. Identifying topics or aspects of conversation and inferring sentiment in online course forum posts can enable instructor interventions to meet the needs of the students, rapidly address course-related issues, and increase student retention. Labeled aspect-sentiment data for MOOCs are expensive to obtain and may not be transferable between courses, suggesting the need for an unsupervised approach. I develop an unsupervised 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.

I now discuss my plan for the rest of my dissertation. Mining discussion forum content provided many insights on online courses. My work on mining discussion forums for aspect and sentiment focuses on posts that report logistic problems with the course. But, I observed that forums also house posts that pose conceptual doubts and misunderstandings of course material. Automatically identifying conceptual doubts and aligning them with the topics in the syllabus could help instructors identify and correct possible misconceptions. I also plan to extend the hierarchical framework developed for studying course aspects in discussion forums to model multi-scale socio-behavioral interactions. Most networks exhibit multi-scale organization. I plan to investigate how to efficiently model multi-scale interactions and use that in prediction tasks such as link prediction, recommendations, modeling group behavior.

The contribution 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, 2) Second, I demonstrate how to use the data-driven formulation for predicting student success in MOOCs, 3) I demonstrate the utility of content analysis of discussion forums, using it to predict student course completion, and 4) I develop methods for predicting fine-grained problems and student opinion in discussion forums.

The rest of the thesis is organized as follows: Chapter 2 reviews work in the five areas – Learning Analytics, Structured Prediction, Latent Variable models, Temporal models, and Sentiment and Aspect mining and describes the tools and methods
I use in my work. In Chapter 3, I present my first work on student engagement in MOOCs. In Chapter 4, I present work on forum content analysis. In Chapter 5, I present extensions to our approach to incorporate hierarchical and multi-scale reasoning of social networks as proposed research and the timeline for this thesis.
Chapter 2

Background

In this section, I provide background for the rest of the proposal. I review work on four areas related to my thesis – 1) Learning analytics, 2) Structured Prediction and Latent Variable methods, and 3) Sentiment Analysis and Topic Modeling, delving into more detail on models and frameworks that I use in my work.

2.1 Related Work

2.1.1 Learning Analytics

Prior work [Kuh, 2003; Carini et al., 2006] has studied 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.

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 survival. 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 a latent variables and learn to differentiate among the engagement types from data.

2.1.2 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 [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 methods, directly models uncertainty over unknowns. Probabilistic graphical models enables compact representations of joint distributions over interdependent unknowns through graphical structures. The PGM formalism enabled the development of reusable algorithms applicable to different distributions.

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., 2013b]. In the following section, I provide an overview of HL-MRFs and a templating language for HL-MRFs—Probabilistic Soft Logic (PSL) [Kimmig et al., 2012].
Hinge-loss Markov random fields (HL-MRF) and Probabilistic Soft Logic (PSL)

Hinge-loss Markov random fields (HL-MRFs) are a scalable class of continuous, conditional graphical models [Bach et al., 2013b]. These models can be specified using Probabilistic Soft Logic (PSL) [Kimmig et al., 2012], a weighted first order logical templating language. An example of a PSL rule is

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

where $P$, $Q$, and $R$ are predicates, $a$ and $b$ are variables, and $\lambda$ is the weight associated with the rule. The weight of the rule indicates its importance in the HL-MRF probabilistic model, which defines a probability density function of the form

$$P(Y|X) \propto \exp \left( - \sum_{r=1}^{M} \lambda_r \phi_r(Y, X) \right)$$

$$\phi_r(Y, X) = (\max\{l_r(Y, X), 0\})^{\rho_r}, \quad (2.1)$$

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

Probabilistic Soft Logic

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 [Getoor and Taskar, 2007], 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 a hinge-loss Markov random field (HL-MRF) [Bach et al., 2013b]. Inference in HL-MRFs is a convex optimization problem, which makes working with PSL very efficient in comparison to relational modeling tools that use discrete representations.

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., 2013a]. In our models, we use this capability to represent student engagement as a latent variable.

Latent Variable Models using Probabilistic Soft Logic

Latent variables can improve the quality of probabilistic models in many ways. Using latent variables to mediate probabilistic interactions can improve general-
ization by simplifying models. Assignments to continuous latent variables can express more nuanced information than assignments to discrete variables, such as any interpolation between true and false, not just the extreme values.

HL-MRFs trained with hard EM are accurate and scalable for three reasons: (1) the continuous variables of HL-MRFs can express rich, latent phenomena, such as mixed group membership, in a hard EM framework, (2) fast, exact MPE inference for HL-MRFs can quickly identify the most probable assignments to variables, and (3) HL-MRFs can easily express complex, yet interpretable, dependency structures among latent variables.

2.1.3 Sentiment Analysis and Topic Modeling

Latent Dirichlet Allocation is the most popular method used for finding topics in documents. Several improvements to LDA have been proposed. Traditional LDA is unsupervised and is applied on the document collection without specifying the nature and type of documents. SeededLDA [Jagarlamudi et al., 2012] is a variation of LDA, which allows the user to specify sample words from the document collection that helps LDA to understand the nature of the documents and which topics occur together.

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., 2003] and variants of it for detecting aspect and sentiment [Lu et al., 2011; Lin and He, 2009]. Liu and Zhang provide a comprehensive survey of techniques for aspect and sentiment analysis.

Titov and McDonald emphasize the importance of an unsupervised approach for aspect detection. However, the authors also indicate that standard LDA [Blei et al., 2003] methods capture global topics and not necessarily pertinent aspects — a challenge that we address in this work. Brody and Elhadad, Titov and McDonald [2008], and Jo and Oh [2011] apply variations of LDA at the sentence level for online reviews.

Most previous approaches for sentiment rely on manually constructed lexicons of strongly positive and negative words [Fahrni and Klenner, 2008; Brody and Elhadad, 2010].

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]. [2013] use a hierarchical aspect-sentiment model and evaluate it for online reviews. Mukherjee and Liu use seed words for discovering aspect-based sentiment topics.
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 can help improve 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 work, we study the different aspects of online student behavior in MOOCs, develop a large-scale, data-driven approach for modeling student engagement. We identify two course success indicators for online courses—1) performance: whether the student earned a certificate in the course, and 2) survival: whether the student followed the course till completion. We demonstrate the construction of a holistic model incorporating content (e.g., language), structure (e.g., social interactions), 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 probabilistic soft logic (PSL) [Broecheler et al., 2010], a framework that provides an easy means to represent and combine behavioral, linguistic, and structural features in a concise manner. We analyze students’ online behavior to identify how they engage with course materials and investigate
how engagement can be helpful in predicting successful completion of the course and performance in the course. 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]. Our work is a step toward helping educators understand how students interact with 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 model engagement using latent variables, which take into account the observed behaviors of online students and their resulting survival and performance in the class. Uncovering this latent information 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 a model that uses the different behavioral aspects to distinguish between forms of engagement: passive or active. We use these engagement types to predict student performance and survival, and reason about their behavior over time.

We apply our model to real data collected from several Coursera\(^1\) courses and empirically show its ability to capture behavioral patterns of students and predict student performance and course completion. 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 complete the class. We show that our model is able to make meaningful class completion predictions using data obtained at an early stage in the class. These predictions can help provide a basis for instructor intervention at an early stage in a course, helping to improve student retention rates.

\(^1\)https://www.coursera.org
3.2 Modeling Student Course Success Indicators: 
Course Performance and Survival

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 describe how we model course success indicators—1) performance: whether the student earned a certificate in the course, and 2) survival: whether the student completed the course, connecting it to the various behavioral, linguistic features of these student interactions. We will refer to both these course success indicators as survival, for ease of understanding.

To model the interactions between these features and student survival, we use probabilistic soft logic (PSL), a system for relational probabilistic modeling. 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 survival.

3.2.1 Modeling MOOC Student Activity

The MOOC online environment mainly consists of two resources: video lectures and forums. Students can watch lectures multiple times and respond to on-demand quizzes during the lectures\(^2\). 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, which are used to create features. We categorize these predicates as either behavioral, interaction-based, or temporal, and describe them in the following sections.

**Behavioral Features** Behavioral features are attributes that the student exhibits while on the MOOC website. We consider two types of behavioral features: aggre-

---

\(^2\)These quizzes are generally not used to calculate the final evaluation.
Aggregate and non-aggregate.

Aggregate features describe the student’s behavior, relative to others. The predicates postActivity(USER), voteActivity(USER), and viewActivity(USER) capture user activity in the forums. The student’s reputation is captured using reputation(USER). These are calculated for each user by assessing if the value of the feature is more than the median value considering all users. The aggregate predicates take Boolean values.

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.

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., 2005] 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 sameThread(POST_1, POST_2) and sameForum(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 survival and engagement values among them.

Temporal Features  
Student activity level changes 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 lastQuiz, lastLecture, lastPost, lastView and lastVote 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.

Constructing Complex Rules  
We use the features above to construct meaningful PSL rules using logical connectives, as demonstrated in Table 3.1\(^3\). The PSL model associates these rules with student survival, either directly or indirectly using latent variables. We explain this process in the following section.

<table>
<thead>
<tr>
<th>• Behavioral Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>postActivity(U) (\wedge) reputation(U)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>• Forum Content Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>posts(U, P) (\wedge) polarity(P)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>• Forum Interaction Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>posts(U_1, P_1) (\wedge) posts(U_2, P_2) (\wedge) sameThread(P_1, P_2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>• Temporal Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>lastQuiz(U, T_1) (\wedge) lastLecture(U, T_1) (\wedge) lastPost(U, T_1)</td>
</tr>
</tbody>
</table>

Table 3.1: Constructing complex rules in PSL

3.2.2 PSL Student Survival Models

Probabilistic relational modeling is a popular approach for capturing structural dependencies such as the one above, and have been applied to a wide range of
problems. We are interested in predicting two facets of student success in online courses—1) performance: whether the student earned a certificate, and 2) survival: whether the student takes the last few quizzes/assignments in the class. In both these prediction scenarios, performance and survival are both calculated as Boolean value. For example, performance = 1 if student earns a certificate from the course and performance = 0 if the student does not. Similarly, survival = 1, if the student takes the last few quizzes/assignments, 0 otherwise.

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

\[
\begin{align*}
\text{postActivity}(U) \land \text{reputation}(U) & \rightarrow \text{survival}(U) \\
\text{voteActivity}(U) \land \text{reputation}(U) & \rightarrow \text{survival}(U) \\
\text{posts}(U, P) \land \text{polarity}(P) & \rightarrow \text{survival}(U) \\
\text{posts}(U, P) \land \neg \text{polarity}(P) & \rightarrow \neg \text{survival}(U) \\
\text{posts}(U, P) \land \text{upvote}(P) & \rightarrow \text{survival}(U) \\
\text{posts}(U_1, P_1) \land \text{posts}(U_2, P_2) \land \text{survival}(U_1) \land \text{sameThread}(P_1, P_2) & \rightarrow \text{survival}(U_2)
\end{align*}
\]

Table 3.2: Rules for the DIRECT model.

\[
\begin{align*}
\text{postActivity}(U) \land \text{reputation}(U) & \rightarrow \text{eActive}(U) \\
\text{voteActivity}(U) \land \text{reputation}(U) & \rightarrow \text{ePassive}(U) \\
\text{posts}(U, P) \land \text{polarity}(P) & \rightarrow \text{eActive}(U) \\
\text{votes}(U, P) \land \text{polarity}(P) & \rightarrow \text{ePassive}(U) \\
\text{posts}(U, P) \land \text{upvote}(P) & \rightarrow \text{eActive}(U) \\
\text{posts}(U_1, P_1) \land \text{posts}(U_2, P_2) \land \text{eActive}(U_1) \land \text{sameThread}(P_1, P_2) & \rightarrow \text{eActive}(U_2) \\
\text{eActive}(U) \land \text{ePassive}(U) & \rightarrow \text{survival}(U)
\end{align*}
\]

Table 3.3: Rules for the LATENT model.

**PSL-DIRECT**

In our DIRECT PSL model, we model student survival by using the observable behavioral features exhibited by the student, linguistic features corresponding to the content of posts, and structural features derived from forum interactions. Meaningful combinations of one or more observable behavioral features (described in the Features section) are used to predict survival. Table 5.1 contains a subset of rules
used in this model (U and P in tables 5.1 and 3.3 refer to USER and POST respectively). As evident in these examples, the simple model contains rules that allow observable features to directly imply student survival.

**PSL-LATENT**

In our second model, we enhance this type of reasoning by including latent variables semantically based on concepts of student engagement. These variables cannot be directly measured from the data, and we therefore treat student engagement as a latent variable and associate various observed features to one or more forms of engagement.

We define three types of engagement variables, denoted `engagement_active`, `engagement_passive` and `disengagement` to capture three types of student engagement in MOOCs. `engagement_active` represents students actively engaged in the course by participating in the forums, `engagement_passive` 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** Submitting lectures, posting on discussion forums and giving quizzes are signs of active engagement by the student.
- **Passive Engagement** Submitting lectures, viewing/voting/subscribing to posts on discussion forums are signs of passive engagement.
- **Disengagement** Temporal features, indicating the last point of user’s activity, capture signs of disengagement.

We connect the latent engagement variables to student survival by introducing PSL rules. In this model, some of the observable features (e.g, `postActivity`, `voteActivity`, `viewActivity`) are used to classify students into one or more forms of engagement or disengagement. Then, the engagement predicates—conjoined with other observable features that are not used to imply user engagement, such as `reputation`—are modeled to imply student survival. For example, in Table 3.3, conjunction of `postActivity` and `reputation` implies `engagement_active`; conjunction of `voteActivity` and `reputation` implies `engagement_passive`; while `engagement_active` and `engagement_passive` implies survival. (Note that `engagement_active` and `engagement_passive` are abbreviated to `eActive` and `ePassive` in the table).

We train the weights for the model by performing expectation maximization with `survival` as the target variable. The resulting model with latent engagement suggests which forms of engagement are good indicators of student survival. Thus, not only does the latent model produce better predictive performance, but it can
provide more insight into MOOC user behavior than a simpler model. In our experiments, we consider some meaningful combinations of data from different phases. The following section provides more details about the student survival experiments.

3.3 Empirical Evaluation

We conducted experiments to answer the following questions. First, how effective are our models at predicting student survival? Second, what are the key factors influencing student survival in an online setting?

3.3.1 Datasets and Experimental Setup

We evaluate our models on three Coursera MOOCs at University of Maryland: Surviving Disruptive Technologies, Women and the Civil Rights Movement, and Gene and the Human Condition. In discussion below, we refer to these courses as DISR-TECH, WOMEN-CIVIL, and GENE, respectively. Our data consists of anonymized student records, grades, and online behavior recorded during the seven week duration of each course.

Figure 3.2 plots 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-CIVIL, and around 14% in GENE, complete the final exam. We use this to define course survival. In all the three 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 quiz submission and viewing forum content. Observing the statistics, DISR-TECH and WOMEN-CIVIL have a higher percentage of total registered students participating in forums compared to GENE.

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.

3.3.2 Student Performance Analysis

We conduct prediction experiments to assess how effective our models are in predicting learner performance, as measured both by their official grade and by whether they complete the course requirements. We also look at the key factors influencing learner performance in the online setting as determined by our model. Our next set
of experiments analyzes our model’s ability to make predictions at an early stage, where detection of disengaged students may help instructors plan interventions.

**Student Performance Results**

For the learner performance models, we filter the dataset to include only learners that attempted one or more quizzes or assignments in the course and earned a non-zero score. For these learners, we label the ones who earned a statement of accomplishment as positive instances (performance 1.0) and learners that did not as negative (performance 0.0). These labels are used as ground truth to train and test the models.

Our experimental results for the two courses are summarized in Tables 4.9 and 3.5, and show performance values for the PSL models discussed earlier. We observe that the LATENT PSL model performs better at predicting learner performance.

To better understand which behavioral factors provide more predictive information, we examine the weights our model assigned the different engagement variables during learning. The rules involving predicates `postActivity`, `voteActivity`, `viewActivity`, and `reputation` have higher weight in the simple learned model. In our engagement model, we observe that rules involving `engagement_active` predict performance, while rules containing `disengagement` have higher weight for predicting $\neg$performance.
### Table 3.4: Performance of FLAT-DIRECT and LATENT PSL performance models in Disruptive Technologies course.

<table>
<thead>
<tr>
<th></th>
<th>AUC-PR Pos.</th>
<th>AUC-PR Neg.</th>
<th>AUC-ROC</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLAT-DIRECT</td>
<td>0.7393</td>
<td>0.5462</td>
<td>0.6673</td>
<td>0.5815</td>
</tr>
<tr>
<td>LATENT</td>
<td>0.7492</td>
<td>0.5748</td>
<td>0.6923</td>
<td>0.6033</td>
</tr>
</tbody>
</table>

### Table 3.5: Performance of FLAT-DIRECT and LATENT PSL performance models in Women Civil Rights course.

<table>
<thead>
<tr>
<th></th>
<th>AUC-PR Pos.</th>
<th>AUC-PR Neg.</th>
<th>AUC-ROC</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLAT-DIRECT</td>
<td>0.7938</td>
<td>0.9012</td>
<td>0.8619</td>
<td>0.6907</td>
</tr>
<tr>
<td>LATENT</td>
<td>0.9222</td>
<td>0.9501</td>
<td>0.9487</td>
<td>0.7229</td>
</tr>
</tbody>
</table>

#### 3.3.3 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 using different time periods of the course.

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 choose a baseline by ranking students on number of lectures viewed. The scores for our baseline and the two models are listed in Table 3.6. The baseline using just the lecture submission feature 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. Because of the class imbalance and the high proportion of students who drop out, models that can identify students who will complete the course are more valuable. The strength of our model 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.
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 parts according to the duration of the course, as mentioned in the PSL Survival Models section.

Table 3.7 lists the performance metrics for our two models using different splits in the data. Similarly to the results in Table 4.9, 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. We refer to the three phases of each course by start, mid, and end. start-mid refers to data collected by combining time spans, and start-end refers to data collected over the entire course.

Early prediction scores, in Table 3.7 under start, mid, and start-mid (i.e., survival prediction using partial data), indicate that our model can indeed make these predictions reliably. As the data available is closer to the end of the course, models make better predictions. Just as in the previous experimental setting, the latent engagement model achieves the highest prediction quality. The LATENT model for start outperforms DIRECT model on all time-periods in WOMEN-CIVIL, including the ones which contain more data (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 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.

3.3.4 Feature Analysis

We evaluate the contribution of each feature by leaving each feature out and observing the resulting change in prediction performance values. The features considered
Table 3.6: Performance of LECTURE-RANK, DIRECT and LATENT models in predicting student survival

<table>
<thead>
<tr>
<th>COURSE</th>
<th>MODEL</th>
<th>AUC-PR Pos.</th>
<th>AUC-PR Neg.</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISR-TECH</td>
<td>LECTURE-RANK</td>
<td>0.333</td>
<td>0.998</td>
<td>0.957</td>
</tr>
<tr>
<td></td>
<td>DIRECT</td>
<td>0.393</td>
<td>0.997</td>
<td>0.936</td>
</tr>
<tr>
<td></td>
<td>LATENT</td>
<td>0.546</td>
<td>0.998</td>
<td>0.969</td>
</tr>
<tr>
<td>WOMEN-CIVIL</td>
<td>LECTURE-RANK</td>
<td>0.508</td>
<td>0.995</td>
<td>0.946</td>
</tr>
<tr>
<td></td>
<td>DIRECT</td>
<td>0.565</td>
<td>0.995</td>
<td>0.940</td>
</tr>
<tr>
<td></td>
<td>LATENT</td>
<td>0.816</td>
<td>0.998</td>
<td>0.983</td>
</tr>
<tr>
<td>GENE</td>
<td>LECTURE-RANK</td>
<td>0.688</td>
<td>0.984</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>DIRECT</td>
<td>0.757</td>
<td>0.985</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td>LATENT</td>
<td>0.818</td>
<td>0.985</td>
<td>0.944</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>COURSE</th>
<th>MODEL</th>
<th>start</th>
<th>mid</th>
<th>end</th>
<th>start-mid</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISR-TECH</td>
<td>LECTURE-RANK</td>
<td>0.204</td>
<td>0.280</td>
<td>0.324</td>
<td>0.269</td>
</tr>
<tr>
<td></td>
<td>DIRECT</td>
<td>0.304</td>
<td>0.400</td>
<td>0.470</td>
<td>0.372</td>
</tr>
<tr>
<td></td>
<td>LATENT</td>
<td>0.417</td>
<td>0.454</td>
<td>0.629</td>
<td>0.451</td>
</tr>
<tr>
<td>WOMEN-CIVIL</td>
<td>LECTURE-RANK</td>
<td>0.538</td>
<td>0.518</td>
<td>0.415</td>
<td>0.533</td>
</tr>
<tr>
<td></td>
<td>DIRECT</td>
<td>0.593</td>
<td>0.647</td>
<td>0.492</td>
<td>0.596</td>
</tr>
<tr>
<td></td>
<td>LATENT</td>
<td>0.674</td>
<td>0.722</td>
<td>0.733</td>
<td>0.699</td>
</tr>
<tr>
<td>GENE</td>
<td>LECTURE-RANK</td>
<td>0.552</td>
<td>0.648</td>
<td>0.677</td>
<td>0.650</td>
</tr>
<tr>
<td></td>
<td>DIRECT</td>
<td>0.647</td>
<td>0.755</td>
<td>0.784</td>
<td>0.692</td>
</tr>
<tr>
<td></td>
<td>LATENT</td>
<td>0.705</td>
<td>0.755</td>
<td>0.789</td>
<td>0.778</td>
</tr>
</tbody>
</table>

are: posting in forums (post), viewing forum content (view), time period of last quiz submitted by user (quiz), temporal features (temporal), and viewing lectures (lecture). 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 and end. The lecture feature is consistently important for predicting student survival. 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. 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. While quiz is a strong feature in most phases, it can be observed that it is not a strong feature in the end phase of the course. The data suggests that this effect is because quiz taking gradually drops as the course progresses, leading to fewer quiz takers in the end phase. Hence, temporal and quiz are not very predictive features in the end phase.

### 3.3.5 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 learners. The second application provides a qualitative analysis of forum activity by observing representative forum content posted by learners with different engagement assignments.

#### Analyzing Engagement Pattern Dynamics

In this section we take a first step toward understanding how learners engagement and interests change throughout the course. We track the changes in engagement assignments patterns for several interesting student populations and discuss potential explanations for these changes.
Engagement-Passive (EP)  Engagement-Active (EA)  Disengagement (D)

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

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

(c) Engagement patterns in learners 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.

We analyze the learner engagement values predicted by the model for three classes of learners—(1) learners dropping out in the middle phase, (2) learners dropping out in the end phase, and (3) learners continuing until course completion.
Learners dropping out in the *middle* phase stop taking quizzes sometime during the middle phase. Similarly, learners dropping out in the *end* phase stop taking quizzes sometime during the *end* phase. We infer engagement assignments for these groups by training on data from the start and middle respectively.

The learners 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 learners, uncovering their movement from one engagement type to another and how engagement-mobility patterns relate to learner survival.

Figure 4 describes the learner engagement values predicted by the model for the three classes of learners. The labels D, EA and EP refer to disengagement, engagement active and engagement passive, 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 learners start to move toward disengagement in the middle phase. While some EP learners, who are not taking quizzes in middle phase, still follow the course passively, placing them in EP rather than D. We hypothesize that these learners may be more likely to respond to intervention than the already disengaged learners.

In Figure 3.4(b), it can be seen that, out of the learners that drop out eventually in the *end* phase, about half of them are in EP.

Finally, Figure 3.4(c) suggests that most engaged learners only exhibit passive forms of engagement in the *start* and *middle* phases of the course. While in the *end* phase, learners 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 learners, passive engagement is a more prevalent type of engagement than active, stressing the fact that careful observation of passive engagement can help MOOCs assess learner survival health.

**Using Engagement for Qualitative Language Analysis**

In addition to predicting performance, the engagement variables are helpful in interpreting the different facets of learner participation in the course. Of particular interest is the content of the posts made by learners and how it corresponds to the values predicted by the our model. Table 3.8 shows some examples of posts made by users and the engagement and performance scores predicted by the model. It is interesting to note that engaged learners post content with negative sentiment on
the course, which may be perceived as participating in the course (for example, the first user in the Table 3.8). While, disengaged learners may post similar content with negative sentiment, but a change in activity will help discern them from the engaged ones (e.g., the third user in Table 3.8).

<table>
<thead>
<tr>
<th>User Type</th>
<th>Performance</th>
<th>Disengagement</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engaged learner (positive sentiment)</td>
<td>0.7508</td>
<td>0.0</td>
<td>Prof. Lucas, Thank you for a great course! And thank you Coursera!</td>
</tr>
<tr>
<td>Engaged learner (negative sentiment)</td>
<td>0.8032</td>
<td>0.0</td>
<td>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.</td>
</tr>
<tr>
<td>Disengaged Learner (negative sentiment)</td>
<td>0.5</td>
<td>0.675</td>
<td>I agree completely. I used a lot of time on my assignment and got 7.5, think the evaluation criteria were wrong, it shouldn’t be rated on whether you have 3 or 4 innovations in your description but on a subjective measure (which is also flawed). Generally you should pass if it’s obvious that you have followed the course and that you have tried to use the theories that the course has (to a certain degree of course).</td>
</tr>
<tr>
<td>Disengaged Learner (negative sentiment)</td>
<td>0.327</td>
<td>1.0</td>
<td>The grades I received are ridiculous! One pointed out that the good point in my assignment was my fluency in English!!! Oh God! Another one told me that I used ”general knowledge” for asking the questions. Yes, I’ve tried to use ”general knowledge” for applying the theory so that it is useful (and not just a theory!). I’ve re-read my assignment and I still can’t believe in my grade. Is it really fair?.</td>
</tr>
<tr>
<td>Auditor</td>
<td>0.0</td>
<td>1.0</td>
<td>This has been an otherwise fantastic course. Too bad the potential for success is so heavily weighted on two assignments.</td>
</tr>
</tbody>
</table>

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

### 3.4 Conclusion

In this work, we take a step toward understanding student engagement in MOOCs using data-driven methods. We formalize, using PSL, intuitions 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 model constructs an interpretation 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

Content Analysis of MOOC Forums

4.1 Introduction

MOOC discussion forums provide a platform for exchange of ideas, course administration and logistics questions, reporting errors in lectures, and discussion 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 section, I will explore the importance of mining content in MOOC discussion forums. In the first part, I will present analysis of MOOC discussion content and demonstrate that analyzing discussion forum content is helpful in predicting student course completion [Ramesh et al., 2014b]. In this analysis, we observe that posts discussing course logistics correlate well with student course completion. In the second part of this section, we delve more in detail into automatically finding the fine-grained problems reported in the logistics posts [Ramesh et al., 2015].

Our analysis of MOOC forum posts reveals that logistics posts often mention specific problems about the course 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 the later part of this section, 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 different courses will face different problems so labels from one course will not necessarily be useful for another. An unsupervised system which has the provision of incorporating domain knowledge about each course will therefore be especially beneficial. Each MOOC course has a specific set of aspects associated with it, but to be able identify all the possible MOOC aspects, we need to develop a system that is not fine-tuned to any particular course, but can adapt seamlessly across courses.

4.2 MOOC Forum content analysis for Modeling Student Survival

Previous work analyzing discussion forum content tried manually labeling posts by categories of interest [Stump et al., 2013]. 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 discussions about course material (COURSE), b) meta-level discussions about the course, including feedback and course logistics (LOGISTICS), and c) other general discussions, which include student introductions, discussions about online courses (GENERAL). In order to capture these categories automatically we provide seed words for each category. For example, we extract seed words for the COURSE topic from each course’s syllabus.

In addition to the automatic topic assignment, we capture the sentiment polarity using Opinionfinder [Wilson et al., 2005]. We use features derived from topic
assignments and sentiment to predict student course completion (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 COURSE topic and receiving good feedback (i.e., votes) is an indicator of student success and might contribute to survival. 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.

Our work builds on work by Ramesh et al. [2013] and [2014a] on modeling student survival using Probabilistic Soft Logic (PSL). The authors included behavioral features, such as lecture views, posting/voting/viewing discussion forum content, linguistic features, such as sentiment and subjectivity of posts, and social interaction features derived from forum interaction. The authors looked at indication of sentiment without modeling the context in which the sentiment was expressed: positive sentiment implying survival and negative sentiment implying drop-out. In this work, we tackle this problem by adding topics, enabling reasoning about 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 4.2.6, 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 and the Civil Rights Movement 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 4.2.2, we show how these are translated into rules in our model.

4.2.1 Background

Probabilistic Soft Logic

Section 2 gives the technical details behind PSL. Table 4.1 lists some PSL rules from our model. The predicate posts captures the relationship between a post and the user who posted it. Predicate polarity(P) represents sentiment via its truth value in [0, 1], where 1.0 signifies positive sentiment, and 0.0 signifies negative senti-
ment. $upvote(P)$ is 1.0 if the post has positive feedback and 0.0 if the post had negative or no feedback. $U$ and $P$ refer to user and post respectively. These features can be combined to produce rules in Table 4.1. For example, the first rule captures the idea that posts with positive sentiment imply student survival.

- $posts(U, P) \land polarity(P) \rightarrow survival(U)$
- $posts(U, P) \land \neg polarity(P) \rightarrow \neg survival(U)$
- $posts(U, P) \land upvote(P) \rightarrow survival(U)$

Table 4.1: Example rules in PSL

### 4.2.2 Enhancing Student Survival Models with Topic Modeling

| Topic 1: kodak, management, great, innovation, post, film, understand, problem, businesses, changes, needs |
| Topic 2: good, change, publishing, brand, companies, publishers, history, marketing, traditional, authors |
| Topic 3: think, work, technologies, newspaper, content, paper, business, disruptive, print, media, course, assignment |
| Topic 4: digital, kodak, company, camera, market, quality, phone, development, future, failed, high, right, old, |
| Topic 5: amazon, books, netflix, blockbuster, stores, online, experience, products, apple, nook, strategy, video, service |
| 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 4.2: Topics identified by LDA

| Topic 1: thank, professor, lectures, assignments, concept, love, thanks, learned, enjoyed, forums, subject, question, hard, time, grading, peer, lower, low |
| Topic 2: learning, education, moocs, courses, students, online, university, classroom, teaching, coursera |

Table 4.3: Seed words in LOGISTICS and GENERAL for DISR-TECH, WOMEN and GENE courses

| Topic 3a: disruptive, technology, innovation, survival, digital, disruption, survivor |
| Topic 3b: women, civil, rights, movement, american, black, struggle, political, protests, african, status, citizenship |
| Topic 3c: genomics, genome, egg, living, ancestors, genes, nature, epigenitics, behavior, genetic, biotechnology |

Table 4.4: Seed words for COURSE topic for DISR-TECH, WOMEN and GENE courses

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 necessity of using computational linguistics methods to classify the content in discussion forums.

In this work, we—1) use topic models to map posts to topics in an unsupervised way, and 2) employ background knowledge from the course syllabus and manual inspection of discussion forum posts to seed topic models to get better separated topics. We use data from three Coursera MOOCs: Surviving Disruptive Technologies, Women and the Civil Rights Movement, and Gene and the Human Condition.
for our analysis. In discussion below, we refer to these courses as DISR-TECH, WOMEN, and GENE, respectively.

**Latent Dirichlet Allocation**

Table 4.2 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 SeededLDA [Jagarlamudi et al., 2012], which allows one to specify seed words that can influence the discovered topics toward our desired three categories.

**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 4.3 and 4.4. The seed words for LOGISTICS and GENERAL are common across all the three courses. The seed words for the COURSE 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 3a, 3b, and 3c denote the course specific seed words for DISR-TECH, 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, only including the seeded topics (i.e., run 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 GENERAL topics. Thus, the k extra topics help represent COURSE topics that do not directly correspond to the course seeds. 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 were well separated for k = 3. Thus, we run SeededLDA with 7 total topics. Tables 4.5, 4.6, and 4.7 give the topics identified for DISR-TECH, 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 GENERAL and the rest are summed up to get the topic assignment for COURSE.

**Using topic assignments in PSL**

We construct two models—a) DIRECT model, including all features except features from topic modeling, and b) DIRECT+TOPIC model, including the topic as-
signments as features in the model. Our DIRECT model is borrowed from Ramesh [2014a]. We refer the reader to [Ramesh et al., 2013] and [Ramesh et al., 2014a] for a complete list of features and rules in this model.

Table 4.8 contains examples of rules in the DIRECT model and the corresponding rules including topic assignments in DIRECT+TOPIC model. The first and second rules 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 changed to capture that students that post a lot of general stuff only 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 COURSE. This captures the significance of posting course-related content that gets positive feedback as opposed to logistics or general content in the forums. This rule helps us discern posts in general/logistic category that can get a lot of positive votes (upvote), but do not necessarily indicate student survival. For example, some introduction threads have a lot of positive votes, but do not necessarily signify student survival.

### 4.2.3 Empirical Evaluation

We conducted 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.

### 4.2.4 Datasets and Experimental Setup

We evaluate our models on three Coursera MOOCs: DISR-TECH, WOMEN-CIVIL, and GENE, respectively. Our data consists of anonymized student records, grades, and online behavior recorded during the seven week duration of each course.
Table 4.9: Performance of DIRECT and DIRECT+TOPIC models in predicting student survival. Statistically significant scores typed in bold.

<table>
<thead>
<tr>
<th>COURSE</th>
<th>MODEL</th>
<th>AUC-PR POS</th>
<th>AUC-PR NEG</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISR-TECH</td>
<td>DIRECT</td>
<td>0.764</td>
<td>0.628</td>
<td>0.688</td>
</tr>
<tr>
<td></td>
<td>DIRECT+TOPIC</td>
<td><strong>0.794</strong></td>
<td><strong>0.638</strong></td>
<td><strong>0.708</strong></td>
</tr>
<tr>
<td>WOMEN</td>
<td>DIRECT</td>
<td>0.654</td>
<td>0.899</td>
<td>0.820</td>
</tr>
<tr>
<td></td>
<td>DIRECT+TOPIC</td>
<td><strong>0.674</strong></td>
<td><strong>0.900</strong></td>
<td><strong>0.834</strong></td>
</tr>
<tr>
<td>GENE</td>
<td>DIRECT</td>
<td>0.874</td>
<td>0.780</td>
<td>0.860</td>
</tr>
<tr>
<td></td>
<td>DIRECT+TOPIC</td>
<td><strong>0.894</strong></td>
<td><strong>0.791</strong></td>
<td><strong>0.873</strong></td>
</tr>
</tbody>
</table>

Table 4.9 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.

4.2.5 Survival Prediction using topic features

Table 4.9 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.

4.2.6 Discussion topic analysis using topic features

Table 4.10: Logistics posts containing negative sentiment for dropped-out students

Table 4.10 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 might drop out if their course-logistics questions are not answered. Table 4.11 gives
I was just looking at the topics for the second essay assignments. The thing is I don't 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 isn't allowing me to see the whole assignment! Someone please help me out and let me know that the options are.

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.

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".

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.

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 racist 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: Example of change in sentiment in a course logistic thread

<table>
<thead>
<tr>
<th>survival</th>
<th>polarity</th>
<th>logistics</th>
<th>general</th>
<th>course</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.0</td>
<td>0.67</td>
<td>0.067</td>
<td>0.267</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.372</td>
<td>0.163</td>
<td>0.465</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.769</td>
<td>0.091</td>
<td>0.179</td>
</tr>
<tr>
<td></td>
<td>0.78</td>
<td>0.67</td>
<td>0.067</td>
<td>0.267</td>
</tr>
</tbody>
</table>

I was just looking at the topics for the second essay assignments. The thing is I don't 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 isn't allowing me to see the whole assignment! Someone please help me out and let me know that the options are.

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.

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".

Table 4.12: Posts talking about COURSE content

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.

Tables 4.10 and 4.11 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 4.12 shows posts flagged as COURSE by the SeededLDA. The polarity scores in the COURSE posts indicate opinions and attitude toward course specific material. For example, post #3 in Table 4.12 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 COURSE posts can mislead survival and justify our using topic predictions to focus sentiment analysis on LOGISTICS posts.

4.3 Fine-grained Aspect-Sentiment models for MOOC Discussion Forums

In the previous section, we demonstrated the importance of modeling content information in MOOC forum posts. The examples from Tables 4.10, and 4.11 indicate a strong correlation between logistic posts and course completion. We observe that the logistic posts usually report 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 work, 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 different courses will face different problems so labels from one course will not necessarily be useful for another. An unsupervised system which has the provision of incorporating domain knowledge about each course will therefore be especially beneficial. Each MOOC course has a specific set of aspects associated with it, but to be able identify all the possible MOOC aspects, 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 an unsupervised 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 are as follows:

- We show how to extract course-specific features for unsupervised aspect and sentiment prediction 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 can predict fine-grained aspect and sentiment and outperforms the model based only on SeededLDA.

### 4.3.1 Problem Setting and Data

MOOC participants primarily communicate through discussion forums, consisting of posts, which are short pieces of text. Table 4.13 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.

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 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 unsupervised prediction. 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.
Table 4.14: Descriptions of coarse and fine aspects.

<table>
<thead>
<tr>
<th>COARSE-ASPECT</th>
<th>FINE-ASPECT</th>
<th>Description</th>
<th># of posts</th>
</tr>
</thead>
<tbody>
<tr>
<td>LECTURE</td>
<td>LECTURE-CONTENT</td>
<td>Content of lectures.</td>
<td>559</td>
</tr>
<tr>
<td></td>
<td>LECTURE-VIDEO</td>
<td>Video of lectures.</td>
<td>215</td>
</tr>
<tr>
<td></td>
<td>LECTURE-SUBTITLES</td>
<td>Subtitles of lecture.</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>LECTURE-AUDIO</td>
<td>Audio of lecture.</td>
<td>136</td>
</tr>
<tr>
<td></td>
<td>LECTURE-LECTURER</td>
<td>Delivery of instructor.</td>
<td>69</td>
</tr>
<tr>
<td>QUIZ</td>
<td>QUIZ-CONTENT</td>
<td>Content in quizzes.</td>
<td>439</td>
</tr>
<tr>
<td></td>
<td>QUIZ-GRADING</td>
<td>Grading of quizzes.</td>
<td>360</td>
</tr>
<tr>
<td></td>
<td>QUIZ-SUBMISSION</td>
<td>Quiz submission.</td>
<td>329</td>
</tr>
<tr>
<td></td>
<td>QUIZ-DEADLINE</td>
<td>Deadline of quizzes.</td>
<td>142</td>
</tr>
<tr>
<td>CERTIFICATE</td>
<td></td>
<td>Course certificates.</td>
<td>194</td>
</tr>
<tr>
<td>SOCIAL</td>
<td></td>
<td>Social interaction posts.</td>
<td>1187</td>
</tr>
</tbody>
</table>

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. 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 4.14 provides a description of each of the aspects and also gives the number of posts in each aspect category after annotation.

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 identified 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.
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 unsupervised prediction approach, 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.

4.3.2 Aspect-Sentiment Prediction Models

In this section, we develop models and feature-extraction techniques to address the challenges of unsupervised aspect-sentiment prediction for MOOC forums. We first employ a seeded topic modeling approach [Jagarlamudi et al., 2012] to identify aspects and sentiment. 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.

Seeded LDA Model

Topic models [Blei et al., 2003], 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 4.14. We also seed topics for positive, negative and neutral sentiment polarities. The seed words for

1http://www.crowdflower.com/
coarse topics are provided in Table 4.15, and fine aspects in Table 4.16. For the sentiment topics (Table 4.17), 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 4.18. Finally, having inferred the SeededLDA model from the data set, for each post
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 $\alpha = 0.01$, $\beta = 0.01$ in our experiments. For SeededLDA models corresponding to the seed sets in Tables 4.15, 4.16, and 4.17, the number of topics is equal to the number of seeded topics. For SeededLDA models corresponding to the seed words in Tables 4.18 and 4.15, we use 10 topics, allowing for some unseeded topics that are not captured by the seed words.

**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), a scalable class of continuous, conditional graphical models [Bach et al., 2013b].

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.

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

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.

**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 4.19: Representative rules from PSL-Joint model**

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.

**Combining Features** The first set of rules in Table 4.19 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 4.19 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 4.17 and 4.18. The third rule states that certain fine aspects occur together with 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.

**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 4.19, 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.

4.3.3 Empirical Evaluation

In this section, we present the quantitative and qualitative results of our models on the annotated MOOC dataset. Our models are unsupervised, and we use the label annotations only for evaluation. Tables 4.20 – 4.23 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>LECTURE-CONTENT</th>
<th>LECTURE-VIDEO</th>
<th>LECTURE-AUDIO</th>
<th>LECTURE-LECTURER</th>
<th>LECTURE-SUBTITLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeededLDA</td>
<td>0.137</td>
<td>0.057</td>
<td>0.08</td>
<td>0.156</td>
<td>0.256</td>
</tr>
<tr>
<td>PSL-Joint</td>
<td>0.407</td>
<td>0.413</td>
<td>0.410</td>
<td>0.411</td>
<td>0.591</td>
</tr>
</tbody>
</table>

Table 4.20: Precision, recall and F1 scores for LECTURE fine aspects

<table>
<thead>
<tr>
<th>Model</th>
<th>QUIZ-CONTENT</th>
<th>QUIZ-SUBMISSION</th>
<th>QUIZ-DEADLINES</th>
<th>QUIZ-GRADING</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeededLDA</td>
<td>0.042</td>
<td>0.006</td>
<td>0.011</td>
<td>0.485</td>
</tr>
<tr>
<td>PSL-Joint</td>
<td>0.524</td>
<td>0.405</td>
<td>0.36</td>
<td>0.521</td>
</tr>
</tbody>
</table>

Table 4.21: Precision, recall and F1 scores for QUIZ fine aspects

<table>
<thead>
<tr>
<th>Model</th>
<th>LECTURE</th>
<th>QUIZ</th>
<th>CERTIFICATE</th>
<th>SOCIAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec.</td>
<td>Rec.</td>
<td>F1</td>
<td>Prec.</td>
</tr>
<tr>
<td>SeededLDA</td>
<td>0.597</td>
<td>0.673</td>
<td>0.632</td>
<td>0.752</td>
</tr>
<tr>
<td>PSL-Joint</td>
<td>0.563</td>
<td>0.715</td>
<td>0.630</td>
<td>0.724</td>
</tr>
</tbody>
</table>

Table 4.22: Precision, recall and F1 scores for coarse aspects
Table 4.23: Precision, recall and F1 scores for sentiment

<table>
<thead>
<tr>
<th>Model</th>
<th>POSITIVE</th>
<th></th>
<th></th>
<th>NEGATIVE</th>
<th></th>
<th></th>
<th>NEUTRAL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec.</td>
<td>Rec.</td>
<td>F1</td>
<td>Prec.</td>
<td>Rec.</td>
<td>F1</td>
<td>Prec.</td>
<td>Rec.</td>
</tr>
<tr>
<td>SEEDEDLDA</td>
<td>0.104</td>
<td>0.721</td>
<td>0.182</td>
<td>0.650</td>
<td>0.429</td>
<td>0.517</td>
<td>0.483</td>
<td>0.282</td>
</tr>
<tr>
<td>PSL-JOINT</td>
<td>0.114</td>
<td>0.544</td>
<td>0.189</td>
<td>0.571</td>
<td>0.666</td>
<td>0.615</td>
<td>0.664</td>
<td>0.322</td>
</tr>
</tbody>
</table>

SeededLDA for Aspect-Sentiment

For SeededLDA, we use the seed words for coarse, fine, and sentiment given in Tables 4.15 – 4.17. 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.

PSL for Joint Aspect-Sentiment (PSL-Joint)

Tables 4.20 and 4.21 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, both 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 SEEDELDA misclassified but were predicted correctly by PSL-Joint.

Table 4.22 presents results for the coarse-aspects. We observe that PSL-Joint performs 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 4.23).
There is a typo or other mistake in the assignment instructions (e.g. essential information omitted) Type ID: programming-mistake
Browser: Chrome 32 OS: Windows 7

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?

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%.

Let’s start a group for discussing the lecture videos.

Table 4.24: Example posts that PSL-Joint predicted correctly, but were misclassified by SeededLDA

<table>
<thead>
<tr>
<th>Correct Label</th>
<th>Predicted Label</th>
<th>Second Prediction</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUIZ-CONTENT</td>
<td>QUIZ-CONTENT</td>
<td>LECTURE-CONTENT</td>
<td>There is a typo or other mistake in the assignment instructions (e.g. essential information omitted) Type ID: programming-mistake</td>
</tr>
<tr>
<td>QUIZ-CONTENT</td>
<td>QUIZ-CONTENT</td>
<td>LECTURE-CONTENT</td>
<td>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?</td>
</tr>
<tr>
<td>LECTURE-AUDIO</td>
<td>LECTURE-AUDIO</td>
<td>LECTURE-SUBTITLES</td>
<td>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%.</td>
</tr>
<tr>
<td>SOCIAL</td>
<td>SOCIAL</td>
<td>LECTURE-VIDEO</td>
<td>Let’s start a group for discussing the lecture videos.</td>
</tr>
</tbody>
</table>

Interpreting PSL-Joint Predictions

Table 4.24 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 am-
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 4.25 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.

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, making sure that all categories of problems receive attention.

4.3.4 Conclusion

In this work, we developed an unsupervised 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 bet-
ter 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 potentially be used for identifying posts for instructor intervention, understanding student requirements and issues, increasing student retention, and improving future iterations of the course.
Chapter 5

Proposed Research

In this section, I propose extensions to the framework designed in the earlier sections. I plan to delve in detail into two topics: 1) Extend the framework for understanding MOOC forums presented in Chapter 4 to predict course-specific difficulties reported in the forums, 2) Extend the hierarchical framework in Chapter 4 toward a framework for performing multi-scale reasoning in social networks. In the following sections, I present ongoing progress toward each of these two topics.

5.1 Predicting difficult course topics using Forum activity

In Section 3, I explored the pertinent problem of understanding student engagement in MOOCs and using that to predict student course completion and student performance in MOOCs. In Section 4, I analyzed the importance of mining forum content in MOOCs by using features from forum content for predicting student course completion. I also devised methods to automatically detect problem-reporting posts and the fine-grained problems reported in them. Our explorations of MOOC forum data unearth other important research questions. Our work on detecting aspects helps us segregate posts into course-specific aspects. In the analysis in Section 4, I demonstrated the importance of mining logistic posts and developed methods for understanding logistics posts in greater detail. Mining discussions on course content can unravel information about which topics were difficult to understand and when the students needed more help in understanding them automatically, without the instructor having to pore through each and every post. Aligning this with the class schedule will help instructors deal with the avalanche effect of some course topics in the forums by predicting their occurrence. It will also help in retaining more students and aid instructor interventions.
5.2 Multi-scale Network Modeling

In Chapter 4, I developed a framework for representing hierarchical and relational label structure and demonstrated that exploiting the relational structure among the labels is useful in predicting fine-grained aspect and sentiment for MOOC forums posts. I plan to extend the hierarchical framework to create multi-scale abstractions of networks.

In Section 4, I investigated the utility of hierarchical and relational label structure connecting aspect and sentiment labels to predict fine-grained aspect labels for MOOC forum posts. Extending our findings, I plan to investigate multi-scale network representations for graphical data.

Networks are organized at multiple scales. Studies in network science specify that networks often exhibit multi-scale organization, represented as hierarchies over the vertices. Vertices are organized into groups, that are further part of larger groups and so on. Previous work has showed that multi-scale organization of networks account well for various network statistics such as scale-invariance, short path lengths, and a high degree of clustering. Modeling networks in terms of hierarchies helps in understanding evolution of networks at different scales. This is particularly useful in understanding influence, social hierarchy in social networks and helps in a wide variety of social network problems such as link prediction, entity resolution, inferring organization structure, and influence. In Section 5.2.2, I will formally define multi-scale networks and identify example networks in which a multi-scale network model will be beneficial.

In Sections 5.2.3 and 5.2.4, I will identify scenarios in which a multi-scale model of networks can be helpful.

5.2.1 Related Work

Multiscale network modeling is a growing sub-field in social network analysis. In the recent past, there has been a revival in interest in hierarchical and multi-scale network modeling. Ho et al. model multi-scale community structure using a Chinese Restaurant Process, and draw the memberships from a stick-breaking process [2011b]. They distinguish between inter-community and intra-community links in their model, but they restrict inter-community links to the same level in the hierarchy. Also, they use a single hierarchical generative process to model all the granularities existing in a network, while this is not generally true for real-world networks. A recent study shows that large social networks often do not contain large well-defined clusters, which suggest a single generative process cannot explain all clustering phenomenon in social networks [Leskovec et al., 2009]. And often the links at coarse-grained levels are observed through the links connecting
the individual components of the group. But modeling the affinities at coarse levels can unravel more information about the network and also help in predicting tasks at finer levels. Shin et al. [2012] use low rank representations of graphs for predicting links at the finer levels. All these methods make limiting assumptions about the links in the network. They also assume that the network connections at coarse levels are present in the network. Often, the connections and group memberships are implicit and not visible as part of the network structure.

5.2.2 Problem Setting
Consider a network with $V$ vertices and $E$ edges. Each of the vertices have a set of attributes $\{A\}$ associated with them. Each element $a_i$ in $A$ is drawn from a set of all possible attributes for the particular network. $|A| << |V|$, the attribute values repeat across the vertices. Further, the attribute values also connected to one another by relationships such as similarity, relatedness, and coarse-to-fine relationships. Abstractions of the graph can be constructed by grouping the vertices on attribute values and their relationships. These abstractions can then be used for understanding the graph and predicting links at the finer levels.

5.2.3 Multi-scale Behavioral Modeling in MOOCs
So far, in this proposal, I have considered each MOOC participant individually. Students can be classified in different ways according to their different characteristics, such as demographics, education-level, background. Other classifications include their level of interest and participation in the course—some students are interested in certificate, while some are interested in the video lectures, and others in the social interaction element of MOOCs. Some of these classifications are hierarchical—for example, english-speaking countries and non-english speaking countries is one way to classify the students. This can be further divided into the respective countries, for a finer classification. This hierarchical organization of students will help predict characteristics and interactions between groups of students, allowing one to know more about the student population as a whole.

5.2.4 Multi-scale Network Modeling in Social Networks
As an example, lets consider the network from a popular social network—LinkedIn. LinkedIn is a popular professional networking site, which helps find individuals in your professional network. Individuals in LinkedIn specify their employers, past employers, skills, education, and other personal information. The people in this social network correspond to vertices, connections between them are edges, and
information about these individuals are attributes. As it can be seen, the people are connected in the form of social network. The attributes are also connected—for example, a person in the Computer Science department at UMD is also a member of UMD, member of academic institutions in the United States, and so on. Using these groupings and their connectedness at coarser levels can be helpful in reasoning about the network at the finer levels. In the next sections, I present problems that can utilize multiscale modeling of networks.

5.2.5 Multiscale Link Prediction

Estimation of proximity (degree of “closeness”) of nodes is an important problem in social networks. Proximity measures quantify the interaction between users based on the structural properties of a graph, such as the number of common friends. An important application of proximity estimation in social networks is link prediction, which is a key problem in social network analysis [Liben-Nowell and Kleinberg, 2007]. Proximity estimation assumes that a pair of nodes in the network with a high proximity score indicates that they have a high degree of social relatedness. For example, if these nodes are people, there is a good chance that they may become friends in future.

[Shin et al., 2012] propose a multi-scale approximation of the graph to obtain multiple granular views of the network in order to perform link prediction. But, they make use of a single hierarchical clustering to cluster the graph to create the granular views. This can be problematic for two reasons—firstly, using a single hierarchical clustering only gives the possible granular views for the particular clustering technique. Secondly, the clustering technique could ignore meaningful structural connections that could exist in coarser granularities, which could help in predicting links at lower levels.

Now, let us consider how to do this using HL-MRFs. A simple way to approach this problem is assume that the groups are known in advance. We can construct groups using one or more of the attributes that are common to users. First, let us consider a problem setting where the groups are constructed in advance. For example, let us consider the example from LinkedIn mentioned in Section 5.2.4. The affinity of each person toward the group can be calculated. For example, is the group is “country = USA”, then depending on the amount of time spent by the person in the country, it is possible can calculate the affinity of the person toward the group. Table 5.1 gives some rules that make use of the group affinity for predictions at social network level. Rule 1 in the table captures that a predicting a group’s preferred jobs is related to the group’s members’ affinity to the group and their individual job preferences. This rule makes the recommendations for jobs at the group level agree with the recommendation at the individual level. Similarly,
the second rule captures that skills of individual people, their jobs are related to the skills exhibited by the group.

\[
\text{job}(G_1, J_1) \wedge \text{affinity}(P_1, G_1) \rightarrow \text{job}(P_1, J_1)
\]

\[
\text{skill}(G_1, S_1) \wedge \text{job-skill}(J_1, S_1) \wedge \text{skill}(P_1, S_1) \rightarrow \text{job}(P_1, J_1)
\]

Table 5.1: Rules for the DIRECT model.

While the above-mentioned approach is possible for small number of well-defined groups, usually social networks have a large number of localized groups. To detect groups and links between groups, I plan to use clustering approaches to cluster the ground markov network together with the underlying social network to detect clusters in the social network and detect patterns of groundings.

Latent Group Membership

In Chapter 3, we use a latent variable approach to understanding student engagement in MOOCs. This can also be viewed as group membership—active engagement, passive engagement, and disengagement being three groups. Bach et al. [2013a] use a similar latent variable approach to detect group membership. Both these approaches model a small number of groups and do not model the structural relationships between the groups. For example, consider the previous example of modeling behavior of groups of MOOC students, where students can be grouped by different metrics into groups that are related. For example, ethnicity and nationality have a hierarchical relationship. Different cohorts within the MOOC have both inter-group connections and intra-group connections with people communicating within and between groups. To model this, we need to create provision for multiple latent variables and model connections between them.

Membership models using HL-MRFs

Mixed membership block models (MMSB) [Airoldi et al., 2008] propose a generative framework for modeling mixed memberships for relational data. They recover the mixed membership and latent block structure from relational data. Ho et al. [2011a] extend this framework to model multi scale membership in relational data. They use a nested chinese restaurant process as a nonparametric structural prior to generate the hierarchy. Both these approaches use Dirichlet priors that make independence assumptions and hence cannot encode complex dependencies. Further, Ho et al. constrain multi scale membership of each actor in the network only super and sub-communities along the actor’s path. So, this prohibits an actor to have mixed membership across communities in multiple hierarchies. While this may be
realistic in some domains, such as food webs discussed in the paper, it is limiting in domains such as social networks, where these hierarchies could be generated using multiple attributes. They also restrict the links in the network to be only between siblings in the hierarchy, thus not allowing links between coarse and fine levels of the hierarchy. Thus, a general purpose framework for building mixed membership models that has the provision to encode complex dependencies will be very helpful.

I plan to explore replacing the Dirichlet prior with an HL-MRF prior and use a scalable probabilistic programming framework for designing custom mixed membership models. As mentioned before, HL-MRFs are a tractable class of graphical models over continuous variables. Tractable MAP inference provided by HL-MRFs, regardless of the graph structure of the graphical model, allow for representing complex dependencies. Since, HL-MRFs operate on continuous random variables and the potential functions used to encode dependencies are convex, MAP inference is a convex optimization problem in these models. Thus, replacing the model parameters with HL-MRFs, we obtain a class of customizable mixed-membership and multi-scale membership models. In the resulting framework, the model can be specified using the first-order logic based probabilistic programming framework discussed in this proposal—probabilistic soft logic (PSL). The model will be capable of generalizing mixed-membership models, encoding complex dependencies as priors using HL-MRFs.

5.3 Timeline

Phase 1: Project: Predict Student Success Indicators using Latent Engagement Patterns.
Publication goal: JAIR (submission: Spring 2015).

Phase 2: Project: Content Analysis of MOOC Discussion Forums.
Publication goal: TACL (submission: Fall 2015).

Phase 3: Project: Multi-scale Analysis of Networks
Publication goal: AAAI/WWW 2015.

Phase 4: Project: Multi-scale Analysis of Networks

Phase 5: Project: Multi-Scale analysis of Real-world networks
Project goal: Summer Internship 2015 and 2016.
Phase 6: Project: Probabilistic Relational Models for Socio-behavioral data

Publication goal: Dissertation (submission: Fall 2016).
Bibliography


