UNIVERSITY OF CALIFORNIA
SANTA CRUZ

PROBABILISTIC METHODS FOR DATA-DRIVEN SOCIAL GOOD

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

TECHNOLOGY AND INFORMATION MANAGEMENT

by

Sabina Tomkins

September 2018

The Dissertation of Sabina Tomkins
is approved:

Lise Getoor, Chair

John Musacchio

Brent Haddad

Lori Kletzer
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

To my grandmother Leslie Israelsky, who asked *what really changes the world.*

Abstract

Probabilistic Methods for Data-Driven Social Good

by

Sabina Tomkins

Computational techniques have much to offer in addressing questions of societal significance. Many such question can be framed as prediction problems, and approached with data-driven methods. In addition to prediction, understanding human behavior is a distinguishing goal in societally-relevant domains. In this work, I describe societally-significant problems which can be solved with a collective probabilistic approach.

These problems pose many challenges to techniques which assume data independence, homogeneity and scale. In settings of societal importance, dependencies can define the data in question; from complex relationships between people, to continuity between consecutive events. Rather than being generated by single, uniform sources, data in these domains can be derived and described by heterogenous sources. Finally, though many data-driven methods depend on large amounts of observations and high-quality labels in order to guarantee quality results, in domains of critical social value it is often infeasible to gather such quantities. These challenges demand methods which can utilize data-dependencies, incorporate diverse forms of information and reason over small numbers of instances with potentially ambiguous labels.

There are also many opportunities in these domains. Models concerned with societally relevant problems can draw from the knowledge established by existing academic disciplines, from the social to the natural sciences. Such knowledge can serve to inform each step of research from choosing an appropriate problem to putting results into perspective. Furthermore, there are opportunities to obtain new insights into human behavior with the abundance of data generated by virtual and online activity, and mobile and sensor networks. The scale of this data necessitates computational methods. Methods which can leverage prior knowledge and remain efficient even with large datasets can offer much in these domains.

In my work I utilize a collective probabilistic approach for data-driven social good. This approach can capitalize on structure between data instances, rather than flattening it. Furthermore, it can readily incorporate domain knowledge which, especially when combined with a collective approach, is instrumental in learning from small datasets. When datasets are large, this approach leverages a class of probabilistic graphical model which offers efficient inference. Finally, this approach can be extended to model unobserved phenomena with latent-variable representations.

I demonstrate the benefits of this approach in three societally-relevant domains, sustainability, education and malicious behavior. While these domains are diverse, the problems they present share several commonalities which are critical in data-driven modeling. For example, modeling data structure, from spatial relationships to social interactions, can reduce issues of sparsity and noise. Domain knowledge can also combat these issues, in addition to improving model interpretability. I show the benefits of domain knowledge in discovering sustainable products, predicting course performance and detecting cyberbullying. In both the domain of sustainability and malicious behavior, I demonstrate how to utilize spatio-temporal structure in the seemingly distinct tasks of disaggregating appliances and predicting the movements of human traffickers. In education and malicious behavior, I show how unobserved social structure is instrumental in not only modeling learning and aggression, but in interpreting these dynamics in groups. In all three domains I show how to model, represent and interpret latent structure. Thus, while making contributions to each problem setting and domain, I also contribute to the broader goal of data-driven modeling for social good.

Acknowledgments

I'd like to think this work originates from the many seemingly crossable yet frustratingly insurmountable gaps between what the world is and what it could be. One can be overwhelmed by the big questions, the large problems, the ills and ailments that perhaps are, and hopefully aren't, an unremittable part of human societies. And let's not forget, the everyday tedium of failure and small acts of unkindness. The world needs a lot of work. If there is any reason to hope, it is that there are so many people who are working on it. I would like to thank all of the people who guided me to try.

I give the greatest thanks to my advisor Lise Getoor. From the explicit advice, to the assimilated standards, to the values she embedded in every aspect of our work. I owe her a deep debt of gratitude for the tireless feedback and support. I know that I will be guided my many of the lessons she imparted but if I have absorbed anything, I hope it is her ability to appraise and pursue research which matters.

Next, I would like to thank everyone who taught me to observe, to wonder, to inquire and to experiment. In short, and with any luck, to think. I have been lucky to meet so many generous instructors. I would like to thank the instructors, professors, and TAs who shared their time and energy to teach us fortunate students something. Of all the educators, Edgar Jasso stands out. Edgar is limitless in what he gives students. I, and so many students, have benefited from his patience, kindness and open door for questions.

I would like to thank the people in my undergraduate career who taught me how to do research. Sometimes through patient exposition, sometimes by questions and feedback, but always by demonstrating the many traits which mark philosophers, and which have served as inspiration. For their patience, kindness, creativity, standards, critique and curiosity, and for encouraging me when I needed it the most, I would like to thank Seyyed Ali Pourmousavi Kani, Hashem Nehrir, Clem Izurieta, Tom Lagatta, Todd Gureckis and John McDonnell.

I heartily thank Rayid Ghani for sharing his counsel time and again and for creating the unique and exciting environment of DSSG. The summer of 2014 was for me, full of role models and mentors; everyone had something to teach and I would like to thank each and every one for the friendship and inspiration. I especially thank

Varun Chandola and Matthew Gee for their mentorship and the energy team for their collaboration.

In my graduate career, I would like to first thank Brent Haddad for his constant support, belief and encouragement. I'm grateful to Cameron Speir for showing me the world of fisheries economics, and my first whales in Santa Cruz. I would like to thank Suresh Lodah for his instruction and mentorship. Cynthia McCarley has made so many things so much easier for me, and I thank her profoundly for all of the help and friendship.

I am very grateful to the many inspiring people with whom I've had the honor to collaborate, and whose enthusiasm and knowledge have served to guide and cement my own passion for research. I would like to deeply thank: Zhe Lui, Yufan Guo, Steve Isleys, Steven Minton, Brian Amanatullah, Yi Zhang and Yunfei Chen. For their wonderful collaborations, as well as for being kind and skillful mentors I would like to thank Anbang Xu, Ben London and Amina Shabbeer.

I would like to give a special thanks to my LINQS collaborators Arti Ramesh, Jay Pujara, Golnoosh Farnadi and Ben London. Their kindness and feedback were constant sources of motivation and they always demonstrated thoughtfulness, positivity and excitement. Above and beyond the individual projects, the brainstorming sessions and the rounds of editing, they shared attitudes and best practices I hope to follow.

I thank the alchemists who turned the moments of anguish into joy, or at least tempered them with common sense and optimism. For listening and for laughing and for all the acts of kindness, for the camaraderie and commiseration, I would like to thank Dhanya Sridhar and Pigi Kouki, Sarah Tan, Natalia Díaz Rodríguez, Shobeir Fakhraei, Holakou Rahmanian and Michał Dereziński.

Finally, I would like to thank the community who taught me to care. My parents for giving me a reckless amount of freedom and trust. My brother for being my most constant friend. The many family members whose conversations taught me the language of discourse, the respective tastes of fact and fiction, and the feeling of love. Papa Tom for poetry and Vonnegut, Uncle Henry for fractions and physics, Luke for history and Nanny for art. I would like to thank Nana, Auntie Tutu, Jo, Bernadette, Elizabeth, Shawn, Jilala and Parra for being constant role models of strength and hu-

manity. Steve and Ellen for their open home. And Todrich for his support. Sometimes, love is all you need.

This research is partially supported by the National Science Foundation(NSF) under Grant Numbers CCF-1740850 and IIS-1703331 and by the US Army Corps of Engineers Engineer Research and Development Center under Contract Number W912HZ-17-P-0101 and the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under contract number FA8750-14-C-0240. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, the US Army Corps of Engineers Engineer Research and Development Center or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Government. I would also like to especially thank the NSF for funding many of the activities and programs I have participated in throughout both my undergraduate and graduate career.

Part I

Preliminaries

Chapter 1

Introduction

Technological advances have much to offer in addressing complex questions of societal significance. However, how to harness these advances for social good is an open question. Computational techniques are one promising approach for addressing questions into human behavior at new depth and scale. Furthermore, when driven by data, statistical-computational models can be utilized to approach a variety of tasks of societal importance. For example, determining which students in a high school will graduate, assessing the impact of environmental events on malicious behavior and detecting aggressive behavior on social media. Each of these tasks can be defined as an prediction problem and approached within an appropriate data-driven framework. Additionally, a distinguishing goal is to improve *understanding* of social phenomena. In addition to predicting outcomes, data-driven techniques can uncover hidden structure in complex problems.

In this thesis, I describe social good problems which can be solved with a collective probabilistic approach. Here, I define social good broadly to refer to these complex problems which are significant to societal well-being. In each problem setting, I demonstrate how introducing, representing and modeling relational structure can improve predictive performance for this class of problems. When relevant, these models can also provide insight into underlying and unseen phenomena. This work spans three large areas of societally relevant problems: sustainability, education and malicious behavior. While different in many ways, each area offers a wealth of problems which benefit from this collective probabilistic approach.

Bringing data to bear on social science questions provides opportunities to contribute to these fields while drawing on their scholarship. Yet applying data-driven approaches requires recognizing and addressing the unique opportunities and challenges of these domains. Additionally, it is necessary to identify advantages and shortcomings of societally-relevant datasets, which might be small, heterogenous and defined by relational structure; thus requiring specialized approaches.

Problems of societal-significance are defined by a rich history of investigation by the social sciences. In working within social good domains, modelers have the opportunity to draw on the substantial body of work from established academic disciplines, as well as on expert knowledge. Models which can incorporate domain knowledge can benefit from the insights of a diverse range of disciplines. For example, insights from psychology can inform models of social behavior, learning sciences can and should inform the study of novel educational environments and fields of natural science can contribute to the design of spatio-temporal models. Incorporating external knowledge can improve model performance, robustness and interpretability, and, as we will see, is especially advantageous when data is sparse.

People now interact in virtual environments, forming and maintaining relationships with others they have never physically met. As people increasing interact in virtual environments, these new forms of interaction prompt new sets of questions. Data-driven approaches are uniquely suited to handle these new online interaction data, whose scale renders non-computational methods obsolete.

However, there are also challenges with social good domains. One issue is that the data necessary to answer relevant questions may not exist or be possible to gather within ethical bounds. When it is possible to gather relevant samples, the sample size may not be sufficiently large to train standard machine learning algorithms. Additionally, it is not always clear to what extent the gathered data can answer relevant questions of societal importance. For example, how one behaves online is not necessarily representative or indicative of one's offline behavior or psychological state. Thus, in utilizing online data to understand human behavior, a critical question is how relevant online behavior can be in investigating issues whose primary sphere is the offline world. Even more difficult to answer, is how online behavior and attitudes influence offline activities

and latent psychological states.

Another opportunity and challenge with social good application data is that it can have rich relational structure. This structure is advantageous in that it can lead to better understanding of relational domains. However, there are fewer methods available for modeling structured data. My work addresses this gap, and introduces models for social good domains with richly structured data.

I propose a collective probabilistic approach to data-driven social good. I utilize a probabilistic programming framework which is capable of leveraging the opportunities and addressing the challenges inherent in social good problems and which is especially suited to handling relational data. These models are able to flexibly incorporate domain knowledge with intuitive first-order logic. By directly modeling structure between observed *and* unobserved random variables they can benefit from collective inference[1]. Embedding domain knowledge and leveraging collective inference are instrumental in overcoming issues of sparsity in small datasets. Finally, these models can lend insight into hidden phenomena through latent-variable representations.

I build these models with Probabilistic Soft Logic (PSL) [11], a recently introduced highly scalable probabilistic modeling framework. PSL offers several advantages across domains. First and foremost, PSL infers the values of unobserved target variables *jointly*, allowing it to take advantage of rich structure between these target variables.This is especially advantageous for the richly structured social good problems I address here. While such a joint formulation can increase computational demand, PSL's convex formulation results in efficient inference, improving the feasibility of otherwise impractical problems. PSL templates probabilistic graphical models with first-order logical rules. Thus, domain expertise can be flexibly encoded with intuitive logic. Finally, PSL also offers a latent-variable formulation which is critical to my work, in which I demonstrate how latent-variables can be utilized to discover relevant structure in complex problems.

Here, I present work from three domains: sustainability, education and malicious behavior. In each of these domains I work with richly structured data. As mentioned, this relational data requires a unique approach, one capable of recognizing rather than ignoring structure between data instances. Not only is a collective

[1] One kind of unobserved random variable is a *target variable*, which might be the exam score of a high school student or the presence of bullying content in text.

probabilistic approach ideal for treating structured data, it is also especially useful in exploiting the opportunities and surmounting the challenges of social good applications.

By employing a consistent class of models across domains I make several contributions to the study of computational methods for social good problems. One dominant shortcoming of data in social good domains is that it can be sparse. Even in extensive time series data, events of interest may be few and far between. In textual data, the most informative words might be the most infrequent. Across domains, I demonstrate that utilizing the inherent structure in data, and creating collective representations can combat these issues of sparsity and noise. Critically, domain knowledge is abundantly available in social good domains, and its incorporation can also combat data shortcomings while improving model interpretability. I utilize domain knowledge in designing recommender systems, modeling student interactions and detecting cyberbullying. Finally, I explore the use of latent-variable models in explaining human behavior. These latent-variables serve several purposes across domains. They can uncover hidden phenomena, contributing to domain knowledge, and utilize this hidden structure to improve model performance. In the next three sections I outline how these contributions apply in three domains of societal significance.

1.1 Sustainability

Increasing the ability of humanity to sustain itself on Earth is a matter of great societal importance. Living *sustainably* has become a goal for many. However, interest in sustainability does not necessarily translate into sustainable actions. The difference between one's intentions and ability to act inline with them is referred to as the *intention-behavior*, *attitude-behavior*, or *value-action gap* [107]. This psychological phenomenon has repercussions on many aspects of behavior, from transportation to food choices.

Household energy consumption can suffer from the intention-behavior gap [33], where consumers with an expressed desire to reduce consumption fail to do so. Similarly, when shopping, consumers can fail to make sustainable choices. A significant barrier to acting sustainably is knowledge, people do not always know which actions to take. I address this challenge with two projects: energy disaggregation and sustainable

recommender systems.

My collective probabilistic approach for energy disaggregation utilizes temporal structure, as well as relationships between appliances, to achieve state-of-the-art performance on the task of detecting individual appliance power usage from aggregate power readings. Additionally, I construct a recommender system which can discover sustainable products and sustainability-minded customers, while making recommendations. These two approaches take coarse information and leverage relational structure to arrive at finer-grained insights which can inform behavior.

1.1.1 Energy Disaggregation

Current greenhouse gas emissions are at dangerous levels and projected to rise still. In order to reduce these emissions, many actions need to be taken at all levels of society, from the political to the personal. Households in the United States account for roughly a third of all US emissions and lowering the energy consumption of this demographic alone would have substantial benefits.

Energy disaggregation is the process of taking an aggregate energy reading and determining which appliances consumed how much energy when. A successful disaggregation algorithm takes meter readings from a household account and returns appliance usage profiles. This precise and actionable feedback can change the way in which consumers think about their appliance usage, potentially enabling them to reduce consumption.

In my work I demonstrate the usefulness of a collective probabilistic approach. I introduce two structured temporal representations of energy readings data and demonstrate the advantages of an interval representation. Rather than inferring only over appliances, I also infer over appliance sets. By utilizing an appliance set representation, this approach benefits from the relationship between sets of appliances and observed energy readings. Thus, in this work there are clear advantages to representing the structure inherent in temporal series and between individual entities bounded by a shared sum. My energy disaggregation work offers several contributions: a comparison of two temporal representations for disaggregation, a flexible framework for incorporating contextual information and a scalable approach for constraining feasible sets of appliances. Finally,

this work contributes to the understanding of how collective inference can be useful in understanding energy consumption systems.

1.1.2 Sustainable Recommender Systems

Barriers to sustainable shopping include: cost, availability, skepticism of labels and insufficient marketing [51]. Unlike cost and availability, skepticism and marketing are issues that might be overcome with better communication about the sustainability profile of available products. Furthermore, knowledge about what makes a product sustainable can be a barrier to consumers [95], while Tanner and Wölfing-Kast [125] found that possessing actionable knowledge is correlated with taking action. Vermeir and Verbeke [139] found positive desires to purchase sustainable food items were impeded by a lack of certainty about those items' sustainable characteristics. Thus, there is a strong need to improve communication and provide actionable information about sustainable products.

Given knowledge of sustainable products, a good recommender system might offer products with these characteristics to interested customers. However, defining what makes a product sustainable is not straightforward. For those who value sustainability, products can be appraised according to complex factors including: environmental impact, impact on the local economy, animal welfare considerations and benefit to the consumer [90].

In this work, I address this behavior-intention gap with a method for jointly discovering sustainable products and sustainability-minded customers. Recommender systems can play a significant role in the process of discovering sustainable products and customers. To this end, I propose a probabilistic recommender system approach which fuses multiple weak signals to infer the sustainability of products. Here I investigate three types of signals: freely available domain knowledge about sustainable companies and certifications, product metadata and the purchasing patterns of customers predicted to be sustainability-minded.

This recommender and discovery model offers many contributions. It is the first known approach to directly model sustainability as a factor in purchasing decisions. This explicit encoding of sustainability allows for the discovery of sustainable products

and sustainability-minded customers. By fusing multiple sources of information into a joint formulation over purchases, products and customers, as well as sustainability information, this approach outperforms established recommender system approaches.

1.2 Education

Massive Open Online Courses (MOOCs) are online courses designed to be available to large audiences. Courses upload material online and students can go through this material independently. In many cases, student access to instructors is limited or absent, unlike typical classroom environments with frequent student-instructor interactions. While promising to create new learning opportunities, the effectiveness of these courses is unknown and can be difficult to evaluate.

One promising opportunity of MOOCs is to increase access to high school courses which suffer from a scarcity of qualified instructors. For example, teaching positions in high school computer science and other science, technology, engineering and mathematics courses can be difficult to staff. However, it is exactly these courses which are projected to be most vital in preparing graduates with the skills they will need to succeed.

An open question is if MOOCs can be effective for high school students, a group which may not have the self-study skills to excel in an independently driven environment. In my work, I evaluate the effectiveness of a high school computer science MOOC. This MOOC is unique in that it provides both online and offline instruction. Furthermore, it prepares students for an offline end-of-year evaluation. This evaluation is administered by a third party and can be used as a post test. As post tests occur at the end of a course and are comprehensive of all relevant topics, they are critical in evaluating student learning and course success. A contribution of my work is a thorough analysis of which students pass the post test and the effect of in-person instruction on success.

1.2.1 Predicting Performance

Post-test performance can be useful as an indicator of learning. By analyzing which students pass the post test one can assess the educational value of a course. Furthermore,

by predicting if students will pass the post test one can determine which students might be in need of intervention.

In this work, I employ a data-driven approach to predict which students will pass a post-test exam. These predictions can then be used to analyze student success. Additionally, I analyze which course attributes are most indicative of success. Such information can be used to modify the course, for example by increasing content in areas with strong correlations with success.

Additionally, I use this predictive model to group students according to their behavior. For example, we can differentiate students who are likely to succeed or not according to which activities they do, and if and what they post on the student forum. This analysis prompted the discovery of unexpected learners, those students whose course performance is misaligned with their post-test performance. In this data, unexpected low learners, students who earned high course scores but low post-test scores were predominantly students who received some in-person instruction. Thus, to further explore this group, I also modeled student interactions and their effect on learning.

1.2.2 Student Interactions and Learning

A question in novel learning environments is how peer interaction can shape learning. Arguello et al [7], develop targeted user studies to observe how different group dynamics can influence learning in MOOCs. They find that particular group structures can provide benefits to learning.

This MOOC includes an in-person option for some students. Thus, these coached students might experience somewhat typical classroom environments where learning can be effected by peer interactions [108, 43]. Hence both in-person and virtual interactions between students which affect performance in this MOOC.

Here, we inspect observational data to see if forum interactions can be correlated to post-test performance. Additionally, we ask if offline interactions, such as studying together can be detected from online data in such a way that they can improve the predictions of post-test performance. In answering this question we employ a collective probabilistic approach which can model the interactions of students as well as students' and classroom sections' latent strength.

1.3 Malicious Behavior

Malicious behavior affects society at many levels from the emotional well-being of individuals to the economy. Detecting and understanding malicious behavior is an important goal in many diverse applications, from promoting healthy societal norms to enforcing law and safety. Here, I introduce work in two related research areas: cyberbullying and spatio-temporal models of human trafficking.

Cyberbullying is a serious threat to both the short and long-term well-being of social media users. For example, targets of cyberbullying attacks experience anxiety and depression [70]. In addition to emotional and psychological effects, cyberbullying can impact school performance and attendance [70]. Human trafficking is another serious threat, imposing great emotional, psychological and physical harm to victims and their families. Any efforts towards the reduction of trafficking have the potential to drastically improve the quality of life of these victims.

Obtaining quality data on malicious behavior is difficult, as aggressors obscure their tracks to avoid detection and complex social power structures prevent victims from reporting their plight. Online data can provide further insights into these interactions, as perceived online anonymity can create environments where aggressors are more open. However, this data can be of poor quality which enhances the difficulty of extracting informative signals into human behavior. My work addresses these concerns by utilizing the inherent structure in these tasks to overcome data quality issues.

In the setting of cyberbullying, I utilize social and linguistic structure to better detect cyberbullying. To assess the effect of exogenous events on human trafficking, I utilize a spatio-temporal approach which draws on both geographic and temporal relationships. In each problem setting, I utilize latent-variables to better understand malicious behavior.

Addressing aggression in online environments demands the ability to automatically detect cyberbullying and to identify the roles that participants assume in social interactions. As cyberbullying occurs within online communities, it is also vital to understand the group dynamics that support bullying behavior. To this end, I propose a socio-linguistic model which jointly detects cyberbullying content in messages, discovers latent text categories, identifies participant roles and exploits social interactions. While

this method makes use of content that is labeled as bullying, it does not require category, role or relationship labels. Furthermore, as bullying labels are often subjective, noisy and inconsistent, an important contribution of this work is effective methods for leveraging inconsistent labels.

Additionally, I investigate the impact of environmental stressors on human trafficking. In this work, I combine online data with offline knowledge of extreme weather events. Here I investigate the open question of how extreme weather events might impact trafficking. Deepening our understanding of this relationship can assist efforts in apprehending traffickers, especially in the aftermath of such events. Furthermore, this approach can be generalized to a variety of other situations in which environmental stressors impact malicious behavior.

1.3.1 Cyberbullying Detection

Bullying has long presented physical, emotional and psychological risks to children, youth and adults. As such, there is an extensive body of knowledge aimed at understanding and preventing bullying. Far less is known about cyberbullying, the newest form of interpersonal aggression. Cyberbullying occurs in an electronic environment [77], from online forums to social media platforms such as Twitter and Facebook. As it can occur at any time or location, cyberbullying poses new psychological risks, while also influencing physical well being [77, 56, 59]. It also introduces new questions of governance and enforcement, as it is less clear in an online environment who can and should police harmful behavior.

A necessary first step in understanding and preventing cyberbullying is detecting it, and here our goal is to automatically flag potentially harmful *social media* messages. These messages introduce unique challenges for natural language processing (NLP) techniques. As they are unusually short and rife with misspellings and slang, when treated with traditional text pre-processing, these messages can be stripped to only one or two words. This sparsity makes cyberbullying messages especially ill-suited for methods which depend on sufficiently large training corpora to generalize well. Solutions which can augment poor textual data with domain knowledge or social data might outperform those which rely on text alone.

Not only is labeled data costly, but it can be error-prone as annotators are generally third parties who are not directly involved with the incidents of cyberbullying. Thus their labeling is subjective, and even labels with high inter-annotator agreement may be incorrect. Rather than throwing out annotations with low inter-annotator agreement, I propose a series of probabilistic models which can directly incorporate uncertainty. Furthermore, I show that modeling uncertainty in the training data can improve the performance of all models, demonstrating that a probabilistic approach is well suited for this domain.

I develop a series of probabilistic models of increasing sophistication. My first model makes use of text, sentiment and collective reasoning. Next, I incorporate seed-words and latent representations of text categories. Finally, I make use of social information by inferring relational ties and social roles. My contributions in this domain include: a joint approach for inferring bullying content, determining types of attacks, assigning participant roles and discovering relational ties. Additionally, I demonstrate the utility of directly modeling annotator uncertainty.

1.3.2 Environmental impacts on Security outcomes

Another malicious behavior which I investigate is human trafficking. Human trafficking effects 20.9 million estimated victims [96]. Of these, an estimated 2 million are children. An open question, and concern, is how changing climates will exacerbate vulnerabilities for potential victims of trafficking. Global temperatures are projected to rise an estimated 8-11°F over the course of the next century [141]. This overall temperature increase will be accompanied by changing climates, resulting in extreme weather and changes to stable ecosystems. While climate change is a serious environmental threat, it also poses significant risks to human well-being and social systems [1, 81]. One projected impact of climate change is on security outcomes, altering and potentially increasing opportunities for crime [101, 3].

Understanding the relationship between environmental stressors and criminal activity requires utilizing multiple heterogenous data, from both online and offline sources. I model the relationship between environmental stressors and crime though a probabilistic approach which can fuse multiple heterogenous signals and model spatial

and temporal relationships. I evaluate the feasibility of this approach by analyzing the relationship between extreme weather events and human trafficking.

Additionally, I propose spatio-temporal models for predicting traffickers' movements. With such models, one can predict where traffickers go in the presence and absence of environmental events. Thus, I introduce two models: one which is *event aware* and one which is *event agnostic*. I expand on both these models to be able to discover and utilize trafficking routes, with a latent-variable model. Thus, my contributions in this domain can be summarized as: a novel exploration of the relationship between trafficking and environmental stressors, a spatio-temporal models for predicting future movements in the presence and absence of extreme events, and a latent-variable model which can discover trafficking routes.

1.4 Contributions and Organization

In my work, I demonstrate how data-driven probabilistic models can capitalize on the opportunities and overcome the challenges presented by societally relevant problems. One particularity of these problems is that they create data with natural structure. However, few computational methods are equipped to sufficiently model this structure. Thus, I explore the question of how to utilize probabilistic programming in approaching problems of societal importance. By approaching distinct problems in three large domains of social interest (sustainability, education and malicious behavior), I thoroughly analyze a particular class of data-driven probabilistic models and demonstrate their appropriateness for these domains. In Chapter 2 I introduce the particular class of models used in my work. That these domains are distinct, covering diverse aspects of human behavior, as well as the natural world, demonstrates the generality of my work. For example, I employ a collective probabilistic approach in modeling positive behaviors, such as sustainable actions and learning, as well as the negative behaviors of cyberbullying and human trafficking. A contribution of my work is in identifying and exploiting the commonalities across these diverse domains. For example, social structure is advantageous in modeling both positive and negative behaviors.

My work both contributes to the general area of data science for social good and to the specific domains I studied. In the first domain that I studied, sustainabil-

ity, my contribution focuses on the action-value gap. A contribution of this work is in solving two prediction tasks to address this problem, energy disaggregation and sustainable recommender systems. My work on energy disaggregation is published at the International Joint Conference on Artificial Intelligence [131]. The work on sustainable recommender systems is published at Recommender Systems [130].

The second social good domain area I address is education. MOOCs have been heralded as a revolutionizing force in education, but their effectiveness has been underevaluated. A contribution of my work is my evaluation of the effectiveness of a high school MOOC for computer science. In Chapter 5 I propose the problem of predicting offline exam performance from online course data, much of this chapter is included in the publication published at the International Conference on Educational Data Mining (EDM)[132]. Next, I expand this work to explore how peer-interactions influence learning, in Chapter 6, which contains work in submission to the Journal of Educational Data Mining.

The third social good domain I study is malicious behavior. A contribution of my work is to introduce *relational* approaches for learning from poor-quality online data in malicious behavior domains. In Chapter 7 I detail my work on detecting bullying behavior from social media data and demonstrate how social and linguistic structure can be useful in overcoming data sparsity. This work is included is published at the conference on Advances in Social Network Analysis and Mining [129]. Chapter 8 outlines my work on assessing the impact of environmental events on human trafficking and describes the task of predicting the movements of human traffickers. In this task, I demonstrate that spatial and temporal structure are useful in predicting traffickers' future locations. This work is published at the Beyond Online Data workshop, and was awarded the Best Paper prize there [127]. It will also appear in the proceedings of the International Conference on Data Mining [128]. As an additional contribution, I introduce latent-variable models to discover informative hidden structure in malicious domains. In Chapter 7, I categorize types of attacks, assign roles to participants and discover relational ties. In Chapter 8, I discover route-segments and links between locations. These latent-variables not only improve predictive performance, but can be used to study unobserved phenomena in these problems.

In this thesis I propose a collective probabilistic approach for richly structured data in social good domains. I demonstrate the benefits of this approach in three domains: sustainability, education and malicious behavior. Towards improving access to information on sustainable behavior, I have developed a state-of-the-art energy disaggregation algorithm [131], as well as a sustainability-aware recommender system [130]. In the domain of education, I have both assessed the efficacy of a high school computer science MOOC and built predictive models of student success [132]. Towards detecting and understanding malicious behavior in data-poor online settings, I have developed two collective probabilistic models in two different problem settings [127, 129, 128]. By analyzing these diverse problems, I present cohesive and domain independent conclusions which span across domains. For example, I demonstrate the utility of modeling participant interactions both in online classrooms and cyberbullying incidents. I show how to include latent-structure in a diverse range of prediction problems, from the future purchases of online shoppers to the future movements of human traffickers. Throughout each problem setting I show that an approach which can model dependencies between random variables, as well as domain knowledge and contextual priors, can yield fresh insights into the nature of complex societal challenges; even when data is limited. I present my cross-domain conclusions in Part V.

Chapter 2

Inference with Probabilistic Graphical Models

2.1 Introduction

In my work I investigate problems of societal-significance where both input and output data have relational structure. For example, in social settings, relationships between individuals can influence their behavior. In spatio-temporal settings, geographical structure can influence predictions which might also be evolving over time. Natural language is highly structured, and text data defines relationships at multiple levels of abstraction, from characters to works of literature.

A defining characteristic of social good domains is that expert-knowledge is plentiful. However, it is not always possible to convert this knowledge into the kinds of feature vectors that most machine learning models expect. It is an open question how to most effectively translate human knowledge into formats most utilizable by computational methods. Here, I propose logical rules as an intuitive representation of human expertise.

Probabilistic Graphical Models (PGMs) satisfy the requirement of expressing the complex and structured knowledge representations of social good domains. PGMs compactly represent joint distributions of data, such that inference can be performed collectively, fully utilizing the statistical dependencies between variables. Furthermore, a popular approach for specifying a PGM is with first-order logic, allowing one to

intuitively model real-world relationships between attributes of instances, and between the instances themselves.

However, inference over complex graphical structures with large numbers of variables and rich dependencies is computationally expensive. When variables are limited to binary values, as is the case with logical atoms, the search space of possible solutions is exponential and inference quickly becomes intractable. Alternatively, one can model atoms with continuous values in the $[0, 1]$ interval and define a PGM over continuous, rather than discrete, random variables. One such class of PGM, is Hinge-Loss Markov Random Fields (HL-MRFs). By expressing dependencies between continuous random variables in the $[0, 1]$ interval, and formulating feature functions as linear hinge-losses, HL-MRFs admit highly scalable inference. To specify a HL-MRF one can use Probabilistic Soft Logic (PSL), a probabilistic programming language which defines a HL-MRF with a set of weighted first-order logical rules. In the next section I will introduce HL-MRF's before providing more detail on PSL in 2.3.

2.2 Hinge-Loss Markov Random Fields

Markov Random Fields (MRFs) are a class of highly expressive probabilistic graphical models. The caveat of this expressivity is scalability, as complexity grows so do the computational costs of learning and inference. Hinge-Loss MRFs are a general class of conditional, continuous probabilistic models, designed to retain the expressivity and overcome the scalability issues of discrete MRFs. Consider the setting common in machine learning, where we have both observed and unobserved variables, and we would like to infer the values of the unknown unobserved variables. One common goal is to infer the maximum a posteriori (MAP) assignments to the unobserved variables. MAP inference in a HL-MRF is a convex optimization problem, which can be solved efficiently with robust optimization procedures for joint distributions.

A HL-MRF is an undirected graphical model, whose probability density function is defined by weighted linear functions over continuous random variables in $[0, 1]$. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be vectors of observed and unobserved continuous random variables in $[0, 1]$, respectively. Formally, a HL-MRF describes the following conditional probability

density function over $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$P(\boldsymbol{y}|\boldsymbol{x}) \propto exp\left(-\sum_{j=1}^{m} w_j \phi_j(\boldsymbol{y}, \boldsymbol{x})\right) \tag{2.1}$$

Where ϕ_j is a *hinge-loss* potential, $\phi_j = \max\{l_j(\boldsymbol{x}, \boldsymbol{y}), 0\}^p, p \in \{1, 2\}$, l_j is a linear function of $\boldsymbol{x}$ and $\boldsymbol{y}$ and w_j is the positive weight associated with ϕ_j. Each ϕ_j is a function which assigns a score to the current assignments to $\boldsymbol{x}$ and $\boldsymbol{y}$. When assignments are more probable, they receive higher scores. Thus these ϕ_j functions appropriately reward likely assignments and penalize those which are improbable.

2.3 Probabilistic Soft Logic

Probabilistic Soft Logic (PSL), is a probabilistic programming language which can be used to template HL-MRFs. PSL has been successfully deployed in a diverse range of settings, from recommender systems [69] to stance prediction in online forums [119]. PSL models are specified through weighted logical rules which capture dependencies between variables.

For a simple example of PSL syntax let us consider a rule which says that if two students are friends, and one student passes a class than the other student also passes:

$$w_{pass} : \textsc{Friends}(S_1, S_2) \wedge \textsc{Passes}(C, S_1) \Rightarrow \textsc{Passes}(C, S_2) \tag{2.2}$$

Here FRIENDS, and PASSES are predicates, S_1, S_2, and C are variables, and w_p is a weight. By substituting constants, s_1, s_2, and c for the variables, S_1, S_2, and C respectively, one obtains three ground atoms: $\textsc{Friends}(s_1, s_2)$, $\textsc{Passes}(c, s_1)$, and $\textsc{Passes}(c, s_2)$, such that each ground atom takes a value in $[0, 1]$. These PSL atoms correspond to random variables in a HL-MRF, for example let $\textsc{Friends}(s_1, s_2)$, $\textsc{Passes}(c, s_1)$, and $\textsc{Passes}(c, s_2)$ correspond to three random variables: $f_1, p_1,$ and p_2 respectively. Given two continuous truth values $q, r \in [0, 1]$ a conjunction of q and r is defined as $q \wedge r = max\{q + r - 1, 0\}$. Finally, using the formula $q \rightarrow r = \neg q \vee r$, we arrive at a weighted *hinge-loss* potential

$$w_p \cdot max\{f_1 + p_1 - p_2 - 1, 0\}. \tag{2.3}$$

The value of $w_p \cdot max\{f_1 + p_1 - p_2 - 1, 0\}$ is the loss associated with Rule 2.2. When the assignments to $f_1, p_1,$ and p_2 are such that this loss is zero, the rule is satisfied. When the loss is greater than zero, the loss is called the distance to satisfaction associated with the rule under the current variable assignments. For example, if $f_1$andp_1 are observed, taking values 1.0, and .8 respectively, we can now infer the value of the remaining unknown variable p_2. Substituting the known values for f_1 and p_1, we obtain Equation 2.4 and see that as long as $p_2 \geq 0.8$ the loss associated with this rule under these assignments will contribute nothing to the overall loss. In a HL-MRF, jointly minimizing the distance to satisfaction of all rules results in MAP assignments to all unknown variables.

$$w_p \cdot max\{1 + 0.8 - p_2 - 1, 0\}. \tag{2.4}$$

HL-MRFs are highly scalable and the efficiency with which they perform inference makes them an optimal choice for prediction tasks involving richly structured data. PSL is an intuitive interface for tempting these HL-MRFs which can express a broad range of settings. One advantage of PSL is that it can readily model latent variables.

In this thesis, I use latent variables to model many unobserved phenomena across domains, from latent speech categories to product sustainability scores. In the latent setting, the joint probability distribution shown in Equation 2.1 is defined over $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ where $\boldsymbol{z}$ is a vector of latent variables. To perform weight-learning in the presence of latent variables, we use a version of expectation maximization described by Bach et al. [12]. I will next discuss the projects which have used PSL, and demonstrate my contributions.

Part II

Sustainability

There are many roles computational methods can play in benefiting environmental sustainability. Living *sustainably* has become a lifestyle goal for many. However, interest in sustainability does not necessarily translate into sustainable purchases. The difference between one's intentions and ability to act in line with them is referred to as the *intention-behavior*, *attitude-behavior*, or *value-action gap* [107].

Reducing household energy use in the United States can have a large and positive impact on the environment. However, deciding how to save energy can flummox many consumers. In a pivotal study, Attari et al. [8] found that when surveyed about energy-saving strategies most participants mentioned *curtailment* (e.g. using less resources), rather than the expert recommended choice of *efficiency improvements* (e.g. replacing an old appliance with a more efficient model). Furthermore, respondents underestimated the usage of the most energy-intensive appliances. Clearly, there is a need to improve consumers understanding of energy-saving strategies.

Similarly, consumers face information difficulties when shopping sustainably. Defra et al. found that roughly 30% of consumers in the UK would like to act sustainably but are not able to translate their environmental concern into sustainable purchases. Hughner et al [51] further expound the value-action gap by showing that though 46-67% of surveyed consumers expressed support for organic food, only 4-10% of those same respondents expressed an intention to purchase organic.

Barriers to sustainable shopping include: cost, availability, skepticism of labels and insufficient marketing [51]. Unlike cost and availability, skepticism and marketing are issues that might be overcome with better communication about the sustainability profile of available products. Furthermore, knowledge about what makes a product sustainable can be a barrier to consumers [95], while Tanner and Wölfing-Kast [125] found that possessing actionable knowledge is correlated with taking action. Vermeir and Verbeke [139] found positive desires to purchase sustainable food items were impeded by a lack of certainty about those items' sustainable characteristics. Thus, there is a strong need to improve communication and provide actionable information about sustainable products.

Here, I am concerned with how to provide consumers with actionable information so that they may make informed and sustainable decisions. I address two challenges

in delivering quality information to consumers: appliance-level feedback on energy consumption and discovering sustainable products in a large unlabeled product catalog.

There is evidence that people benefit from specific and timely feedback on their energy consumption [20, 120, 26, 137]. However, in order to provide this feedback one must be able to disaggregate the aggregate power reading collected by energy meters. I propose a collective probabilistic method for doing so.

In addition, people may be better able to shop sustainably given detailed sustainability profiles of products they are interested in [122]. However, defining what makes a product sustainable is not straightforward. For those who value sustainability, products can be appraised according to complex factors including: environmental impact, impact on the local economy, animal welfare considerations and benefit to the consumer [90]. Finally, there are no ground truth labels of sustainability, and when making recommendations from large product catalogs, it is crucial to be able to scalably score products according to their sustainability. I present a recommender system which can discover sustainable products and customers and use these categorizations to make sustainable recommendations.

Reducing consumption levels of developed countries is imperative for sustaining life on earth and for improving global equality. While many consumers may wish to consume less, they do not always know how to do so. In Chapter 3 I present a collective probabilistic approach for disaggregating individual appliances signals from aggregate readings. Such an approach can be used to provided detailed energy feedback to consumers. In Chapter 4 I present a recommender system approach for discovering sustainable products and customers. This method can be used to both discover and analyze sustainable shopping habits and to make sustainable recommendations.

Each of these approaches benefits by utilizing the inherent structure in these problems. In Chapter 3, I introduce appliance sets which are useful in relating the power usage of groups of appliances to observed usage. Additionally, I explore temporal representations and make use of temporal dependencies. In Chapter 4, my proposed recommender system considers similarities between consumers and between products as a form of structure. Recommendations are made collectively utilizing these similarities.

Together, this work addresses the information deficit which might increase the

value-action gap. However, information is not the only factor of making sustainable consumption decisions. For example, conveying social norms can have a large effect on sustainable behavior [27, 113]. Expanding these methods to take account of social behavior is a potential next step.

Furthermore, the usefulness of these methods will ultimately be determined through their ability to convey actionable information to real people, in real time. In future work I will these methods in real-life experiments. An open question is how best to convey the information, e.g. a descriptive norm framing might be more effective in improving sustainability than a clear presentation of the sustainability profiles of appliances and products.

Chapter 3

Actionable Energy Insights: Disambiguating Household Appliances

3.1 Introduction

Households consume over one third of all electricity in the United States [2], and opportunities for decreasing this share abound [28]. However, these opportunities are impeded by the lack of information available to residential consumers. Consumers are often uninformed as to which appliances consume the most energy [35, 8], and which actions have the greatest savings potential. Furthermore, there is growing evidence that detailed feedback about energy use can reduce consumption [20, 120, 26, 137]. Currently, this reduction is hampered by the fact that residents receive only aggregate energy information. As a simple analogy, consider receiving a shopping bill with a single figure and being asked to spend less on the next shopping trip. Based on this information alone, it would be difficult to discern how to adjust purchasing habits.

Advanced Metering Infrastructure (AMI), such as smart meters, measure aggregate energy consumption at defined intervals and transmit measurements wirelessly. Smart meters offer a unique opportunity to gather real-time data, learn energy consumption patterns, and ultimately offer actionable insights to consumers. Energy disaggregation (also referred to as non-intrusive load monitoring (NILM)) is the process of determining the energy consumption of individual appliances, given only an aggregated energy reading. A successful disaggregation algorithm can give consumers an

itemized energy bill, displaying how much energy is consumed by each appliance, rather than the aggregate monthly reading they currently receive. While there are existing approaches to this problem [61, 73, 92], none have been deployed in a real-world setting with low-frequency smart-meter readings.

Here, we propose a probabilistic energy disaggregation framework which determines the most likely collection of appliances that are in use. Our framework uses (soft) constraints and context to disambiguate between similar appliances. By introducing appliance sets we can infer appliance states collectively, allowing us to benefit from the inherent structure in this problem. We also introduce a representation for energy readings which encodes state duration directly. To formulate the inference task we use a *hinge-loss* Markov random field (HL-MRF) [10], which allows a flexible probabilistic formulation and admits efficient inference. We evaluate our proposed framework on two real-world data sets. Empirical results demonstrate that our proposed probabilistic model significantly outperforms existing state-of-the-art techniques. In addition to our novel probabilistic formulation, we show the effectiveness of our proposed interval representation and the benefit of jointly modeling appliance states. Finally, we catalog several situations where contextual information is helpful for disambiguating active appliances.

3.2 Related Work

Hart's seminal paper [44] on non-intrusive load monitoring (NILM) introduced the problem of energy disaggregation. Broadly, there are two classes of solutions, those that require additional hardware and those that do not. Here we review only those approaches which do not require any additional equipment installation and thus are truly non-intrusive.

While alternative approaches exist [61, 73], factorial hidden Markov models (FHMM)s [36] and variants thereof, have been a popular choice for disaggregation algorithms [66, 63, 92, 55]. In this setting each appliance is represented with a single HMM, where the discrete hidden state variables correspond to the state of the appliance, and the observed continuous random variables correspond to the power readings. FHMMs allow multiple HMMs to be joined through a single observed variable in such a way that

approximate inference is tractable. However, as the inference is approximate, it is not optimal, a shortcoming addressed by a number of recent papers [79, 115].

Our method is similar in spirit to the FHMM line of work as we also employ a structured probabilistic framework. However our HL-MRF structure is more flexible and can model a host of constraints without being restricted by the generative assumptions of an FHMM. Furthermore, we build on the work of Kim et al. [[63]] and Li and Zha [[75]], by integrating non-traditional context features. Like Shaloudegi et al. [[115]], our approach exploits state-of-the-art optimization techniques such as ADMM [19], however our approach ends up being significantly more scalable, taking just a few minutes on homes from the REDD dataset versus over an hour for ADMM-RR.

3.3 Problem Definition

A disaggregation algorithm, $\mathbf{r} \rightarrow \mathbf{A}$, produces a mapping from a sequence of energy readings, $\mathbf{r}$, to a corresponding sequence of appliance states, $\mathbf{A}$. In the problem setting we consider, we define a sequence of n energy readings, $\mathbf{r} = \langle R_1, \ldots, R_n \rangle$. In addition, we are provided with a set of m appliances, $A = \{a_1, \ldots, a_m\}$. Each appliance a_i is associated with a set of k_i possible states. The number of states varies by appliance, some may have two states (*on*, *off*), while others may have multiple modes (*off*, *low*, *medium*, *high*). For each appliance state, we denote the energy consumed by appliance i in state k as c_i^k. We introduce an indicator $a_{i,j}^s$ that specifies that appliance i is in state s during reading j and define $\mathbf{A}$ as a matrix of these indicators. Each column is a vector of indicators corresponding to a particular reading, such that the j^{th} column has the form $\{a_{1,j}^{s_1}, \ldots, a_{1,j}^{s_{k_1}}, \ldots, a_{m,j}^{s_1}, \ldots, a_{m,j}^{s_{k_m}}\}$.

Using this model formalization, we introduce a probabilistic formulation for the disaggregation problem. We introduce a binary random variable $y_{i,j}^s$ corresponding to each indicator variable, $a_{1,j}^s$. The variable $y_{i,j}^s$ takes value 1 when appliance a_i is in state s during reading j, and 0 otherwise. The goal of disaggregation is to estimate the probability, $P(y_{i,j}^s)$, of each possible appliance state for each reading in the sequence, and determine the most probable state for each appliance, $y_{i,j}^{\hat{s}} = \underset{s \in \{1 \ldots k_i\}}{\operatorname{argmax}} P(y_{i,j}^s)$. This probability estimate should obey the constraint that expected consumption equals the

measured usage in the reading, or $R_j = \sum_{i=1}^{m} c_i^s y_{i,j}^{\hat{s}}$.

Frequently, the task of estimating these probabilities is under constrained, such that many potential configurations of appliance states may yield the same observed energy consumption. Thus, disaggregation algorithms can improve identifiability by modeling the joint probability distribution over all appliance states for reading j, $P(Y_j)$. Many possible probabilistic models can be used to characterize this probability distribution. One common example is using previous appliance states to model current appliance states, e.g. estimating the conditional probability $P(Y_j|Y_{j-1})$. In the next section, we discuss how to build a robust probabilistic model of appliance states.

3.4 Modeling Approach

We propose a flexible framework which can disaggregate individual appliances from aggregate power readings. The framework is designed with real-world applicability as the end goal; it can adapt to multiple categories of information and is able to disaggregate even with coarse power readings. One of the framework's central goals is disambiguating between appliances which have similar power demands, but are used in different contexts for differing lengths of time. We introduce two conceptual representations: appliance sets and an interval temporal formulation.

3.4.1 Appliance Sets

Rather than predict the states of appliances independently, our model captures a joint configuration of appliances that we refer to as an *appliance set*. The energy consumption of an appliance set can easily be determined by aggregating the energy usage of each appliance. By grouping appliances together we are able to reason jointly about the relative likelihood of sets rather than individual appliances, for example, a heater and an air conditioner is an unlikely pair, while an air conditioner and pool pump is not. Here we only infer the states of those sets which are within a reasonable distance to the observed total power. Thus the input of target appliances to be disaggregated can be split into J unique subsets, corresponding to collections of feasible appliances. Then for any reading R_t there is a single set S_j such that all appliances in S_j are on at time t.

3.4.2 Interval Representation

In addition to modeling appliance sets, we also explore aggregating several instantaneous readings into an interval representation. If two consecutive readings are similar, it is likely that they correspond to the same appliance set. In this case, inferring the most likely appliance set for an interval rather than for each individual reading in the interval can improve the efficiency as well as the accuracy of the model. To define intervals, we coalesce readings where total power has only minimal fluctuations. When the difference in consumption of two consecutive readings exceeds a threshold δ, our model establishes a new interval. The sequence of power readings is now indexed by intervals instead of time, so that we have V intervals to disaggregate, rather than N readings, $\langle R_1, \ldots, R_V \rangle$, where $V \ll N$.

3.5 Probabilistic Disaggregation Framework

Our framework integrates diverse sources of information into a joint probability distribution over active appliance sets. We model this probability distribution as a *hinge-loss* Markov random field (HL-MRF)[10]. Finding the most probable appliance set for each reading corresponds to maximum a-posteriori (MAP) inference in the HL-MRF. To specify an HL-MRF, we use the templating language PSL (see Chapter 2).

As a reminder of how PSL uses logical rules to template HL-MRFs, here we show an example rule in the disaggregation setting:

$$w_a : \textsc{ActiveInSet}(A_i, S_j) \wedge \textsc{AppSet}(R_l, S_j) \Rightarrow \textsc{IsOn}(R_l, A_i).$$

Here the rule has weight w_a and predicates: AppSet, IsOn, and ActiveInSet capture the relationships between the variables: A_i, S_j, and R_l. A predicate and its arguments (either variables or constants), constitute an atom. Atoms in PSL rules have a truth value in the continuous interval $[0, 1]$, allowing a relaxation of Boolean logic.

We outline the rules which define our model, and in Section 3.5.6 we describe how these rules are used to define a HL-MRF which captures the probabilistic dependencies and constraints in our domain.

3.5.1 Energy Disaggregation Template

We introduce several logical predicates to capture important elements of our model. Here we consider two appliance states, on and off, however the extension to additional states is straightforward. We capture that appliance i is on during reading l via the atom $\textsc{IsOn}(R_l, A_i)$. We define a mapping between appliances and appliance sets using $\textsc{ActiveInSet}(A_i, S_j)$, which is true when appliance i is active in set j. To capture if the appliance set j is active during reading l, we use the atom $\textsc{AppSet}(R_l, S_j)$, where only one appliance set can be active at a time. In our model we consider all appliance sets, however, in practice we restrict the number of feasible appliance sets considered for any reading as explained in Section 3.6.1. Accordingly, we introduce the following constraint:

$$\sum_{j=1}^{J} \textsc{AppSet}(R_l, S_j) = 1.0.$$

3.5.2 Interval Duration

Appliance usage often follows consistent patterns, and understanding the length of time an appliance spends in each particular state provides a powerful disambiguating signal. In our model, we define the discrete duration classes: *very short*, *short*, *medium*, and *long*. The length of each duration is estimated from data and differs for each dataset, but for example, a very short duration is less than 4 minutes, while a long duration would be more than 19 minutes. Duration classes were found from quartiles of all appliance durations in the training data, for example the threshold under which a duration would be labeled *very short* was the 25% percentile of all appliance on-state durations.

Using the atom $\textsc{Duration}(R_l, \mathit{length}(R_l))$, we can specify the duration of interval R_l. We can then learn a duration-specific prior for each appliance:

$$w_{dur} : \textsc{Duration}(R_l, \mathit{length}(R_l)) \Rightarrow \neg\textsc{IsOn}(R_l, A_i).$$

These learned priors can capture patterns such as microwaves rarely being on for long durations and dishwashers rarely being on for very short durations.

3.5.3 Observed Consumption

The difference between an appliance set's consumption and the observed power reading is modeled with the atom $\textsc{CloseToConsumption}(R_l, S_j)$. Let the expected consumption of S_j, be $E_j = \sum_{i \in S_j} \mu_i$, $\textsc{reading}(R_l)$, be the value of reading R_l in watts. Then, $\textsc{CloseToConsumption}(R_l, S_j)$ equals,

$$1 - \min\left(1, \frac{|\textsc{reading}(R_l) - E_j|}{\max(E_j, \textsc{reading}(R_l))}\right).$$

The weighted rule below expresses that the truth value of an appliance set depends on its distance to consumption.

$$w_c : \textsc{CloseToConsumption}(R_l, S) \Rightarrow \textsc{AppSet}(R_l, S)$$

To express the relationship between appliances and appliance sets we introduce the predicate $\textsc{ActiveInSet}(A_i, S_j)$, which is 1 if A_i is active in S_j, and 0 otherwise. Thus to propagate information about appliances to appliance sets, and vice versa we use two rules:

$$w_{app} : \textsc{ActiveInSet}(A_i, S_j) \wedge \textsc{AppSet}(R_l, S_j) \Rightarrow \textsc{IsOn}(R_l, A_i)$$

$$w_{as} : \textsc{ActiveInSet}(A_i, S_j) \wedge \neg\textsc{IsOn}(R_l, A_i) \Rightarrow \neg\textsc{AppSet}(R_l, S_j)$$

3.5.4 Capturing State Changes

One expectation of energy readings is that changes in observed energy usage correspond to a single appliance changing state, rather than a significant change in the active appliances. Implementing this intuition using appliance sets requires capturing the relationships between appliance sets more directly.

We specify the difference between two consecutive readings using the atom, $\textsc{Diff}(R_1, R_2, \mathrm{D})$, and apply a threshold $D > \delta$ to generate only meaningful differences. We introduce the predicate $\textsc{Positive}$ for differences greater than 0, and the atom $\textsc{CloseToDiff}(A, \mathrm{D})$ to capture the distance between the energy consumption of appliance A and the observed difference D. Finally $\textsc{Toggle}$ states that the difference between two appliances sets is exactly appliance A, that is $(S_1 \cup S_2) \setminus (S_1 \cap S_2) = A$. We then put these into two final rules:

$$
\begin{aligned}
w_{ton}: \quad & \textsc{Diff}(R_1, R_2, \mathrm{D}) \wedge \textsc{Precedes}(R_1, R_2) \\
\wedge \quad & \textsc{CloseToDiff}(A, \mathrm{D}) \wedge \neg\textsc{IsOn}(R_1, A) \wedge \textsc{Positive}(\mathrm{D}) \\
\wedge \quad & \textsc{Toggle}(S_1, S_2, A) \wedge \textsc{AppSet}(R_1, S_1) \Rightarrow \textsc{AppSet}(R_2, S_2)
\end{aligned}
$$

$$
\begin{aligned}
w_{toff}: \quad & \textsc{Diff}(R_1, R_2, \mathrm{D}) \wedge \textsc{Precedes}(R_1, R_2) \\
\wedge \quad & \textsc{CloseToDiff}(A, \mathrm{D}) \wedge \textsc{IsOn}(R_1, A) \wedge \neg\textsc{Positive}(\mathrm{D}) \\
\wedge \quad & \textsc{Toggle}(S_1, S_2, A) \wedge \textsc{AppSet}(R_1, S_2) \Rightarrow \textsc{AppSet}(R_2, S_1)
\end{aligned}
$$

Additionally we create a set of rules to capture the persistence of an appliance being on. To do so we use a predicate $\textsc{Precedes}(R_i, R_j)$, which is true if index value i, directly precedes index value j, or $i = j - 1$. The following rules allow us to express the probability that if an appliance contributes to R_l it will contribute to R_{l+1}.

$$
\begin{aligned}
& w_{stay_on}: \textsc{IsOn}(R_l, X) \wedge \textsc{Precedes}(R_l, R_{l+1}) \Rightarrow \textsc{IsOn}(R_{l+1}, A_i) \\
& w_{turn_off}: \textsc{IsOn}(R_l, X) \wedge \textsc{Precedes}(R_l, R_{l+1}) \Rightarrow \neg\textsc{IsOn}(R_{l+1}, A_i) \\
& w_{turn_on}: \neg\textsc{IsOn}(R_l, X) \wedge \textsc{Precedes}(R_l, R_{l+1}) \Rightarrow \textsc{IsOn}(R_{l+1}, A_i)1) \\
& w_{stay_off}: \neg\textsc{IsOn}(R_l, X) \wedge \textsc{Precedes}(R_l, R_{l+1}) \Rightarrow \neg\textsc{IsOn}(R_{l+1}, A_i)
\end{aligned}
$$

3.5.5 Contextual Rules

Contextual rules are designed to capture the context in which a resident uses a given appliance. By developing a rich sense of context we reduce reliance on ground truth data, and introduce information which reduces the variance across appliances. Here we introduce two types of contextual information, temporal and temperature.

3.5.5.1 Temporal Rules

Appliance usage often depends on the time of day, and day of the week. For example, it is more likely that a cooking appliance, such as a microwave, will be used in the evening, than in the middle of the night. Thus we introduce two predicates which state the hour and day of the week at which a reading occurred: $\textsc{Hour}(R_l, H)$ and

DayOfWeek(R_l, D), where $H \in \{0, 23\}$ and $D \in \{Sunday...Monday\}$. We then learn the relationships between hour, day of the week, and appliance.

$$w_{day} : \text{DayOfWeek}(R_l, D) \Rightarrow \text{IsOn}(R_l, A_i)$$

$$w_{hour} : \text{Hour}(R_l, H) \Rightarrow \text{IsOn}(R_l, A_i)$$

3.5.5.2 Temperature Rules

We also model the relationship between temperature and appliance usage. Such a relationship should be particularly strong with heating and cooling appliances such as an air conditioner. To incorporate temperature into the model we introduce the predicate Temperature(R_i, Temp), where Temp is either *cold*, *mild* or *hot*. Thus we relate appliances to temperature with the following rule:

$$w_{temp} : \text{Temperature}(R_l, \text{Temp}) \Rightarrow \text{IsOn}(R_l, A_i)$$

This concludes our overview of the energy disaggregation rules, we now explain how these rules can define a HL-MRF.

3.5.6 From Disaggregation Templates to HL-MRFs

HL-MRFS are a general class of conditional, continuous probabilistic models, parametrized with a set of weighted *hinge-loss* functions. Hinge-loss functions can model a rich diversity of relationships, and critically, admit highly scalable inference. We now specify how a HL-MRF is defined from a set of weighted logical rules, such as those defined in the probabilistic disaggregation framework. Let us now turn to an example rule from the previous section:

$$\lambda : \text{ActiveInSet}(A_i, S_j) \wedge \text{AppSet}(R_l, S_j) \Rightarrow \text{IsOn}(R_l, A_i)$$

where λ is a weight, ActiveInSet, AppSet, and IsOn are predicates, and A_i, S_j, and R_l are all variables. By substituting constants, a_i, s_j, and r_l for the variables, A_i, S_j, and R_l respectively, one obtains three ground atoms: ActiveInSet(a_i, s_j), AppSet(r_l, s_j), and IsOn(r_l, a_i), such that each ground atom takes a value in $[0, 1]$. Suppose that we let

the atoms ActiveInSet(a_i, s_j), AppSet(r_l, s_j), and IsOn(r_l, a_i) correspond to three random variables: $y_1, y_2,$ AND y_3 respectively.

Let $\mathbf{y}$ be a vector of $IsOn(R_l, A_i)$ variables for all V intervals and all N appliances, and let $\mathbf{s}$ be a vector of $AppSet(R_l, S_j)$ variables for all V intervals and all J appliance sets. Let x denote all atoms which are observed, and for which the truth values are thus known. We then formulate our task as finding the MAP assignments to $\mathbf{y}$ and $\mathbf{s}$ under the probability distribution defined by the HL-MRF formulation of the rules described in 3.5.1:

$$\underset{\mathbf{y},\mathbf{s}}{\operatorname{argmax}} P(\mathbf{y}, \mathbf{s}|x).$$

In the next section we evaluate the performance of this framework on two real-world datasets.

3.6 Empirical Evaluation

We evaluate our proposed disaggregation framework[1] on two real-world datasets. We demonstrate the effectiveness of our framework compared to ADMM-RR [115], a recent, state-of-the-art approach. We also explore the situations under which contextual information can help improve model performance.

3.6.1 Data

We evaluate on two public datasets: The Reference Energy Disaggregation Dataset (Redd) [67] and Pecan Street Inc. (Dataport) [[54]].

REDD: This is the most widely used dataset for energy disaggregation. The dataset describes six homes, and each home has an average of 21 days. The dataset contains fine-grained meter readings at approximately six second intervals. Following Makonin et al. [[79]] and Johnson et al. [[55]], we evaluate our model using homes 1, 2, 3, and 6 from the dataset and likewise, we disaggregate the refrigerator, lights, microwave, and dishwasher. Unlike previous work, we omit the furnace as it appears in only a single home, and for home 6 we omit the dishwasher.

[1] Code available: `https://bitbucket.org/linqs/appliance_disambiguation`

DataPort: We also evaluate on the Pecan Street dataset which describes meter readings from Austin, Texas. This data set is much larger than REDD (hundreds of homes over several years) and is less well-studied in the context of disaggregation. The readings are coarser, available at either 1 minute or 1 hour intervals. We evaluated on the eight most common appliances: air conditioner, furnace, refrigerator, dishwasher, kitchen outlet, dryer, microwave and clotheswasher. As not all homes have perfect data records; here we choose the five homes (2859,3413,6990,7951,8292) for which there was at least one year of complete data (no records missing for any of the target appliances) and which had completed surveys describing demographics and household features. The models for each home (including the weights) are trained using the first 50% of the data, the next 25% of the data is used as a validation set for model parameters, and the model was evaluated on the final 25% of the data. With DATAPORT this training, validation, and testing is done for each month separately. For both datasets, we estimate the mean and standard deviation of the power consumption for each appliance based on only the training data. Evaluation estimates used both training and validation data. To find the thresholds to partition duration lengths into *very short*, *short*, *medium*, and *long*, we found quartiles for the interval lengths, such that 25% of all duration lengths were assigned to each duration. Similarly, each temperature was uniformly partitioned into one of three categorical labels, corresponding to *cold*, *mild*, and *hot*.

To partition the data into intervals, for each home we let δ be the difference between the average draw of a small appliance less its standard deviation. For REDD homes the small appliance was either the lights or the refrigerator, and for DATAPORT it was always the refrigerator. To assign feasible appliance sets for each interval, we compute the absolute difference between the mean consumption for each appliance set and the observed power, scaled to be in $[0, 1]$ and we then retain only those sets which are within 0.5 of the actual power consumption. If there are no such sets, we select the top three closest sets.

3.6.2 Results

To estimate appliance consumption, both the appliance state and the power consumption must be inferred. In order to evaluate performance, we measure both at how well

an algorithm can predict the appliance states (precision, recall and F-measure), and how well it predicts actual consumption (Mean Absolute Error (MAE)). We evaluate three methods:

Instance: This model treats each instance independently, models the appliance sets jointly and does not use duration information.

Interval: This is the model described in Section 3.5 where we model both appliance sets and their durations.

ADMM-RR: This is the current state of the art method which uses a factorial HMM model [115]. We use the code they provide online.[2]

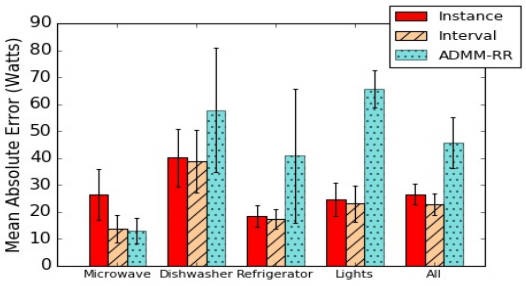


Figure 3.1: MAE on REDD data.

Fig. 3.1 shows the MAE of the three methods on the REDD data set, and Table 3.1 shows the precision, recall and F-measure, for each appliance using the interval model. Fig. 3.2 provides a detailed view of the difference between observed and estimated energy usage for one specific home. The interval model performs the best overall, reducing the MAE of ADMM-RR by 50%.

The next set of results are on the DATAPORT data set. This dataset includes contextual features not available in the REDD dataset, so in addition to comparing the instance and interval approaches, we explore a model which captures contextual temporal and temperature features. Fig. 3.3 compares the MAE of three PSL-based methods with the ADMM-RR baseline on the DATAPORT dataset, and Table 3.2 shows

[2]To learn the required parameters we used the Matlab HMM toolbox [83] and the first 75% of the data for each home.

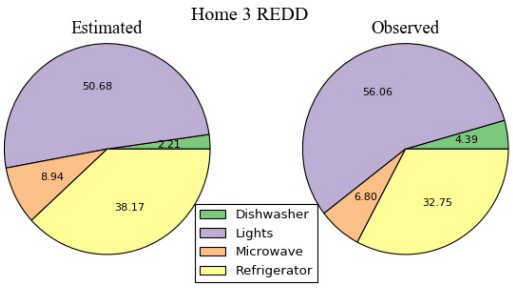


Figure 3.2: Percentage of total energy consumption of each appliance for a representative REDD home; predictions from the Interval model.

	Precision	Recall	F-Measure
Dishwasher	0.518	0.598	0.555
Lights	0.708	0.813	0.757
Microwave	0.707	0.712	0.709
Refrigerator	0.851	0.879	0.865
Average	0.696	0.751	0.722

Table 3.1: Performance of Interval Model on REDD data.

	Precision		Recall		F-Measure	
	Interval	+Context	Interval	+Context	Interval	+Context
Air Conditioner	0.901	0.899	0.815	0.823	0.856	0.859
Clotheswasher	0.226	0.228	0.333	0.274	0.269	0.249
Dishwasher	0.063	0.072	0.360	0.368	0.108	0.121
Dryer	0.591	0.571	0.731	0.749	0.653	0.648
Furnace	0.844	0.829	0.621	0.645	0.716	0.726
Kitchen Appliance	0.045	0.046	0.426	0.358	0.081	0.082
Microwave	0.330	0.354	0.394	0.394	0.359	0.373
Refrigerator	0.675	0.675	0.806	0.828	0.735	0.744
Average	0.459	0.459	0.561	0.555	0.505	0.503

Table 3.2: Performance of the Interval and Context Models on DATAPORT data. The effect of context varies by home and appliance.

the precision, recall, and F-measure for each appliance for the interval and context-based models. A paired t-test demonstrates the context-based PSL model provides a

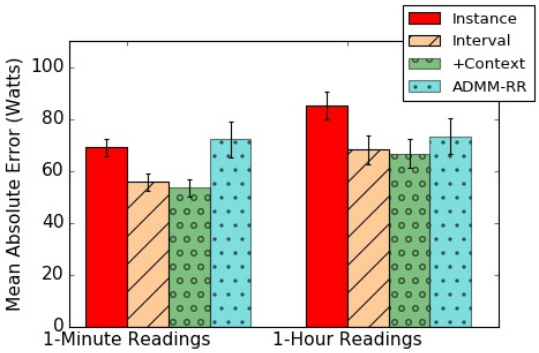


Figure 3.3: The MAE for Instance, Interval and Context Models on DATAPORT data, with both 1-minute and 1-hour readings. The contextual information provides a statistically significant improvement for readings at both time resolutions.

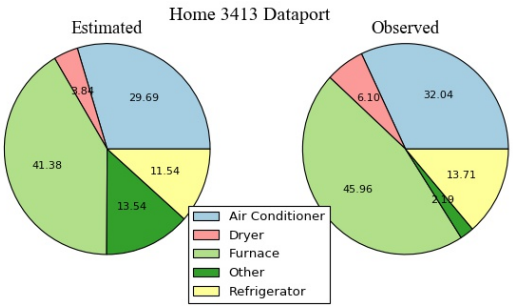


Figure 3.4: Percentage of total energy consumption of each appliance for a representative REDD home; predictions from the Interval model.

statistically significant improvement over the interval model on MAE. Table 3.2 reveals that the effect of context differs by appliance, with the contextual models performing best on the appliances with high energy consumption. Fig. 3.4 provides a more detailed view of the difference between observed usage and model estimates for one specific home.

To understand the value of different contextual signals, we examine appliances for which temperature and temporal information has the greatest benefit. Contextual rules on temperature improved the predictions for heating and cooling appliances, with statistically significant improvements for the refrigerator and air conditioner. Temporal information improved average predictions for appliances with periodic usage patterns, such as the clotheswasher, dishwasher, dryer, and microwave.

3.7 Discussion

Our proposed disaggregation framework achieves state-of-the-art performance on two real-world datasets, reducing error by 50% and 25% on REDD and DATAPORT datasets, respectively. Our algorithm adapts to different granularities of data, from the 6 second samples of REDD, to the 1-minute and 1-hour samples from DATAPORT. Not surprisingly, the DATAPORT dataset is more challenging, as we disaggregate more appliances at a rougher resolution, yet the F-measure for the most energy-intensive appliances remains above 0.7. Across datasets, we demonstrate that the interval representation, which aggregates readings into usage events, improves performance over a purely instance-level representation even when readings have low granularity. Beyond representation, one strength of our approach is the ability to incorporate different types of contextual information. Contextual information provides a small reduction in error and targeted improvements in predictive performance, however we had anticipated more pronounced improvements. One potential explanation is the limited range of contextual information due to the small temperature range in the dataset. In future work, we hope to pursue a deeper exploration of contextual information and its efficacy at generalizing predictions across homes.

While smart meters have been installed in homes across the United States, their potential in reducing consumer energy consumption is far from realized. Improved energy disaggregation algorithms can help to reach that potential by discovering appliance-level consumption patterns that consumers need to make informed decisions about their energy usage. In the paper, we propose such an algorithm, which is efficient, scalable, and readily adapted to new sources of information. By performing inference over intervals, and modeling collections of feasible appliance sets, we reduce the complexity of the problem while retaining the advantages of a structured probabilistic formulation. A key advantage of our framework is that additional information can be incorporated easily. For example, we could enrich our models by incorporating richer user and building profiles.

3.8 Conclusion

While smart meters have been installed in homes across the United States, their potential in reducing consumer energy consumption is far from realized. Improved energy disaggregation algorithms can help to reach that potential by discovering appliance-level consumption patterns that consumers need to make informed decisions about their energy usage. Here, we propose such an algorithm, which is efficient, scalable, and readily adapted to new sources of information. By performing inference over intervals, and modeling collections of feasible appliance sets, we reduce the complexity of the problem while retaining the advantages of a structured probabilistic formulation. A key advantage of our framework is that additional information can be incorporated easily. For example, we could enrich our models by incorporating richer user and building profiles.

Another important aspect of sustainability is product consumption. Reducing, and adjusting consumption towards more sustainable products can have a positive environmental impact. In the next chapter we propose a similar approach in designing a sustainable recommender system. As in the energy disaggregation framework, our recommender system can also easily incorporate heterogenous data, such as sustainability indicators and past consumption history. This sustainable recommender system can be used to discover and recommend sustainable products, which ultimately might increase sustainable consumption.

Chapter 4

Sustainability at Scale: Bridging the Intention-Behavior Gap with Sustainable Recommendations

4.1 Introduction

Consumers with expressed intent in shopping sustainably do not always act on these intentions. Consumers may not trust product labeling, or understand the differences between various certifications [88]. Additionally, pricing and availability can prevent people from purchasing sustainably.

Given knowledge of sustainable products, a good recommender system might offer products with these characteristics to interested customers. However, defining what makes a product sustainable is not straightforward. For those who value sustainability, products can be appraised according to complex factors including: environmental impact, impact on the local economy, animal welfare considerations and benefit to the consumer [90].

For this research, we have purposefully used a broad definition of sustainability that covers all these aspects. A product which scores strongly according to any one dimension might be labeled as sustainable, and multiple sustainability-related features will increase its likelihood of being labeled as sustainable. Future research could use similar methods with more narrow definitions of sustainability.

We contend that recommender systems can play a significant role in the process of discovering sustainable products and customers. To this end, we propose a probabilistic approach which fuses multiple weak signals to infer the sustainability of products. Here we investigate three types of signals: freely available domain knowledge, product metadata and the purchasing patterns of customers predicted to be sustainability-minded.

This approach offers several advantages: we are able to flexibly incorporate prior knowledge about what might imply the sustainability of products or customers; no sustainability ground truth labels are required; and multiple recommender system inputs can be used to improve predictions. Additionally, we are able to benefit from the joint formulation of all three tasks of: discovering product sustainability scores, discovering the sustainability-mindedness of customers and inferring future purchases. We demonstrate these benefits by showing improvements in predicting future purchases of 80.8% in precision@5 over a SVD++ [68] implementation.

4.2 Problem Definition

We are given customer-item purchase data $X = (x_{i,j,d})$, where $x_{i,j,d}$ is 1 if customer c_i purchased item p_j on date d, and 0 otherwise. Our goal is to infer future purchases Y. This is an implicit feedback setting where the only information we have is which items a customer bought and when.

Our goal is to predict which items are sustainable and which customers are sustainably-minded. To that end, we will predict scores s^c and s^p, respectively, to each customer and item, corresponding to their degree of sustainability. For a customer, this score represents the extent to which they might be interested in sustainable products. For a product, the score is the extent to which this product might be environmentally sustainable. Unlike the purchases, there are no true labels for s^c or s^p, the extent to which a customer is interested in sustainability is ultimately an unknown quantity. Thus we characterize these variables with a latent formulation and infer their values without ground-truth scores.

4.3 Our Approach

We propose a collective probabilistic model which allows us to both predict future purchases $\boldsymbol{y}$ and jointly discover sustainability scores, s^c and s^p. We construct this model with PSL which offers several advantages for this setting: logical rules can intuitively capture domain knowledge, collective inference of both purchases and scores is highly efficient and latent variables facilitate the fusing of multiple signals. Next, we introduce describe PSL in this setting, in more detail.

4.3.1 Probabilistic Soft Logic Recommender System

To demonstrate how PSL can be used in a recommender system setting, consider a rule which states that if two customers are similar, and one customer is sustainable, the other one is as well. We introduce a predicate Similar, which takes two customer IDs as arguments and which expresses the similarity between these two customers as a value between 0 and 1. There are many ways to express similarity, and it is possible to use multiple definitions in a single PSL model. Here we calculate similarity from latent factors learned from a SVD++ model. To express the sustainability proclivity of a customer, we introduce the predicate SustainableCustomer, which takes a customer ID, c_i, as an argument, and whose truth value corresponds to the sustainability score s^{c_i}. With these predicates, and a weight w_{sim} which reflects the relative importance of this rule, we define our rule in PSL as follows:

$$w_{sim} : \textsc{Similar}(c_i, c_j) \wedge \textsc{SustainableCustomer}(c_i) \Rightarrow \textsc{SustainableCustomer}(c_j).$$

Combined with data, a PSL model defines a joint probability distribution over scores and purchases. Here, we incorporate latent variables, such as s^c, s^p, and purchase predictions. In this case, the joint probability distribution is defined over $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$, where $\boldsymbol{z}$ is a vector of latent variables. PSL is adept at fusing multiple sources of information. Next, we introduce our three sources of sustainability signals.

4.3.2 Three Signals of Sustainability

We propose three sources of information about what makes a product sustainable. One source is freely available information which can be collected from published material on or offline, as well as from domain experts. Another source is the product metadata. Finally, customers' purchasing patterns can be leveraged to identify additional products.

Domain Knowledge With the abundance of online publishing dedicated to sustainability, from magazines such as Mother Earth to consumer services such as GoodGuide, there are multiple sources for identifying potentially sustainable products. We incorporate knowledge of which brands are sustainable into our model.

Metadata One source of information is product metadata. This category can include a broad range of features, from product descriptions to the number of organic ingredients listed in a prepared food item. Here we consider two sources: certifications and specialities. Certifications are third party assessment of some sustainability-related attributes. For example, the USDA Organic program certifies that agricultural products have been produced using approved methods. Specialties are tags that provide product filters, for example: Organic, or Gluten-Free.

Sustainability-Minded Shoppers If we can identify a group of sustainability-minded shoppers then we can learn from their purchasing patterns. These customers may purchase products which are not officially certified, but which they believe to be sustainable, and which we would like to score as sustainable. Thus, we use purchasing patterns to identify sustainable customers and additional sustainable products from their shopping.

Each type of information suffers from its own drawbacks. Domain knowledge will inevitably be sparse, covering only small subsets of large datasets. Product metadata can contain errors or omissions that make it unreliable at times. Finally, purchasing patterns are also weak signals. From an analysis point of view, it would be ideal for sustainability-minded shoppers to only purchase sustainably. However, this is unrealistic [38], and we cannot expect the purchase history of sustainability-minded shoppers to be perfectly sustainable. By fusing these signals, we overcome the drawbacks of each and gain a more complete understanding of what makes a product sustainable. Next, we show how to do so in PSL.

4.3.3 PSL Sustainable Discovery Model

We infer the values of three latent variables: SustainableCustomer(C), SustainableItem(P) and PredictPurchase(C, P, D). We also infer the value of the target variable Purchase(C, P, D). The values of SustainableCustomer(C), and SustainableItem(P), correspond to s^c and s^p, and reflect the sustainability scores of customers and items. To predict if a customer will purchase a product on a given day, we use PredictPurchase. We consider each purchase event as a separate random variable, Purchase(C, P, D) which takes a value in [0,1], and is 1 if customer C purchased product P on date D.

w_{np} :¬Purchase(C, P, D)
w_{nsc} :¬SustainableCustomer(C)
w_{nsp} :¬SustainableProduct(P)

Table 4.1: Prior beliefs.

In large catalogs it is common that most products won't be purchased. We reflect this with the first rule in Table 4.1. We also encode that we expect most customers and products to not be sustainable in the next two lines. These rules encode our prior beliefs. However, the weights to these initial priors can be updated with data.

w_i : Certification($P, Cert_i$) $\Rightarrow$ SustainableProduct(P)
w_j : Specialty($P, Spec_j$) $\Rightarrow$ SustainableProduct(P)
w_b : SustainableBrand(B) $\wedge$ Brand(P, B) $\Rightarrow$ SustainableProduct(P)
w_s : SustainableProduct(P_1) $\wedge$ SimilarProducts(P_1, P_2) $\Rightarrow$ SustainableProduct(P_2)

Table 4.2: Sustainable Products

In Table 4.2 we introduce the rules by which we infer the sustainability of products. In the first rule, we relate certifications to products. A certification is awarded by an external service, for example a product can be USDA certified organic. For each certification i, we instantiate a new rule with weight w_i so that we can learn the relative value of each certification in predicting sustainability. Similarly, each product can potentially

be described with a Specialty, and for each specialty j, we instantiate a unique rule and learn a weight w_j. Additionally, if we have domain knowledge about the sustainability of companies, we can use this to infer which products are sustainable with the third rule in Table 4.2, which states that products offered by sustainable brands are themselves sustainable. Finally, we propagate information about similar products with the rule that says if two products are similar, and one is sustainable, the other one will be as well.

$w_{sc} : \textsc{SustainableCustomer}(C_1) \wedge \textsc{SimilarCustomers}(C_1, C_2)$
$\Rightarrow \textsc{SustainableCustomer}(C_2)$
$w_{shc} : \textsc{SustainableProduct}(P) \wedge \textsc{HasPurchased}(C, P)$
$\Rightarrow \textsc{SustainableCustomer}(C)$
$w_{spc} : \textsc{SustainableProduct}(P) \wedge \textsc{PredictPurchase}(C, P, D)$
$\Rightarrow \textsc{SustainableCustomer}(C)$

Table 4.3: Sustainable Customers

In Table 4.3, we introduce the rules to identify sustainability-minded shoppers. In the first rule, we leverage similarities across customers; if two customers are similar and one is sustainable, the other one likely is as well. Additionally, if at any time a customer has purchased a sustainable product, that customer might be sustainability-minded. Finally, we model that if a customer will purchase a sustainable product, they are likely sustainablility-minded.

Next, we fuse existing knowledge into predictions of what a customer will purchase. With the first two rules in Table 4.4 we again leverage customer and product similarities. In the next rule, we predict that a purchased item will be purchased again. This rule is very context specific, in many recommender system settings it would not be desirable to recommend a previously purchased item. However, as our analysis focuses on food, a category where repeat purchases are common, this rule is appropriate. We can also incorporate predictions from other recommender systems. For example, here we utilize predictions made by a Singular Value Decomposition (SVD) algorithm, with the predicate SVDPredicts, which assumes a value between 0 and 1 according to the

SVD algorithm. We then apply the predictions to the target variable $\textsc{Purchase}(C, P, D)$, with the last rule in Table 4.4.

$w_{tda} : \textsc{PredictPurchase}(C_1, P, D) \wedge \textsc{SimilarCustomers}(C_1, C_2)$ $\Rightarrow \textsc{PredictPurchase}(C_2, P, D)$
$w_{spp} : \textsc{PredictPurchase}(C, P_1, D) \wedge \textsc{SimilarItems}(P_1, P_2)$ $\Rightarrow \textsc{PredictPurchase}(C, P_2, D)$
$w_{pp} : \textsc{HasPurchased}(C, P) \Rightarrow \textsc{PredictPurchase}(C, P, D)$
$w_{svd} : \text{SVD}\textsc{Predicts}(C, P) \Rightarrow \textsc{PredictPurchase}(C, P, D)$
$w_{ppp} : \textsc{PredictPurchase}(C, P, D) \Rightarrow \textsc{Purchase}(C, P, D)$

Table 4.4: Predicting Purchases

4.4 Quantitative Evaluation

In the quantitative evaluation, we assess our framework by its ability to correctly predict customer purchases. To do so we compare our approach to one baseline and one state-of-the-art approach. Our approach is also notable in its ability to discover sustainable products and customers, as we have no ground truth labels for this task, we present a qualitative evaluation in Section 4.5.

4.4.1 Data

In these experiments we consider customer purchase data from Amazon.com. We focus the experiments on the Grocery category. This excludes related products such as Amazon Pantry. We choose food as a first exploration of sustainability, as there are clear sustainable metadata, such as organic and fair-trade.

We create training and test sets with 10,000 customers in training, 5,000 customers in the validation set and 5,000 customers in the test set. For each customer, all purchases in the validation set occur *after* all purchases in the training set, and all purchases in the test set occur after all purchases in the validation set. Since we have implicit preference feedback (i.e., we only know what was purchased, not what was con-

sidered but not purchased), we sample 100 negative purchases for training as in Said and Bellogín [109]. To further refine the problem, we only considered products purchased below a given threshold across a large number of customers. The total number of products was approximately 21,000. In the validation and test set, for each customer, we infer purchases on a single date D.

Our external information about which companies are sustainable was collected from two online sources.[1] When filtering for the companies present in our dataset we were left with sixteen sustainable companies. These were largely food companies, with the exception of Seventh Generation which sells household products.

The specialities and certifications are provided by Amazon. We consider the following specialties: organic, organic & whole grain, all natural, gluten free, wheat free, dairy free, natural ingredients only, sustainably caught, biodegradable and not-tested-on-animals. We expect these to indicate sustainability to varying degrees, for example, *organic* should be a stronger indicator than *all natural*. The certifications included in this analysis were: non-GMO, Rainforest Alliance, all organic certifications and all fair trade certifications.

4.4.2 Experiments

We compare our approach to two common baselines: a nearest neighbor (NN) search and SVD++ [68] which is preferable in the implicit feedback setting. Both methods are implemented in the python package Surprise [49]. We perform a grid search on hyper-parameters for SVD++ on the validation set. In Table 4.5 we show the percent improvement of the NN and our approach (PSL) over the SVD++ method, on the task of predicting purchases.

Here we see that PSL achieves the best performance according to the precision@k and Mean Average Precision (MAP). We expect SVD++ to outperform the NN baseline, however, it does so only according to the MAP. Next we inspect the sustainable signals: domain knowledge, product metadata and purchasing patterns.

[1] `https://eating-made-easy.com/sustainable-food-companies/`, `https://www.motherearthliving.com/food-for-health/sustainable-food-companies-zmoz12jazmel`

	SVD++	PSL	NN
p@5	0	**80.8**	55.5
p@20	0	**48.6**	30.5
p@50	0	**14.9**	7.46
MAP	0	**63.7**	-5.80

Table 4.5: Percent relative improvement w.r.t. SVD++, with largest statistically significant improvements bolded.

4.4.3 Effects of each signal

Our model fuses three types of information: domain knowledge, product metadata, and purchasing patterns. Thus we ask, how many products were discovered with each source of information? As our model is collective, and infers sustainability jointly, to resolve this question we ask how many products were discovered without each type of information. That is, how many products without metadata were determined to be sustainable, how many products without domain knowledge, and how many products without either were determined to be sustainable.

	Total	Domain, No Metadata	Metadata, No Domain	With Neither
Found	$\frac{862}{21000} \approx 4.1\%$	$\frac{39}{862} \approx 4.5\%$	$\frac{26}{862} \approx 3.0\%$	$\frac{13}{862} \approx 1.5\%$

Table 4.6: Most products were discovered using all three signals.

In Table 4.6, we see that 862, or roughly 4% of all products considered, were found to be sustainable. Comparing the importance of domain knowledge and metadata, we see that they are relatively similar. Of the discovered sustainable items, roughly 4.5% and 3.0% were missing metadata and domain knowledge respectively. This suggests that domain knowledge may be a stronger signal than metadata. However, only 1.5% were missing both. This shows that the combined sources of information are complementing each other rather than finding distinct sets of products. However, there is still a question of the quality of these discovered products. Next, we turn to a human evaluation to assess the discovered products.

4.5 Qualitative Evaluation

To assess the sustainability of discovered products, we ask a domain expert to evaluate a subset of selected products with high sustainability scores. In this evaluation, each item is scored as: Reasonably Sustainable, Partially Sustainable, or Not Sustainable. In this assessment we do not ask if it would be more sustainable to not purchase this item. For example, we assess the sustainability of a particular product of bottled water and if the bottle is made of recycled material it may obtain a high score. However, it would be more sustainable to buy a reusable water bottle. Thus, the evaluations of the products are made in isolation, without considering if a better alternative would be to not purchase at all.

We split the evaluation so that some products from each category in Table 4.6 are explored. Twenty products missing no information, ten products with no metadata, ten products with no domain knowledge, and ten products with neither domain knowledge or metadata are evaluated. We show the evaluations in Fig. 4.1.

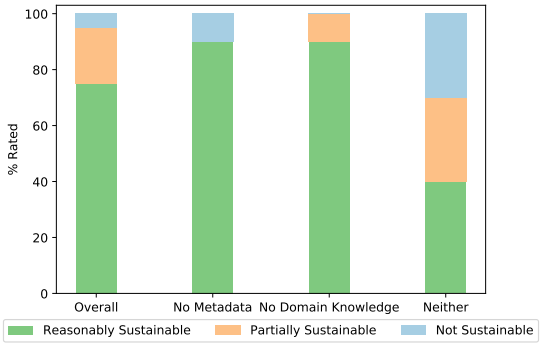


Figure 4.1: Human ratings for a subset of discovered products. Discovered products are rated as reasonably sustainable overall.

We see that the majority of evaluated products are found to be reasonably sustainable. Both metadata and domain knowledge are useful in detecting sustainable products. However, these two categories do not cover the entire dataset. Of the products without either metadata or domain knowledge, 40% of those predicted to be sustainable were deemed reasonably sustainable by an expert and 30% were deemed partially sustainable. While improvements can be made to better discover products with incomplete

information, the majority of these products were somewhat sustainable.

4.6 Discussion

We propose this work as a first step towards sustainable recommender systems. To the best of our knowledge, ours is the first recommender system to take sustainability into account. Our proposed model is able to find a reasonable number of sustainable products and the majority of these products are deemed reasonably sustainable by a domain expert. There are many ways in which this work could be expanded. In the current implementation, a small number of companies proved to be helpful in discovering sustainable products. Thus, expanding this number might improve results. Furthermore, our model would surely benefit from incorporating additional predictions from other recommender systems, which has proved useful in related work [69]. Another next step would be to utilize temporal information such as seasonal purchasing habits.

Our approach could be considered a hybrid recommender system [21, 39]. Thus, our work is similar to approaches which incorporate item features and domain knowledge with collaborative filtering [135, 133]. One consideration in our setting is that sustainability-minded customers may have constraints about what they will not buy; for example, products containing toxic pollutants. Like constraint-based methods [31], our probabilistic method can also incorporate hard constraints.

4.7 Conclusion

The intention-behavior gap expresses that people do not always behave in accordance to their intentions. This gap is partly due to the lack of easily accessible and credible information on product sustainability. To address this issue, we propose a model which utilizes three sources of information to discover sustainable products and make sustainability informed recommendations. We demonstrate that this approach leads to better recommendations than baselines. Furthermore, 74% of the discovered products are deemed reasonably sustainable by an expert human evaluator.

Part III

Education

Educational data mining, or learning analytics, is a young field, whose goal is to be able to advance student learning by analyzing new kinds of, and unprecedented amounts of data. New kinds of data include: interactions with automatic tutors, activity on online courses, and use of digital education applications. Whereas a study of a specific intervention or curriculum in a classroom would require physical monitoring, and thus be impossible to conduct at scale, digital interactions can be automatically recorded for large numbers of students simultaneously. This explosion of data from massive open online classrooms, (MOOCs) creates an unprecedented opportunity to answer basic educational research questions. Simultaneously this new environment poses many questions which demand exploration, for example, can students successfully learn in an online environment, which kinds of students is this environment best suited for, and how does learning differ from a physical to online environment.

One particular setting which has not seen sufficient attention, is high school MOOCs. Colleges, and universities have been most aggressive in offering MOOCs, for many different audiences, from college students, to life-long learners, to professionals seeking job-skill certificates. Comparatively, high schools have less resources which enable them to offer MOOCs, and there has been a dearth of data due to their relatively small deployment.

I address this dearth in my work. Here, I analyze a novel high school MOOC. This MOOC is unique in that it offers a hybrid teaching format where schools can elect to have student *coaches* who guide students through the online material. One offering of the course is training and support for these coaches. Thus, it provides a novel combination of offline interactions and online instruction.

Furthermore, this course prepares students for the Advanced Placement (AP) exam in computer science. This provides a natural post test for students. A post test evaluates long-term learning by assessing students' knowledge shortly after the conclusion of a course, i.e. a post test should determine how much a student learned from a course. That the AP exam is administered by a third party, using material which was not created by a course instructor, ensures that it assesses students not only on course mastery but on mastery of the essential computer science concepts taught by the course.

In this work, I utilize post-test performance to analyze student learning in this high school computer science MOOC. In Chapter 5, I determine attributes of different types of students and use these attributes to build a predictive model of student success. Such a model could be used to determine which students are on track to succeed and which students might need more help. Additionally, in Chapter 5, I inspect the effectiveness of in-person coaching using causal methods.

In Chapter 5, I discover two unexpected groups of students whose course scores do not align with their post-test performance. Unexpected high learners are those students who do not participate much or perform strongly in the course, but who score highly on the post test. Similarly, unexpected low learners achieve unexpectedly low post-test scores. One possible explanation for the occurrence of unexpected low learners is that students share answers amongst each other, and perhaps otherwise game the system. This would explain high course performance which fails to translate into high post-test scores.

Answer sharing is only one form of potential peer interactions, in the coached setting where classmates likely have offline relationships they may also have diverse course interactions. For example, in some cohorts course specific memes have been generated and shared on the student forums. Additionally, students may work together collaboratively such that weaker students learn from stronger students rather than copying their solutions.

In Chapter 6 I explore how peer interactions might influence student learning. To do so I model the unobserved phenomena of *working together*. Students who work together might benefit or suffer from collective knowledge and perform similarly overall. Furthermore, working together might be a form of engagement which elicits increased motivation to learn and succeed. Here, I model the relationship between working together and performance.

Taken together Chapter 5 and Chapter 6 inspect the ability of high school students to learn effectively in an online environment. One question that remains is the effectiveness of coaching. In a setting with only one student per coach, we found some evidence of coaching resulting in stronger performance. However, the number of students analyzed in this case was very small and future work might explore this finding more.

Additionally, coaches have diverse skill levels and backgrounds and further analyzing how these differences impact learning is a potential next step.

Chapter 5

Effectiveness of Online Education

5.1 Introduction

Massive Open Online Courses (MOOCs) have emerged as a powerful mode of instruction, enabling access around the world to high quality education. Particularly for college curricula, MOOCs have become a popular education platform, offering a variety of courses across many disciplines. Now open online education is being deployed to high schools worldwide, exposing students to vast amounts of content, and new methods of learning. Even as the popularity of high school MOOCs increases, their efficacy is debated [82]. One challenge is that the large amount of self direction MOOCs require may be lacking in the average high school student.

To understand the applicability of the MOOC model to high schoolers, we analyze student behavior in a year-long high school MOOC on Advanced Placement (AP) Computer Science. This course is distinguished from traditional college-level MOOCs in several ways. First it is a year-long course, while college MOOCs average 8-10 weeks in duration. This provides ample opportunity to mine student interactions for an extended period of time. Secondly, while traditional MOOCs have no student-instructor interaction, the high school MOOC that we consider incorporates instructor intervention in the form of coaching and online forum instructor responses. Evaluating the effectiveness of this hybrid model allows us to investigate the effect of human instruction on high school students, a group which may particularly benefit from supervision.

Finally, we introduce a post test as a comprehensive assessment occurring after

the termination of the course. A valid post test should assess students' knowledge on critical course concepts, such that students' course mastery is reflected in their post-test score. We treat the Advanced Placement (AP) exam as a post test and consider students' performance on this test as being indicative of long term learning. Previous MOOC research evaluates students on course performance [62]. While course performance can be a good metric for evaluating student learning in the short term, post-test performance is a more informative metric for evaluating long-term mastery.

We propose and address the following research questions, aimed at evaluating the success of MOOCs at the high school level.

1. Can high school students learn from a MOOC, as evidenced here by their post-test (AP exam) performance?

2. How does coaching help students achieve better course performance and learning?

3. How can we predict student's post test performance from course performance, forum data, and learning environment?

5.2 Related Work

Research on online student engagement and learning, is extensive and still growing [65], Anderson et al. [5], and Ramesh et al. [100] develop models for understanding student engagement in online courses. Tucker et al. [134] mine text data in forums and examine their effects on student performance and learning outcomes. Lorenzo and Clayphan [140] analyze the effects of course design and teaching effect on students' pace through online courses. They conclude that both the course design and the mode of teaching influence the way in which students progress through and complete the course. Simon et al. [116] analyze the impact of peer instruction in student learning.

Particularly relevant to our findings is the impact of gaming the system on long-term learning. Baker et al. [13] investigate the effect of students gaming an intelligent tutor system on post-test performance. In the high school MOOC setting, we observe a similar behavior in some students achieving high course performance, but low post-test performance. We identify plausible ways in which these students can be gaming

the system to achieve high course performance and present analysis that is potentially useful for MOOC designers to prevent this behavior.

There is limited work on analyzing student behavior in high school MOOCs. Kurhila and Vihavainen [72] analyze Finnish high school students' behavior in a computer science MOOC to understand whether MOOCs can be used to supplement traditional classroom education. Najafi et al. [85] perform a study on 29 participating students by splitting them into two groups: one group participating only in the MOOC, and another group is a blended-MOOC that has some instructor interactions in addition to the MOOC. The report that students in the blended group showed more persistence in the course, but there was no statistically significant difference between the groups' performance in a post-test. In our work, we focus on empirically analyzing different elements of a high school MOOC that contribute to student learning in an online setting. We use post-test scores to capture student learning in the course and examine the interaction of different modes of course participation with post-test performance. Our analysis reveals course design insights which are helpful to MOOC educators.

5.3 Data

This data is from a two-semester high school Computer Science MOOC, offered by a for-profit education company. The course prepares students for College Board's Advanced Placement Computer Science A exam and is equivalent to a semester long college introductory course on computer science. In this work, we consider data from the 2014-2015 school year for which 5692 students were enrolled.

The course is structured by terms, units, and lessons. Lessons provide instruction on a single topic, and consist of video lectures and activities. The lessons progress in difficulty beginning with printing output in Java, and ending with designing algorithms. Each lesson is accompanied with activities. These activities are not graded, instead students receive credit for attempting them. Students take *assessments* in three forms: assignments, quizzes, and exams, each released every two weeks.

At the end of the year students take an Advanced Placement (AP) exam. Students can use their AP exam performance exam as a substitution for a single introductory college course. The AP exam score ranges from 1 to 5. In all, we have data for

1613 students who take the AP exam. This number is a lower limit on the total number of students who may have taken the course and the AP. The course provides a forum service for students, which is staffed with paid course instructors. Approximately, 30% of all students who created course accounts also created forum accounts, 1728 students in all.

This course is unique in that it provides a coach service which high schools can purchase. This option requires that the school appoint a coach, who is responsible for overseeing the students at their school. The coach is provided with additional offline resources, and has access to a forum exclusive to coaches and course instructors. The average classroom size is approximately 9 students with a standard deviation of approximately 12 students. The largest classroom size coached by a single coach is 72, while some coaches supervise a single student. Of all students who have enrolled in the course, approximately, 23% (1290) are coached and 77% (4402) are independent. From here on we refer to the students enrolled with a coach as *coached students*.

We summarize the class statistics in Figure 5.1 below. The majority of coached students sign up for the student forum, and many persist with the course to take the final AP exam at the end of the year.

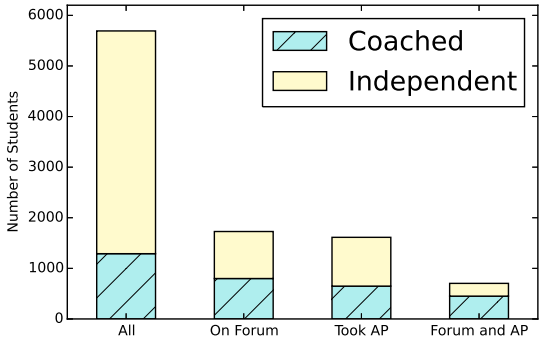


Figure 5.1: Student participation varies between coached and independent students.

5.4 Empirically Characterizing Success of a High-School MOOC

In this section, we use post-test performance and course performance to question the success of MOOCs for high school students. With an empirical analysis, we provide insights on how to adapt high school MOOCs to benefit different groups of students. To investigate this question, we focus on the subset of students for whom we have post-test data. To evaluate student success in the course, we identify three measures of course participation in MOOCs that are relevant to the high school population: *overall score*, *course completion*, and *post-test score*.

Overall Score The overall score captures the combined score across course assignments, quizzes, exams, and activities, each of which contributes to the final score with some weight. We maintain the same weights as those assigned by the course, exams are weighted most heavily, activities the least.

Overall Score = .3*(Assignment Score + Quiz Score) + .6*Exam Score + .1*Activity Score.

Course Completion The second success measure we use is course completion. Course completion measures the total number of course activities and assessments completed by the student.

$$\textit{Course Completion} = \frac{\text{Total Activities and Assessments Attempted}}{\text{Total Number of Activities and Assessments}}$$

Post-Test Score This score captures student scores in the post test that is conducted 2 weeks after the end of the course. The score ranges from 1 to 5. This score captures the advance placement (AP) score, hence we also refer to it as the AP score.

To evaluate the effectiveness of the high school MOOC on student performance, we first examine the relationship between course completion and course performance. We hypothesize that as students complete a higher percentage of the course, they should do better in the course assessments leading to higher course performance scores and post-test scores. Examining the correlation of course completion to post-test performance, we find that they are positively correlated. This suggests that the course indeed helps

students in achieving good performance in the assessments. However, we find that of the students that achieve an overall score of 90 or greater, only 70% pass the post test. Similarly, of the students who complete 90% of the course, only 63% pass the post test. These initial observations indicate the need to perform a more detailed study in order to understand the different student populations in the course.

Next, we examine the relationship between overall score and post-test score, captured in Figure 5.2. From this plot, we see a positive linear relationship between course performance and post-test score. Notably, we observe that the average post-test score of the students who achieve a 90% or higher in the course is above a 4.0, and well above a passing score.

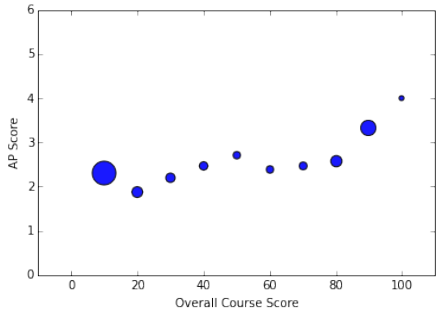


Figure 5.2: The dot sizes are proportional to the number of students achieving the overall score.

Students regularly complete three kinds of assessments: assignments, quizzes, and exams. Assignments are programming exercises, testing students' coding abilities. Programming assignments are submitted online through an interface capable of compiling programs and displaying error messages. Quizzes are multiple choice assessments on course material, with an emphasis on recently covered topics. Exams have a similar format to quizzes but are slightly longer. Both quizzes and exams are timed and students cannot change their answers once they submit them. In all, there are 15 assignments, 8 quizzes and 6 exams in the course. We will refer to them as $A_{1:15}$, $Q_{1:8}$, and $E_{1:6}$, in the discussion below.

In Figure 5.4, we present results of student performance across assessments. Figures 5.4(a), 5.4(b), and 5.4(c) present average student assignment, quiz, and exam scores for students who passed/failed the post test, respectively. We find that students

who pass the post test do better on assessments. We also observe that the scores across all assessments show a decreasing trend as the course progresses. This signals that the assessments get harder for both groups of students as the course progresses. Another important observation is the increase in scores for both groups at assignment 8, quiz 5, and exam 4; these assessments are at the start of the second term in the course, indicating that students may have higher motivation at the start of a term.

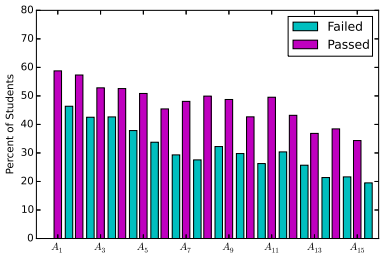
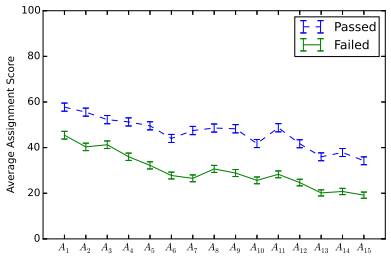
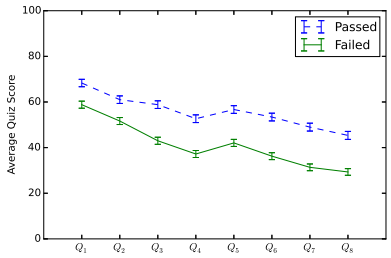


Figure 5.3: Students who pass are more likely to attempt assignments than students who fail.

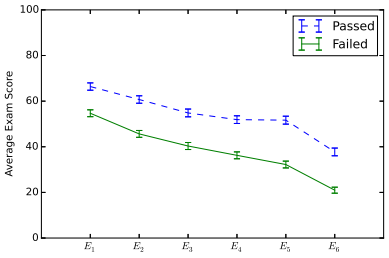
Additionally, some assessments show a greater difference between the two groups of students, and performance on these assessments are more informative of student learning. In Figure 5.4(c), we observe that for both passed and failed students, we see the greatest dip in performance in the final exam. As the final exam is the most comprehensive exam, and possibly most related to the post test, analyzing why students do so poorly on this exam is a worthwhile direction of study in its own right.



(a) Average assignment scores of passed and failed students



(b) Average quiz scores of passed and failed students



(c) Average exam scores of passed and failed students

Figure 5.4: Passed students have higher average scores across all assessments than failed students.

Another important dimension is considering assignment completion rate of these two groups of students. In Figure 5.3, we examine the relationship between attempting assignments and course performance and find that students passing the post test also attempt more assignments. This implies that the high scores of these students are not only the product of strong prior knowledge, but are also the result of learning from the course.

5.5 Forum Participation and Post-Test Performance

In this section, we analyze forum participation of students and examine it's effect on course success. To do so, we answer the following questions:

- Does participation in forums impact post-test performance and learning?
- What are the key differences between participation styles of students who pass the course and students who do not?

We first look at the average score of students who use the forum compared to the average score of students who do not use the forum. Students who use the forum have a statistically higher post test performance score of **2.77**, whereas students who do not use the forum obtain a score of **2.34**, (p ¡ .001). It is not clear if the forum impacts learning, or if instead, students with a high desire to learn are more likely to use the forum.

To accurately evaluate forum participation of the two sub-populations, we analyze them on different types of forum participation. Forum participation comprises of different types of student interactions: asking questions, answering other student questions, viewing posts, and contributing to conversation threads. Table 5.1 gives the comparison of students who pass the post test against student who do not across the various forum participation types. The different types of forum participation types are referred to as: Questions, Answers, Post Views, and Contributions. We also consider the number of days that a student was logged into the forum, which is denoted by Days Online.

On average, students who pass the course make more contributions than students failing in the course. They also answer more questions. Both groups seem to spend roughly the same amount of time online, to view the same number of posts, and to ask the same number of questions. What most distinguishes a student who passes, from one who fails is whether they are answering questions and contributing to conversations.

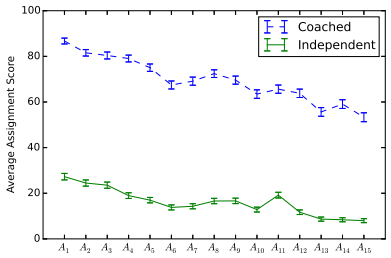
Forum Behavior	Failed Mean	Passed Mean	Failed Median	Passed Median
Questions	3	4	0	1
Answers	1	4	0	0
Post Views	147	140	73	62
Contributions	9	16	1	2
Days Online	19	21	11	13

Table 5.1: The average forum participation is significantly more for students that pass the course. The behavior for which there was a statistical significance difference between the groups are highlighted in bold.

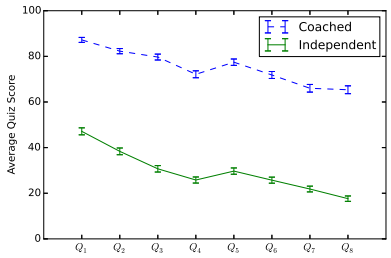
This analysis further demonstrates the importance of forums to MOOCs. Answering questions and contributing to conversations are two behaviors indicative of strong post-test performance. We hope that MOOC designers can use this information to create appropriate intervention and incentive strategies for students.

5.6 Coaching

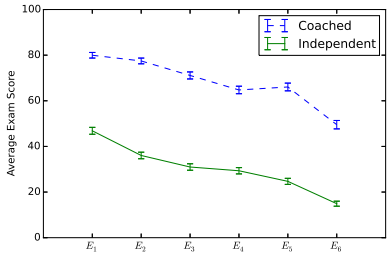
In this section, we evaluate the effect of coaching on student learning. We compare coached students to independent students using their participation in course assessments and forums. We conclude this section by looking at the subset of students who have only one coach, in order to isolate the effect of coaching from other classroom effects.



(a) Average assignment scores of coached and independent students



(b) Average quiz scores of coached and independent students



(c) Average exam scores of coached and independent students

Figure 5.5: Coached students have higher average scores than independent students.

5.6.1 Course Behavior

We inspect the average assessment scores of coached and independent students in Figure 5.5. Observing scores across assignments, quizzes, and exams in Figures 5.5(a), 5.5(b), and 5.5(c), respectively, we find that coached students perform better than independent students across all assessments.

Such differentially high performance in the course should indicate higher performance in the AP exam for coached students. However, we see that coached students fail to get a high post-test score. The average post-test score for a coached student is 2.43, while it is 2.59 for an independent student. We test statistical significance using a t-test with a rejection threshold of $p < 0.05$. In Section 5.6.2, we analyze forum participation of students to understand this difference in scores.

5.6.2 Forum Participation of Coached and Independent Students

Analyzing forum participation of coached and independent students, we find that there is a significant difference in forum participation between coached and independent students. Table 5.2 gives the comparison between coached and independent students in forum participation. On average, coached students ask more questions and answer fewer questions on the forums when compared to independent students. Coached students exhibit more passive behavior by predominantly viewing posts rather than writing posts, when compared to independent students. This can be particularly dangerous if the posts which are viewed contain assignment code.

Forum Behavior	Coached Mean	Independent Mean
Questions	2.81	1.90
Answers	1.45	1.72
Post Views	145.49	81.50
Contributions	8.10	7.33
Days Online	20.64	12.55

Table 5.2: Coached students view more posts and ask more questions. The behavior for which there was a statistical significance difference between the groups are highlighted in bold.

In Table 5.3, we compare coached students who pass to coached students who fail and see the same differences as those observed between all students who pass, and all students who fail. Students who pass are more likely to answer questions, and contribute to conversations.

Forum Behavior	Passed Mean Coached	Failed Mean Coached
Questions	3.97	2.87
Answers	3.04	0.56
Post Views	141.56	164.14
Contributions	14.19	5.93
Days Online	22.71	21.53

Table 5.3: The differences in forum behavior between coached students who pass and who fail follow the same trends in forum behavior exhibited by the general population, and shown in Section 5. The behavioral features for which there was a statistical significance difference between the groups are highlighted in bold.

5.6.3 Coaches with Only One Student

To examine the effect of coaching class size on coached students' post-test performance, we examine coached students in a classroom size of one. Comparing average post test scores of coached students who are singly advised by their coaches (classroom size of one) with independent students, we find that the average post-test score for the coached students is 3.6, while it is 3.2 for independent students. We hypothesize that the lower score of coached students in classroom size greater than one is due to the possibility of sharing answers when students study together. This explains their high overall score but lower post-test scores. This analysis further suggests that the effect of coaching is confounded by the effects of learning in a classroom with peers. To fully understand the effect of a coach guiding a student through the learning process, the peer-effects of classmates should be better understood and isolated. In Section 5.7, we take first steps in this direction by proposing student types.

5.7 Inspecting Unexpected Student Types

In this section, we identify and analyze various types of students in the course based on their performance in the assessments. We classify students into two broad types based

on whether the overall scores and post-test scores are correlated. Figure 5.6 gives the relationship between overall score and post test score for all students. Two groups of students emerge, students who exhibit a correlation between overall scores and post test scores, and students who do not. These two groups can be further broken down based on whether they obtain a high score on the post test, yielding four groups of students.

- *Low learners*: These students have low values for both overall scores and post test scores.
- *High learners*: These student obtain high values for both overall scores and post test scores.
- *Unexpected low learners*: These students obtain high overall scores, but low post test scores.
- *Unexpected high learners*: These students obtain high post test scores, but low overall scores.

Among these, the unexpected low learners and unexpected high learners deviate from the rest of the students. To analyze these two groups, we delve deeper into other aspects of the course such as forum participation and coaching.

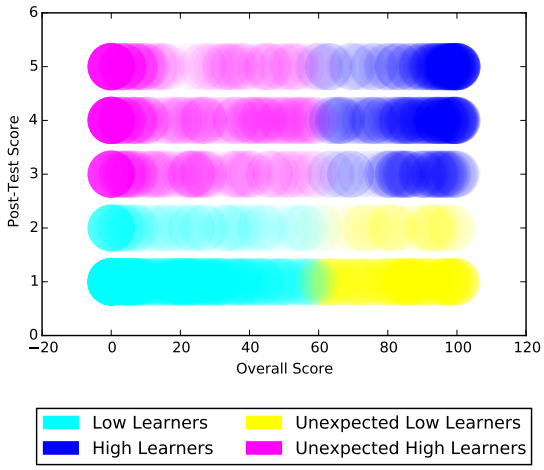


Figure 5.6: Four groups of students emerge: low learners, high learners, unexpected low and high learners. For high course performance we choose a threshold of 60% as a passing grade.

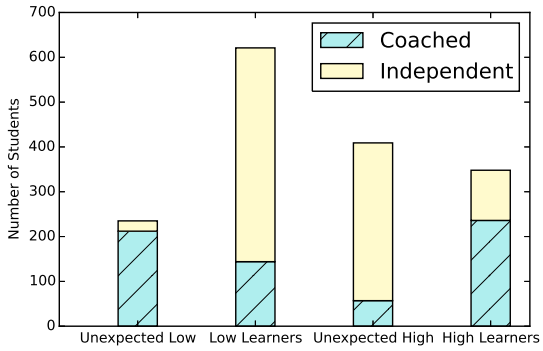


Figure 5.7: The majority of unexpected low learners are coached, while the majority of unexpected high learners are independent.

5.7.1 Unexpected Low Learners

Unexpected low learners are those students who perform well on the course assessments (with an overall score of over 60%) but who do not earn a passing post-test score. We hypothesize that this might be due to their not retaining information from the course, or not arriving at high overall course scores on their own. To understand their low post-test performance, we examine their forum behavior and coaching environment.

As can be seen in Figure 5.7, approximately 91% of unexpected low learners are coached students. Most of these students are part of large classrooms coached by the same coach, increasing the possibility of getting answers from their peers/coach. Plagiarism is a significant challenge in online courses as proctoring students online is not as efficient as in classroom courses.

Further, analyzing forum performance, we find that approximately 76% of unexpected low learners use the forum. Of those who use the forum, 91% are coached. Table 5.4 gives the forum participation of coached and independent unexpected low learners. The forum participation of these students have a strong similarity to failing students in Table 5.1, participating passively in the course by viewing forum posts and contributing to less answers. The coached students are less active than the independent students on the forum in every way, even in post views. While it was posited before that active forum participation is indicative of learning and high AP exam performance, this may not be the case in all groups. For example, the small number of independent

students may be using the forum for social, rather than learning purposes.

Forum Behavior	Coached Mean	Independent Mean
Questions	3.5	9.2
Answers	0.5	15.0
Post views	195.0	293.0
Contributions	7.1	67.0
Days Online	25.6	35.2

Table 5.4: Forum behaviors for which there is a statistical significance between groups are highlighted in bold.

5.7.2 Unexpected High Learners

Unexpected high performers earn an overall course score of less than 60% but pass the AP exam with a 3 or above. Approximately 86% (357 out of 409) of unexpected high leaners are independent and approximately 80% of the unexpected high learners (323 out of 409) are not on the forums. That this group can do so well on the post test, without either a high amount of course or forum participation strongly suggests that either these students have prior knowledge in computer science or that they are not being primarily exposed to computer science through this course but are instead using it to supplement another mode of instruction. A pre test of students' prior computer science knowledge would provide further clarity.

5.8 Predicting Performance from Student Behavior

In Sections 5.4 and 5.5, we see that students' post-test performance is affected by their course and forum behavior. We construct features with which to model these different characteristics of student behavior. These student models are then used to predict post-test scores. By discovering the relative rank of the student model features, we draw insights about student behavior relevant to learning, and to course design.

5.8.1 Student Model Features

We group the course features from student interactions into four broad categories: 1) course behavior, 2) forum behavior, 3) coaching environment, and 4) topic analysis of forum posts. We extract features from student course behavior and forum behavior, which we describe in Sections 5.4 and 5.5. The two other feature categories are described below.

5.8.1.1 Coaching Environment

Students in the online course are either coached or independent. Coaches are provided a separate discussion forum, apart from the student forum, where they can interact with other coaches and instructors of the course. We extract features that capture coaches' prior knowledge and their involvement in guiding students. Table 5.5 gives the list of coaching related features extracted from the discussion forum for coaches.

Feature	Explanation
Coached	Boolean feature capturing whether a student is coached or independent
Coach Views	# posts viewed by the coach
Coach Questions	# questions posted by the coach
Coach Answers	# answers posted by the coach
Coach Contributions	# contributions in the forum

Table 5.5: Coaching related features

5.8.1.2 Posts Topic Distribution

For extracting topics of the post, we explore the topic modeling framework using Latent Dirichlet Allocation (LDA) [17]. Before using LDA we clean the text data by removing stop words, stemming certain words, and removing all common course words, such as code. To obtain the topic distribution of posts, we use the Machine Learning for Language Toolkit (MALLET) [80]. We use the following parameters for the topic model: number of topics = 150, and optimize-interval = 100, where the hyper-parameters required by LDA, α and β, are set to the default values.

5.8.2 Predictive Model

We incorporate extracted features in a linear kernel Support Vector Machines (SVM) model, using the python package Scikit-learn [93]. Comparing this model with other machine learning algorithms such as logistic regression, decision trees, and Naive Bayes we found the results to be comparable. We filter our student pool to those who participated in the forums and took the post test (approximately 16% of all students who completed the post test). A subset of features that are predictive of post-test performance were selected using recursive feature elimination in Scikit-learn [93]. Recursive feature elimination works by training a classifier which weighs features and then trims all features with the lowest weights; this trimming allowed us to obtain the best predictions, and to understand which features are most predictive of student success.

5.8.3 Empirical Results

In this section, we present empirical results using the SVM model defined above to predict post-test performance. To evaluate the effectiveness of this model we compute the F-measure, which is the harmonic mean of precision and recall. F-measure is an optimal metric for a setting with unbalanced classes such as ours, where accuracy may appear to be deceptively high if a classifier reliably predicts the majority class. Our model gives an F-measure of 0.81 for predicting post-test performance. We validate our results with 10-fold cross validation. In the next sections, we analyze the attributes of student behavior which are most predictive of performance.

5.8.3.1 Topics and Performance

The topics discovered by the topic model fall into four broad categories: help requests, assignments, course material, and course activities. In Table 5.6, we present the ten topics which are most predictive of post-test performance. The first three topics in the table fall into the help requests category. They include words such as *trouble*, *help*, and *fail*. Four of the top ten topics correspond to assignments, with top words which are descriptive of assignments from the course. For example, in assignment A_4 students are asked to write a program to count the number of hashtags, links, and attributions in a tweet, and in the topic associated with this assignment we see the words: *hashtag*,

tweet, *attributions*, *mentions*, and *links*. Two topics represent the concepts discussed in the course: object oriented programming, and hash maps. The hash maps topic is particularly interesting as hash maps are not introduced in the course, but students still use them in their projects, and discuss them on the forum. The other prominent topics are topics related to course activities. For example, the activity topic in the the table is an activity given to students to print the location of a vehicle. This is the most elaborate activity that students undertake in the course, hence it appears in the top predictive topics for predicting post-test performance.

Topic Label	Top Words
Help requests	trouble, don't, perfectly, won, updated
Help requests	hope, helps, change, find
Help requests	fail, expected, updated, supposed
Assignment content (A_4)	hashtag, tweet, attributions, mentions, links
Lecture (hashmaps)	Map, key, Getvalue, Hashmap, entry
Course Activity	vehicle, location, backward, forward, GetLocation
Assignment content (A_6)	ArrayList, words, remove, equals, size
Assignment content (A_{10})	strand, size, TurnOn, green, BurntOut
Assignment content (A_{14})	sort, insertion, swap, insert, algorithm
Lecture (OOP and Methods)	object, constructor, methods, parameter, returns

Table 5.6: Top predictive topics and the words in these topics

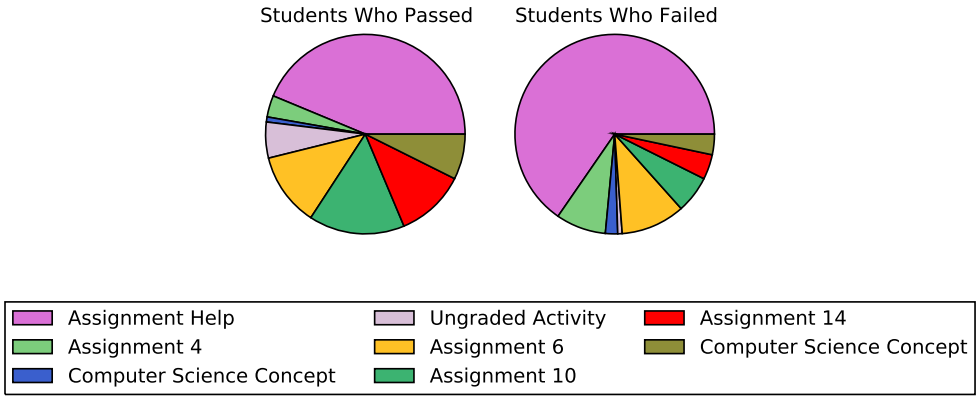


Figure 5.8: Students who pass post about different topics than students who fail.

Figure 5.8 gives the distribution of passed and failed students across the different ten most predictive topics given in Table 5.6. We observe that passing students post about the course activity on vehicles more than failing students. Since activities only contribute to a small portion of their grade, participation in activities is a good measure for students' level of motivation and learning.

Additionally, we observe that failing students are far more likely to write posts which fall in the help category. Looking at some of the posts in this category, we find that these posts are often short and use help words, but do not contain detailed information about the specific assignment problem in question. This finding suggests that analyzing the posts for linguistic cues is helpful in understanding students' motivation.

The third important take away from this analysis is that this topic distribution can help discover patterns in student behavior. For example, passing students post about assignment A_{10} more than failing students. But, failing students post more about assignment A_4. As assignments tend to get harder as the course progresses, the difference in behavior can be attributed to failing students needing help on the easier assignments, while the savvier students focus on the harder assignments.

5.8.3.2 Critical Assessments

Here, we describe the most predictive assignments, quizzes and exams that we use in the predictive model. We find that assignments A_4, A_8, A_9, and A_{10} are the most predictive assignments. These assignments are on core concepts and hence may be the most critical assignments in the course. This observation is bolstered by the fact that these assignments are referenced in the forums more than other assignments. Two of these assignments feature in the top ten predictive topics given in Table 5.6. Pinpointing the moment when a student needs help is not only predictive of their success, but also critical in maintaining engagement and understanding. Understanding which assignments are discussed more in the forums can reveal important information for initiating instructor interventions.

5.9 Conclusion

From this analysis we conclude that MOOCs are a viable option for high school students. Forty-seven percent of students who took the post test passed it. Four hundred and sixty four of these students were to the best of our knowledge self-directed. While we can say that MOOCs work for some high school students, the particularities of this group must be understood. It is not clear, for example, how the students who achieve high course scores, but low AP exam scores are able to do so. Are they receiving answers from other students, or have they truly mastered the course content, but lack the ability to demonstrate this mastery on a test? High school MOOC students are a unique group with particular modeling demands.

We have developed models of these students, characterizing high and low learners by their course and forum behavior, as well as by the topics that they post about. These models have allowed us to differentiate the behavior of students who pass from that of students who fail. In this case study post-test performance was correlated with course-performance, such that students who earned a high course score also earned a high post-test score. Students who performed well on the post test were more likely to contribute to conversations, and to answer questions on the student forum. They were also more likely to post about ungraded activities, and less likely to write posts asking for help. Coached students were more likely to perform well in the course, and spent more time on the forum. Understanding the differences between students who excel and those who do not is crucial in developing the courses that students, and particularly high school students need.

In the next chapter, continue our analysis of this high school MOOC. In Chapter 6 we introduce latent variables to explicitly model student types in predicting course performance. Additionally, we incorporate social factors in predicting performance, such as classroom membership and collaborative activity.

Chapter 6

Peer interactions and learning

6.1 Introduction

In online environments, social behavior has been shown to affect attrition [106], support knowledge construction [60, 9] and improve course enjoyment [76]. Forum interactions have proven invaluable as a source of information on how students form online relationships, share knowledge and engage interactively. Critically, forum engagement has shown to be indicative of course success [132].

However, not all interactions are automatically beneficial to student success. For example, Rosé et al. [105] report the negative effects of antisocial behavior. Furthermore, when working together to learn course content, not all student groups should expect the same benefits from collective efforts [118].

An important aspect of collective knowledge construction is the organizational structure of the group. However, there are many open questions into how various group structures might interact with student types and overall learning. For example, Wen et al. [143] found that when organized in groups, students with strong group leaders performed best and other group characteristics can impact both individual and collective performance [78]. Additionally, group dynamics, such as degree centrality and competitiveness, can impact course performance [118].

We inspect the same MOOC as in Chapter 5. A striking characteristic of this MOOC is that it has a live-instruction component. Some students enroll in the course through their high school and have access to a instructor referred to as a coach. Each

coach created different classroom environments and aspects of these environments were correlated to performance. Here, we use classroom membership to to collectively predict individual performance and infer latent classroom strength. This allows us to model some unobserved aspects of group learning, where students in the same classroom might benefit from unseen collaboration and collective knowledge construction.

Unexpectedly, in Chapter 5, we discovered two groups of surprising students whose course performance did not align with their post-test performance. One potential problem in an online environment is that students can freely share answers with each other, where each can submit the same answer to the online system. Many different dynamics can arise from this simple scenario each with its own implications for learning. For example, if one strong student is solving the problem on their own, and sharing the answer with their peers, these answer receivers may miss learning opportunities. Alternatively, if the students are working cooperatively together, their joint answer might be stronger and they may benefit from the group effort. To analyze these different scenarios, we infer latent co-working ties which reflect the extent to which two students are working together on a given assignment. We then use coworking dynamics in predicting student strength. This collective performance model not only allows us to discover and exploit collaboration dynamics, but improves upon the model predictions presented in Chapter 5.

6.2 Related Work

Understanding the social processes of student learning is an active area of research [6]. In online courses, student forums are essential in lending insight to how peer behavior impacts motivation, engagement and other performance metrics [106, 60, 9, 47]. Other work has inspected the impacts of physical colocation [76, 18] and social search [123] on learning and the meaningfulness of social networks outside of course forums [138]. As social interactions are found to have large positive impacts on student experiences much work has focused on building tools to improve collaborative learning. Rosé et al. have proposed supportive technologies to enhance forum participation and course collaboration [105, 71]. As well as on automatically analyzing both online networks and forum corpora [104].

Additionally, some studies have focused on the effects of peer interactions on learning. Wang et al. [142], investigated the relationships between types of MOOC forum behavior and learning, and found interactions to be indicative of learning gains in some cases. For example, the extent to which a student actively recounts course material on the forum correlates to post-test performance. Wen et al. [143] analyzed various team aspects in an online course to discover which correlated with performance. Their work demonstrates the potential positive benefits of learning teams, especially when team leaders are active in team building.

In our work, we infer unobserved offline collaboration through online interactions. These inferred offline interactions are then used in modeling the influence of collaboration on performance. In addition, we model latent student strength. Thus, we can model how interactions between students of varying strength levels can impact performance.

6.3 Probabilistic Performance Prediction

Here, we extend the prediction model in Chapter 5. In our new model, INTERACTIONS, in addition to predicting if a student will pass the post test or not, we would also like to predict their latent strength and collaboration partnerships. We propose several potential collaboration dynamics. As each of these might be appropriate for this setting, we learn weights for each potential dynamic. For coached students, we also predict the strength of the section in which they are enrolled. We construct our models with PSL.

To illustrate PSL in the online course context, consider a rule which says that if two students exhibit similar course performance and one passes the post test, the other will be likely to as well. To express this rule we introduce the predicate SIMILARCOURSEPERFORMANCE, which takes two students as arguments and which expresses their course performance similarity as a value between 0 and 1. Additionally, we introduce the predicate PASS, which takes a student as an argument and whose truth value indicates whether this student passes the post test. With these predicates, we define our rule in PSL as follows:

$$w_{sim} : \textsc{SimilarCoursePerformance}(S_a, S_b) \wedge \textsc{Pass}(S_a) \Rightarrow \textsc{Pass}(S_b).$$

Next we demonstrate how PSL can be used to template a predictive model for post-test performance. For each student, S_i, we would like to predict if this student will pass or not, that is we would like to infer the truth value of $\textsc{Pass}(S_i)$. We would also like to assign each student a strength value, such that strong students are likely to pass. To do so, we introduce $\textsc{Type}(S_i, T)$, where T is either *Strong* or *Weak*.

As slightly more students fail than pass, we include a prior with small weight which indicates that most students will not pass (shown as first rule in Table 6.1). To express that a student's strengths are related, that is a strong student cannot simultaneously be a weak student, we impose a functional constraint. This constraint ensures that the truth values of all potential types for a given student sums to 1 (shown as final rule in Table 6.1).

$$\begin{aligned} w_{neg} &: \neg \textsc{Pass}(S) \\ \infty &: \sum_{t \in Types} \textsc{Type}(S, t) = 1 \end{aligned}$$

Table 6.1: Priors and Constraints

We make use of the predictions from the model explained in Chapter 5 to infer students' latent strengths. In the first two rules in Table 6.2, we learn weights to express the dependence between the SVM predictions and students' types. We also use the topics of the posts each student contributes to the student forum. For each topic which is considered by feature selection to be predictive, we learn a weight for the rule which states that mentioning this topic is indicative of a student being *Strong*. This is shown with the rule template shown in the third line of Table 6.2, where the value of $\textsc{PostTopic}(P, t_k)$ is the topic distribution assignment of topic t_k and $\textsc{Author}(S_i, P)$ is 1 if S_i authored post P. Similarly, we would like to express the relationship between forum contributions and student strength. To do so, we introduce the predicates $\textsc{Answers}(S_i, A)$, $\textsc{Contributions}(S_i, A)$, $\textsc{DaysOnline}(S_i, A)$ and $\textsc{Views}(S_i, A)$, where A can be either *low* or *high*. We then express the relationships between these behaviors and strength with the remaining rules in Table 6.2.

$$
\begin{aligned}
w_{svms} &: \text{SVMPredicts}(S) \Rightarrow \text{Type}(S, \textit{Strong}) \\
w_{svmw} &: \text{SVMPredicts}(S) \Rightarrow \text{Type}(S, \textit{Weak}) \\
w_{t_j} &: \text{PostTopic}(P, t_k) \wedge \text{Author}(S, P) \Rightarrow \text{Type}(S, \textit{Strong}) \\
w_{a_h} &: \text{Answers}(S, \textit{high}) \Rightarrow \text{Type}(S, \textit{Strong}) \\
w_{a_l} &: \text{Answers}(S, \textit{low}) \Rightarrow \text{Type}(S, \textit{Weak}) \\
w_{v_h} &: \text{Views}(S, \textit{low}) \Rightarrow \text{Type}(S, \textit{Strong})
\end{aligned}
$$

Table 6.2: Inferring Student Strength

We are also interested in how membership in particular classrooms might impact performance. To express the strengths of various learning environments, we introduce $\text{SectionStrength}(C, T)$, where C refers to a class or section ID and T is either *Strong* or *Weak*. As for students, we constrain sections' types with the final rule in Table 6.3, which expresses that a section cannot be simultaneously *Strong* and *Weak*.

$$
\begin{aligned}
w_{nwt} &: \neg\text{WorkingTogether}(S_i, S_j) \\
\infty &: \sum_{t \in Types} \text{SectionStrength}(C, t) = 1
\end{aligned}
$$

Table 6.3: Collaborative Priors and Constraints

To infer section strength, we use features discovered in Chapter 5. In addition to the student forum, the course includes a forum where coaches can interact with each other and with course instructors. To model the behavior of coaches on this forum, we introduce the predicates $\text{CoachAnswers}(C, A)$ and $\text{CoachOnline}(C, A)$, where A can be either *low* or *high*. $CoachAnswers$ refers to the number of questions a coach answers on the coach forum and CoachOnline refers to the number of days the coach logs into the coach forum. Additionally, we consider the number of students in a classroom with $\text{CoachClassSize}(C, A)$. The rules in Table 6.5 relate coach forum behavior to section strength.

For each rule in Table 6.2, we also have a rule which propagates these student behaviors into section strengths. These rules express that if a student is predicted to be strong and is enrolled in a given section, then that section should also be strong. We make this clear with the following template:

$\textsc{Behavior}(S) \wedge \textsc{BehaviorCorrelated}(T) \wedge \textsc{InSection}(S, C) \Rightarrow \textsc{SectionStrength}(C, T)$

where an example behavior correlated with strength is answering questions on the student forum. Thus, each rule in Table 6.2 which infers student strength, has a version following this template which instead infers section strength.

Additionally, we observed that coached and students exhibit different behavior on the forum. Moreover, these behaviors are differently correlated with success. For example, when viewing coached students view a high number of posts this behavior is associated with low performance. However, when independent students view a high number of posts it is correlated with high performance. In order to capture these differences, we introduce two predicates $\textsc{Coached}(S)$ and $\textsc{Independent}(S)$ where each return binary values according to whether a student, S, is coached or independent.

$w_{ec_x} : \textsc{AssessmentScore}(S, A_x) \wedge \textsc{Coached}(S) \Rightarrow \textsc{Type}(S, \textit{Strong})$
$w_{ei_x} : \textsc{AssessmentScore}(S, A_x) \wedge \textsc{Independent}(S) \Rightarrow \textsc{Type}(S, \textit{Strong})$
$w_{v_h c} : \textsc{Views}(S, \textit{high}) \wedge \textsc{Coached}(S) \Rightarrow \textsc{Type}(S, \textit{Weak})$
$w_{v_h i} : \textsc{Views}(S, \textit{high}) \wedge \textsc{Independent}(S) \Rightarrow \textsc{Type}(S, \textit{Strong})$
$w_{d_i} : \textsc{DaysOnline}(S, \textit{high}) \wedge \textsc{Independent}(S) \Rightarrow \textsc{Type}(S, \textit{Strong})$
$w_{ct_c} : \textsc{Contributions}(S, \textit{high}) \wedge \textsc{Coached}(S) \Rightarrow \textsc{Type}(S, \textit{Strong})$
$w_{ct_i} : \textsc{Contributions}(S, \textit{high}) \wedge \textsc{Independent}(S) \Rightarrow \textsc{Type}(S, \textit{Weak})$

Table 6.4: Student Types, Behavior and Performance

Furthermore, we introduce these type-specific rules in Table 6.4. In the first rule, we learn the association between student strength and performing well on course assessments. The $\textsc{AssessmentScore}(S, A_x)$ of student S on exam A_x is their true score scaled to be between 0 and 1. In the following rules, we model the relationship between types of forum behavior and student strength, when these relationships differ by student instruction type. As in Table 6.2, the rules which concern coached students also have a duplicate for inferring section strength, following the template above.

Next, we would like to uncover collaborative behavior and utilize it to improve predictions. Here, we are interested in one form of collaborative behavior, when students work closely together on the same assignment. We model this behavior with the

predicate WorkingTogether(S_i, S_j), which takes two students as arguments and expresses the extent to which these students might be inferred to be working together. As we do not know if two students are working together or not, we model working together behavior as a latent variable and infer the values to WorkingTogether(S_i, S_j) for all potential S_i, S_j pairs. We express the prior belief that most pairs of students do not work together with the first rule in Table 6.3.

w_{cah} : CoachAnswers$(C, high) \Rightarrow$ SectionStrength$(C, Strong)$
w_{cao} : CoachOnline$(C, low) \Rightarrow$ SectionStrength$(C, Weak)$
w_{ccs} : CoachClassSize$(C, low) \Rightarrow$ SectionStrength$(C, Strong)$

Table 6.5: Inferring Section Strength

We then use section strength to inform the predictions of students' strengths. This is shown in Table 6.6. This rule expresses that if a student is enrolled in a strong section, they may also be a strong student. The next rule in this table expresses that students in the same section will have similar strengths. The final rule restates this, but modulates this dynamic by section strength.

w_{sec_s} : SectionStrength$(C, Strong) \wedge$ InSection$(S, C) \Rightarrow$ Type$(S, Strong)$
w_{col} : SameSection$(S_i, S_j) \wedge$ Type$(S_i, T) \Rightarrow$ Type(S_j, T)
w_{col_s} : InSection$(S_i, C) \wedge$ InSection$(S_j, C) \wedge$ SectionStrength$(C, T) \wedge$ Type(S_i, T) $\Rightarrow$ Type(S_j, T)

Table 6.6: Section Strength and Student Strength

w_{wt} : Author$(S_i, P_1) \wedge$ Author$(S_j, P_2) \wedge$ SimPost$(P_1, P_2) \Rightarrow$ WorkingTogether(S_i, S_j)
w_{wta} : WorkingTogether$(S_j, S_i) \Rightarrow$ WorkingTogether(S_i, S_j)

Table 6.7: Working Together

To infer that students may be working together, we inspect forum post content. We define a post similarity function, PostSim(P_1, P_2), which takes two posts as arguments and returns their similarity as a value between 0 and 1, where 1 indicates

equality. If two students have similar posts, we predict that they are working together, as shown in the first rule in Table 6.7.

$$
\begin{aligned}
&w_{cs} : \textsc{InSection}(S_i, C) \wedge \textsc{InSection}(S_j, C) \wedge \textsc{SectionStrength}(C, \textit{Strong}) \\
&\qquad \wedge \textsc{WorkingTogether}(S_i, S_j) \wedge \textsc{Type}(S_i, \textit{Strong}) \Rightarrow \textsc{Type}(S_j, \textit{Strong}) \\
&w_{cw} : \textsc{InSection}(S_i, C) \wedge \textsc{InSection}(S_j, C) \wedge \textsc{SectionStrength}(C, \textit{Weak}) \\
&\qquad \wedge \textsc{WorkingTogether}(S_i, S_j) \wedge \textsc{Type}(S_i, \textit{Weak}) \Rightarrow \textsc{Type}(S_j, \textit{Weak}) \\
&w_{csw} : \textsc{InSection}(S_i, C) \wedge \textsc{InSection}(S_j, C) \wedge \textsc{SectionStrength}(C, \textit{Weak}) \\
&\qquad \wedge \textsc{WorkingTogether}(S_i, S_j) \wedge \textsc{Type}(S_i, \textit{Strong}) \Rightarrow \textsc{Type}(S_j, \textit{Weak})
\end{aligned}
$$

Table 6.8: Collaboration and Student Types

Now, we express how working together might relate to student and section strength in the rules in Table 6.8. We model that a student is in a given section, C, with $\textsc{InSection}(S_i, C)$. The first rule in Table 6.8 expresses that if two students are in a strong section and they are working together, if one of them is a strong student, the other is likely to be as well. Similarly, if two students are in the same weak section and they are working together, if one is weak, the other one likely is as well. The last rule in Table 6.8 expresses a different dynamic. Here, if two students are in a weak section and one of them is strong, they may be working with a weaker student. This rule captures the potential dynamic of answer sharing. Coaches of strong sections might be better able to prevent such behavior; consequently we model this dynamic occurring in weaker sections.

Finally, we predict whether a student will pass with the two rules in Table 6.9, which connect a students' strength to passing. Together, the rules presented in this section make up the collective probabilistic model Interactions. Next, we inspect the results of this model and compare it to the SVM model in Chapter 5. Additionally, we analyze the discovered collaborations.

$$
\begin{aligned}
&w_{sp} : \textsc{Type}(S_i, \textit{Strong}) \Rightarrow \textsc{Pass}(S_i) \\
&w_{wp} : \textsc{Type}(S_i, \textit{Weak}) \Rightarrow \neg\textsc{Pass}(S_i)
\end{aligned}
$$

Table 6.9: Predicting Performance

6.4 Empirical Evaluation

Here, we evaluate our model's ability to predict students' post-test performance. We compare a Support Vector Machine (SVM) to our collaborative PSL model. For each model we perform 3-fold cross validation.

There are 196 potential features for each model. Here, we use Recursive Feature Elimination (RFE) to select the most useful features. Twenty features were found to be the best, and the SVM was trained on these twenty features. We additionally performed feature standardization before the RFE procedure.

We show the predictive results in Table 6.10. We see that the PSL model outperforms the SVM overall, and for each student group as well. Both models are better able to predict the performance of coached than independent students.

	Precision	Recall	F-Measure
	All		
SVM	77.3	75.9	76.6
INTERACTIONS	**82.0**	**82.4**	**82.2**
	Coached		
SVM	72.6	74.2	73.4
INTERACTIONS	**78.4**	**80.5**	**79.5**
	Independent		
SVM	86.0	79.0	82.5
INTERACTIONS	**88.6**	**85.7**	**87.1**

Table 6.10: Predicting post-test performance. In this table, statistically significant improvements are shown in bold.

6.4.1 Inferring Class Strength

In addition to inferring if students will pass the post test, we also infer the latent strength of each class or section with a coach. This inference allows us to utilize section strength in predicting students' performance and in modeling collaboration dynamics. While true section strength is an unobserved variable, here we introduce a proxy score to evaluate our model.

To do so, we introduce two proxy scores and evaluation metrics. For each class we calculate the percentage of students who pass the post test. We then calculate both binary classification metrics and error metrics. In the binary case, we treat each section with a pass rate greater than 50% as a strong section, and round our predicted strength values with a .5 threshold. Otherwise, we compare the raw percentages to the predicted strength values and calculate the Mean Square Error (MSE). These results are shown in Table 6.11.

	Precision	Recall	F-Measure	MSE
Section Strength Classification	67.9	1.00	80.9	.280

Table 6.11: Inferring class strength.

In Table 6.11, we see that we are reasonably able to infer section strength. The F-Measure is quite high at 80.9. However, recall is much higher than precision, and precision can be further optimized. Additionally, we see that the MSE is fairly low.

A section's strength is most likely a combination of the abilities and achievements of its constituent students and of the coach leading the section. Coaches can influence students' learning and determining what aspects of coaching are most effective can have a large impact on learning. Here, we model several potential indicators of a coach's strength and show that these offer reasonable performance. While these indicators might help course designers in choosing coaches to approach with extra resources, they are not informative enough to lend insight into successful classroom strategies. For example, we do not know how often strong instructors decided to meet with their students, nor do we know their prior knowledge.

6.4.2 Collaboration Dynamics

We also modeled working together on specific assignments as one form of collaboration dynamic. As we saw in Chapter 5, there was a group of coached students who achieved high course scores, yet failed to pass the AP exam. We posited that they may have received answers from successful peers, and thus been able to pass the course without sufficiently learning the material to pass the post-test. This behavior can be summarized with the term *answer sharing*. Additionally, we know that students benefit from

collaboration, and we might see successful students working together.

In inferring these collaboration dynamics, we introduced a post similarity function. Given two messages, the first criterion is that these messages have high text similarity. Text similarity is calculated as the cosine similarity between term frequency-inverse document frequency (TF-IDF) vectors. In constructing the TF-IDF vectors we consider the entire corpus. Additionally, we only consider interactions between students in the same coached section.

Next, we inspect which kinds of students work together. To do so, we consider all students who pass the AP as *Strong* students, and all students who do not pass as *Weak* students. In total, we find 19 pairs of students out of 119 potential pairs. Of these, the largest interactions are between weak students.

	Weak	*Strong*
Weak	9	5
Strong	5	5

Table 6.12: Number of interactions of each type. We do not distinguish between weak-strong and strong-weak.

The absolute numbers here are very small. We do see that there is potential evidence of *answer sharing*, in the strong-weak interactions. Of the weak students in these interactions, they are all unexpected low learners with high course scores and low post-test scores. Furthermore, of the weak-weak interactions the majority are also between unexpected low learners. However, we also see interactions between strong students.

6.5 Discussion

In this chapter we investigated whether we might improve predictions by modeling collaboration dynamics. Indeed, this was the case. We saw improvements in F-Measure over the SVM model of 7.4%, 8.3% and 5.7% overall, for coached students and for independent students, respectively. As the collaborative model can take advantage of

classroom information and interactions between coached students, we expect to see the largest increase with regards to coached students.

In this collaborative model, we are also able to analyze section strength and pairs of students inferred to be working on assignments together. By doing so, we see that we are reasonably able to predict section strength, and that this is useful in predicting student's performance as well. Additionally, we saw that the most common coworking interaction was between weak students and that the majority of all interactions involved one unexpected low learner. While collaborations can be beneficial for learning, high school students may need more guidance in forming successful teams. However, we only uncovered a small total number of such interactions. To better measure the effect of collaboration, it would be useful to observe changes in students' strength after working together. For example, do we see stronger students aiding weaker students to the extent that they can also pass the post-test? Our inferred coworking pairs are a static snapshot of evolving relationships. In future work we will take temporal aspects into account when studying these effects.

6.6 Conclusion

By modeling collaborative dynamics and inferring target values collectively, we were better able to predict students' exam performance. We also introduce latent variables to uncover hidden structure in this domain. By inferring latent section strength, we could better predict the performance of coached students and by inferring pairs of students working on assignments together, we could better address questions left open in Chapter 5.

In the next two chapters we continue this work of utilizing inherent structure in collectively inferring target variables. We also introduce latent-variables to uncover hidden phenomena in domains concerning malicious behavior. For example, in detecting cyberbullying behavior we infer latent relational ties, and while predicting the future movements of human traffickers we discover latent links between locations. Next, we explore this work in more detail.

Part IV

Malicious Behavior

Online activity captures traces of human behavior. As people form, strengthen and shatter relational ties on social media their publicly posted comments, likes, and friendship links provide rich sources of information into how they behave both online, and to a certain extent, offline. Online market places reveal what people seek and purchase, sometimes for offline purposes. In Part II I demonstrated how online shopping data can be used to discover sustainable products and customers. In Part III I analyzed online course data to draw conclusions about learning. This part is concerned with how online data can be used to detect, and model malicious behavior.

Malicious behavior is in some ways emboldened in online environments. Social media and online forums can provide a veil of anonymity where posters assume pseudonyms and can post without fear of offline repercussions. Furthermore, public social media, such as Twitter, can provide bullies with large audience. As youth are reported to perceive anonymous attacks as worse than non-anonymous attacks, and public attacks as worse than private ones [121], online environments may worsen circumstances for victims. Similarly, criminal activity can be facilitated by online markets where new customers can be reached more easily via electronic methods [50, 29]. In regards to sexual services, Cooper outlines how the internet increases access, affordability and anonymity [24].

In these settings, online data can provide a wealth of information about otherwise difficult to track activities. However, such data, while plentiful, can be of poor quality. When subjected to standard preprocessing, online messages can stripped to a small number of words. In malicious domains, criminals may attempt to obfuscate critical identifying information, leaving slight amounts of feasibly extractable pieces of information.

Modeling the structure inherent in these domains can alleviate issues arising from poor data quality. My work centers around this goal. In Chapter 7 I introduce a socio-linguistic approach for detecting cyberbullying. This method makes use of social and linguistic structure to detect cyberbullying, and uses a number of latent representations to categorize types of attacks, assign roles to participants and discover relational ties. In Chapter 8 I introduce a spatio-temporal model for predicting the future movements of human traffickers. Similar to the cyberbullying work, I also introduce latent

variables to better model the problem, and to provide some interpretation to the results. Here, I introduce latent route-segments and links between locations which can describe multi-step movements.

Each chapter addresses questions into human behavior through online data. In Chapter 7, I inspect how social status relates to participants' roles. For example, are people with low-status more likely to be victims or bullies? Additionally, I explore the different ways that aggressors form attacks. I also inspect gender differences among participant roles.

Chapter 8 evaluates how environment events impact human trafficking. As climate change escalates the impact of environmental stressors on human activity is theorized to intensify [101, 3]. I propose three potential impacts of catastrophic events on human trafficking, and explore each through online data.

Together, this work demonstrates the benefits of incorporating structure in malicious domains, where online data can be sparse and of low quality. Additionally, including latent variables can provide many benefits. These variables can uncover hidden structure and which can be used to improve predictive performance. They can also be used to explore unobserved phenomena. For example, in the cyberbullying case I analyze different types of attacks and the connection between relational ties and participant roles.

A persistent open question in this work is to what extent online activity represents offline psychological states, behavior and relationships. For example, in the case of cyberbullying an open question is how representative online interactions are of offline relationships. In the case of human trafficking, a question is the accuracy of the information in online ads. For example, ads often contain mentions of the provider's ethnicity, though this might be not be the true ethnicity of the victim, but rather an ethnicity that the poster expects to create a strong response rate. Future work in this area might further explore the relationship between online and offline behavior.

Chapter 7

A Socio-linguistic Approach for Cyberbullying Detection

7.1 Introduction

Bullying has long presented physical, emotional and psychological risks to children, youth and adults. As such, there is an extensive body of knowledge aimed at understanding and preventing bullying. Far less is known about cyberbullying, the newest form of interpersonal aggression. Cyberbullying occurs in an electronic environment [77], from online forums to social media platforms such as Twitter and Facebook. As it can occur at any time or location, cyberbullying poses new psychological risks, while also influencing physical well being [77, 56, 59]. It also introduces new questions of governance and enforcement, as it is less clear in an online environment who can and should police harmful behavior.

A necessary first step in understanding and preventing cyberbullying is detecting it, and here our goal is to automatically flag potentially harmful *social media* messages. These messages introduce unique challenges for natural language processing (NLP) techniques.

As they are unusually short and rife with misspellings and slang, when treated with traditional text pre-processing, these messages can be stripped to only one or two words. This sparsity makes cyberbullying messages especially ill-suited for methods which depend on sufficiently large training corpora to generalize well. Solutions which

can augment poor textual data with domain knowledge or social data might outperform those which rely on text alone.

Not only is labeled data costly, but it can be error-prone as annotators are generally third parties who are not directly involved with the incidents of cyberbullying. Thus their labeling is subjective, and even labels with high inter-annotator agreement may be incorrect. Rather than throwing out annotations with low inter-annotator agreement, we propose a series of probabilistic models which can directly incorporate uncertainty. Furthermore, we show that modeling uncertainty in the training data can improve the performance of all models, demonstrating that a probabilistic approach is well suited for this domain.

We develop a series of probabilistic models of increasing sophistication and implement these models in PSL. Our first model makes use of text, sentiment and collective reasoning. Next, we incorporate seed-words and latent representations of text categories. Finally, we make use of social information by inferring relational ties and social roles.

Our models are evaluated on a dataset of youth interactions on the social media platform Twitter. Twitter has emerged as a fertile environment for bullying [15]. Twitter's ability to provide a veil of anonymity can facilitate bullying [34, 126]. Whittaker and Kowalski [144] found that though fewer survey participants used Twitter (69.4%) compared to Facebook (86.5%), a higher percentage of participants experienced cyberbullying on Twitter (45.5%) than Facebook (38.6%) (and other platforms). We compare our models to a baseline N-Grams model. This model is comparable to standard bag-of-n-grams approaches. Additionally, we compare to an implementation of a state-of-the-art approach.

Our contributions include strategies for learning from uncertain annotations and linguistic models which demonstrate the utility of domain knowledge and collective reasoning. In addition, we introduce two novel latent-variable models which categorize attacks and identify user roles by exploiting inferred relational ties. These models are advantageous in that they combat the sparsity of textual data and provide interpretation of latent phenomena, such as relational power dynamics, in cyberbullying.

7.2 Related Work

Hodas et. al. analyzed network dynamics of Twitter users to address questions of friendship [45] and found that Twitter dynamics reaffirm well-studied principles of friendship. Kim et al. [64] studied 2193 pairs of users and found that their online interactions were impacted by offline friendships.

Current methods for detecting cyberbullying have largely relied on linguistic classifiers [22, 146, 103, 136, 145]. Using TF-IDF features and contextual features, Yin et al. [146] predict bullying with high accuracy. Chen et al. [22] develop a novel framework which incorporates unique style features and structure. Zhao et al. [147] also make use of word embeddings.

Similar to our work, Raisi and Huang [99] use a small seed vocabulary to indicate bullying events. They leverage participant roles to expand the set of candidate bullying terms and better predict bullying. Their approach corroborates the idea that seed indicators can be successful in this task. Their work is also similar to ours in that they jointly infer roles and message labels. Critically, their work differs in that they learn from *unlabeled* training data. Also in a similar vein to our work, Reynolds et al. [103] compare a rule-based approach to a SVM and find that the rule-based model obtains higher accuracy. Like Huang et al. [48], we use social network information. The authors demonstrate that social features, such as network edge centrality, can improve classification. Our work differs from theirs in that we jointly infer relationship status while detecting messages, rather than constructing social features as a pre-processing step. Finally, our work has been highly motivated by Van Hee et al. [136] who go beyond binary classification to label events as belonging to textual categories. We compare our models to an SVM implementation of Van Hee's approach.

Similar to existing linguistic models, we build rich textual features to detect cyberbullying. Additionally, we exploit dependencies and similarities between words, documents and categories using a collective approach to better classify messages and discover latent categories. A critical difference from other fine-grained approaches is that we do not need labels to detect fine-grained bullying categories. Like existing work, we infer participant roles and describe messages with textual categories. Our work differs from previous state-of-the-art in that we address four distinct challenges in

one model which jointly infers: participant roles, latent textual categories and relational ties, while detecting cyberbullying messages.

7.3 Modeling Preliminaries

In order to model the intricate dependencies between language and participants in social interactions, we propose a collective probabilistic approach. We construct our models with PSL.

To illustrate PSL in the cyberbullying context, consider a rule which says that if two messages are similar, and one contains bullying content, then the other one may be likely to as well. To express this rule we introduce the predicate SIMILAR, which takes two messages as arguments and which expresses their similarity as a value between 0 and 1. Similarity can be described with many measures; for example, as the cosine distance between document embeddings or as the Jaccard distance between one-hot representations of the documents. In our linguistic models, we define similarity as the cosine distance between documents in a learned embedding space, while in our latent models, we introduce a range of similarity-like measures. Additionally, we introduce the predicate BULLYINGCONTENT, which takes a message as an argument and whose truth value indicates bullying. With these predicates, we define our rule in PSL as follows:

$$w_{sim} : \textsc{Similar}(T_a, T_b) \wedge \textsc{BullyingContent}(T_a) \Rightarrow \textsc{BullyingContent}(T_b).$$

Next we demonstrate how PSL can be used to template models for cyberbullying detection.

7.4 Probabilistic Cyberbullying Detection Models

We propose a series of five probabilistic models. The first three models: N-GRAMS, N-GRAMS++ and SEEDS++, employ linguistic information to detect cyberbullying in text. The next two models: LATENT LINGUISTIC and LATENT SOCIO-LINGUISTIC, utilize latent abstractions to categorize text and label roles. Additionally, LATENT SOCIO-LINGUISTIC infers relational ties.

7.4.1 Linguistic Models

In each model, our goal is to find the MAP assignment to $\boldsymbol{y}$. To do so, we introduce BullyTweet(T) which takes a tweet T, as an argument and whose truth value reflects the extent to which this tweet contains bullying content. By inferring the truth values of BullyTweet(T) for all tweets, we are finding the MAP assignment to $\boldsymbol{y}$.

N-Grams: The n-grams model consists of the rules in Table 7.1. To address class imbalance, we introduce a prior in all models, $\neg$BullyTweet(T) which captures that the majority of messages do not contain bullying content. For each n-gram, we instantiate a weighted rule correlating the presence of this n-gram within a tweet to the extent to which it is a bullying tweet. By training these weights, we learn which n-grams indicate bullying content. Here we consider n-grams to be unigrams and bigrams.

N-Grams++: The n-grams++ model contains the rules in Table 7.1 and Table 7.2. To overcome the sparsity of the feature vectors in the n-grams model, we propose two advanced features: sentiment and a document embedding similarity. Sentiment is assessed at the tweet level and provides some signal even from otherwise uninformative tweets. As bullying messages can be highly charged, we model the valence of a tweet with three predicates: NegTweet, PosTweet and NeuTweet, respectively expressing the negativity, positivity, and neutrality of a tweet. Additionally, we learn distributed representations of each tweet, allowing us to abate some issues of sparsity. By mapping tweets to an embedding space, using any common text embedding method, we can encode the relationship that documents which are close to each other in that space should have similar labels. To do so, we introduce the predicate Similar(T_i, T_j) whose truth value is the cosine similarity between the embeddings of T_i and T_j, respectively, scaled to be in [0,1]. By modeling similarity, we can explicitly express dependencies between our target variables, thus benefiting from collective inference.

w_i : HasWord(T, ng_i) $\Rightarrow$ BullyTweet(T)

Table 7.1: N-Grams

$$
\begin{aligned}
w_{neg} &: \textsc{NegTweet}(T \Rightarrow \textsc{BullyTweet}(T) \\
w_{pos} &: \textsc{PosTweet}(T) \Rightarrow \neg\textsc{BullyTweet}(T) \\
w_{neu} &: \textsc{NeuTweet}(T) \Rightarrow \neg\textsc{BullyTweet}(T) \\
w_c &: \textsc{BullyTweet}(T_i) \wedge \textsc{Similar}(T_i, T_j) \Rightarrow \textsc{BullyTweet}(T)
\end{aligned}
$$

Table 7.2: Sentiment and Document Similarity

The disadvantage of N-Grams and N-Grams++ is that the weight relating each word to bullying must be tuned with training data, which is commonly sparse because many words appear in only a few documents. This sparsity can make learning from a small corpus difficult and can hamper generalization to unseen data. While our N-Grams++ model mitigates this to some extent, a model which exploits domain knowledge to reduce the number of free parameters may be better suited to this task.

$$
\begin{aligned}
w_n &: \textsc{Insult}(T) && \Rightarrow \textsc{BullyTweet}(T) \\
w_t &: \textsc{Threat}(T) && \Rightarrow \textsc{BullyTweet}(T) \\
w_x &: \textsc{Sexual}(T) && \Rightarrow \textsc{BullyTweet}(T) \\
w_s &: \textsc{Silencing}(T) && \Rightarrow \textsc{BullyTweet}(T) \\
w_{m2} &: \textsc{HasWord}(T, W) \wedge \textsc{DirectMention}(W) \wedge \textsc{Insult}(T) \Rightarrow \textsc{BullyTweet}(T) \\
w_{m3} &: \textsc{HasWord}(T, W) \wedge \textsc{IndirectMention}(W) \wedge \textsc{Insult}(T) \Rightarrow \neg\textsc{BullyTweet}(T)
\end{aligned}
$$

Table 7.3: Seed Phrases

Seeds++: We propose Seeds++ to address the limitations in correlating n-grams to bullying content with short social media messages. Seeds++ consists of the rules in Table 7.2 and Table 7.3, which relates certain phrases to bullying. Van Hee et al. [136] found seven different categories of bullying messages. In our data we found evidence of three: name calling, threatening and sexual remarks, as well as a fourth category we refer to as silencing. Each category contains a small set of phrases.

Furthermore, we differentiate between attacks directed at individuals and third parties with two rules. These rules specify if a message contains a direct second or indirect third person reference. For example, "you" is a second person reference and 'they" is a third person reference. If a word W, (which can also be an n-gram), is a

second person reference DirectionMention(W) will be 1, and if it is a third person reference IndirectMention(W) will be 1.

7.4.2 Latent Variable Models

To model unobserved phenomena, we introduce two latent variable models. In the latent-linguistic model, the textual categories of messages and the roles of participants are modeled as latent variables. By describing messages with categories, and users with roles, we can relate roles to types of attacks, rather than expressly connecting roles to low-level linguistic cues.

Next, the socio-linguistic model expresses unobserved relational ties between participants and connects these relationships to participant roles and textual categories. As the true strength of any relationship between two people is an unobservable quantity, we model relational ties with latent variables. Here, we treat these relational ties as indicators of friendship (although they could also represent other types of social ties) and predict the presence and strength of these ties both with social network and linguistic signals.

Like the linguistic models described earlier, these latent variable PSL models template HL-MRFs. One difference is that in the latent setting the joint probability distribution is defined over $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$, where $\boldsymbol{z}$ is a vector of latent variables. To perform weight-learning in the presence of latent variables, we use the method described by Bach et al. [12].

In the following models we make use of functional constraints in PSL. Consider the predicate Category(T, C) in $[0, 1]$ which is 1 if tweet T can be described by category C. We allow partial membership to categories while ensuring that a tweet cannot fully belong to more than one category. This is expressed in PSL with the following functional constraint:

$$\infty : \sum_{c \in Categories} \textsc{Category}(T, c) = 1.$$

This constraint has infinite weight and ensures that a tweet cannot completely belong to two categories simultaneously. In the following latent models, we use functional constraints in modeling users' roles in tweets, the text categories of tweets and the membership of words in text categories.

Latent-Linguistic: In this model we propose that each tweet can be described with a latent category; these categories can be thought of as speech or dialog acts [114] or as sub-topics. For example, we define three act types: *Attack*, *Teasing* and *Other*. Furthermore, we define four kinds of attacks: name calling (N), sexual name calling (Sn), silencing (S) and threatening (Th). We use $Attack_i$, where i indexes into $\{N, Sn, S, Th\}$, to indicate the occurrence of each type of attack. We use $NotAttack_j$ to indicate the remaining categories, where j indicates one of teasing (Ts) or other (Ot).

$\infty : \sum_{c \in Categories} \textsc{TextCat}(T, c) = 1$
$\infty : \textsc{TextCat}(T, Ot) + \textsc{TextCat}(T, Ts) = 1 - \textsc{BullyTweet}(T)$
$w_{a_i b} : \textsc{TextCat}(T, Attack_i) \Rightarrow \textsc{BullyTweet}(T)$

Table 7.4: Inferring Text Categories

$\infty : \sum_{c \in Categories} \textsc{WordInCat}(W, c) = 1$
$w_{wtc_i} : \textsc{Category}_i(W) \Rightarrow \textsc{WordInCat}(W, C)$
$w_{tda} : \textsc{WordInCat}(W, C) \wedge \textsc{HasWord}(T, W) \Rightarrow \textsc{TextCat}(T, C)$

Table 7.5: Words to Categories

Let $\textsc{TextCat}(T, C)$ be a random variable in $[0, 1]$, such that it is 1 if the text category of T is C. In Table 7.4 we introduce constraints which enforce useful dynamics for this task. The first rule is a functional constraint. This forms a probability vector for each tweet where the entries of the vectors are the text categories. This constraint allows mixed membership while ensuring that a tweet cannot completely belong to more than one category. In the second rule in Table 7.4, we relate the categories of *teasing* and *other* to non-bullying messages. This rule also sets the truth value of a bullying message to the combined truth values of each of the types of bullying attacks. This is useful as messages can be combinations of categories, for example, a message may include name calling and threats. In the final rule, we relate each type of attack to bullying messages.

As in Seeds++, the latent models make use of seed words; however, these words now connect specific categories. Additionally, rather than correlating each word

to bullying content, these categories are related to bullying. For each word, W, and text category C, we model the extent to which this word belongs in this category, $\textsc{WordInCat}(W, C)$. For each word, we learn a probability vector over categories which is encoded with the functional constraint in Table 7.5. The second rule in Table 7.5 relates known seed words to certain categories, where $\textsc{Category}_i(W)$ is 1 if a seed word is known to be in $\textsc{Category}_i$. Instead of modeling that messages with seed words indicate bullying content, with the final rule in Table 7.5, we model the categories of messages according to their constituent words.

$w_{wsc} : \textsc{WordInCat}(W_i, C) \wedge \textsc{WordSim}(W_i, W_j) \Rightarrow \textsc{WordInCat}(W_j, C)$
$w_{wscc} : \textsc{WordInCat}(W_i, C) \wedge \textsc{SameCluster}(W_i, W_j) \Rightarrow \textsc{WordInCat}(W_j, C)$
$w_{wcoc} : \textsc{WordInCat}(W_i, C) \wedge \textsc{CoOccur}(W_i, W_j) \Rightarrow \textsc{WordInCat}(W_j, C)$
$w_{wac} : \textsc{WordInCat}(W_i, C) \wedge \textsc{Associated}(W_i, W_j) \Rightarrow \textsc{WordInCat}(W_j, C)$

Table 7.6: Word Associations

To find potential words which may belong in a given category, we have four collective word rules, shown in Table 7.6. The first two rules find replacement words for seed words where replacements may have the same semantics as the seeds. For example, we consider the similarity between words in an embedding space with WordSim. We also utilize word clusters and consider words in the same cluster as seed words with SameCluster. We also consider related non-replacement words which may be commonly used *with* seed words. Each pair of words is assigned a co-occurence score, CoOccur, which is the number of documents this pair occurs in, scaled to be in $[0, 1]$. Word associations can capture both replacements and associations, as words with similar conceptual meaning may be recalled in free association tasks [86]. Thus, we further expand potential category words to include words associated with seed words.

$w_{gda_i} : \textsc{NegTweet}(T) \Rightarrow \textsc{TextCat}(T, Attack_i)$
$w_{nda_j} : \textsc{NeuTweet}(T) \Rightarrow \textsc{TextCat}(T, NotAttack_j)$
$w_{pda_j} : \textsc{PosTweet}(T) \Rightarrow \textsc{TextCat}(T, NotAttack_j)$

Table 7.7: Sentiment of Text Categories

Exactly as we used sentiment to suggest bullying content, we now model the relationship between sentiment and categories. In Table 7.7 the argument $Attack_i$ refers to any of the four attack categories and we express that tweets with negatively charged content can be described as attacks, while those with positive or neutral content are either in the category *Other* or *Teasing*.

$w_{m2_i} : \textsc{HasWord}(T, W_i) \wedge \textsc{DirectMention}(W_i) \wedge \textsc{HasWord}(T, W_j)$
$\wedge \textsc{WordInCat}(W_j, Attack_i) \Rightarrow \textsc{DialogAct}(T, Attack_i)$
$w_{m3} : \textsc{HasWord}(T, W_i) \wedge \textsc{InDirectMention}(W_i) \wedge \textsc{HasWord}(T, W_j)$
$\wedge \textsc{WordInCat}(W_j, Attack_i) \Rightarrow \textsc{DialogAct}(T, Ot)$

Table 7.8: Subjects and Text Categories

As in Seeds++, we differentiate between attacks with second and third person mentions. As shown in Table 7.8, those tweets with bullying words and second person references may be attacks. Alternatively, messages with third person references are less likely to be attacks.

In addition to predicting the bullying content of messages, we infer the participants roles. We refer to users targeted in bullying tweets as victims and authors of those tweets as bullies. Let U be a user who is either mentioned in, or authors, a tweet. We infer each user's role in a tweet, $\textsc{RoleInTweet}(T, U, R)$, where R can be either a victim V, bully B, or other O. We focus on a user's role in a particular tweet as user's roles are often flexible and depend on the context [110]. For the following rules, consider the template:

$$w_{ac} : \textsc{Author}(T, U) \wedge \textsc{BullyAttribute}(T) \Rightarrow \textsc{RoleInTweet}(T, U, B).$$

For all rules which follow the template above, the model also includes a rule derived from the corresponding template:

$$w_{mc} : \textsc{Mentions}(T, U) \wedge \textsc{BullyAttribute}(T) \Rightarrow \textsc{RoleInTweet}(T, U, V).$$

$$\infty : \sum_{x \in roles} \textsc{RoleInTweet}(T, U, x) = 1$$

$$\infty : \textsc{RoleInTweet}(T, U_{author}, B) = \textsc{TextCat}(T, N) + \textsc{TextCat}(T, Sn) + \textsc{TextCat}(T, Th)$$

$$\infty : \textsc{RoleInTweet}(T, U_{author}, B) = \textsc{TextCat}(T, N) + \textsc{TextCat}(T, Sn) + \textsc{TextCat}(T, S)$$

$$w_{dab_i} : \textsc{Author}(T, U) \wedge \textsc{TextCat}(T, Attack_i) \Rightarrow \textsc{RoleInTweet}(T, U, B)$$

$$w_{dao_j} : \textsc{Author}(T, U) \wedge \textsc{TextCat}(T, NotAttack_j) \Rightarrow \textsc{RoleInTweet}(T, U, O)$$

$$w_{po} : \textsc{Author}(T, U) \wedge \textsc{PosUser}(U) \Rightarrow \textsc{RoleInTweet}(T, U, O)$$

$$w_{nb} : \textsc{Author}(T, U) \wedge \textsc{NegUser}(U) \Rightarrow \textsc{RoleInTweet}(T, U, B)$$

$$w_s : \textsc{RoleInTweet}(T, U_a, B) \wedge \textsc{Mentions}(T, U_b) \wedge \textsc{Author}(T, U_a) \Rightarrow \textsc{RoleInTweet}(T, U_b, V)$$

Table 7.9: User Roles

Rather than restrict each user to take exactly one role in each tweet, we allow users to assume roles to varying degrees. We then ensure that the degree to which a user assumes a given role is related to the extent to which they take any other role. For example, if there is strong evidence suggesting that a user is a bully, they cannot also be a victim with high certainty. This hard constraint is expressed in the first line of Table 7.9. The second and third rules in Table 7.9 are also functional constraints. These rules aggregate the categories of name calling and sexual name calling with threatening and silencing. This ensures that even if the certainty in any one of these categories is weak, if the aggregate certainty is high, we can detect bullying behavior. In the next two rules, we describe the relationship between each text category and role. When a tweet category is a kind of attack, the author is a bully and otherwise the author takes another role.

With the next two rules in Table 7.9, we correlate a user's propensity to be a bully or a victim with their past history. This decision is inspired by the findings of Hosseinmardi et al. [46]. For each user, we model the sentiment of their aggregate messages such that the truth value of $\textsc{PosUser}(U)$ will be close to 1 if U is generally positive in their messages. When a user U is generally negative, the truth value of

NegUser(U) will be high. The final rule states that if a bully author mentions a user in a tweet, that user is a victim.

Socio-Linguistic: In the final model, we infer relational ties between users and express a number of intuitions about how these ties might influence bullying behavior. We treat ties as positive relationships. For example, a tie between two users might indicate friendship or some other positive association. We model these ties with latent variables as their ground truth strength and nature is unobserved. We express the extent to which U_a is tied to U_b with Tie(U_a, U_b).

The rules in Table 7.10 describe how we infer relational ties. In the first rule, we express that most users in a social network are likely to not have ties. Next, we propose that a user following another indicates a tie. The following rule states that users may have ties to the ties of their ties. Next, we use linguistic information to predict ties with four rules. When a tweet's author mentions another user in a tweet that is teasing, or a category other than bullying, then those two users may be related. Additionally, if the tweet contains positive or neutral sentiment then the author and mentioned user may be related.

$w_{nf} : \neg\textsc{Tie}(U_a, U_b)$
$w_{fo} : \textsc{Follows}(U_a, U_b) \Rightarrow \textsc{Tie}(U_a, U_b)$
$w_{as} : \textsc{Tie}(U_a, U_b) \wedge \textsc{Tie}(U_b, U_c) \Rightarrow \textsc{Tie}(U_a, U_c)$
$w_{tf} : \textsc{Author}(T, U_a) \wedge \textsc{Mentions}(T, U_b) \wedge \textsc{DialogAct}(T, Ts) \Rightarrow \textsc{Tie}(U_a, U_b)$
$w_{of} : \textsc{Author}(T, U_a) \wedge \textsc{Mentions}(T, U_b) \wedge \textsc{DialogAct}(T, Ot) \Rightarrow \textsc{Tie}(U_a, U_b)$
$w_{pf} : \textsc{Author}(T, U_a) \wedge \textsc{Mentions}(T, U_b) \wedge \textsc{PosSent}(T) \Rightarrow \textsc{Tie}(U_a, U_b)$
$w_{nf} : \textsc{Author}(T, U_a) \wedge \textsc{Mentions}(T, U_b) \wedge \textsc{NeuSent}(T) \Rightarrow \textsc{Tie}(U_a, U_b)$

Table 7.10: Inferring Relational Ties

We also model social conversational patterns by learning the extent to which users with relational ties use certain text categories. Additionally, we use these ties to discover teasing. Whenever both bullying and teasing terms occur in a message between users with ties, we suggest that that is a teasing, rather than a bullying message.

$$w_{fc_i} : \textsc{Tie}(U_a, U_b) \wedge \textsc{Author}(T, U_a) \wedge \textsc{Mentions}(T, U_b) \Rightarrow \textsc{TextCat}(T, C_i)$$

$$w_{ft_i} : \textsc{Tie}(U_a, U_b) \wedge \textsc{Author}(T, U_a) \wedge \textsc{Mentions}(T, U_b) \wedge \textsc{HasWord}(T, W_x)$$
$$\wedge \textsc{HasWord}(T, W_y) \wedge \textsc{WordInCat}(W_x, \mathit{Attack}_i) \wedge \textsc{WordInCat}(W_y, Ts)$$
$$\Rightarrow \textsc{TextCat}(T, Ts)$$

Table 7.11: Ties and Conversation

$$w_{fv} : \textsc{Tie}(U_a, U_b) \wedge \textsc{Mentions}(T_a, U_a) \wedge \textsc{Mentions}(T_b, U_b)$$
$$\wedge \textsc{RoleInTweet}(T_a, U_a, X) \Rightarrow \textsc{RoleInTweet}(T_b, U_b, X)$$

$$w_{fb} : \textsc{Tie}(U_a, U_b) \wedge \textsc{Author}(T_a, U_a) \wedge \textsc{Author}(T_b, U_b)$$
$$\wedge \textsc{RoleInTweet}(T_a, U_a, X) \Rightarrow \textsc{RoleInTweet}(T_b, U_b, X)$$

$$w_{nb} : \textsc{Author}(T, U) \wedge \textsc{HighPopularity}(U) \Rightarrow \textsc{RoleInTweet}(T, U, B)$$

$$w_{po} : \textsc{Mentions}(T, U) \wedge \textsc{HighPopularity}(U) \Rightarrow \neg\textsc{RoleInTweet}(T, U, V)$$

$$w_{gu} : \textsc{Tie}(U_a, U_b) \wedge \textsc{Author}(T_a, U_a) \wedge \textsc{Author}(T_b, U_b) \wedge \textsc{RoleInTweet}(T_a, U_c, V)$$
$$\Rightarrow \textsc{RoleInTweet}(T_b, U_c, V)$$

Table 7.12: Social Behavior

Finally, peer pressure can have a large influence on youth, who are likely to adopt the behavior of their peers [14, 32]. Here we describe normative behavior with two rules which state that users who share a relational tie are likely to assume similar roles. Moreover, the role a participant takes may depend on their position within an exchange; e.g., whether they are mentioned in or are authoring messages. We restrain the association to friends assuming the same authorship role. To model that users with higher popularity may target users with low popularity, we introduce the predicates HighPopularity. Additionally, we capture ganging-up behavior with the last rule in Table 7.12, where users who share a relational tie may target the same victim.

7.5 Learning From Uncertain Annotations

As it is difficult and costly to acquire high-quality annotations in the cyberbullying domain, we explore the possible benefits of learning from low certainty annotations. There are two possible forms of uncertainty in this setting: disagreement between annotators

and the uncertainty of individual annotators. Here we have three annotators who label the tweets with a 0 (not bully), 1 (maybe bully) and 2 (bully).

Let a be the average normalized value, e.g. $\frac{1}{6}\sum_{i=1}^{3} a_i$, where a_i is the label of the i-th annotator and $\tilde{y}$ be the final label used in training. We explore three methods for determining $\tilde{y}$. In the *Discrete* method, we discard all labels without high inter-annotator agreement. That is, we round all labels such that if $a \geq \frac{2}{3}$, $\tilde{y} = 1$ and $a \leq \frac{1}{3}$, $\tilde{y} = 0$ and all $\frac{2}{3} > a > \frac{1}{3}$ are discarded. We also introduce an alternate method, *Soft*, where $\tilde{y} = a$, which allows us to use all of the information provided by the annotation. Finally, we introduce a *Hybrid* method. In this method, when there is high inter-annotator agreement, we treat the label as discrete and set $\tilde{y}$ exactly as in the discrete method. When all annotators are uncertain, or when all three disagree, such that $\frac{2}{3} > a > \frac{1}{3}$, we set $\tilde{y} = a$.

7.6 Empirical Evaluation

Here we explore cyberbullying on Twitter. We focus on youth and adolescents as they represent a particularly at-risk group [117], and collect tweets by or referencing users under the age of 18. In total, we collect approximately 4.5 million tweets. Of these, a subset were sampled for labeling according to the procedure of Chen et al. [23]. This resulted in a total of 1798 tweets which were labeled by three annotators. The annotators were asked to mark whether each message was a bullying message with 0, 1 and 2 indicating no, maybe and yes. The Fleiss inter-annotator agreement was .14. Approximately 27% of all tweets were labeled as bullying.

All tweets, including those which were not labeled, were used to construct social features, such as if a user mentioned another user, and to build global user features, such as the overall sentiment across a user's tweets. Each user's positive or negative score is assigned by computing the compound score of their entire message history. All users with a compound score less than -0.5 were counted as negative, while all users with a compound score greater than 0.5 were positive. The HighPopularity scores are a function of the total number of times a user is mentioned. All users with a high mention count (>5) are considered to have high popularity. Independently from the tweet collection process, we additionally collected user demographics for users in the

dataset. For each user, we query using their screen name to collect their followers and followees. By using non-labeled tweets to construct these features, we avoid the potential biases introduced in the selection process for labeling.

The tweets were cleaned according to standard text pre-processing practices. Stop words were removed, as were numbers and non-English words. Words with more than two repeat characters were trimmed, for example "haaappy" became "happy". Only those words which appeared in at least 5 tweets were retained. Sentiment was assigned with the open-source python tool VADER [52].

To calculate the similarity between two tweets, we compute the cosine similarity according to a trained Doc2Vec model [74]. There is a trade-off in document embedding models where a small domain-specific corpus may not have enough content to properly learn the embedding space, yet a publicly available corpus may not fully capture the nuances of a specific domain. For this reason, we train a Doc2Vec model on our corpus using Gensim [102] but seed it with pre-trained word vectors[1]. To seed the model, we used the openly available Glove [94] Word2Vecs which were trained on Twitter and are thus appropriate for this domain. The word associations were found using Nelson et al.'s [86] free association norms. Any word which had a positive forward associative strength (FSG) value, was added to the candidate pool of associated words for a given seed word. Clustered words were found using the hierarchical word clusters which were trained on Twitter data, as described in [91] and are published online[2]. To calculate the co-occurrence score, we count the frequency of word-pairs across all documents. For each word we calculate its co-occurrence score for all other words by dividing by the maximum co-occurrence count.

All models are trained using 5-fold cross validation on 80% of the data. In all folds we maintain a distribution of 30% bully tweets. The reported results are on the final held-out test set of 20% of the data. Here we compare five PSL models: the N-Grams, N-Grams++, Seeds++, Latent-Linguistic and Socio-Linguistic.

Additionally, we compare these to an implementation of Van Hee's state-of-the-art approach [136]. Unlike Van Hee, we do not include character level trigrams among the final features, as we did not find their inclusion to be helpful on the validation set.

[1] To initialize the Doc2Vec model with Word2Vecs we use: `https://github.com/jhlau/doc2vec`

[2] `http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html`

Also, we used VADER to calculate document level sentiment, rather than combining single word scores. Like Van Hee we included: positive, negative, neutral and a combined (compound) score. The classifier is implemented as a support vector machine (SVM) [25] using the Python package scikit-learn [93].

Bullies and Gender Li [77] found that young men are more likely to engage in cyberbullying, both as aggressors and victims. We obtained users' genders from their profiles. Using this information we explore the bullying differences between genders. For example, in Fig. 7.1, we see that the majority of our data is collected on young

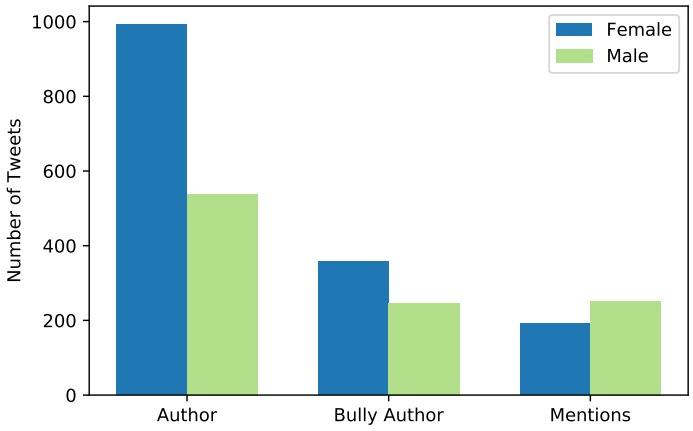


Figure 7.1: The dataset has more female than male authors. The percentage of males who bully is higher than the percentage of females.

women. Though there are more female than male authors of bullying tweets, a given male is more likely to author a bullying tweet than a given female, as ∼49.5% of the male authored tweets are labeled as bully tweets, compared to ∼36.0% for females.

7.6.1 Results

The first evaluation we present is on the ability of the models to detect cyberbullying content in messages, evaluated with F-Measure[3]. Another comparison is between the three labeling strategies. Additionally, we evaluate the ability of the latent variable

[3]Our code will be published with the final version of the paper.

models to assign participant roles. We also discuss the textual categories discovered by the latent models and the traits of the discovered relational ties.

Detecting Cyberbullying: We compare the five PSL models to Van Hee's approach. To report the detection results, we round $\boldsymbol{y}$ to 0 or 1 values. In Fig. 7.2, we report the average F-Measure across all labeling strategies for each model (the SVM uses only the discrete strategy). We see that adding collective rules and sentiment, in N-Grams++, improves the performance of N-Grams, while the seed phrases in Seeds++ are more powerful than n-grams. The latent models are best at detecting cyberbullying, with Socio-Linguistic achieving the highest F-Measure of 63.2.

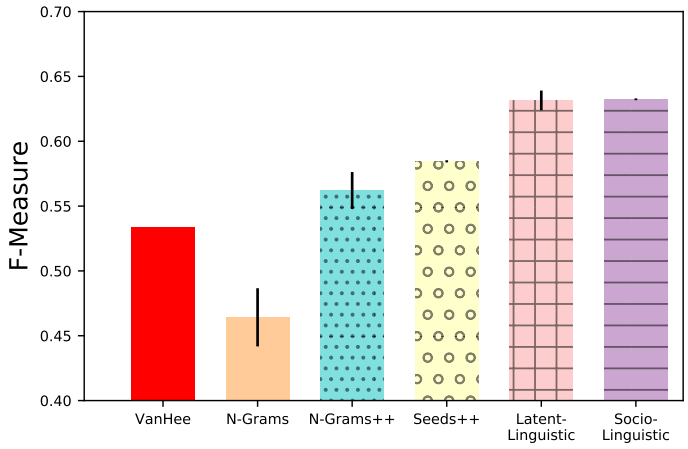


Figure 7.2: Collective rules improve the N-Grams model, and Socio-Linguistic achieves the best performance (bars are standard error).

In Table 7.13, we average the results for each labeling strategy across all PSL models. Utilizing uncertainty in any form is beneficial and the Hybrid approach yields the best F-Measure.

Role Assignment: Here we consider the effect of social information on detecting roles. To do so, we compare Latent-Linguistic to Socio-Linguistic. For the ground truth, each author of a bullying tweet is labeled as a bully; additionally each user who is mentioned in a bullying tweet is considered to be a victim. In Table 7.14, we see that the socio-linguistic model achieves the best F-Measure with predicting bullies and victims.

	Precision	Recall	F-Measure
Discrete	48.0	70.7	56.0
Soft	47.1	**75.0**	**57.8**
Hybrid	46.9	**79.6**	**58.7**

Table 7.13: The Soft and Hybrid methods provide statistically significant improvements (shown in bold) in F-Measure and recall over the discrete method.

	Precision	Recall	F-Measure
	Bullies		
Latent Linguistic	55.4	58.8	57.2
Socio-Linguistic	54.1	**61.5**	**57.7**
	Victims		
Latent Linguistic	81.6	64.4	72.0
Socio-Linguistic	76.7	**70.2**	**73.3**

Table 7.14: When assigning roles, Socio-Linguistic achieves statistically significantly higher F-measure and recall than Latent-Linguistic according to a paired t-test.

Roles and Text Categories: Here we inspect the relative frequencies of the bullying categories. To do so, we consider all tweets for which the sum of truth values for the bullying categories exceeds the sum of the truth values of the non-bullying categories. The most common category is name calling, with an average of 70.3% of all tweets predicted to be bullying belonging to this category. Sexual name calling was the second most common category with 24.6%. The remainder of the tweets were predominantly in the threatening category (4.5%), with 0.6% labeled as silencing.

Relational Ties: We analyze ties predicted with Socio-Linguistic. We first look at the number of predicted ties. Next, we ask whether bullies have more ties on average and if bullies share ties with attack targets. Ties are discovered by Socio-Linguistic using each of the three labeling strategies, and we report the results averaged across the three methods. We inferred 131 ties among 421 users in the test set. The average number of ties per user was .40, while it was 1.5 for those users with

at least one tie.

Here we consider authors of bullying tweets as bullies and users mentioned in those tweets as victims. Though this approach has shortcomings, it allows us to inspect characteristics of ties between users who have bullied and been victimized. In this approach we do not find that bullies have more ties. The average number of ties for bullies is .19, while it is .20 for victims. Of the ties, 9 were between bullies and victims, where victims had ties to bullies. In contrast, only 4.33 bullies (averaged across the three labeling strategies) had ties to victims, potentially reflecting that power dynamics influence relationships.

7.7 Discussion

There is a clear benefit to leveraging document representations and classifying bullying content collectively, as seen in the improvement of N-Grams++ over N-Grams. Furthermore, when the feature vectors representing short messages are sparse, we see an advantage to using seed phrases (with Seeds++) rather than n-grams.

In Latent-Linguistic we describe messages with fine-grained categories. These categories are seeded as in Seeds++ and through a number of collective word rules, these categories can adaptively expand. This abstraction of messages to textual categories is useful in predicting cyberbullying messages and this is shown by the clear improvement from Seeds++ to Latent-Linguistic. It is also useful for being able to interpret cyberbullying in more detail. For example, we see that name-calling is the primary form of cyberbullying in this dataset, followed by sexual name-calling.

Cyberbullying is a social activity, thus in Socio-Linguistic we predict relational ties as well as participant roles. A primary question of this model is if we can infer these ties with limited social media data which does not contain explicit relationship links. We find that leveraging social information provides an improvement in performance over Latent-Linguistic. This suggests that the inferred ties might be meaningful. In denser social networks this improvement might be more pronounced.

By inferring relational ties we are able to better interpret group dynamics. For example, there is support for the theory that bullies are popular within peer-networks, although we did not see this theory reflected in the predicted number of bully ties. We

did, however, see that victims were more likely to have ties to bullies than bullies to victims. This supports the idea that it may be socially advantageous to act positively towards bullies, who can hold positions of high social-status, while it is less socially advantageous to exhibit positive behavior towards members with low social-status, such as targets of bullying attacks. One persisting question is how different network structures might impact bully-victim relationships. We have found one example of the influence of social-status within a small discovered social network. Larger datasets might afford more discoveries into the nature of these complex relationships.

One limitation of our work is that we only investigated these dynamics on one social media platform. Yet, online interactions can be influenced by the platform which hosts them [37]. In future work we will broaden this investigation to additional social media platforms. Additionally, while the size of our dataset was comparable to those in published literature [15], we would surely learn more from a larger dataset. Finally, here we only consider the roles of participants in individual tweets. An important next step is to contextualize how users' roles vary and persist across situations.

7.8 Conclusion

Machine learning models face inherent challenges from social media data which can consist of short messages with misspellings and slang. Cyberbullying detection is made all the more difficult by a reliance on third-party annotators to acquire sufficient data for training. We address these concerns with two categories of models: domain-inspired linguistic models and a socio-linguistic model. The domain-inspired models combat sparsity by reducing the number of parameters which must be learned and by exploiting relations between words and documents collectively. Our socio-linguistic model is capable of inferring relationship ties from limited social media data while detecting cyberbullying. To the best of our knowledge, it is the first model in this domain which jointly infers bullying content, textual categories, participant roles and relationship links. By formulating these tasks jointly, we can learn from social dynamics to provide a statistically significant improvement in both cyberbullying detection and role assignment.

Similarly, in Chapter 8 we use online data, in the form of advertisements, to

study the illicit behavior of human trafficking. As in this work on cyberbullying, in studying human trafficking we use attributes of victims to predict aggressive behavior, in this case, the behavior of traffickers. In this chapter, we used social and linguistic structure to overcome issues of sparsity. In the following chapter we use spatio-temporal structure to predict where traffickers will travel next.

Chapter 8

A Spatio-temporal Approach for Tracking Traffickers

8.1 Introduction

Global temperatures are projected to rise an estimated 8-11°F over the course of the next century [141]. This overall temperature increase will be accompanied by changing climates, resulting in extreme weather and changes to stable ecosystems. While climate change is a serious environmental threat, it also poses significant risks to human well-being and social systems [1, 81]. One projected impact of climate change is on security outcomes, altering and potentially increasing opportunities for crime [101, 3].

Understanding the relationship between environmental stressors and criminal activity requires utilizing multiple heterogenous data, from both online and offline sources. We propose to model the relationship between environmental stressors and crime though a probabilistic approach which can fuse multiple heterogenous signals and model spatial and temporal relationships. We evaluate the feasibility of this approach by analyzing the relationship between extreme weather events and human trafficking.

An estimated 20.9 million people are victims of human trafficking [96]. Of these, an estimated 2 million are children. Any efforts towards the reduction of trafficking have the potential to drastically improve the quality of life of these victims.

Studying the dynamics of human trafficking, from victim characteristics to risk factors, is complicated by the difficulty of acquiring reliable data on a population which

strives to avoid detection [57, 40, 41]. Victims face many barriers to sharing information with law enforcement, from possible repercussions from abusers, to lack of access to and mistrust of law enforcement. Typical sources of information are victims' testimonies from non-governmental organizations and printed materials.

Alternatively, there is a wealth of information available from online sources. Human traffickers post advertisements for sexual services on public websites, such as Backpage.com. Thus, it is possible to collect large amounts of advertisements where the service provider may be a victim of trafficking. However, traffickers take care to obfuscate identifiable features, such as phone numbers, and advanced tools are needed to extract relevant information. Knowledge graphs are one tool which have proven successful at capturing relevant details from online advertisements [124]. These extracted entities, such as physical characteristics of potential victims, can be used to better study trafficking [98, 30]. For example, when it is possible to extract them, the phone numbers from ads can be used to reconstruct trafficking routes or circuits [53].

We have collected online advertisements posted on websites where traffickers advertise the services of their victims. By applying entity extraction techniques to these ads, we can extract relevant information such as: phone numbers, location information, and details on the demographics of the service providers. Using the phone numbers which identify ad posters and the locations of ads, we can track the movements of ad-posters over time. For the remainder of the paper we use the term trafficker to refer to these ad-posters.

In our work, we combine online data with offline knowledge of extreme weather events. We investigate the open question of how extreme weather events might impact trafficking. Deepening our understanding of this relationship can assist efforts in apprehending traffickers, especially in the aftermath of such events. Furthermore, our approach can be generalized to a variety of other situations in which environmental stressors impact security outcomes.

As a preliminary study of this complex relationship, we propose three potential effects of catastrophic storms on human trafficking: change in vulnerability of effected populations, change in attraction of effected areas and change in trafficking routes. We summarize these potential effects with the following research questions:

1. Is there a measurable change in the vulnerability of effected populations?

2. After an extreme event, how does the volume of trafficking change in effected areas?

3. Do extreme events disrupt trafficking routes?

We investigate two environmental events: *Hurricane Matthew* and *Typhoon Goni*. Hurricane Matthew inflicted heavy damages throughout the Caribbean and southeastern United States. Here, we have the opportunity to inspect the diverse effects of a wide-ranging natural disaster. Typhoon Goni affected many in East Asia. By studying these two environmental events, we can compare their effect on human trafficking across the globe.

We propose a series of three probabilistic spatio-temporal models to predict traffickers' movements. The first model predicts where traffickers will go next without any knowledge of extreme events. We demonstrate that this simple model achieves a reasonable F-Measure at predicting future locations.

Additionally, we propose a model which infers structure between locations as route segments. Such a model can capture longer-term dependencies between movements. For example, knowing where an ad-poster was two time-steps ago may help differentiate between two otherwise ambiguous future locations. This model also predicts movements while allowing us to learn which locations are linked. By incorporating spatial structure, we can augment cases where evidence of travel is lacking. We extend this model with an event-aware model which predicts where traffickers will go in the aftermath of an extreme event. We implement our models in PSL [11], in which we can model this task's spatio-temporal structure.

We present a novel analysis of how environmental events effect human trafficking. Before addressing this relationship we discuss related work in Section 8.2 and introduce some background on human trafficking in Section 8.3. We then address R1 and R2 in Section 8.4. In Section 8.5 we introduce our probabilistic models and in Section 8.6 we evaluate their ability to predict future locations and to discover route segments while addressing R3.

8.2 Related Work

A critical task in understanding human trafficking is to extract relevant information from online advertisements. This task has been addressed by a number of publications. Szekely et al. [124] propose a knowledge graph approach to extracting attributes from ads, and deployed this system with law enforcement agencies. Portnoff et al. [98] address issues of authorship, utilizing Bitcoin as well as the posting service Backpage.com, to identify true post authors. Also addressing whether posts originate from the same authors Nagpal et al. [84] use a support vector machine classifier. To confront the challenge of successfully determining advertisement locations, Kapoor et al. [58] propose a constraint-based approach.

Another important task is determining which advertisements are truly cases of human trafficking. Alvari et al. [4] use an semi-supervised approach to detect ads with high risk of trafficking. Addressing both the challenge of extracting and utilizing data from online advertisements, Dubrawski et al. [30] use extracted textual features to detect incidences of trafficking.

There has been limited prior work on the question of how events can disrupt trafficking. Dubrawski et al. [30] establish a positive correlation between the Super Bowl and ads for trafficking. They also attribute a rise in trafficking to a population boom in North Dakota. Their work demonstrates the need to further understand how events can transform trafficking dynamics.

Similarly, the question of predicting traffickers' movements has been under-explored. Ibanez and Suthers [53] collected online advertisements and analyzed phone number patterns to extract circuits. They discovered several regional circuits, such as between locations along the West Coast of the United States. This work highlights both the spatial, and habitual nature of trafficker travel.

In our work, we collect ethnicities, phone numbers, and locations from online advertisements. These entities can then be used to address relevant questions and tasks in this domain. Here, we address the novel question of how environmental events affect trafficking activity. As one approach to investigating this question, we propose a series of collective probabilistic models for predicting where traffickers will travel next.

8.3 Sources, Destinations, and Transit Hubs

Nations with human trafficking activity can be categorized as sources, destinations, transit hubs or some combination thereof [16]. Source countries are countries of origin for victims of trafficking. Destination countries are those where victims are taken. Transit countries are those which victims pass through. These distinctions hold ramifications for how traffickers travel within, to and from various countries.

In our data we investigate trafficking primarily in two countries: the United States and the Philippines. The United States is both a source, a destination and a transit country [89]. Furthermore, it is the second largest destination country in the world [112] and understanding trafficking behavior in the United States has global ramifications. The Philippines is primarily a source country, although it also has destination locations, and serves as a transit hub [89]. Furthermore, the Philippines is one of the top three source destinations for victims in the United States [89] . As the largest number of female and children victims are trafficked either within Asia, or are from Asian countries [87] and an estimated 3% of the Philippine population is at risk of being trafficked at any time [42], this is an important area to study. IThese characteristics that differentiate source and destination countries might influence how environmental events effect trafficking in each.

8.4 Impacts on Trafficking

In this section we address research questions one and two. First, we begin with a description of the data. To produce the results of this study, we analyzed a dataset collected over several years from websites with prostitution-related ads and/or reviews. The ads in this dataset include many noisy attributes, including the city the ad was posted in, phone numbers, and personal attributes such as ethnicity, weight, eye color, etc.

A multi-phase process was used to collect this data. Web pages on the sites were retrieved by crawling the sites on a regular basis. To extract the data in ads on the site, two types of extraction techniques were utilized. The first technique identifies semi-structured data on the site, that is, data that is displayed on the site in a template-

style structure. Most commercial web sites that post ads for sex providers have such a template structure, so that data items such as phone numbers, prices, and physical attributes of the victim are displayed in the same place on each advertisement page. Our extraction software automatically identifies this type of template structure on a site through the use of a machine learning approach. This approach first clusters the pages based on the overall similarity of the pages, so that (ideally) each cluster has a single type of template on the page. Then, each cluster of pages is analyzed more carefully to identify a grammar that describes the template used on those pages, allowing the system to automatically extract the data fields from the template. Finally, a human user curates the data the first time the site is processed, in order to confirm the labels provided by the system, and then the process operates automatically on subsequent crawls. Because some of the data fields in the templates often contain free text (such

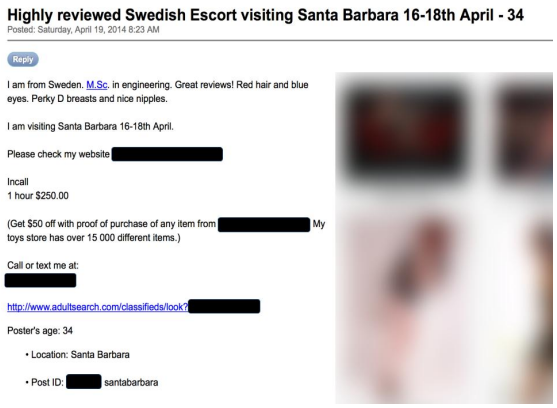


Figure 8.1: Sample online ad.

as a free text description of the service provider's "expertise"), we also use a second technique to extract data. In particular, we use site-independent extraction rules to look for specific types of data, such as ethnicities (e.g., Asian, Swedish) that have a distinctive vocabulary, as well as fields that have a distinctive internal structure (such as phone numbers). Used together, these two techniques provide higher recall than if a single technique is used alone.

Fig. 8.1 shows a real online ad. As is typical, the ad describes physical characteristics such as ethnicity. Also to note is the term 'visiting', and the short time span of the visit. This ad includes a phone number which can be extracted to track movements,

though it has been blacked out here.

8.4.1 Vulnerability Assessment

Environmental events can expose populations to risk, forcing people to lose their homes and their livelihoods. The Polaris report [97] identified recent migration or relocation as the top risk factor for rescued victims of human trafficking. Vulnerable populations may then become targets for traffickers in a number of ways. For example, seeking new opportunities, victims of natural disasters may put trust in traffickers who claim to be offering well-paying jobs. Additionally, as law enforcement divert resources to rescue and repair, it may be easier for traffickers to abduct victims. The end result of this increased vulnerability may be that we see increases in postings which advertise ethnicities associated with a vulnerable population. Alternatively, environmental stressors may increase the difficulty of traveling to effected areas, and we may observe a decreases in the vulnerability of effected populations.

In assessing vulnerability we consider two events: Hurricane Matthew and Typhoon Goni. Matthew was a category 5 hurricane, which struck the Carribbean and the southeastern United States in October 2016. This hurricane resulted in 603 fatalities and an estimated 15.09 billion in damages. Typhoon Goni was a large tropical storm which struck much of East Asia in August 2015. Goni resulted in 74 confirmed deaths and an estimated 831.7 million in damages.

To quantify changes in vulnerability of affected populations, we inspect the ad mentions of ethnicities from effected areas. To control for annual fluctuations in the total number of posted ads, we consider relative changes in the percentage of ads mentioning a given ethnicity. In doing so, we also need to control for seasonal patterns which can influence the ad volume at specific times of the year. To do so we take the following approach: for the year of a hurricane we calculate the change in ad volume between the month preceding and the month following a hurricane. We also calculate this ratio in the year before the hurricane. Let relative ad volume RAV(time, ethnicity) be a function which returns the percentage of ads posted in a period of time which mention a given ethnicity, e, and let m_i^j be a time period of month i during year j. We then define,

$$\delta_h = \frac{RAV(m^j_{i+1}, e)}{RAV(m^j_i, e)}, \quad \delta_p = \frac{RAV(m^{j-1}_{i+1}, e)}{RAV(m^{j-1}_i), e},$$
$$d_h = \frac{RAV(m^j_{i+1}, e) - RAV(m^j_i, e)}{RAV(m^j_i, e)}.$$

Thus δ_h and δ_p capture the change in volume between two periods, where δ_h refers to the year of the hurricane and δ_p refers to the previous year. By comparing these two measures we can determine if a greater change occurred in the year of the hurricane or the previous year. We also look at the difference in subsequent periods with d_h.

Hurricane Matthew: Matthew had a devastating effect on the Caribbean, and consequently, we inspect whether Caribbean populations experienced an increased vulnerability to human trafficking. To measure changes in vulnerability, we inspect the relative number of ads mentioning Caribbean ethnicities. In Fig. 8.2, we present d_h for the time period of Hurricane Matthew. We show only those cities where the difference between δ_h and δ_p was statistically significant according to a chi-squared contingency test. We see that several Florida cities witnessed statistically significant increases in ads mentioning Caribbean ethnicities. The largest increase is seen in Jacksonville. We observe statistically significant increases in mentions of Jamaican ethnicities in Jacksonville (of 160%).

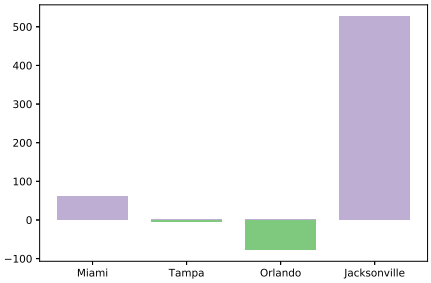


Figure 8.2: The relative percent change in ad mentions d_h for Caribbean ethnicity for cities in Florida after Hurricane Matthew.

Typhoon Goni: Additionally, we consider a different kind of environmental event in a different part of the world. Here we investigate the impact of Typhoon Goni. Goni had a large impact on the Philippines, and we inspect whether any locations saw an increase in postings mentioning Philippine ethnicities. We found that two locations had statistically significant increases: Dubai (United Arab Emirates) and New York

(United States). Dubai's increase by far dwarfed that of New York at 249% to .9%. Additionally, the majority of these postings were made with new phone numbers: 90% new numbers in Dubai and 28% in New York. Historically, roughly 14% of postings are listed with new phone numbers. This supports the theory that new victims of 'filipino' descent were advertised in these locations after Goni, lending credence to the hypothesis that these victims were only recently trafficked after the typhoon.

8.4.2 Trafficker Movements

Next, we ask if environmental stressors force traffickers to leave or enter affected areas. To do so, we consider the phone numbers which are included in posted advertisements. It is not always possible to successfully extract phone numbers from ads, and not all ads include them in the first place. Thus, these findings represent a sample of what might be found with full phone number data. We first focus this analysis on the United States. In the state of Florida we inspect all locations where the ads on the day of the hurricane are a small fraction of the historical average. Doing so, we find that the locations with the smallest ad fractions are on the Eastern coast of Florida in the projected path of Hurricane Matthew.

Location	**Ads on Day of Hurricane**	**Following Week Daily Average**	**Historical Average**
Jacksonville	25	187	193 $\pm$12
West Palm Beach	82	171	267 $\pm$20

Table 8.1: Ad activity for affected Florida cities before and after Hurricane Mathew. The historical average also shows the standard error of the mean.

We first assess the possibility of traffickers leaving affected areas. In this analysis, we inspect how many phone numbers posted in the week before each event remain in the following week. In Jacksonville, we see an average of 149 ads/day in the week preceding the hurricane. On the day of the hurricane this drops to 25. Of the numbers posted in the week preceding, 66.3% are missing in the next week (compared to a national historical average of 22%). In West Palm Beach, 56% of numbers are missing in the following week. Thus, we do see that the hurricane forces an abnormally large

number of traffickers to decrease their posting activity. That the majority of phone numbers in the next week are new to the area might indicate that traffickers leave environmentally effected areas.

Alternatively, after the immediate chaos of the event, traffickers might increase their activity in affected areas. An increase in trafficking could be attributed to a decreased risk of detection, as law enforcement diverts attention to rescue and repair. To investigate this, we look at the number of ads from three days after the hurricane to ten days. As we see decreased activity in the days immediately following the hurricane, by considering the activity from three days after we can isolate some of the incoming from the outgoing traffic. However, we continue to see depleted numbers in the following week in West Palm Beach, while Jacksonville sees maintained levels.

In the Philippines we observe a different trend. Here, in the effected area of Santa Ana, we see increases in ads in the timeframe of the hurricane. On August 22nd, 2015, we see 125 ads posted in Santa Ana. This compares to a historical average of 43 ($\pm$3.5) ads per day. There are many possible explanations for this increase which deserve exploring. For example, one explanation is that traffickers are drawn to an area with reduced law enforcement. Another explanation is that Santa Ana may be relatively more stable than surrounding areas. In this data we do not have comprehensive postings for the entirety of the Philippines, and locations with small postings are left out. However, given postings of more municipalities, we may learn more about how the effects of environmental events can vary by population and centrality to the event.

Together, both the influx and outflux of traffickers hold important insights for law enforcement. Traffickers may leave affected areas, disrupting any current plans for apprehension. Simultaneously, or with some small delay, new groups of traffickers may enter the same area. Thus any effort to apprehend previous traffickers, such as gathered intelligence, may not apply to these new groups.

8.5 Spatio-Temporal Models

In order to model trafficking routes, we propose three spatial probabilistic models which predict the movements of traffickers. The first model utilizes spatio-temporal relationships and location ad characteristics to predict where traffickers will go next. The second

1. $w_{l_i l_j}$: Location$(PNumID, t_k, l_i) \wedge$ Precedes(t_k, t_{k+1})
$\Rightarrow$ NextLocation$(PNumID, t_{k+1}, l_j)$

2. w_v : Location$(PNumID, t_k, l_i) \wedge$ Precedes$(t_k, t_{k+1}) \wedge$ CloseGreatCircle(l_i, l_j)
$\Rightarrow$ NextLocation$(PNumID, t_{k+1}, l_j)$

3. w_s : Location$(PNumID, t_k, l_i) \wedge$ Precedes$(t_k, t_{k+1}) \wedge$ SameState(l_i, l_j)
$\Rightarrow$ NextLocation$(PNumID, t_{k+1}, l_j)$

4. w_c : Location$(PNumID, t_k, l_i) \wedge$ Precedes$(t_k, t_{k+1}) \wedge$ SimCity(l_i, l_j)
$\Rightarrow$ NextLocation$(PNumID, t_{k+1}, l_j)$

5. $\infty : \sum_{l \in Locations}$ NextLocation$(PNumID, T, l) = 1$

Table 8.2: Rules for the model Spatio-Temporal.

model employs latent-variables to more explicitly model transit relationships between locations by identifying *route-sgements*. This latent formulation allows for additional modeling capabilities of how ad-posters move. Furthermore, the discovered values for latent route-segments variables can be used to better understand spatio-temporal dynamics of trafficking. Finally, we introduce an event-aware route-segment discovery model. This model utilizes knowledge of external events to also predict locations and discover route-segments.

We implement these models using PSL. To illustrate PSL in the human-trafficking context, consider a rule which says that if two locations are geographically close, traffickers are likely to move between them. To express this rule, we introduce the predicate Neighbors, which takes two locations as arguments and which expresses their spatial closeness as a value between 0 and 1. Additionally, we introduce the predicates Location and NextLocation which both take a phone number id $PNumID$, a time, t, and a location id l, as arguments, and whose truth value indicates whether $PNumID$ is at l at time t. We define our rule in PSL as follows:

$$w_{move} : \text{Location}(PNumID, t, l_i) \wedge \text{Neighbors}(l_i, l_j) \Rightarrow \text{NextLocation}(PNumID, t+1, l_j).$$

We propose three models which use PSL to template spatio-temporal relation-

ships and apply these relationships towards predicting the locations of human traffickers. The first model uses spatio-temporal rules to predict traffickers' locations, and the next models use event-information to improve these predictions. The second model expands on the spatio-temporal model to simultaneously discover routes while predicting future movements. Finally, we expand this *route segment discovery* model to make predictions with awareness of environmental events. In each model, the target variable is the next location l of each trafficker, $PNumID$, at time T_{k+1}, denoted NextLocation($PNumID, T_{k+1}, l$).

8.5.1 Location Prediction Model

In this first model, which we refer to as Spatio-Temporal, we predict future movements using spatial relationships, movement patterns between locations, and demographic similarities between locations. We model the tendency to move from each location to any other with rule 1 in Table 8.2. By learning a different weight w_{l_i,l_j} for each location pair, i, j, we can learn the relative frequency with which traffickers move from each location to the next. In rules 2 and 3 in Table 8.2, we express that traffickers are likely to move between locations which are spatially close. Here we use two distance measures. One is the great-circle distance[1], the other is a boolean value of whether two locations are in the same state.

In addition, we introduce a measure of ad similarities between two cities. Such a measure can supplement missing data when traffickers abruptly delete phone numbers and/or acquire new phones. For example, if we know that two cities have highly similar ads it may suggest that they are visited by the same traffickers. As one measure of the similarity of ads, we consider similarities in the distributions of mentioned ethnicities. Rule 4 expresses that a trafficker might next visit a location with an ethnicity distribution similar to that of their current location.

The final rule in Table 8.2 is a hard constraint. Hard constraints relate the truth values of given variables. For example, if there is certainty that a given trafficker is in a certain location at a given time, they cannot also be at another location at the same time. To express that a constraint must be satisfied it is given infinite weight.

[1] `https://en.wikipedia.org/wiki/Great-circle_distance`

1. $w_{npr} : \neg \textsc{OnRouteSg}(X, RS)$
2. $w_{npl} : \neg \textsc{Link}(X, Y)$
3. $w_{plr} : \textsc{LocationID}(X) \wedge \textsc{RouteSgID}(RS) \Rightarrow \textsc{OnRouteSg}(X, RS)$
4. $w_{ltr} :$ $\textsc{Location}(\text{PNumID,T,X}) \wedge \textsc{OnRouteSg}(X, RS)$
$\Rightarrow \textsc{RLocation}(PNumID, X, R, T)$
5. $w_{rp} : \textsc{RLocation}(PNumID, X, RS, T) \Rightarrow \textsc{OnRouteSg}(X, RS)$
6. $w_v : \textsc{CloseGreatCircle}(X, Y) \wedge \textsc{OnRouteSg}(X, RS) \wedge \textsc{OnRouteSg}(Y, RS)$
$\Rightarrow \textsc{Links}(X, Y)$
7. $w_s : \textsc{SameState}(X, Y) \wedge \textsc{OnRouteSg}(X, RS) \wedge \textsc{OnRouteSg}(Y, RS) \Rightarrow \textsc{Links}(X, Y)$
8. $w_{rc} : \textsc{OnRouteSg}(X, RS) \wedge \textsc{OnRouteSg}(Y, RS) \wedge \textsc{Precedes}(T_1, T_2)$
$\wedge \textsc{Location}(PNumID, T_1, X) \wedge \textsc{Location}(PNumID, T_2, Y) \Rightarrow \textsc{Links}(X, Y)$
9. $w_{sc} : \textsc{SimCity}(X, Y) \Rightarrow \textsc{Links}(X, Y)$
10. $w_{rl} : \textsc{OnRouteSg}(Y, RS) \wedge \textsc{Links}(X, Y) \Rightarrow \textsc{OnRouteSg}(Y, RS)$
11. $w_{l_1 l_2} : \textsc{RLocation}(PNumID, X, RS, T_1) \wedge \textsc{Precedes}(T_1, T_2)$
$\Rightarrow \textsc{NextLocation}(PNumID, T_2, Y)$
12. $\infty : \sum_{l \in Locations} \textsc{NextLocation}(PNumID, T, l) = 1$
13. $\infty : \sum_{l \in Locations} \textsc{RLocation}(PNumID, T, RS, l) = 1$
14. $w_{ef} : \textsc{RLocation}(PNumID, X, RS, T_1) \wedge \textsc{Effected}(Y) \wedge \textsc{Links}(Y, Z) \wedge \textsc{PostEvent}(T)$
$\wedge \textsc{Precedes}(T_1, T_2) \Rightarrow \textsc{NextLocation}(PNumID, T_2, Z)$
15. $w_{eb} : \textsc{RLocation}(PNumIDX, RS, T_1) \wedge \textsc{Effected}(Y) \wedge \textsc{Links}(Z, X) \wedge \textsc{PostEvent}(T)$
$\wedge \textsc{Precedes}(T_1, T_2) \Rightarrow \textsc{NextLocation}(PNumID, T_2, Z)$

Table 8.3: Rules for the model Route-Segments and Event-Aware Segments.

8.5.2 Route Segment Discovery Model

In addition to predicting future movements of traffickers, we infer connected route components, which we refer to as route segments. These segments can be links between pairs of cities which are frequently traveled between, or longer paths of destinations which are visited in a sequence. To model these dynamics, we introduce the model

ROUTE-SEGMENTS. We model route segments with a set of latent variables which describe their behavior. To describe if a location is on a route segment, we introduce ONROUTESG(L, RS) and to describe that traffickers move from location l_1 to location l_2, we introduce LINK(l_1, l_2). A LINK expresses a directed relationships and is also a latent variable whose value we infer. To describe where a particular number is along a route segment, we introduce RLOCATION(N, L, RS, T) which is 1 if number N is at location L, on route segment RS, at time T.

Initially, each location is randomly assigned to a route segment. These assignments are then updated according to the rules outlined in Table 8.3 which we explain next. This process essentially clusters the route segments, where route segments are grouped together according to some distance measure. For example, here we introduce rules which state that two locations might be on the same route segment if they are geographically close. Additionally, we constrain that if two locations are on the same route segment then a link should exist between them. However, this constraint is not hard, and it is possible that locations will be assigned to the same route segment, but not linked.

Like the predictive model described earlier, the latent variable PSL models template graphical models. One difference is that in the latent setting the joint probability distribution is defined over the observed variables $\boldsymbol{x}$, target variables $\boldsymbol{y}$ and latent variables $\boldsymbol{z}$, where $\boldsymbol{z}$ is a vector of latent variables. To perform weight-learning in the presence of latent variables, we use the expectation maximization method described by Bach et al. [12].

Rule 1 in Table 8.3 expresses the prior belief that most locations are not on most route segments. Similarly, Rule 2 expresses that most locations are not traveled between. With Rule 3 we assign specific priors to locations being on route segments. For example, we seed each route segment with one seed location, such that w_{plr} is high. For all other locations w_{plr} is relatively low. The choices of seed locations are explained in more detail in Section 8.6. For each location position X and each route segment RS we learn a weight for this location being on this route segment with Rule 4. With Rule 5 we further express that if a phone number ID is at a location on a route segment, then that location is on that route segment.

Rules 6 and 7 transfer the spatial rules from the previous model into this new route segment discovery problem. Here, if two locations are close or in the same state, and each location is on the same segment, then a link exists between the two locations. Additionally, if two locations are frequently traveled between, then there might be a link between them, as expressed by Rule 8. As in the location prediction model we utilize similarities in ethnicity distributions. Here, with Rule 9, we express that if two locations have similar ethnicity distributions, then a link might exist between them. To express that LINKS should connect locations on the same route segment, we include Rule 10.

We infer a phone number ID's next location with Rule 11. For each location pair, we learn a weight $w_{l_1 l_2}$ which expresses the extent to which a number will move from route segment position at location l_1 to the location l_2. This rule has a similar functionality to Rule 1 in Table 8.2.

As in SPATIO-TEMPORAL, here we constrain the values to NEXTLOCATION$(PNumID, T, X)$ such that a phone number ID cannot simultaneously be in two locations at the same time. Additionally, we constrain route segment positions such that a phone number ID cannot simultaneously be in two locations on the same route segment at the same time. The remaining rules in Table 8.3 are relevant for the event-aware route segment discovery model, which we discuss next.

8.5.3 Event-Aware Route Segment Discovery Model

Our final model, which we refer to as EVENT-AWARE SEGMENTS, is an event-aware route segment discovery model which consists of all of the rules in Table 8.3. In this model we incorporate rules which can infer future locations given knowledge of events. Here, we encode the idea that instead of visiting environmentally affected areas, traffickers will visit neighboring locations on the same route segment.

To express that traffickers may skip affected locations and visit the next location on a given route segment, we develop Rule 14. To express that traffickers may reverse course, and visit a previous location, we introduce Rule 15. Here, EFFECTED(Y) is 1 if location Y was effected by a given event. To infer only those movements which occur after the event of interest, we introduce POSTEVENT(T) which is 1 if time T

succeeds the date of the event, and 0 otherwise.

8.6 Empirical Evaluation

In this section, we present our evaluations on two tasks: predicting future movements and discovering routes. We cast the prediction of future movements as a classification task, where we predict if a number ID will visit a given location. Locations are grouped together, such that groups are routes and each location is assigned to one route. Here, a route is a path such that multiple locations are visited in some temporal order. We begin this section with a discussion of the data.

8.6.1 Trafficker Movement Data

We compare our three proposed models, Spatio-Temporal, Route-Segments and Event-Aware Segments, on postings related to Hurricane Mathew and Typhoon Goni. In the case of Hurricane Matthew, we first collect phone numbers from postings made in locations in Florida from July to December 2016. We then collect all additional postings which mention these numbers in a time period from July 2015 to December 2016. This provides us with 17 contiguous months of movement data. In this second stage, these seed numbers can be in postings made in any location. Postings are made at a daily resolution, and the difference between two consecutive postings can range from daily to monthly. We then get a collection of K numbers, $N = \mathbf{n}^1, ..., \mathbf{n}^k$, where each $\mathbf{n}^i$ is posted in a location l at time t, $\mathbf{n}^i_{l,t}$. This process is repeated for Typhoon Goni, where we first collect phone numbers from postings made in the Philippines and surrounding locations from April to November 2015. In the second stage, postings which contain these seed numbers are collected from April 2014 to November 2015, providing us with 19 months of contiguous data.

On average, for the hurricane data extracted as described above, each phone number is posted in three locations. In the Philippines data, the average number of locations is five for each phone number. In Fig. 8.3 we see the number of traffickers which visit certain numbers of locations. As the number of locations increases, the number of traffickers posting in more than that many locations decreases exponentially. This relationship is stable across both datasets. That most numbers visit a small amount

of locations suggests the presence, and perhaps predominance, of smaller organizations. This aligns with other reports which find that the majority of trafficking is conducted by small groups [111, 112].

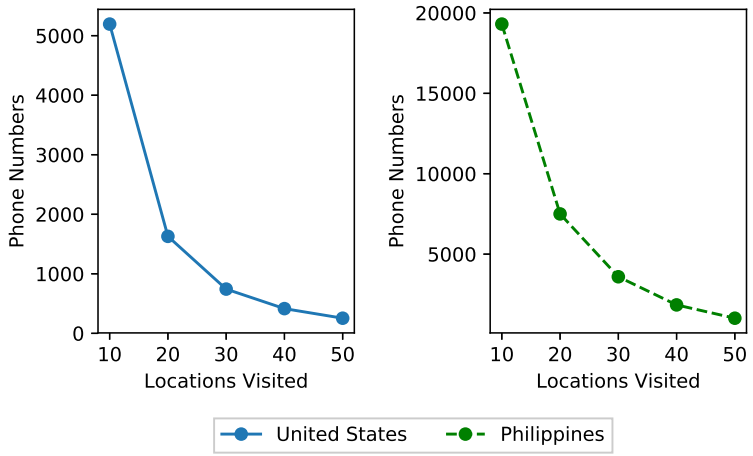


Figure 8.3: Total phone numbers (on the y-axis) which visit more than a certain number of locations (on the x-axis).

Fig. 8.6 shows the frequency of movements between locations in the United States. As can be seen, there are some clear patterns between certain locations. These patterns can be utilized in predicting the next locations. The top source locations are Manhattan, NY, New York, NY[2], San Francisco, CA, San Jose, CA, and Westchester, FL. Travel most often occurred between Fort Lauderdale and West Palm Beach in FL.

In Fig. 8.5, we show the travel patterns in and out of the Philippines. Here, we see a striking difference from the United States data. Unlike in the United States, where there is a large amount of inter-state travel, in the Philippines, there is a lot of traffic between the Philippines and external locations. We also see that the most popular locations in the Philippines are San Mateo and Santa Ana. For each of these locations there are not large differences in the in and out degrees. The most common source and destination location for San Mateo is San Francisco, CA while the most common source and destination for Santa Ana is Los Angeles, CA. It is interesting to see, at least for the subset of data that we have sampled, that it is relatively uncommon to travel between San Mateo and Santa Ana.

[2]If an advertisement differentiated between Manhattan and New York we kept that distinction.

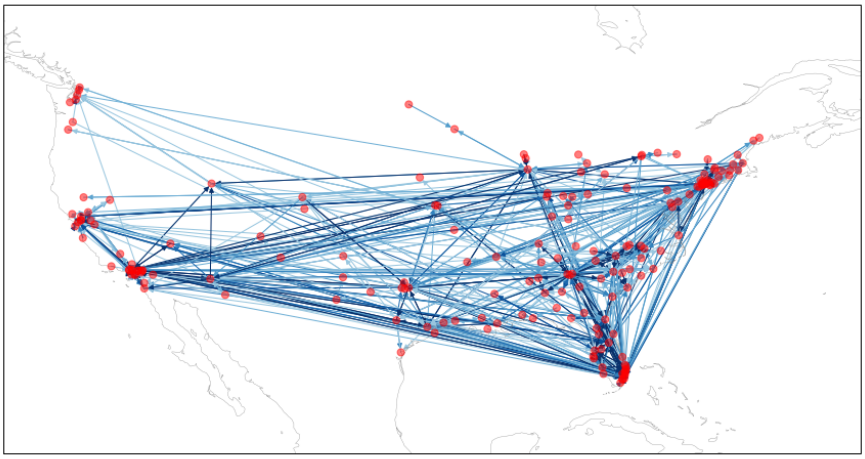


Figure 8.4: Travel to/from locations in the United States. Edges are weighted according to number of trips on that edge.

Figure 8.5: Travel to/from locations in the Philippines. Edges are weighted according to number of trips on that edge.

For the tasks of predicting future locations and discovering routes, we conduct several additional preprocessing steps for both datasets. We only consider postings with phone numbers that can be associated with a single location on the day of the posting. This step helps filter out those numbers which are posted in multiple locations simultaneously and whose location is thus ambiguous. Additionally, we only retain those numbers which are posted in at least three distinct locations at three distinct times. At test time, we infer the location of all phone numbers N at the time of Hurricane Matthew. We thus create a test set from all $\mathbf{n}^i_{l,t_h}$ where t_h falls on October 7th, 2016. We create a validation set from all $\mathbf{n}^i_{l,t_{h-1}}$ where for each $\mathbf{n}^i$, t_{h-1} is the time stamp which immediately precedes t_h. The training set is then all remaining phone numbers, $\mathbf{n}^i_{l,t_p}$ where $t_p < t_{h-1}$. This produces 2090 numbers, and approximately 18,400 postings. We refer to this dataset as MATTHEW. For Typhoon Goni, we choose August 22nd, 2015 as t_h. This results in a dataset of 6086 numbers and approximately 63,000 postings, which we refer to as GONI. In the following experiments we create a train, validate and test set for both MATTHEW and GONI. We split by phone number, such that 25% of all numbers are used as a validation set, and 25% are withheld in the test set and not seen in any stage of training.

We assess the ability of each model to predict the next location. In a real-world setting it is not clear how candidate next locations would be chosen. The simplest scenario would be to infer the most likely next location, out of all possible locations. To reduce the number of locations considered and generate a reasonable candidate set of next locations, we sample 10 potential next locations according to their great-circle distance to the current location. We additionally sample 10 locations from the same state, uniformly. In order to explore possible next locations which might not be geographically close, we randomly sample 10 more locations from the entire set of locations. This provides us with a small but reasonable set of 30 possible next locations for each location.

In inferring routes we initially randomly assign each location a route. These assignments are then updated by the model. For each route of k total routes we choose one seed location which belongs to this route with a high weight value. For these seed locations, w_{plr} in Rule 3 is set to 1000. These seeds are chosen as the top k locations

according to the amount of outgoing traffic. For all remaining locations the initial route assignment is done randomly and for each location we set w_{plr} to 100 for exactly one randomly chosen route id. The value of k is then a hyper-parameter which we explore with validation data.

To measure the similarity in ethnicity mentions between two locations, we compute a histogram of counts for ethnicity mentions for each location, and then compute the *Kullback-Leibler*-divergence between them. We then translate the divergence into a similarity between [0,1]. Similarities are computed between all pairs of cities with non-zero counts of the top ten ethnicities across the entire dataset[3].

8.6.2 Evaluation of Location Prediction

We compare our three spatio-temporal models to a distance-based baseline, which we refer to as SpatialThis baseline simply chooses the closest next location from the same set of potential next locations as the other models. In Table 8.4, we show the F-Measure on the task of predicting next locations. We evaluate the Event-Aware Segments on Matthew, as we stronger evidence of the impact of environmental effects on trafficker movements.

	F-Measure	
	Matthew	**Goni**
Spatial	9.37	8.57
Spatio-Temporal	71.0	74.8
Route-Segments	**82.8**	**94.2**
Event-Aware Segments	**83.4**	**95.1**

Table 8.4: F-Measure of each model on the location-prediction task. Bold signifies statistically significant improvements over both Spatial and Spatio-Temporal.

In addition to the F-Measure we consider the error rate, i.e. the fraction of phone number ids in the test set for which the predicted next location was incorrect. The Route-Segments model improved the error rate from 28.8% (in Spatio-Temporal),

[3]The choice of ten was a hyper-parameter explored with training data.

to 17% for MATTHEW. In GONI, the improvement was even more pronounced, from 25.2% to 5.78%. Additionally, we saw that areas in the potential path of Hurricane Matthew were more difficult to predict, as they obtained an error rate of 29% in ROUTE-SEGMENTS. Incorporating event-aware rules provides a reduction in error, as EVENT-AWARE SEGMENTS achieved an error rate of 27% for these locations.

8.6.3 Evaluation of Discovered Route Segments

Next, we inspect the discovered route segments for each dataset. To determine the number of route segments that best fits the data, we search over different numbers of routes and evaluate with the validation set. For each dataset we evaluated on 3-30 segments, in step sizes of three. Each location can belong to a route-segment to some degree. For each location l, we assigned it to route-segment rs_i where the value of $\text{ONROUTESG}(l, rs_i)$ was higher than all other $\text{ONROUTESG}(l, rs_j)$, $j \neq i$. To construct routes, we consider all links, $l_i - l_j$, where l_i and l_j were on the same route segment and the value of $\text{LINK}(l_i, l_j)$ was greater than .5.

In MATTHEW we found that nine route segments achieved the highest location-prediction performance on the validation set. Fig. 8.6 shows these segments; however, we only display locations which are on paths of at least length three. We form paths by linking locations on the same route segment. The longest route-segment originates in Los Angeles, CA and has 15 locations (shown in dark green). The shortest route-segment originates in West Palm Beach, FL and has three locations. While the Florida route predominantly includes locations on the East Coast of the United States, the California route includes locations across the country. However, unsurprisingly, it includes California locations such as Long Beach and Los Angeles. This was also the most common segment, with 57 numbers taking part in a some portion. The most common route segment (of length three) was spatially concentrated, from Long Beach to Orange County to Los Angeles, CA.

Fig. 8.7 shows an example route-segment in the United States. This segment shows some regional tendencies, as many of the locations are southeastern portion of the country. However, we also see that travel between large and somewhat distant cities is common.

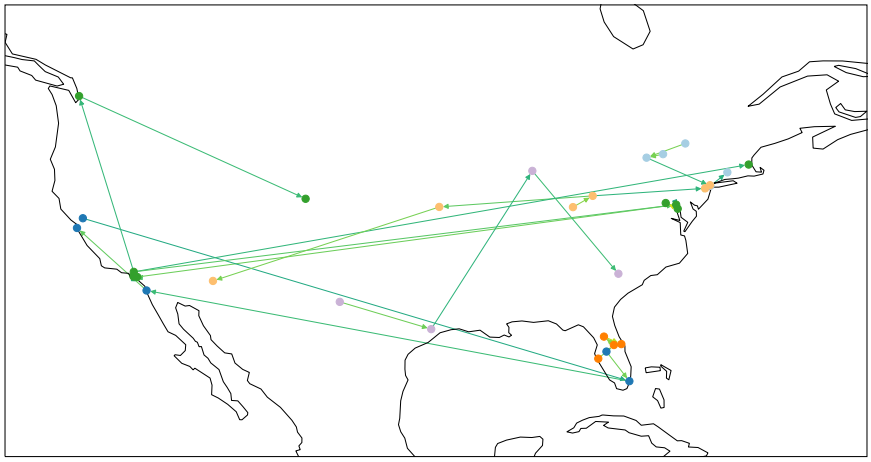


Figure 8.6: Discovered route segments in MATTHEW. The color of each node corresponds to a route-segment id and edge opacity indicates link strength.

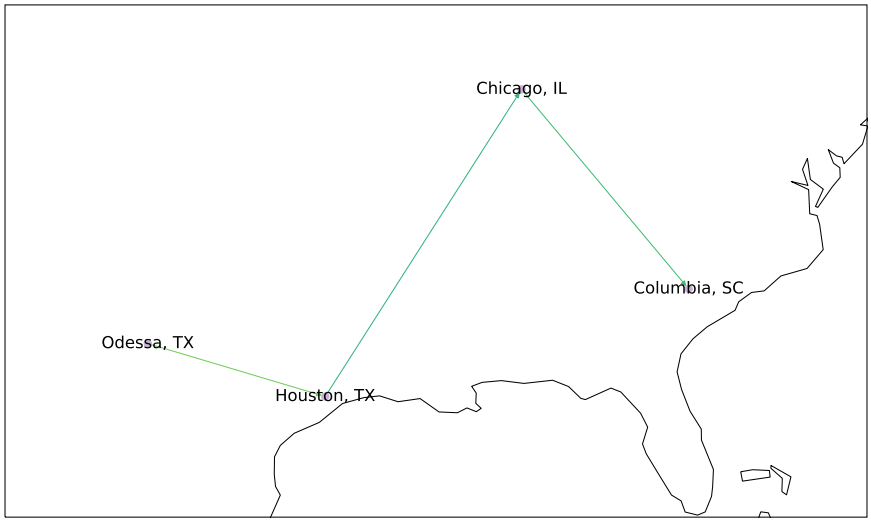


Figure 8.7: A route segment in MATTHEW.

In GONI, the longest route segment originated in San Francisco and had 11 locations. This was also the most common route, with 77 ad-posters traveling along some portion. This dataset revealed more international connections, for example there was a strong link between Dubai and Singapore. However, regional national route segments were also discovered, such as within the state of California and between locations in Canada. In comparison to MATTHEW, in GONI, many of the route segments were

international as can be seen in both examples in Fig. 8.8.

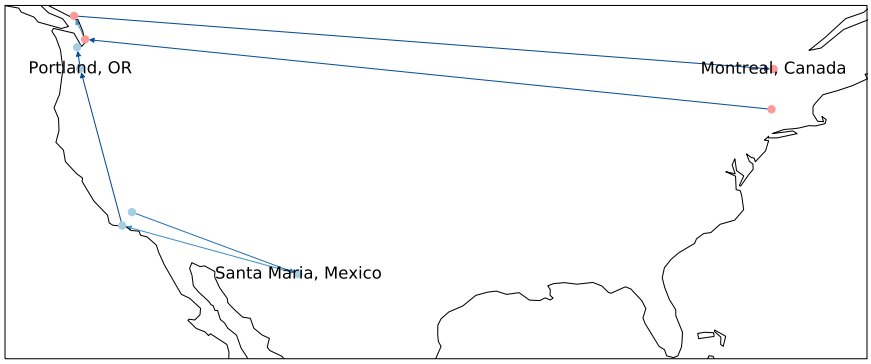


Figure 8.8: Selected discovered route segments in GONI. The color of each node corresponds to a route-segment id and edge opacity indicates link strength.

8.7 Discussion

We have proposed that extreme weather events can impact human trafficking, both through their impact on populations and locations. We saw that post-event there was evidence of an increase in the relative number of advertisements mentioning ethnicities from event-effected areas. Mentions of both Caribbean and Philippine ethnicities increased after Hurricane Matthew and Typhoon Goni, respectively. As the Philippines are a common source destination for trafficking victims, that we see substantial increases in Filipino ads in two locations suggests that traffickers may have taken advantage of environmentally induced vulnerabilities.

Readily available online data can provide surrogate statistics into the otherwise difficult to trace activity of human trafficking. However, there are many limitations with this data. One hindrance is that phone numbers are not maintained for long periods of time. Another limitation is that not all ads are posted by human traffickers.

Our model can incorporate both spatio-temporal and behavioral knowledge, providing a flexible framework for fusing heterogenous data sources to understand complex events. This framework might prove useful in a number of related tasks. Modeling the impact of extreme weather on human migration is another potential use case of our model.

8.8 Conclusion

Human trafficking is a serious social problem, and understanding this phenomenon is an important social science question. We investigated how environmental events might impact human trafficking in three ways: by changing the vulnerability of environmentally effected populations, by changing the attraction of traffickers to effected locations, and by disrupting trafficking travel. We also proposed a series of spatio-temporal predictive models. These model can incorporate both data-driven approaches, as well as domain knowledge such as how events might disrupt movement. More generally, this approach can be used to model environmental impacts on many types of human behavior.

Part V

Concluding Remarks

Richly structured and probabilistic approaches can fully exploit the social, dynamic and uncertain nature of human behavior. In this thesis I have introduced relational models which can better capture the richness of temporally and contextually interconnected appliance use, customers and products in an online shopping catalog, socially connected students in an online classroom, youth engaged in cyber-bullying within a social network and human traffickers' movements. Neglecting the temporal, contextual, and social relations in these problem settings hinders the understanding of human behavior in these domains. To advance state-of-the-art in sustainability, education and malicious behavior I have demonstrated the efficacy of a collective probabilistic approach.

Predictive models can solve a large range of tasks relevant in these social good domains. However, these models should be built with a careful understanding both of the available data and of the specifics of the domain. In each problem I have presented findings from exploratory data analysis which are useful apart from predictive results. For example, in the education setting I found that two groups of unexpected student types and showed that each group is defined by very different characteristics. These groups have implications for course design, demonstrating the different needs and expectations of different types of students. Additionally, I used causal techniques to investigate the effect of in-person instruction on student learning. In the cyberbullying setting I explored the relationship between gender and attacks. Finally, in the human trafficking domain, I assessed the question of how environmental events impact human trafficking. Thus, data analysis methods are one form of computational technique that can provide insight into problems of societal significance and which should be deployed given new data sets.

Electronic devices, from mobile phones to dishwashers, are increasingly equipped with the ability to communicate with remote servers, allowing for the large-scale tracking of human behavior in the physical world. This has resulted in a wealth of spatio-temporal data. However, such data requires advanced methods which can scalabley reason over structured data. Here, I have proposed a collective probabilistic approach for spatio-temporal settings and demonstrated its benefits in two different problem settings: energy disaggregation and predicting the movements of human traffickers. My approach

for disaggregating individual appliances from aggregate energy readings achieves state-of-the-art performance through multiple temporal representations and the use of soft constraints in relational energy usage. In predicting the future movements of human traffickers my approach achieves a high F-Measure and I demonstrate the utility of considering second-order dependencies in predicting future movements. Together, this work shows the benefits of flexible spatio-temporal models, which can adapt to multiple representations and orders of information.

Online data offers many opportunities to study human behavior at new depth and scale. However, though plentiful, it is not always of high quality. Here, I show how structure can overcome issues of data sparsity in two problem settings: cyberbullying detection and human trafficker location prediction. In the cyberbullying setting, I use both social and linguistic structure to predict cyberbullying from social media messages that once cleaned, are often no more than three words. This approach outperforms state-of-the-art while lending insight into bullying in online settings. In the human trafficking domain I show how spatio-temporal structure can be used to predict movements of traffickers from obscure online ads.

Much of human behavior is influenced by others. Our actions, beliefs and wellbeing can be effected by our friends, colleagues and family. Here, I explore social structure in two problem settings, classrooms in a high school computer science MOOC and friend networks in cyberbullying incidents. In each setting I show that social structure can be useful in predicting aspects of human behavior. In the education setting, modeling students who are in the same classroom and/or working together can improve predictions of post-test performance. While inferred relational ties are helpful when determining participants' roles in cyberbullying incidents.

Unobserved phenomena can influence much of our behavior. For example, psychological states are often unobserved, yet they dictate what we do and how we think and feel. In data-driven research, the data can fail to capture the true richness of the problem setting, where unobserved factors can effect the behavior of target variables. Across domains, I investigate how latent variables can be used to improve modeling and predictive performance. In sustainability, I model the latent sustainability of products and customers. In education, I model latent collaborative behavior and classroom and

student strength. In malicious behavior, I model many latent aspects of cyberbullying behavior, as well as latent route-segments in human trafficking travel. This work offers one strategy in improving interpretation in these domains. By analyzing the discovered values of latent variables one might improve understanding of the problem. For example, by analyzing discovered route segments one can trace potential routes traveled by traffickers. These routes might be useful in criminal apprehension. However, *understanding* human behavior in social good domains is critical and this work confronts only one piece of that large goal. One area for improvement in regards to these models is evaluation. Throughout this work I have proposed different methods of evaluating inferred values for unobserved phenomena, but this work could be improved, for example, with the acquisition of high-quality labels.

There are many directions in which this work might be expanded. Throughout these problems a persistent question is how to apply trained models to unseen datasets. Especially when training data is small, it is unclear to what extent trained methods will generalize. Much of this work can be expanded to adapt to different data settings and levels of supervision. Another question which I left largely unaddressed is how to model hierarchical structure in these settings. Such structure could be relevant in many of these problems, for example spatial and temporal hierarchies might be useful in many of these problems.

Computational methods can offer much to problems of societal significance. Here, I have shown the crucial role of utilizing structure in a collective probabilistic approach to problems in sustainability, education, and malicious behavior. However, if computational methods are to be broadly successful in providing human decision makers with data-driven insights there is much work to be done. For such methods to be able to aid society in addressing its greatest ills, the members of society must reach a greater capacity for trusting and understanding what these methods can and should do. If data is collected in a manner which subjects deem unacceptable, their discomfort will be a limiting factor in the benefits of any gleaned insights. Furthermore, the potentials of model transparency will be limited by the technical literacy of intended audiences. I hope to explore these concerns in my future work. Both in conducting research which improves method interpretability, and in implementing user studies to determine which

kinds of results are most useful in decision making. Ultimately, such work should not only improve the practical applicability of computational methods, but provide insights to guide their improvement towards societally beneficial ends.

Bibliography

[1] Climate change, human security and violent conflict. *Political Geography*, 26(6), 2007.

[2] Energy Information Administration. Electric power annual. 2016. `http://www.eia.gov/electricity/annual/html/epa_01_01.html`.

[3] Robert Agnew. Dire forecast: A theoretical model of the impact of climate change on crime. *Theoretical Criminology*, 16(1), 2012.

[4] Hamidreza Alvari, Paulo Shakarian, and JE Kelly Snyder. A non-parametric learning approach to identify online human trafficking. In *Conference on Intelligence and Security Informatics (ISI)*, 2016.

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

[6] Terry Anderson. Getting the mix right again: An updated and theoretical rationale for interaction. *The International Review of Research in Open and Distributed Learning*, 4(2), 2003.

[7] Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. Talk to me: Foundations for successful individual-group interactions in online communities. In *Conference on Human Factors in Computing Systems (CHI)*, 2006.

[8] Shahzeen Z. Attari, Michael L. DeKay, Cliff I. Davidson, and Wndi Bruine de

Bruin. Public perceptions of energy consumption and savings. *Proceedings of the National Academy of Sciences*, 107(37), 2010.

[9] Reuven Aviv, Zippy Erlich, Gilad Ravid, and Aviva Geva. Network analysis of knowledge construction in asynchronous learning networks. *Journal of Asynchronous Learning Networks*, 7(3), 2003.

[10] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. 2015.

[11] Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research (JMLR)*, 2017.

[12] Stephen H. Bach, Bert Huang, Jordan Boyd-Graber, and Lise Getoor. Paired-dual learning for fast training of latent variable hinge-loss mrfs. In *International Conference on Machine Learning (ICML)*, 2015.

[13] Ryan Baker, Albert Corbett, Kenneth Koedinger, and Angela Wagner. Off-task behavior in the cognitive tutor classroom: When students "game the system". In *Conference on Human Factors in Computing Systems (CHI)*, 2004.

[14] Karl E. Bauman and Ennett Susan T. On the importance of peer influence for adolescent drug use: commonly neglected considerations. *Addiction*, 91(2), 1996.

[15] Amy Bellmore, Angela J. Calvin, Jun-Ming Xu, and Xiaojin Zhu. The five Ws of bullying on Twitter: Who, What, Why, Where, and When. *Computers in Human Behavior*, 44(C), 2015.

[16] Frances P. Bernat and Tatyana Zhilina. Human trafficking: The local becomes global. *Women & Criminal Justice*, 20(1-2), 2010.

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

[18] Jan Blom, Himanshu Verma, Nan Li, Afroditi Skevi, and Pierre Dillenbourg. Moocs are more social than you believe. Technical report, eLearning Papers, 2013.

[19] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 2011.

[20] Gwendolyn Brandon and Alan Lewis. Reducing household energy consumption: A qualitative and quantitative field study. *Journal of Environmental Psychology*, 19(1), 1999.

[21] Robin Burke. *Hybrid Web Recommender Systems*. 2007.

[22] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *IEEE International Conferences on Privacy, Security, Risk and Trust and on Social Computing (SOCIALCOM-PASSAT)*, 2012.

[23] Yunfei Chen, Lanbo Zhang, Aaron Michelony, and Yi Zhang. 4is of social bully filtering: identity, inference, influence, and intervention. In *International Conference on Information and Knowledge Management (CIKM)*, 2012.

[24] Al Cooper. Sexuality and the internet: Surfing into the new millennium. *CyberPsychology & Behavior*, 1(2), 1998.

[25] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3), 1995.

[26] Sarah Darby. The effectiveness of feedback on energy consumption. A Review for DERFA of the Literature on Metering, Billing, and direct Displays. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays*, 486(2006), 2006.

[27] Christophe Demarque, Laetitia Charalambides, Denis J Hilton, and Laurent Waroquier. Nudging sustainable consumption: The use of descriptive norms to promote a minority behavior in a realistic online shopping environment. *Journal of Environmental Psychology*, 43, 2015.

[28] Thomas Dietz, Gerald T. Gardner, Jonathan Gilligan, Paul C. Stern, and Michael P. Vandenbergh. Household actions can provide a behavioral wedge to

rapidly reduce us carbon emissions. *Proceedings of the National Academy of Sciences*, 106(44), 2009.

[29] Herbert B Dixon Jr. Human trafficking and the internet (and other technologies, too). *Judges J.*, 52(1), 2013.

[30] Artur Dubrawski, Kyle Miller, Matthew Barnes, Benedikt Boecking, and Emily Kennedy. Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking*, 1(1), 2015.

[31] A. Felfernig and R. Burke. Constraint-based recommender systems: Technologies and research issues. In *International Conference on Electronic Commerce (ICEC)*, 2008.

[32] Christopher J. Ferguson, Claudia San Miguel, and Richard D. Hartley. A multivariate analysis of youth violence and aggression: The influence of family, peers, depression, and media violence. *The Journal of Pediatrics*, 155(6), 2009.

[33] Elisha R. Frederiks, Karen Stenner, and Elizabeth V. Hobman. Household energy use: Applying behavioural economics to understand consumer decision-making and behaviour. *Renewable and Sustainable Energy Reviews*, 41, 2015.

[34] Patxi Galán-García, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying. 2014.

[35] Gerald T. Gardner and Paul C. Stern. The short list: The most effective actions U.S. households can take to curb climate change. *Environment: Science and Policy for Sustainable Development*, 50(5), 2008.

[36] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden markov models. *Machine Learning*, 1997.

[37] Fabio Giglietto, Luca Rossi, and Davide Bennato. The open laboratory: Limits and possibilities of using facebook, twitter, and youtube as a research data source. *Journal of Technology in Human Services*, 30(3-4), 2012.

[38] Andrew Gilg, Stewart Barr, and Nicholas Ford. Green consumption or sustainable lifestyles? identifying the sustainable consumer. *Futures*, 37(6), 2005.

[39] F. S. Gohari and M.J. Tarokh. Classification and comparison of the hybrid collaborative filtering systems. *International Journal of Research in Industrial Engineering*, 6(2), 2017.

[40] Jo Goodey. Human trafficking: Sketchy data and policy responses. *Criminology & Criminal Justice*, 8(4), 2008.

[41] Elzbieta M. Gozdziak and Elizabeth A. Collett. Research on human trafficking in north america: A review of literature. *International Migration*, 43(1/2), 2005.

[42] Andrew P. Guth. Human trafficking in the Philippines: the need for an effective anti-corruption program. *Trends in Organized Crime*, 13(2), 2010.

[43] Allyson Fiona Hadwin, Sanna Järvelä, and Mariel Miller. Self-regulated, co-regulated, and socially shared regulation of learning. *Handbook of self-regulation of learning and performance*, 30:65–84, 2011.

[44] George W. Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 1992.

[45] Nathan Oken Hodas, Farshad Kooti, and Kristina Lerman. Friendship paradox redux: Your friends are more interesting than you. In *International Conference On Web And Social Media (ICWSM)*, 2013.

[46] Homa Hosseinmardi, Amir Ghasemianlangroodi, Richard Han, Qin Lv, and Shivakant Mishra. Towards understanding cyberbullying behavior in a semi-anonymous social network. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2014.

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

[48] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. Cyber bullying detection using social and textual analysis. In *International Workshop on Socially-Aware Multimedia (SAM)*, 2014.

[49] Nicolas Hug. Surprise, a Python library for recommender systems. `http://surpriselib.com`, 2017.

[50] Donna M. Hughes. Trafficking in human beings in the european union: Gender, sexual exploitation, and digital communication technologies. *SAGE Open*, 4(4), 2014.

[51] Renée Shaw Hughner, Pierre McDonagh, Andrea Prothero, Clifford J. Shultz, and Julie Stanton. Who are organic food consumers? a compilation and review of why people purchase organic food. *Journal of Consumer Behaviour*, 6(2-3), 2007.

[52] Clayton J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *International Conference On Web And Social Media (ICWSM)*, 2014.

[53] Michelle Ibanez and Daniel D. Suthers. Detection of domestic human trafficking indicators and movement trends using content available on open internet sources. In *Hawaii International Conference on System Sciences (HICSS)*, 2014.

[54] Pecan Street Inc. Dataport. Accessed: 2016.

[55] Matthew J. Johnson and Alan S. Willsky. Bayesian nonparametric hidden semi-Markov models. *Journal of Machine Learning Research*, 14(Feb), 2013.

[56] Jaana Juvonen and Elisheva F. Gross. Extending the school grounds? – Bullying experiences in cyberspace. *Journal of School Health*, 78(9), 2008.

[57] Kristiina Kangaspunta. *Collecting Data on Human Trafficking: Availability, Reliability and Comparability of Trafficking Data*. 2007.

[58] Rahul Kapoor, Mayank Kejriwal, and Pedro Szekely. Using contexts and constraints for improved geotagging of human trafficking webpages. In *Workshop on Managing and Mining Enriched Geo-Spatial Data*, 2017.

[59] Catarina Katzer, Detlef Fetchenhauer, and Frank Belschak. Cyberbullying: Who Are the Victims? *Journal of Media Psychology*, 21(1), 2009.

[60] Shaun Kellogg, Sherry Booth, and Kevin Oliver. A social network perspective on peer supported learning in moocs for educators. *The International Review of Research in Open and Distributed Learning*, 15(5), 2014.

[61] Jack Kelly and William Knottenbelt. Neural NILM: Deep neural networks applied to energy disaggregation. In *International Conference on Embedded Systems for Energy-Efficient Built Environments (BuildSys)*, 2015.

[62] Gregor Kennedy, Carleton Coffrin, Paula de Barba, and Linda Corrin. Predicting success: How learners' prior knowledge, skills and activities predict MOOC performance. In *Conference on Learning Analytics And Knowledge (LAK)*, 2015.

[63] Hyungsul Kim, Manish Marwah, Martin Arlitt, Geoff Lyon, and Jiawei Han. Unsupervised disaggregation of low frequency power measurements. In *International Conference on Data Mining*. 2011.

[64] Youngsoo Kim, Felicia Natali, Feida Zhu, and Eepeng Lim. Investigating the influence of offline friendship on twitter networking behaviors. In *Hawaii International Conference on System Sciences (HICSS)*, 2016.

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

[66] J. Zico. Kolter and Tommi Jaakkola. Approximate inference in additive factorial HMMs with application to energy disaggregation. In *International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, 2012.

[67] J Zico Kolter and Matthew J Johnson. Redd: A public data set for energy disaggregation research. *KDD Workshop on Data Mining Applications in Sustainability*, 2011.

[68] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Conference on Knowledge discovery and data mining (SIGKDD)*, 2008.

[69] Pigi Kouki, Shobeir Fakhraei, James Foulds, Magdalini Eirinaki, and Lise Getoor. HyPER: A flexible and extensible probabilistic framework for hybrid recommender systems. In *Conference on Recommender Systems (RecSys)*, 2015.

[70] Robin M. Kowalski and Susan P. Limber. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health*, 53(1, Supplement), 2013.

[71] Rohit Kumar and Carolyn P. Rosé. Architecture for building conversational agents that support collaborative learning. *IEEE Transactions on Learning Technologies*, 4(1), 2011.

[72] Jaakko Kurhila and Arto Vihavainen. A purposeful MOOC to alleviate insufficient cs education in finnish schools. *Transactions of Computing Education*, 15(2), 2015.

[73] Henning Lange and Mario Bergés. BOLT: Energy disaggregation by online binary matrix factorization of current waveforms. In *International Conference on Systems for Energy-Efficient Built Environments (BuildSys)*, 2016.

[74] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning (ICML)*, 2014.

[75] Liangda Li and Hongyuan Zha. Household structure analysis via Hawkes processes for enhancing energy disaggregation. In *International Conference on Artificial Intelligence (IJCAI)*. 2016.

[76] Nan Li, Himanshu Verma, Afroditi Skevi, Guillaume Zufferey, Jan Blom, and Pierre Dillenbourg. Watching moocs together: investigating co-located mooc study groups. *Distance Education*, 35(2), 2014.

[77] Qing Li. New bottle but old wine: A research of cyberbullying in schools. *Computers in Human Behavior*, 23(4), 2007.

[78] Yiping Lou, Philip C. Abrami, and Sylvia d'Apollonia. Small group and individual learning with technology: A meta-analysis. *Review of Educational Research*, 71(3), 2001.

[79] Stephen Makonin, Fred Popowich, Ivan V. Bajic, Bob Gill, and Lyn Bartram. Exploiting HMM sparsity to perform online real-time nonintrusive load monitoring. *IEEE Transactions on Smart Grid*, 2015.

[80] Andrew McCallum. MALLET: A machine learning for language toolkit. 2002.

[81] Anthony J McMichael, Rosalie E Woodruff, and Simon Hales. Climate change and human health: present and future risks. *The Lancet*, 367(9513), 2006.

[82] Bock Mike and O'Dea Victoria. Virtual educators critique value of MOOCs for K-12. 2013.

[83] Kevin Murphy. Hidden Markov Model (HMM) Toolbox for Matlab, 1998.

[84] Chirag Nagpal, Kyle Miller, Benedikt Boecking, and Artur Dubrawski. An entity resolution approach to isolate instances of human trafficking online. *arXiv preprint arXiv:1509.06659*, 2015.

[85] Hedieh Najafi, Rosemary Evans, and Christopher Federico. MOOC integration into secondary school courses. *The International Review of Research in Open and Distributed Learning*, 15(5), 2014.

[86] Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3), 2004.

[87] Piper Nicola. A problem by a different name? a review of research on trafficking in South?East Asia and Oceania. *International Migration*, 43(1-2).

[88] Krittinee Nuttavuthisit and John Thøgersen. The importance of consumer trust for the emergence of a market for green products: The case of organic food. *Journal of Business Ethics*, 140(2), 2017.

[89] The United States Department of State. Trafficking in persons report. https://www.state.gov/j/tip/rls/tiprpt/, 2017. Accessed: 05-12-2018.

[90] Yuko Onozaka, Gretchen Nurse, and Dawn Thilmany McFadden. Defining sustainable food market segments: Do motivations and values vary by shopping locale? *American Journal of Agricultural Economics*, 93(2), 2011.

[91] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, and Nathan Schneider. Part-of-speech tagging for twitter: Word clusters and other advances. *Technical Report, Machine Learning Department. CMU-ML-12-107.*, 2012.

[92] Oliver Parson, Siddhartha Ghosh, Mark Weal, and Alex Rogers. Non-intrusive load monitoring using prior models of general appliance types. In *Artificial Intelligence (AAAI-12)*, 2012.

[93] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[94] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2014.

[95] Josephine Pickett-Baker and Ritsuko Ozaki. Pro-environmental products: marketing influence on consumer purchase decision. *Journal of Consumer Marketing*, 25(5), 2008.

[96] Polaris. The facts. https://polarisproject.org/human-trafficking/facts, 2016. Accessed: 04-12-2018.

[97] Polaris. Hotline statistics. https://polarisproject.org/2017statistics, 2017. Accessed: 04-12-2018.

[98] Rebecca S Portnoff, Danny Yuxing Huang, Periwinkle Doerfler, Sadia Afroz, and Damon McCoy. Backpage and bitcoin: Uncovering human traffickers. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2017.

[99] Elaheh Raisi and Bert Huang. Cyberbullying detection with weakly supervised machine learning. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2017.

[100] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. Learning latent engagement patterns of students in online courses. In *International Conference on Artificial Intelligence (AAAI)*, 2014.

[101] Matthew Ranson. Crime, weather, and climate change. *Journal of Environmental Economics and Management*, 67(3), 2014.

[102] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *LREC Workshop on New Challenges for NLP Frameworks*, 2010.

[103] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In *Conference on Machine Learning and Applications (ICMLA)*, 2011.

[104] Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 3(3), 2008.

[105] Carolyn P Rosé, Pam Goldman, Jennifer Zoltners Sherer, and Lauren Resnick. Supportive technologies for group discussion in moocs. *Current Issues in Emerging eLearning*, 2(1), 2015.

[106] Carolyn Penstein Rosé, Ryan Carlson, Diyi Yang, Miaomiao Wen, Lauren Resnick, Pam Goldman, and Jennifer Sherer. Social factors that contribute to attrition in moocs. In *Conference on Learning @ Scale Conference(L@S)*, 2014.

[107] Sustainable Consumption Roundtable. I will if you will: Towards sustainable consumption, 2006.

[108] Allison M Ryan and Helen Patrick. The classroom social environment and changes in adolescents motivation and engagement during middle school. *American Educational Research Journal*, 38(2), 2001.

[109] Alan Said and Alejandro Bellogín. Comparative recommender system evaluation: Benchmarking recommendation frameworks. In *Conference on Recommender Systems (RecSys)*, 2014.

[110] Christina Salmivalli. Participant role approach to school bullying: implications for interventions. *Journal of Adolescence*, 22(4), 1999.

[111] Jyoti Sanghera. Unpacking the trafficking discourse. In *Trafficking and prostitution reconsidered*. 2017.

[112] Edward J. Schauer and Elizabeth M. Wheaton. Sex trafficking into the United States: A literature review. *Criminal Justice Review*, 31(2), 2006.

[113] P Wesley Schultz, Jessica M Nolan, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius. The constructive, destructive, and reconstructive power of social norms. *Psychological science*, 18(5), 2007.

[114] John R Searle. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press, 1985.

[115] Kiarash Shaloudegi, András György, Csaba Szepesvari, and Wilsun Xu. Sdp relaxation with randomized rounding for energy disaggregation. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[116] Beth Simon, Julian Parris, and Jaime Spacco. How we teach impacts student learning: Peer instruction vs. lecture in cs0. In *Technical Symposium on Computer Science Education (SIGCSE)*, 2013.

[117] Robert Slonje, Peter K Smith, and Ann Frisén. The nature of cyberbullying, and strategies for prevention. *Computers in Human Behavior*, 29(1), 2013.

[118] Goggins S.P., Galyen K.D., Petakovic E., and Laffey J.M. Connecting performance to social structure and pedagogy as a pathway to scaling learning analytics in moocs: an exploratory study. *Journal of Computer Assisted Learning*, 32(3).

[119] Dhanya Sridhar, James Foulds, Marilyn Walker, Bert Huang, and Lise Getoor. Joint models of disagreement and stance in online debate. In *Association for Computational Linguistics (ACL)*, 2015.

[120] Paul C. Stern. Contributions of psychology to limiting climate change. *American Psychologist*, 66(4), 1989.

[121] Fabio Sticca and Sonja Perren. Is cyberbullying worse than traditional bullying? examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. *Journal of Youth and Adolescence*, 42(5), 2013.

[122] Gerrit St́okigt, Johannes Schiebener, and Matthias Brand. Providing sustainability information in shopping situations contributes to sustainable decision making: An empirical study with choice-based conjoint analyses. *Journal of Retailing and Consumer Services*, 43, 2018.

[123] Yu-Sheng Su, Chester S.J. Huang, and Ting-Jou Ding. Examining the effects of moocs learners social searching results on learning behaviors and learning outcomes. *Eurasia Journal of Mathematics, Science and Technology Education*, 12(9), 2016.

[124] Pedro Szekely, Craig A. Knoblock, Jason Slepicka, Andrew Philpot, Amandeep Singh, Chengye Yin, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, David Stallard, Subessware S. Karunamoorthy, Rajagopal Bojanapalli, Steven Minton, Brian Amanatullah, Todd Hughes, Mike Tamayo, David Flynt, Rachel Artiss, Shih-Fu Chang, Tao Chen, Gerald Hiebel, and Lidia Ferreira. Building and using a knowledge graph to combat human trafficking. In *International Semantic Web Conference (ISWC)*, 2015.

[125] Carmen Tanner and Sybille Wölfing Kast. Promoting sustainable consumption: Determinants of green purchases by swiss consumers. *Psychology & Marketing*, 20(10), 2003.

[126] Robert S. Tokunaga. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, 26(3), 2010.

[127] Sabina Tomkins, Golnoosh Farnadi, Brian Amantullah, Lise Getoor, and Steven Minton. The impact of environmental stressors on human trafficking. In *Beyond Online Data ICWSM Workshop*, 2018.

[128] Sabina Tomkins, Golnoosh Farnadi, Brian Amantullah, Lise Getoor, and Steven Minton. The impact of environmental stressors on human trafficking. In *International Conference on Data Mining (ICDM)*, 2018.

[129] Sabina Tomkins, Lise Getoor, Yunfei Chen, and Yi Zhang. A socio-linguistic approach for cyberbullying detection. In *Advances in Social Networks Analysis and Mining (ASONAM)*, 2018.

[130] Sabina Tomkins, Steve Isley, Ben London, and Lise Getoor. Sustainability at scale: Bridging the intention-behavior gap with sustainable recommendations. In *Recommender Systems (RecSys)*, 2018.

[131] Sabina Tomkins, Jay Pujara, and Lise Getoor. Disambiguating energy disaggregation: A collective probabilistic approach. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[132] Sabina Tomkins, Arti Ramesh, and Lise Getoor. Predicting post-test performance from online student behavior: A high school mooc case study. In *International Conference on Educational Data Mining (EDM)*, 2016.

[133] B. Towle and C. Quinn. Knowledge based recommender systems using explicit user models. In *AAAI Technical Report*, 2000.

[134] Conrad Tucker, Barton K. Pursel, and Anna Divinsky. Mining student-generated textual data in MOOCs and quantifying their effects on student performance and learning outcomes. In *Conference of American Society for Engineering Education*, 2014.

[135] Abdulbaki Uzun, Christian Räck, and Fabian Steinert. Targeting more relevant, contextual recommendations by exploiting domain knowledge. In *Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec)*, 2010.

[136] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. Detection and fine-grained classification of cyberbullying events. In *Recent Advances in Natural Language Processing (RANLP)*, 2015.

[137] Jeannet H. Van Houwelingen and W. Fred Van Raaij. The effect of goal-setting and daily electronic feedback on in-home energy use. *Journal of Consumer Research*, 16(1), 1989.

[138] George Veletsianos, Amy Collier, and Emily Schneider. Digging deeper into learners' experiences in moocs: Participation in social networks outside of moocs, notetaking and contexts surrounding content consumption. *British Journal of Educational Technology*, 46(3), 2015.

[139] Iris Vermeir and Wim Verbeke. Sustainable food consumption: Exploring the consumer attitude–behavioral intention gap. *Journal of Agricultural and Environmental ethics*, 19(2), 2006.

[140] Lorenzo Vigentini and Andrew Clayphan. Pacing through MOOCs: course design or teaching effect? In *Conference on Educational Data Mining (EDM)*, 2015.

[141] John Walsh and Donald Wuebbles. National climate assessment. `https://nca2014.globalchange.gov/report`, 2014. Accessed: 04-12-2018.

[142] Xu Wang, Diyi Yang, Miaomiao Wen, Kenneth Koedinger, and Carolyn P Rosé. Investigating how student's cognitive behavior in mooc discussion forums affect learning gains. *International Educational Data Mining Society*, 2015.

[143] Miaomiao Wen, Diyi Yang, and Carolyn Penstein Rosé. Virtual teams in massive open online courses. In *International Conference on Artificial Intelligence in Education*, 2015.

[144] Elizabeth Whittaker and Robin M. Kowalski. Cyberbullying via social media. *Journal of School Violence*, 14(1), 2015.

[145] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), 2012.

[146] Dawei Yin, Brian D. Davison, Zhenzhen Xue, Liangjie Hong, April Kontostathis, and Lynne Edwards. Detection of Harassment on Web 2.0. In *Content Analysis in the Web 2.0 (CAW2.0)*, 2009.

[147] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In *International Conference on Distributed Computing and Networking (ICDCN)*, 2016.