

PERSONALIZED SEARCH AND RECOMMENDATION FOR HEALTH INFORMATION RESOURCES

A Thesis
Presented to
The Academic Faculty

by

Steven P. Crain

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Computational Science and Engineering

Georgia Institute of Technology
December 2012

PERSONALIZED SEARCH AND RECOMMENDATION FOR HEALTH INFORMATION RESOURCES

Approved by:

Professor Hongyuan Zha, Advisor
School of Computational Science and
Engineering
Georgia Institute of Technology

Professor Eugene Agichtein
Mathematics and Computer Science
Department
Emory University

Professor Mark L. Braunstein
School of Interactive Computing
Georgia Institute of Technology

Professor Amy S. Bruckman
School of Interactive Computing
Georgia Institute of Technology

Professor Alexander Gray
School of Computational Science and
Engineering
Georgia Institute of Technology

Date Approved: August 23, 2012

To God,
with Whom all things are possible,
to my wife Andrea,
who has invested her life into me,
and to my daughters Jennifer and Aileen,
who eagerly followed my research.

This work of five years
would not have been possible
without you all.

ACKNOWLEDGEMENTS

Dr. Hongyuan Zha always provided helpful guidance. I would like to thank my collaborators, Dr. Shabbir Syed-Abdul, Dr. Jiang Bian, Dr. Joshua Dillon, Dr. Jian Huang, Luis Fernandez-Luque, Dr. Dan Pelleg, Xin Sun, Dr. Elad Yom-Tov, Dr. Ingmar Weber, Dr. Shuang-Hong Yang, Yanjun Zhao and Ke Zhou. I am also grateful for the mentoring I received as an intern with Dr. Mike Jones at Mitsubishi Electric Research Lab, Dr. Cathy Jiao at Oak Ridge National Lab and Dr. Lei Duan at Microsoft. It has been a joy to work with them.

This research has been partially supported by an Oak Ridge Computational Science and Engineering Fellowship, a Department of Homeland Security Career Development Grant, National Science Foundation grant IIS-1116886, Norwegian Research Council project 174934 and grants from Hewlett-Packard, Microsoft and Tromsø Forskningsstiftelse.

Finally, I am very grateful to Manny Hernandez, president of the Diabetes Hands Foundation, for allowing us to conduct research with his users and for many profitable discussions. Much of this work would not have been possible without his patient support.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS	iv
LIST OF TABLES	viii
LIST OF FIGURES	ix
SUMMARY	x
PROLOGUE	1
I INTRODUCTION	3
1.1 Concepts	8
1.1.1 Health Information Resource	8
1.1.2 Consumer and User	8
1.1.3 Language Gap	8
1.1.4 Topic	9
1.1.5 Quality Health Content	10
1.1.6 Event	10
1.1.7 Group Vitality	11
1.1.8 Social Recommendation	11
1.2 Notation	12
II PREVIOUS WORK	13
2.1 Latent Topic Models	13
2.1.1 LDA Model	15
2.1.2 Collapsed Gibbs	20
2.1.3 Variational Approximation	21
2.1.4 Interpretation and Evaluation	23
2.1.5 Parameter Selection	26
2.1.6 Adapting Topic Models to Applications	26

2.2	Quality of Online Health Information	28
2.3	Survival and Event History Analysis	31
2.3.1	Survival Analysis	31
2.3.2	Event History Analysis	35
2.4	Groups Interactions	36
2.5	Latent Behavior Models	36
III	LANGUAGE GAPS IN HEALTH INFORMATION RETRIEVAL	41
3.1	Introduction	41
3.2	Topic Models	42
3.3	Dialect Topic Models	44
3.3.1	Algorithm	46
3.3.2	Implementation	49
3.4	Topic-Adapted Latent Dirichlet Allocation	50
3.4.1	Algorithm	51
3.5	Experiments	52
3.5.1	Data Sets	52
3.5.2	Dialect Topic Models	53
3.5.3	Topic-Adapted Latent Dirichlet Allocation	57
IV	QUALITY OF MEDICAL CONTENT IN SOCIAL MEDIA . .	60
4.1	Introduction	60
4.2	Anorexia Content on YouTube	61
4.2.1	Methods	61
4.2.2	Results	63
4.3	Anorexia Communities at War	65
4.3.1	Methods	66
4.3.2	Results	69
4.3.3	Discussion	74
4.4	Conclusion	77

V	ONLINE HEALTH DISCUSSION GROUPS	78
5.1	Introduction	78
5.2	Activity in Online Health Communities	81
5.3	Event History Analysis for Social Networks	84
5.3.1	Censoring	86
5.3.2	Interpretation	87
5.4	Experiments	88
5.4.1	Results	89
5.5	Discussion	92
5.6	Conclusion	95
VI	SOCIAL RECOMMENDATION	96
6.1	Introduction	96
6.2	Social Recommendation	100
6.2.1	Encoding Social Events	100
6.3	Experiments	102
6.3.1	Results	105
6.4	Analysis	111
6.5	Conclusion	112
VII	FUTURE WORK	114
VIII	CONCLUSION	117
	PUBLICATIONS	119
	REFERENCES	120
	INDEX	136
	VITA	138

LIST OF TABLES

1	Effect of Dirichlet prior on example topic assignments.	18
2	Example topic models.	43
3	Example topics found by τ LDA.	57
4	Classification results for YouTube anorexia videos.	63
5	Engagement with YouTube anorexia videos.	63
6	Distinguishing tags for pro-recovery and pro-anorexia users on Flickr.	69
7	Cessation of posting anorexia videos following comments.	72
8	Hazard analysis results for cessation of posting anorexia videos.	73
9	Most significant shared features for group event dynamics.	89
10	Most significant features for group event dynamics by event type.	91
11	Group event data by event type.	103

LIST OF FIGURES

1	Diagram of the LDA graphical model.	15
2	Example topic model with three topics.	17
3	Diagram of the LDA variational model.	21
4	Diagram of the diaTM graphical and variational models.	46
5	Diagram of the τ LDA graphical model.	51
6	DiaTM evaluation results.	54
7	τ LDA evaluation results.	59
8	YouTube anorexia videos data extraction.	62
9	Typical pro-anorexia videos.	64
10	Demographics of a YouTube anorexia video.	65
11	Anatomy of a Flickr page.	66
12	Visualizations of network graphs for pro-recovery and pro-anorexia interactions.	71
13	Number of groups in TuDiabetes by week.	80
14	Group event intensity by user age.	84
15	CPMF performance by encoding.	106
16	CPMF-T performance by latent dimensionality.	107
17	PCPMF-T performance by cross-task sharing.	108
18	CPMF and PCPMF performance by task and amount of training data.	109

SUMMARY

Consumers face several challenges using the Internet to fill health-related needs. (1) In many cases, they face a language gap as they look for information that is written in unfamiliar technical language. (2) Medical information in social media is of variable quality and may be appealing even when it is dangerous. (3) Discussion groups provide valuable social support for necessary lifestyle changes, but are variable in their levels of activity. (4) Finding less popular groups is tedious. We present solutions to these challenges.

We use a novel adaptation of topic models to address the language gap. Conventional topic models discover a set of unrelated topics that together explain the combinations of words in a collection of documents. We add additional structure that provides relationships between topics corresponding to relationships between consumer and technical medical topics. This allows us to support search for technical information using informal consumer medical questions.

We also analyze social media related to eating disorders. A third of these videos promote eating disorders and consumers are twice as engaged by these dangerous videos. We study the interactions of two communities in a photo-sharing site. There, a community that encourages recovery from eating disorders interacts with the pro-eating disorder community in an attempt to persuade them, but we found that this attempt entrenches the pro-eating disorder community more firmly in its position.

We study the process by which consumers participate in discussion groups in an online diabetes community. We develop novel event history analysis techniques to identify the characteristics of groups in a diabetes community that are correlated with consumer activity. This analysis reveals that uniformly advertise the popular

groups to all consumers impairs the diversity of the groups and limits their value to the community.

To help consumers find interesting discussion groups, we develop a system for personalized recommendation for social connections. We extend matrix factorization techniques that are effective for product recommendation so that they become suitable for implicit power-law-distributed social ratings. We identify the best approaches for recommendation of a variety of social connections involving consumers, discussion groups and discussions.

Prologue

Jeremy slammed the phone down in frustration and pushed his chair back from the desk, scattering a stack of bills. The earth might as well have opened up and swallowed him when Dr. Jacobs said he had some kind of diabeets [*sic*]. While the doctor droned on about eating something or other, Jeremy had just stared blankly ahead and mechanically assured his doctor he understood it all. What was he going to do? Wasn't diabeets a disease for old women? Troubling thoughts darted around his head like a flock of noisy starlings until he couldn't think anymore.

Melinda appeared at the top of the stairs, "What is it honey?" she asked with the mild southern accent that normally sent his heart racing. Now, however, he couldn't respond. Would life ever be the same again?

Over the next few weeks, Jeremy visited his doctor, who took the time to explain the changes he would need to make in his diet and exercise because of the type II diabetes. He also enrolled in some classes at the hospital that gave him a better understanding of how blood sugar was effected by diet and exercise and helped him feel more comfortable with the disease. Often he would have to work late, and then would find himself hungry and irritable, driving through rush hour traffic to try to get to a class about a disease he should never have. After the regimen of classes ended, the instructor encouraged them to join a support community, but Jeremy decided the hassle was too much.

Melinda was a great help as he learned to cope with his disease. She spent hours searching the Internet for information so that she could adjust the meals she cooked. Sometimes she would find the search frustrating because it was hard to find much in-depth information about diabetes. Most of what was available was very brief, and some of it seemed questionable even though it claimed to be by a doctor. She felt sad that she couldn't help her husband more—diabetes was foreign to her, too.

After a few months, Melinda wrote something about their diabetes struggles on Facebook. A friend suggested that they try an online diabetes community that had really helped her. Melinda read through the recent posts in the public forums, and was surprised to discover how many people were going through the same trials. She excitedly called her husband at work to tell him all about it.

To Jeremy, who was not feeling at all well after eating what should have been a harmless pasta salad at lunch, she seemed way over-excited. Eventually, he started to get involved with the online community and quickly learned many things that hadn't really made sense to him in the classes. He also found that when he was feeling depressed he could write about it online and get back dozens of encouraging replies.

While Jeremy was exploring the website, he noticed that the community had some groups that looked interesting. One about environmental concern caught his attention, but it was disappointing when he realized that the group only had wall posts from a handful of people. Jeremy gave up on the groups with a sigh. It was time to get back to work, anyway.

CHAPTER I

INTRODUCTION

Like Jeremy and Melinda, millions of Americans search for health information online. The TickerSM survey [128], which tracks consumer health-related behavior, found that 20% of Americans tap into on-line social health resources. Consumers with household incomes above \$75,000 were particularly likely to include online social networks in their quest for health information. A Pew survey found that low-income and ethnic Americans were also increasingly seeking health information on-line [138].

Consumers who seek health information on-line improve their health literacy: they become more involved in managing their own health and communicate better with their doctors [90]. They are also more likely to talk about embarrassing health problems in an on-line social group [179]. However, consumers are only able to access the small fraction of available resources that are easy to find, for example the TickerSM study found that 94% of respondents accessed health information through Facebook (<http://www.facebook.com>). In addition, although the group a consumer participates in makes a huge difference in outcome [88, 177, 132, 114], no technology has been developed to help consumers screen the many possible resources for a personal fit. There is a great need for tools to support consumers finding health-related social resources that are currently hard to find.

We make four major contributions in this thesis. (1) We develop structured topic models that are able to bridge the language gap between consumers and the medical information they seek. (2) We analyze the eating-disorder-related content in social media in order to better understand the motivations behind contributing and seeking hazardous medical content. (3) We study the vitality of online diabetes discussion

groups and gain insight into why only a small number of groups have substantial activity. (4) We investigate algorithms for making personalized recommendations for social attachments. In the following, we summarize the major contributions of the thesis chapter by chapter.

Chapter 2. Previous Work. We describe the state-of-the-art for the techniques and algorithms upon which we base our work. (1) Our work on the language gap builds on prior work on topic modeling, especially latent Dirichlet allocation (LDA) and its extensions to support multiple languages. We present the LDA models in detail, discuss how these models are evaluated and interpreted and survey the related research. (2) Our work on the quality of medical content in social media advances work done by other researchers. We survey both research that is specifically related to eating disorders and research addressing similar issues with other diseases. We also discuss other research related to the quality of online information. (3) Event history analysis is an important tool in the analysis of interactions in both the eating disorder-related communities and in the vitality of diabetes groups. We describe the process of event history analysis, including relevant generative models of events. We also discuss their interpretation and survey the related research. (4) Matrix factorization provides the foundation for personalized recommendation in a health community. We describe the state-of-the-art techniques and survey the related research.

Chapter 3. Language Gaps In Consumer Health Information Retrieval. Consumers often find it difficult searching on-line for health information. Some health information is designed with consumers in mind, such as that provided by Medline Plus® (<http://www.nlm.nih.gov/medlineplus>), the Centers for Disease Control and Prevention (<http://www.cdc.gov>) and WebMD® (<http://www.webmd.com>). These sites are designed to be understandable by a wide variety of consumers, but a language gap makes it hard for consumers to find relevant information even on

consumer-targeted sites. Ordinary consumers are usually unfamiliar with the medical concepts, so they use a wide variety of words to attempt to describe their situations when forming queries and questions—consumer-oriented health information may be understandable, but it cannot cover the expressive range of consumers and situations. Often consumers want more detailed information than consumer-targeted sites provide, but relevant technical resources like journal papers use technical language and unfamiliar concepts with subtle distinctions.

Topic models are routinely used when searching for information when the precise words used in the target document are uncertain. A topic model takes a query or document that is represented by words and translates it into a new space of topics that are intended to represent human ideas. There are many algorithms that learn a topic model automatically from an unlabeled collection of training documents.

For example, LDA starts with a random topic model and iteratively refines it to find a model that provides the best explanation for the words that appear together in the training documents. Each topic is associated with a pool of words that can all be used interchangeably. Consequently, LDA works well when each document containing a particular concept shares a common probability for each word. For example, the words “sugar” and “glucose” are essentially interchangeable when talking about diabetes, so LDA would easily learn that they belong in the same topic. On the other hand, LDA does very poorly when there are two different sets of words for the same topic that do not mix together. A document collection in multiple languages presents the extreme case: few documents will mix words across languages even though the concept is the same, so each topic that LDA learns contains only words from one language.

Polylingual LDA (pLDA) introduces parallel topics in multiple languages [119], so that the same topic can be expressed in different words in each language. It is able to learn the model by making use of pairs of training documents in different

languages but in the same topic. The language gap between consumers and medical information has a similar nature, except that there is no ready source of parallel training documents. Instead, we make use of a large collection of documents at a variety of different technicalities. We enforce the requirement that each topic exist at multiple technicalities and the intermediate-technicality documents provide the bridge that connects the topics in the more common extreme-technicality documents. We demonstrate two different models with this characteristic: dialect topic models (diaTM) use three versions of each topic and features that signal which version to use; topic adapted LDA (τ LDA) uses two versions of each topic and minimal supervision.

We train and evaluate these models using documents from consumer question answering forums, consumer-oriented health sites and technical medical documents. DiaTM explained testing documents 60% better than LDA and performed 90% better for finding medical documents related to consumer questions. Meanwhile, τ LDA was 240% better than LDA for finding the medical documents and is also useful for estimating the level of technicality for a document.

Chapter 4. Quality of Medical Content in Social Media. Online health information is very valuable to consumers, but it comes with hazard: dubious and even dangerous health information abounds on the Internet. We examine online communities related to eating disorders. In particular, people who suffer from *anorexia nervosa* avoid eating food even after reaching an unhealthy weight. The “proana” community actively promotes anorexia as a legitimate lifestyle and provides support to young women who want to lose weight at any cost. In contrast, the “pro-recovery” community actively encourages people suffering from eating disorders to get help. We measure the volume of proana and pro-recovery information on several online social sites, finding that about a third of the anorexia-related content on social media promotes anorexia. Moreover, the viewers of proana content are much more engaged with the content, being twice as likely to comment on it, mark it as a favorite or

indicate that they like it. We also analyze the interactions between the proana and pro-recovery communities. The pro-recovery community wages an active campaign to persuade the other through comments left on proana social media. Using survival analysis, we show that the attempts to persuade the proana community actually work to entrench them further in their position.

Chapter 5. Online Health Discussion Groups. We study the interactions that take place in the groups and forums of an online diabetes community. We first use event history analysis to identify the factors that influence the amount of activity in each group. We develop novel extensions to support dynamic, partially observed and highly correlated factors. Our use of singular value decomposition (SVD) to disambiguate the effect of factor correlation may have wide-spread value in event history analysis. We found that making personalized recommendations of discussion groups to consumers would enhance the long-term sustainability of the community.

Chapter 6. Social Recommendation. We extend recommendation technology so that it can make interaction-based recommendations. Conventional matrix factorization uses a model that results in an approximately Gaussian distribution of predicted ratings, which is foreign to the inherently power-law distribution of ratings implicitly derived from social interactions. We develop two formulations that transform between the two probability spaces and demonstrate that this technique is essential for stability and accuracy. We also evaluate a variety of different variants of matrix factorization to identify the most appropriate algorithms for multiple important social recommendation tasks.

Chapter 7. Future Work. The results we present in this thesis provide the opportunity for significant future work.

Chapter 8. Conclusion. We review our contributions to computer-mediated health resource finding and discuss the implications for assisting consumers in their quest for online health information resources.

1.1 Concepts

1.1.1 Health Information Resource

The focus of this research is on facilitating access to any kind of online resource that pertains to *health*, including disease management and prevention. We likewise have an expansive definition of *information* that encompasses the kinds of health-related needs a person could have, including factual content, opinions, encouragement and mentoring. The *resources* we talk about include: Web pages with health content; articles from medical journals; social media, like blog posts, images or videos; comments on social media; other people—individuals and groups. Together, health information resources encompass the variety of online resources that might provide health-related benefits to a person.

1.1.2 Consumer and User

We focus on the needs of *consumers*, by which we mean people with at most limited relevant medical training. The label *consumer* is motivated by the imbalance of specialized knowledge that typically results in a producer-consumer relationship between health professionals and the general public. In order to emphasize the relationship between a consumer and a particular service, we will often refer to a *user* of the service. In this thesis, the user we are talking about is always a consumer.

1.1.3 Language Gap

Just as health professionals know more about health than consumers, they also think more about it and talk about it continuously as part of their work. To facilitate this, they have well-defined *concepts* that they share and associated language that makes it easy to talk about complex medical ideas and distinguish their nuances. Consumers, on the other hand, typically think and talk about health much less. They have correspondingly less developed health concepts that are appropriate for discussing and thinking about the most common and familiar health situations.

When faced with unfamiliar medical concepts, humans can compensate for the lack of specialized concepts and language by improvising from more common ideas and language. For example, a health professional may understand and articulate the concepts related to blepharospasm, including involuntary muscle movement and the muscular structures around the eye. One particular consumer faced this problem late one night and struggled to find words to express the twitching of his eyelids. He asked on a social question answering site, “Why my eyes (lashes) start bitting [*sic*] sometimes?”¹ This is a dramatic example of a common problem: the language that a consumer chooses for expressing a concept is very different from the language a professional would use. We call a systematic departure from how words are normally selected in a language to express the same meaning a *dialect*. A significant difference in dialect that impairs communication, as between the word selections of a consumer and a professional, is a *language gap*. More generally, a *cognitive gap* is any kind of difference at the language, conceptual or reasoning levels.

Synonymy refers to the situation where different words can be used to express the same concept. When a word can have multiple meanings, as is common when consumers express health concepts using common words, it is called *polysemy*.

1.1.4 Topic

When an author prepares a document, she chooses different words depending on the concepts that she writes about. For example, a sentence about diabetes requires different words than one about cancer. A *topic* is such a systematic pattern of word selection that is associated with a concept. For convenience, we sometimes refer to patterns of word selection as topics even without evidence that they correspond to meaningful concepts.

¹We do not further cite this quotation to maintain the privacy of this consumer.

1.1.5 Quality Health Content

The meaning of *quality* for health information is of some debate [118]. On the one hand, the medical profession may be too quick to discount some alternative health theories. On the other hand, some alternative health ideas, including eating disorders, have ample evidence that they are dangerous. In this thesis, we focus on the relatively clear situation around eating disorders, where we can confidently define quality content as content that encourages recovery from eating disorders instead of supporting eating disorders as a healthy alternative.

1.1.6 Event

An event is a potentially significant occurrence. If an event occurs, it occurs at a specific instance in time. If we can detect exactly when the event occurs, we call the event *observable*. Events often influence later events. For example, it is an event when a waiter spills coffee in a patron’s lap. This event is observable (we can hear the shriek), it happens at a specific time and it influences later events, especially the tip-leaving event. Events often have *participants*, like the waiter and the patron.

The events that we deal with are *interactions* between two participants. These events are characterized by the type of event (spill-coffee), the time of the event and the identity of the participants (waiter and patron). An event is naturally represented using a frame [121], with interpretations of the participant slots that depend on the type of event. For events where the time is unknown, we occasionally find it expedient to refer to the combination of event-type and specific participants as an event.

While we treat events as atomic occurrences, in reality they can be decomposed into a sequence of finer events. Our example event could be split into a sequence of events: the waiter reaches to clear a napkin from the table; the waiter tips the tray he is carrying; the coffee spills off the tray onto the patron; the patron emits a shriek. The appropriate granularity for events depends on: what can be observed;

hypotheses about the relationships between events; significance of events; complexity of recording and analyzing events; privacy; and the ability to meaningfully aggregate similar events. The reason that the waiter spilled the coffee is not important to the patron or to the proprietor, so a single event is quite appropriate in the example.

1.1.7 Group Vitality

The amount of benefit that members receive from a group is modulated by the amount and type of activity in the group. For example, a discussion group with substantial amounts of relevant discussions is often more useful than it would be with no discussions. We make the assumption that activity in a group is an investment on the part of the participating user, so that higher levels of activity are reflective of the participating users attaining higher utility from the group. (We can neglect undesirable activity only because it is not substantial in the groups we are studying.)

We thus get two different characteristics of a group that together approximate its usefulness or *vitality*. The characteristics of its active participants give a profile of the users who find the group valuable. This is useful for understanding the nature of the group vitality, even though it cannot be measured. The other characteristics of vitality is the amount of and kinds of activity in the group, which can be directly measured.

1.1.8 Social Recommendation

Traditionally, *recommendation* is suggesting a small set of items to a user, on the hope that some of the items will be useful to the user. This can help the user explore new items that she might otherwise overlook or it can help her focus her attention on the items that matter. In *social recommendation*, we instead suggest social connections or interactions. For example, in *group recommendation*, we suggest groups that a user might join. This should not be confused with suggesting items to a group of users, although we do touch on that in this work.

1.2 *Notation*

We represent matrices with bold uppercase letters, \mathbf{A} , vectors with bold lowercase letters, \mathbf{a} , and scalar parameters and indices with lowercase letters, a . Scalar constants are represented with capital letters, A . However, when we refer to a vector or scalar that is part of a matrix, we will maintain the capitalization, \mathbf{A}_{ij} . The transpose operator is represented using a superscript “T”, \mathbf{a}^T .

CHAPTER II

PREVIOUS WORK

In the later chapters of this thesis, we analyze consumer behavior and enhance their access to health information resources by building on a variety of existing techniques. In this chapter, we describe these techniques in detail and discuss the related research. Section 2.1 describes latent topic models, especially latent Dirichlet allocation. We explain what these models are, what makes them work and how they are implemented, interpreted and extended. Then, in Section 2.2, we discuss work related to the quality of online social content, with a focus on health-related aspects. We next present survival and event history analysis in Section 2.3, including the history, motivation, implementation and interpretation. Section 2.4 describes related work in analyzing and understanding the dynamics of discussion groups. Section 2.5 presents latent behavior models as they are applied to recommendation.

2.1 Latent Topic Models

Consumers searching for health information often face a formidable cognitive gap. Although much effort has been spent addressing cognitive gaps related to consumer health, they remain a crucial research challenge [43, 187, 145, 129, 152, 151]. [100, 22, 21, 171] take a very manual approach, educating clinicians and manually constructing communication scripts tailored for patients. Other researchers have attempted to translate technical documents into simpler language [139, 155] or to make a consumer’s query more technical [188, 39, 108, 33]. [39, 81, 166, 146] try to bridge a conceptual gap using an ontology. Some research has attempted to generate consumer-specific documents based on a detailed model of the consumer and intervention requirements [40]. Often electronic medical records have supplied a user model for

customized medical information [63, 65, 40, 165]. In contrast to other approaches attempting to circumvent language gaps, we bridge the gaps and expose the useful information contained in the gaps through structured topic models.

It is common to represent documents as a *bag of words* (BOW), accounting for the number of occurrences of each word but ignoring the order. This representation balances computational efficiency with the need to retain the document content. It also results in a vector representation that can be analyzed with techniques from applied mathematics and machine learning, notably dimension reduction, a technique that is used to identify a lower-dimensional representation of a set of vectors that preserves important properties. However, the BOW representation does not reflect correspondences between different synonymous words or disambiguate multiple meanings for a given word.

BOW vectors have a very high dimensionality — each dimension corresponding to one word from the language. However, for the task of analyzing the concepts present in documents, a lower-dimensional semantic space is ideal — each dimension corresponding to one concept or one topic. Dimension reduction finds a lower-dimensional space that may recover or approximate this underlying semantic space. The new representation reveals the topical structure of the corpus more clearly than the BOW representation. *Topic models* provide a probabilistic framework for the dimension reduction task.

The earliest form of topic models was latent semantic indexing (LSI) [61] which uses singular value decomposition (SVD) [159] to find a topic space through dimension reduction. Hoffmann provided a probabilistic framework, PLSI, that was motivated by LSI [87]. PLSI provides a good basis for text analysis, but it has two problems. First, it contains a large number of parameters that grows linearly with the number of documents so that it tends to overfit the training data. Second, there is no natural way to compute the probability of a document that was not in the training data. LDA

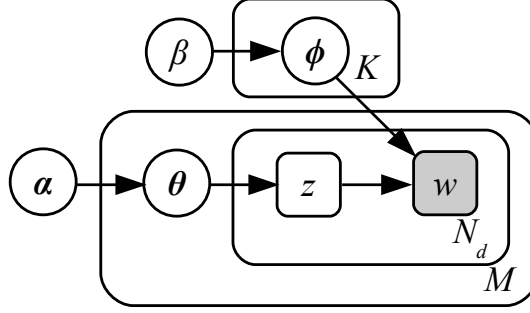


Figure 1: Diagram of the LDA graphical model.

includes a process for generating the topics in each document, thus greatly reducing the number of parameters to be learned and providing a clearly-defined probability for arbitrary documents. Because LDA has a rich generative model, it is also readily adapted to specific application requirements, as we discuss in Section 2.1.6.

2.1.1 LDA Model

A diagram of the graphical model for LDA, showing how the different random variables are related, is shown in Fig. 1. In the diagram, each random variable is represented by a circle (continuous) or square (discrete). A variable that is *observed* (its value is known) is shaded. An arrow is drawn from one random variable to another if the value of the second variable depends on the value of the first variable. A rectangular plate is drawn around a set of variables to show that the set is repeated multiple times, as for example for each document or each token.

The Dirichlet distribution is at the heart of LDA because it reflects the power-law-like distributions of human communication. The Dirichlet distribution is unusual because its outcomes are a discrete probability distribution instead of a simple discrete outcome. To gain an intuition for what a Dirichlet distribution is like, imagine that you start with a stack of cards. Repeatedly draw a card from the deck, but replace the card that was drawn with two of the same card, so that cards that have previously been drawn become more likely in the future. The resulting set of drawn cards will

depend on the cards that were initially in the stack, but a small number of types of cards will typically dominate. The mixture of types of cards that we get from this process is the outcome of the Dirichlet distribution. The parameters of the distribution are the number of each kind of card in the initial stack. If we initially have the same number for each type of card, the distribution is called *symmetric*.

Choose the word probabilities for each topic. The distribution of words for each topic i is represented as a multinomial distribution ϕ_i over the W words, which is drawn from a symmetric Dirichlet distribution with parameter β .

$$\phi_i \sim \mathcal{D}(\beta); \quad \mathbb{P}(\phi_i|\beta) = \frac{\Gamma(W\beta)}{[\Gamma(\beta)]^W} \prod_{v=1}^W \phi_{iv}^{\beta-1}.$$

Choose the topics of the document. The topic distribution for document d is represented as a multinomial distribution θ_d over the K topics, which is drawn from a Dirichlet distribution with parameters α . The Dirichlet distribution captures the document-independent popularity and the within-document burstiness of each topic.

$$\theta_d \sim \mathcal{D}(\alpha); \quad \mathbb{P}(\theta_d|\alpha) = \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_{di}^{\alpha_i-1}.$$

Choose the topic of each token. Document d is a sequence of N_d tokens, indexed by n . The topic z_{dn} is chosen from the document topic distribution.

$$z_{dn} \sim \mathcal{M}(\theta_d); \quad \mathbb{P}(z_{dn} = i|\theta_d) = \theta_{di}.$$

Choose each token. Each token w_{dn} is chosen from the multinomial distribution associated with the selected topic.

$$w_{dn} \sim \mathcal{M}(\phi_{z_{dn}}); \quad \mathbb{P}(w_{dn} = v|z_{dn} = i, \phi_i) = \phi_{iv}.$$



Figure 2: Word cloud representation of an example topic model. The font size is proportional to the probability of seeing a word in a document about diabetes (left cluster), blood types (center) or nutrition (right).

Mechanism

LDA identifies coherent topics from patterns in word co-occurrence. The Dirichlet prior for the topics in a document increases the probability of reusing topics for multiple words, so that rare words are strongly associated with the dominant topics from their contexts and multiple topics for a word are also clarified from the context. To make this clear, we manually constructed a topic model based on a small set of sentences on three topics: diabetes, blood types and nutrition. The word clouds for the three topics are shown in Figure 2. Consider the following paragraph which uses words from two different topics.

Regarding diseases, blood type O’s are predisposed to ulcers and thyroid malfunction. In fact, it has been found that the thyroid gland of people with blood type O produce insufficient amount of thyroid hormone. In addition, type O’s often have insufficient levels of iodine in their blood, a mineral which is essential for thyroid hormone regulation. This has many side effects such as fluid retention, weight gain and fatigue. It is important that people having blood type O have a diet rich in saltwater fish and kelp which are good sources of iodine to help regulate the thyroid gland. [3]

The first step of LDA assigns each word to one or more topics based on the topic models. We ignore the stop words and assign novel words equally to each topic, resulting in 19 words assigned to the blood types topic, 17 to the diabetes topic and

Table 1: Assignments of words from example document before and after applying Dirichlet prior.

Word	Before Prior			After Prior		
	diabetes	blood type	nutrition	diabetes	blood type	nutrition
blood	0.119	0.881	0.000	0.050	0.950	0.000
hormone	1.000	0.000	0.000	0.999	0.001	0.000
type	0.299	0.551	0.150	0.176	0.824	0.000
	0.333	0.333	0.333	0.281	0.718	0.001
Total	17.0	19.1	13.9	14.2	35.8	0.0

14 to the nutrition topic. Next, LDA applies the Dirichlet prior, which accentuates the predominant topics. The effect of the Dirichlet prior is shown in Table 1. Notice that “type” is ambiguous, but the Dirichlet prior helps to resolve that it comes from the blood types topic in this context.

During training, the novel words in this example would be assigned 71.6% to the blood types topic and 28.3% to the diabetes topic. This assignment would be averaged with all of the assignments in other documents for each of these novel words, so that any consistent pattern would effect the topic models but errors would be likely to disappear during the averaging. Thus, LDA results in topics in which the words that are most probable frequently co-occur with each other in documents.

Wallach *et al.* [172] show that the symmetry or asymmetry of the Dirichlet priors strongly influences the mechanism. For the topic-specific word distributions, a symmetric Dirichlet prior provides smoothing so that unseen words will have non-zero probability. However, an asymmetric prior would uniformly affect all topics, making them less distinctive. In contrast, they showed that an asymmetric prior for the document-specific topic distributions made LDA more robust to stop words and less sensitive to the number of topics. The stop words were mainly relegated to a small number of highly probable topics that influence most documents uniformly. The asymmetric prior also results in more stable topics, which means that additional

topics will make small improvements in the model instead of radically altering the topic structure. This is similar to the situation of LSI, where performance is optimal when LSI scales the contribution of each dimension according to its eigenvalue. In the same way, LDA will perform best if α is non-uniform and corresponds to some natural values characteristic of the dataset.

One disadvantage of LDA is that it tends to learn broad topics. Consider the case where a concept has a number of aspects to it. Each of the aspects co-occurs frequently with the main concept, so LDA will favor a topic that includes the concept and all of its aspects. It will further favor adding other concepts to the same topic if they share the same aspects. As this process continues, the topics become more diffuse. When sharper topics are desired, a hierarchical topic model [23] may be more appropriate.

Likelihood

Training an LDA model involves finding the optimal set of parameters, under which the probability of generating the training documents is maximized. The probability of the training documents under a given LDA model is called the *empirical likelihood* \mathbb{L} . It can also be used to identify the optimal model configuration using Bayesian model selection.

$$\begin{aligned}\mathbb{L} &= \prod_{d=1}^M \prod_{n=1}^N \mathbb{P}(w_{dn}|z_{dn}, \phi) \mathbb{P}(z_{dn}|\theta_d) \mathbb{P}(\theta_d|\alpha) \mathbb{P}(\phi|\beta) \\ &= \phi_{zw} \theta_{dz} \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_{di}^{\alpha_i-1} \frac{\Gamma(W\beta)}{[\Gamma(\beta)]^W} \prod_{v=1}^W \phi_{iv}^{\beta-1}.\end{aligned}$$

Unfortunately, the direct optimization of the likelihood is problematic because the topic assignments z_{dn} are not directly observed. Even inference for a single document is intractable. We describe two different approximations for LDA, each of which has

advantages. Collapsed Gibbs sampling samples a value for each z_{dn} in turn, conditioned on the topic assignments for the other tokens. Variational Bayes approximates the model with a series of simpler models that bound the likelihood but neglect the troublesome dependencies.

2.1.2 Collapsed Gibbs

Gibbs sampling is commonly used to estimate the distribution of values for a probability model when exact inference is intractable. First, values are assigned to each variable in the model, either randomly or using a heuristic. Each variable is then sampled in turn, conditioned on the values of the other variables. In the limit of the number of iterations, this process explores all configurations and yields unbiased estimates of the underlying distributions. In practice, Gibbs sampling is implemented by rejecting a large number of samples during an initial burn-in period and then averaging the assignments during an additional large number of samples.

In *collapsed* Gibbs sampling, certain variables are marginalized out of the model. Griffiths and Steyvers [78] propose collapsed Gibbs sampling for LDA, with both θ and ϕ marginalized. Only z_{dn} is sampled, and the sampling is done conditioned on α , β and the topic assignments of other words \bar{z}_{dn} .

$$\mathbb{P}(z_{dn}|\bar{z}_{dn}) \propto (N_{dz} + \alpha_z)(N_{zw} + \beta).$$

The N statistics do not include the contribution from the word being sampled, and must be updated after each sampling.

The equation makes intuitive sense. A topic that is used frequently in the document has a higher probability in θ and so is more likely for the current token also. This characteristic corresponds to the topic burstiness observed in documents [47]. Similarly, a topic that is frequently assigned for the same word corpus-wide is more likely to be correct here also.

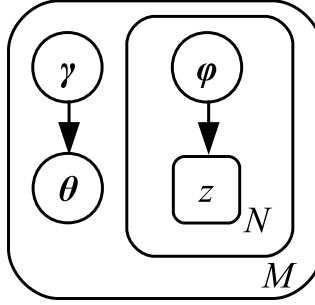


Figure 3: Diagram of the LDA variational model.

After burn-in, the implementation can keep statistics of the number of times each topic is selected for each word. These statistics can then be aggregated and normalized to estimate the topic distributions for each document or word. It is often necessary to apply a topic model to additional documents. This is easily done with little change to the algorithm: the N_{zw} statistic does not need to be updated.

2.1.3 Variational Approximation

Variational approximation provides an alternative algorithm for training an LDA model. Before we explain how the model is trained, we will consider how the topics of a document can be inferred from an existing LDA model. A direct approach for topic inference is to apply Bayes' rule,

$$\mathbb{P}(\boldsymbol{\theta}|\mathbf{w}_d) = \frac{\mathbb{P}(\boldsymbol{\theta}, \mathbf{w}_d)}{\mathbb{P}(\mathbf{w}_d)} = \frac{\int_{\mathbf{Z}} \mathbb{P}(d, \boldsymbol{\theta}, \mathbf{Z}|\boldsymbol{\alpha}, \beta) d\mathbf{Z}}{\int_{\mathbf{Z}, \boldsymbol{\theta}} \mathbb{P}(d, \boldsymbol{\theta}, \mathbf{Z}|\boldsymbol{\alpha}, \beta) d\mathbf{Z} d\boldsymbol{\theta}},$$

where $\mathbf{Z} = \{z_{d1}, z_{d2}, \dots, z_{dN_d}\}$. This kind of calculation, called marginalization, considers the probability of every possible topic assignment for each word in the document in order to derive the probability of each topic distribution, but this calculation involves too many possibilities to be tractable. The *variational Bayesian* approach provides an approximate solution; instead of inferring the latent variables by directly marginalizing the joint distribution $\mathbb{P}(\mathbf{w}_d, \boldsymbol{\theta}, \mathbf{Z}|\boldsymbol{\alpha}, \beta)$, it uses a much simpler distribution as a proxy and performs the inference through optimization.

Variational inference approximates the true posterior distribution of the latent variables by a fully-factorized distribution—this proxy is usually referred to as the variational model, which assumes all the latent variables are independent of each other. For LDA,

$$q(\mathbf{Z}, \boldsymbol{\theta} | \boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\boldsymbol{\theta} | \boldsymbol{\gamma}) \prod_{n=1}^N q(z_n | \boldsymbol{\phi}_n) = \mathcal{D}(\boldsymbol{\theta} | \boldsymbol{\gamma}) \prod_{n=1}^N \mathcal{M}(z_n | \boldsymbol{\phi}_n).$$

Essentially, this variational distribution is a simplification of the original LDA graphical model by removing the edges between the nodes $\boldsymbol{\theta}$ and \mathbf{Z} (Figure 3). The optimal approximation is achieved by optimizing the distance—for example, the Kullback-Leibler divergence (KL-divergence) [101]—between the true model and the variational model:

$$\min_{\boldsymbol{\gamma}, \boldsymbol{\phi}} KL[q(\boldsymbol{\theta}, \mathbf{Z} | \boldsymbol{\gamma}, \boldsymbol{\phi}) || \mathbb{P}(\boldsymbol{\theta}, \mathbf{Z} | \boldsymbol{\alpha}, \beta)].$$

It can be shown that the above KL-divergence is the discrepancy between the true log-likelihood and its variational lower-bound that is used in the variational EM algorithm (described later in this section) for estimating the LDA hyperparameters $\boldsymbol{\alpha}$ and β .

The optimization has no close-form solution but can be implemented through iterative updates,

$$\gamma_i = \alpha_i + \sum_{n=1}^N \phi_{ni}, \quad \phi_{ni} \propto \beta_{iw_n} \exp[\Psi(\gamma_i)],$$

where $\Psi(\cdot)$ is the bi-gamma function.

Variational EM for parameter estimation

We can learn an LDA topic model by maximizing the likelihood of the corpus.

$$\begin{aligned} & \max_{\boldsymbol{\alpha}, \beta} \sum_{d=1}^M \ln \mathbb{P}(\mathbf{w}_d | \boldsymbol{\alpha}, \beta) \\ &= \max_{\boldsymbol{\alpha}, \beta} \sum_{d=1}^M \ln \int_{\boldsymbol{\theta}_d, \mathbf{Z}_d} \mathbb{P}(\mathbf{w}_d, \boldsymbol{\theta}_d, \mathbf{Z}_d | \boldsymbol{\alpha}, \beta) d\boldsymbol{\theta}_d d\mathbf{Z}_d. \end{aligned}$$

Again, it involves intractable computation of the marginal distribution and we therefore resort to variational approximation, which provides a tractable lower bound,

$$\begin{aligned}\mathbb{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}) &= \ln \mathbb{P}(\boldsymbol{w}_d | \boldsymbol{\alpha}, \beta) - KL(q(\boldsymbol{Z}, \boldsymbol{\theta} | \boldsymbol{\gamma}, \boldsymbol{\phi}) || \mathbb{P}(\boldsymbol{Z}, \boldsymbol{\theta} | \boldsymbol{\alpha}, \beta)) \\ &\leq \ln \mathbb{P}(\boldsymbol{w}_d | \boldsymbol{\alpha}, \beta),\end{aligned}$$

where $\mathbb{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}) = \mathbf{E}_q[\ln \mathbb{P}(\boldsymbol{w}_d, \boldsymbol{\theta}, \boldsymbol{Z}) - \ln q(\boldsymbol{w}_d, \boldsymbol{\gamma}, \boldsymbol{\phi})]$ is the variational lower bound for the log-likelihood. The maximum likelihood estimation therefore involves a two-layer optimization,

$$\max_{\boldsymbol{\alpha}, \beta} \sum_{d=1}^M \max_{\boldsymbol{\gamma}_d, \boldsymbol{\phi}_d} \mathbb{L}(\boldsymbol{\gamma}_d, \boldsymbol{\phi}_d).$$

The inner-loop (the optimization with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$, referred to as the variational E-step) goes through the whole corpus and performs variational approximation for each of the documents, which ends up with a tight lower bound for the log-likelihood. Then the M-step updates the model parameters ($\boldsymbol{\alpha}$ and β) by optimizing this lower-bound approximation of the log-likelihood. The E- and M-steps are alternated in an outer loop until convergence.

In the E-step, $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ are alternately optimized for each document—in practice, 20 iterations is adequate for a good fit. The outer loop may need to be repeated hundreds of times for full convergence. For best results, the likelihood of a separate validation corpus controls early stopping.

2.1.4 Interpretation and Evaluation

Intuitively, the topics discovered by topic models correspond to concepts or topics that are meaningful to the authors. In this section, we discuss how to jump from the mathematical representations to meaningful topics, how to evaluate the resulting models and how to apply them to applications.

Interpretation

The common way to interpret topic models is through inspection of the word-topic associations. Typically, practitioners sort the words by the probability of generating each conditioned on the topic and attempt to discern the commonality of the top five to twenty words. This approach was popularized following Blei *et al.* [26], and is generally used to report qualitative topic model results, even though it has many disadvantages. The chief problem is that the top words are often dominated by globally probable words that may not be representative of the topic. Stop word removal and variations on inverse document frequency (IDF) weighting both help substantially, but the characterization is sensitive to the precise method used to order words. Mei *et al.* [116] provide an alternative approach that automatically selects a portion of a document to use as a label for each topic. Buntine and Jakulin [35] provide a more general framework for interpreting topic models.

Evaluation

There are three main approaches to evaluating topic models. The fit of the models to test data is important for understanding how well the models generalize to new data, but application-driven metrics are also essential if the model is to be useful. When it is necessary for a human to interact with the model, interpretability should also be evaluated.

Fit of Test Data A very common approach is to train a model on a portion of the data and to evaluate the fit of the model on another portion of the data, by computing the probability of generating test documents given the model.

Perplexity [16] is the most common way to report this probability. Computed as

$$\exp \left(-\frac{1}{N} \sum_{d=1}^M \sum_{n=1}^{N_d} \ln \mathbb{P}(w_{dn} | \text{model}) \right),$$

the perplexity corresponds to the effective size of the vocabulary. For example, a value of 100 indicates that the probabilities resulting from the model are equivalent to randomly picking each word from a vocabulary of 100 words. Thus smaller values indicate that the model fits the test data better.

Wallach *et al.* [174], evaluate several different ways to compute this probability and recommended the *left-to-right* method, in which the probability of generating each token in a document is conditioned on all previous tokens in the document, so that the interaction between the tokens in the document are properly accounted for. This method is uncommon in practice because it is computationally expensive and requires a departure from the BOW model.

Application Performance Another common approach is to measure the utility of topic models in some application. Whenever the topic modeling is being carried out with a specific application in mind, this is an important evaluation. For example, Wei and Croft [175] discuss the evaluation of LDA models for document search using standard information retrieval metrics. They rank possible result documents according to the probability that the query would be generated from a model of each document, then report the average precision on test data at fixed levels of recall.

Interpretability Topic models are intended to recover a meaningful underlying set of human concepts. Unfortunately, the fit of test data and application performance metrics completely ignore the topical structure. In fact, models with better perplexity are often harder to interpret [44]. This is not surprising, because the task of finding a meaningful model that fits well is more constrained than the task of finding any model that fits well, so the best fitting model is likely to be less meaningful.

Chang *et al.* [44] propose a new evaluation protocol based on a user study. Starting with a list of top words for each topic that has been tainted with an additional word, users are asked to identify the spurious word. User performance on this task is higher

when the topic is coherent so that the extra word stands out. They also conducted a similar experiment to measure the appropriateness of topic assignments to test documents.

2.1.5 Parameter Selection

Asuncion *et al.* [14] compare a variety of different algorithms for the LDA model. They found that with careful selection of the regularization hyperparameters α and β , all of the algorithms had similar perplexity. A grid search over possible values yields the best performance, but interleaving optimization of the hyperparameters with iterations of the algorithm is almost as good with much less computational cost.

2.1.6 Adapting Topic Models to Applications

The graphical model of LDA can be easily extended to match the characteristics of a specific application. Here we survey some of the fruitful approaches.

One important class of extensions to LDA has been the introduction of richer priors for document topic and word distributions. Instead of using a fixed, global Dirichlet hyperparameter α for all the documents in a corpus, Mimno and McCallum use regression from document features to establish a document-specific α [120]. This is a valuable enhancement when other meta-features are available that are expected to influence the selected topics, as, for example, the identity of the author, the publication venue and the dates.

The Bayesian hierarchy of LDA provides a useful modeling pipeline for data with complex structure. The hierarchy can model web-like interconnections and uncertain labels [181, 184]. The *mixed membership stochastic block model* couples two LDA hierarchies to model inter-connected entities [10], which provides a flexible model for network graphs and has proven useful for a variety of applications ranging from role discovery to community detection in social, biological and information networks.

Hierarchical topic models (hLDA) are used to identify subtopics that are increasingly more specific [23]. The hLDA model automatically learns a tree structure hierarchy for topics while they are discovered from the documents. For additional flexibility, hierarchical Dirichlet processes [161] can automatically discover an appropriate number of topics and subtopics. There are also principled ways to learn correlations between topics [24, 106]. Other extensions support richer document representations and contextual information, including bigrams [173], syntactic relationships [29, 79] and product aspects [163].

Multinomial distributions for word occurrences usually have a difficult time modeling the word burstiness in language — if a word appears in a document once, it will likely appear again in the same document. To discount this impact, Doyle and Elkan replace the per-topic Multinomial distribution with a Dirichlet-Compound Multinomial (also called the multivariate Pólya distribution) [64]. Reisinger *et al.* substitutes spherical admixture models [140], which not only incorporate negative correlations among word occurrence but also admit the natural use of cosine similarity to compare topics or documents.

Standard topic models are not appropriate for identifying consistent topics across multiple languages, because the multiple languages do not co-occur in documents frequently enough to be assigned into the same topics. Mimno *et al.* developed an extension that works with loosely *aligned* documents [119]—pairs of documents in different languages that have nearly the same mixture of topics. Boyd-Graber and Blei explore various strategies for discovering multilingual topics from unaligned documents [28]. Similar issues arise with documents in multiple dialects.

Topic models have been used very effectively at improving recall in the face of language gaps [119, 13, 34, 136, 9]. Ahmed and Xing model the ideological gap between political documents, but the extremes overlap only in shared neutral topics [9]. Paul and Girju also model topics in different language variants, supporting a

star architecture (each collection has its own topic models which relate back only to common shared topics). In contrast, our diaTM model (Section 3.3) places all of the collections into a low dimensional space for more robust modeling of the dialects. Our τ LDA model, in Section 3.4, adds topic-specific dialects which have not been previously considered.

2.2 *Quality of Online Health Information*

An increasing percentage of the population in Europe and America use the Internet to find health-related information [102, 70]. Unfortunately, not all the online health-related information is trustworthy or harmless. There is online content opposing vaccinations [6], promoting self-injuries [122] and even teaching self-asphyxiation techniques [107]. Some studies have evaluated the quality and accuracy of the health related information on video web sites [104], typically focusing on a selected topic including myocardial infarction [134], rheumatoid arthritis [154], the H1N1 pandemic [133], prostate cancer [158], and kidney stones [157].

Eating disorders, such as anorexia nervosa and bulimia, are prevalent in most developed countries [86]. These disorders are a major public health concern due to their high mortality and co-morbidities [12]. Anorexia nervosa commonly appears during adolescence and can be devastating to the patient’s well-being and may cause death [86, 12]. Although generally viewed as highly negative by the general public, there is an online movement of people promoting anorexia as a lifestyle choice. Norris *et al.* conducted an in-depth analysis of 12 pro-anorexia web sites and found that the most prevalent themes in them were related to inspiration and assistance in achieving or maintaining anorexia [131]. A more recent study examined 180 pro-anorexia web sites, and found similar results [27]. Mulveen examined 15 discussion threads on a pro-anorexia site to discover the main themes of the discussion and to understand the

reasons for participation in such threads [126]. Members of this community use particular web sites to share photographs and text designed to inspire members to attain or maintain an unusually low weight. They also offer tips on how to lose extreme amounts of weight. These web sites frequently operate forums where members can discuss issues related to their positive view of anorexia and find a *pro-ana buddy* to support each other in maintaining their disease. However, the small number of threads investigated makes it difficult to draw conclusions from Mulveen’s investigation.

It is impossible to ignore the impact the media has on beauty and health standards. The desire to be an anorexic begins with the psychological fight to achieve this ideal “skinny” beauty that has been imprinted in the brain [113]. Online pro-anorexia content has been found to exacerbate the eating disorders and promote the anorexic lifestyle [17, 144]. Anorexia has a huge impact on the health of the patients, including high rates of morbidity and mortality [12]. The Eating Referral website reported in a recent poll by People magazine that 80% of women felt insecure about their appearance because of the women they saw on television, movies and in fashion magazines to the extent that they are willing to try diets that pose health risks (34%) or go “under the knife” (34%). 93% indicated they had made various and repeated attempts to lose weight to measure up to the images [83].

The impact on health that these images have on women is profound. The financial, social, psychological and physical damage from a lifetime pursuing unreasonable thinness are impossible to measure. Depression, despair, depletion of self-esteem, withering and wasting away of physical, psychological and financial resources are unbelievably evident in the proana community. A too low BMI leads to a weakened immune system, low blood pressure, cardiovascular disease, osteoarthritis and opportunistic infections like Tuberculosis [2]. The line between reality and fiction becomes blurred when women are showcased in media with idealized bodies that have been achieved through photoshop manipulation [110]. “The average woman in the US is

5-foot-4 and weighs 140 pounds. The average model, on the other hand, is 5-foot-11 and weighs 117 pounds” [113].

The increase of harmful content is a serious concern, since a consumer searching for trustworthy information is likely to encounter pernicious content. For example, a vulnerable teen searching for information on healthy weight loss is likely to find sites recommending that he make himself vomit. Online content promoting anorexia as a lifestyle is very common [176] and Custers *et al.* found that 12% of Belgian female students in the sixth, ninth and eleventh grades had viewed pro-anorexia content [59]. Consumption of pro-anorexia content has been found to correlate with worsening of anorexia [176, 82, 144]. Rouleau *et al.* in a recent review described three potential risks associated with pro-anorexia sites: prevention of help-seeking, reinforcement of disordered eating and operating under the guise of support [144].

Online pro-anorexia content is a case of societal concern and some health authorities and organizations have been attempting to curb its effects. To the best of our knowledge, three interventions have been suggested so far. (1) Some governments, including Brazil, Israel and Italy [1, 4], ban advertisements containing severely underweight models. (2) Driven by public pressure, social content sharing sites including Tumblr (<http://tumblr.com>), Pinterest (<http://pinterest.com>) and Instagram (<http://instagram.com>) have attempted to ban pro-anorexia content [124]. Lewis *et al.* suggested filtering search results to reduce exposure to pro-anorexia content [105]. (3) Some Internet service providers attempt to add warning labels when users try to access pro-anorexia content [112]. One should also note that all these interventions are likely to be more effective at preventing users from beginning to engage with pro-anorexia content than in dissuading existing users from continuing to consume such content. Only (3) has been substantiated with a controlled study. In addition, there has been a lack of research comparing the characteristics of the pro-anorexia communities with those that recognize anorexia as a disease.

2.3 *Survival and Event History Analysis*

Event history analysis is a powerful statistical technique for understanding how the frequency or timing of events (*event dynamics*) is related to measurable variables. Similarly, studying the dynamics of events that happen in groups provides valuable insight into how consumers interact with the groups. Early group dynamics analysis was rule-based [85]. More recently, Snijders used a Cox-intensity Poisson model with exponential random graphs (ERGM) to model friendship ties [156]. Brandes *et al.* extended this model to more general interactions [30], but its feature model is too restrictive for our application. Backstrom *et al.* studied the interaction of the friendship graph among group members and group growth [15].

Event history analysis is a well-established tool for understanding the factors that predict events [5]. For example, it has often been used to measure the effectiveness of drugs for treating a disease, where use of the drug causes a change in events related to the disease. It has been applied to analyze human actions in emergency situations [38] and to model citations between scholarly works [170]. Event history analysis works from a parametrized probabilistic model for when events will occur. The parameters of the model can then be inferred using common statistical approaches including maximum likelihood estimation or Bayesian inference. We will first discuss survival analysis, which relates to analyzing events like death that can only happen once, and then we will discuss more general event history analysis.

2.3.1 *Survival Analysis*

Suppose that we are interested in understanding the timing of an event like a consumer joining a group that ideally can only happen once. Let us further assume that the consumer was not a member of the group when we started observing at time $t = 0$. Then we define the survival function

$$S(t) = \mathbb{P}(T > t) = 1 - F(t) = 1 - \int_0^t f(s) ds, \quad (1)$$

where T is the time the event occurs and F and f are the cumulative distribution and probability density functions for T . The term *survival* comes from interpreting the event as death, so that the survival function gives the probability that a subject will not have died by time t . In our example, it gives the probability that the consumer will not have joined the group by time t .

We next define the hazard rate

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(T \in [t, t + \Delta t) | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)},$$

which is the conditional probability density that the event will happen near time t given that it had not happened yet. Notice from Eqn. (1) that

$$\begin{aligned} \frac{d}{dt} \ln S(t) &= -\frac{f(t)}{S(t)} = -h(t) \\ -\ln S(t) &= \int_0^t h(s) ds \equiv H(t) \\ S(t) &= \exp(-H(t)), \end{aligned}$$

where we call $H(t)$ the cumulative hazard rate.

Given an event that is observed to have happened at time T , its contribution to the negative log likelihood of a model for the dynamics is $\ln(f(T))$. If we have a whole set of events i , each of which occurs at time T_i , we can compute the negative log likelihood from the whole set as

$$\mathbb{L} = -\sum_i \ln f_i(T_i).$$

The dependence of f on i is parametric; survival analysis proceeds by selecting and interpreting the parameters that minimize the negative log likelihood.

However, in any real set of observations, we will not perfectly observe each of the event times. Some of the events will have not happened by the end of the observation (*right censoring*) and for some we may know that the event occurred, but only know the time up to an interval, $T_i \in [a_i, b_i]$ (*interval censoring*). *Uninformative censoring*,

which is independent of future events, is readily incorporated into survival analysis. Let $g_i(t)$ and $\mathbf{G}_i(t)$ be the probability density and cumulative probability density of the censoring process—that is, the probability density that event i happening near time t will not be observed due to censoring. As long as the censoring is uninformative, the negative log likelihood for an observed event is $-\ln(1 - \mathbf{G}_i(T_i)) - \ln f_i(T_i)$. Right-censored events contribute $S_i(C_i)g_i(C_i)$ where events after censor time C_i are not observed and interval censored events contribute $S_i(b_i) - S_i(a_i)$.

The hazard functions $h_i(t)$ depend on both the event and the time. It is common to decompose this into the product of two parametric functions, $h_i(t) = h_0(t)h(\phi_i)$, one of which depends on the event and the other on the time. ϕ_i is typically a vector of features for event i , which are called *covariates* because the model posits that they are correlated with the event dynamics. In some cases it is useful to define *dynamic covariates* $\phi_i(t)$ that depend on events that happened before time t . For example, a consumer joining a group is plausibly related to the number of members already in the group. Common parametric forms include $h_0(t) = \lambda \rho t^{\rho-1}$, where λ is a scale parameter and ρ is a shape parameter, and the Cox model $h(\phi_i) = \exp(\beta^T \phi_i)$ [50], where β controls the impact of each covariate. Using these models, the survival function is $\exp(-\lambda \rho t^\rho \exp(\beta^T \phi_i))$. Outside of event history analysis, this model has been successful at modeling social events [156].

Normally, the dependence of event rates on time is not very enlightening. Instead, the value comes from estimating how the hazard rate is impacted by changes in the covariates ϕ_i . This is formalized as the *relative risk* (RR_k) of covariate k , the relative rate at which the hazard rate changes as the covariate change,

$$RR_k = \frac{1}{h_i} \frac{dh_i}{d\phi_{ik}}.$$

Using a Cox model, the relative risk of covariate k is simply $\exp(\beta_k)$, which makes the interpretation of a learned Cox model especially straight-forward.

Cox showed that optimizing a partial likelihood function was adequate to recover the relative risks [51]. Suppose that we know that some event occurred at time t . For a Poisson event model, the probability that the event at time t was a particular event is proportional to the hazard rate for that event at that time. For each event i , let $Y_i(t)$ be an indicator of whether the event is *at risk* at time t ; that is, $Y_i(t)$ will be 0 if event i had already occurred by time t and 1 otherwise. Then the likelihood of event i happening at time t is

$$\mathbb{L} = \frac{Y_i(t)h_i(t)}{\sum_j Y_j(t)h_j(t)}.$$

For a decomposable hazard function, the dependence of the hazard rate on time drops out and we get simply

$$\mathbb{L} = \frac{Y_i(t)h(\phi_i)}{\sum_j Y_j(t)h(\phi_j)}.$$

Thus, the only dependence on the time is through the set of events that are at risk because they have not previously occurred.

There are two common practices that are used when multiple events happen at the same time [5]. The first approach ignores the duplicated times, essentially assuming that the events happened in some arbitrary order at slightly different times. The second approach computes an average of likelihoods computed using different orderings of events. Both of these approaches essentially assume that two events cannot really happen at the same instant. In Chapter 5 we develop an alternative approach based on the assumption that multiple events can occur simultaneously.

To avoid overfitting or underfitting, survival models should have at least 10 events per parameter [137]. The most robust method of statistical significance testing is the likelihood ratio test [49, 123]. Suppose that we have two nested models, with parameters β_0 and β_1 that are identical except for in k dimensions indexed by i where $\beta_{0i} = 0$. Then the probability that β_1 is a better explanation for the observed

data is

$$1 - \mathcal{X}(2[\ln \mathbb{L}(\boldsymbol{\beta}_1) - \ln \mathbb{L}(\boldsymbol{\beta}_0)], k),$$

where $\mathbb{L}(\boldsymbol{\beta})$ is the empirical likelihood of $\boldsymbol{\beta}$ conditioned on the data and $\mathcal{X}(x, k)$ is the cumulative distribution function of the χ^2 distribution at x with k degrees of freedom. This is often used to test the merit of adding an additional covariate to a model.

Another of our contributions in Chapter 5 is to use singular value decomposition (SVD) to properly handle correlations between the dimensions of ϕ_i . SVD is routinely used to handle correlated features in linear regression [84]. It has also been used to identify correlated covariates and select one to include in survival models [49]. New issues naturally arise when using SVD to incorporate the full set of correlated features.

2.3.2 Event History Analysis

Survival analysis deals with events that only occur once. Most social interactions, however, occur repeatedly. Event history analysis is very similar to survival analysis except that it deals with events that may repeat. For each event i , we define a variable $Y_i(t)$ that indicates whether the event i is possible (*at risk*) at time t . $Y_i(t)$ may depend on other events that happened before time t . Because the events are repeatable, we work with the number of times the event has occurred, $N_i(t)$. $N_i(t)$ is a submartingale, so it can be written

$$N_i(t) = \int_0^t Y_i(s) \lambda_i(s) ds + \mathcal{M}_i(t),$$

where $\lambda_i(t)$ is the instantaneous rate of occurrences and $\mathcal{M}_i(t)$ is a zero-mean martingale [117]. λ_i can be modeled using the same parametric models as we used for the hazard function h_i , yielding essentially the same results for censoring, interpretation and partial likelihood.

2.4 Groups Interactions

There has been considerable research on the dynamics of group interactions. Researchers have investigated sustainability for large groups where consumers need cognitive support [37, 91], which is not relevant to health communities because of their relatively low discussion volumes. They have also identified several social aspects that influence social dynamics, including homophily, influence, sense of agency, sense of community and resource balance [41, 91, 94, 115, 143, 153].

Diversity arises as an important factor in our analysis. Generally, research finds that homogeneity within a group helps the group to work together well, but limits the amount of creativity [169]. Moreover, Wu and Huberman found that extreme positions are much more likely to arise in homogeneous groups [180]. On the other hand, a dense sharing of users between different groups may help valuable leadership develop and solidify the unique culture of a community [41, 92].

Health forums have unique characteristics that set them apart from general online forums. Discovering that one has a chronic disease and learning to live with it causes a wide range of emotional reactions [60, 162]. For example, depression is three times more common in diabetes patients than in the general population [71] and reoccurs frequently for many years [109]. Depression is not expected to have a significant effect on the volume of online interactions [130], but may affect the types of interactions. Many patients experience resilience which increases the effort they expend in supportive communities [185]. This is expected to increase the sustainable size of the community and the level of involvement from community members.

2.5 Latent Behavior Models

Matrix factorization is very effective for many collaborative filtering applications [7, 19, 98, 135, 141, 147], because it discovers a latent model of the users and groups that gives rise to a model for user behavior. Factorization methods seek to associate both

consumers and items with latent profiles represented by vectors in a low dimensional space. The consumers' ratings for the items can be captured by similarity between these low dimensional factors. Often, a simple dot product is used to measure the similarity. This is called *cosine similarity* because of a mathematical property relating the dot product to the cosine of the angle between the vectors. Probabilistic matrix factorization formulates the matrix factorization in a graphical model in which latent factors are assumed to be generated from multivariate normal distributions [149].

For example, to recommend groups to a user, we could start with a partially-observed matrix \mathbf{R} where R_{ug} indicates the affinity between user u and group g . Matrix factorization then identifies latent profiles for the users and groups that explain the observed affinities, $R_{ug} \approx \mathbf{U}_u \mathbf{G}_g^T$. These latent profiles can be used for recommendation to user u by sorting all groups according to $\mathbf{U}_u \mathbf{G}_g^T$. Matrix factorization is normally solved by optimizing the squared error loss function with regularization,

$$\text{minimize } \frac{1}{2} \sum_{ug} (R_{ug} - \mathbf{U}_u \mathbf{G}_g^T)^2 + \frac{1}{2\sigma_U^2} \|\mathbf{U}\|_{\mathcal{F}}^2 + \frac{1}{2\sigma_G^2} \|\mathbf{G}\|_{\mathcal{F}}^2. \quad (2)$$

The first term in this equation measures the errors that the model makes when it tries to predict R_{ug} using $\mathbf{U}_u \mathbf{G}_g^T$. The other terms use the Frobenius norm to add regularization which prevents overfitting, controls the size of the latent profile vectors and yields a probabilistic interpretation.

Completely Shared Profiles

In many cases, we have information for several different related tasks that may be able to share information. For example, there are many kinds of entities in an online community: users are members of the community; discussions are a message from a user and a set of replies from other users; forums provide a loosely-organized place where discussions can take place; groups provide a more structured place for discussions. Each pair of entity types results in a different recommendation task. For example, we might recommend groups for a user to join or recommend a discussion

to a whole group of users. Multi-task matrix factorization leverages all of the information available about these tasks to improve the performance on the individual tasks.

Collective matrix factorization models multiple recommendation tasks jointly by sharing the latent profile for an entity between all tasks [189]. For example, the same profile is used for a user whether we are recommending a discussion or a group. For the sake of discussion, consider that we have, in addition to R_{ug} , partially observed ratings R_{ud} that indicate how well user u likes the discussion d . Because users participate in many more discussions than groups, the extra information provides a more robust model of user behavior. We associate a latent vector D_d to each discussion, resulting in the optimization problem

$$\begin{aligned} \text{minimize } & \frac{1}{2} \sum_{ug} (R_{ug} - \mathbf{U}_u \mathbf{G}_g^T)^2 \\ & + \frac{1}{2} \sum_{ud} (R_{ud} - \mathbf{U}_u \mathbf{D}_d^T)^2 \\ & + \frac{1}{2\sigma_U^2} \|\mathbf{U}\|_{\mathcal{F}}^2 + \frac{1}{2\sigma_G^2} \|\mathbf{G}\|_{\mathcal{F}}^2 + \frac{1}{2\sigma_D^2} \|\mathbf{D}\|_{\mathcal{F}}^2. \end{aligned}$$

This equation is very similar to Eqn. (2) with just the addition of an error term for the new task and a regularization term for the new type of entity. More generally, we add summations for each type of relation, with some accommodation for differences in scaling between types of relation.

Partially Shared Profiles

Because we have a wide variety of relationships, there may be latent features that are only relevant for certain relationships. For example, users may join groups because of the topic but participate in discussions because of emotional factors. Collective matrix factorization can accomplish this automatically by learning 0 values for emotional features in the groups, but enforcing this sparsity could yield improved performance [80]. Partially shared collective matrix factorization accomplishes this by restricting

the profile columns used for each relation:

$$\begin{aligned}
\text{minimize } & \frac{1}{2} \sum_{ug} (R_{ug} - \mathbf{U}_u \mathbf{T}_{UG} \mathbf{G}_g^T)^2 \\
& + \frac{1}{2} \sum_{ud} (R_{ud} - \mathbf{U}_u \mathbf{T}_{UD} D_d^T)^2 \\
& + \frac{1}{2\sigma_U^2} \|\mathbf{U}\|_{\mathcal{F}}^2 + \frac{1}{2\sigma_G^2} \|\mathbf{G}\|_{\mathcal{F}}^2 + \frac{1}{2\sigma_D^2} \|\mathbf{D}\|_{\mathcal{F}}^2,
\end{aligned} \tag{3}$$

where the diagonal matrices \mathbf{T}_{UG} and \mathbf{T}_{UD} have the value 1 for each latent factor that is relevant to the corresponding relation.

Evaluation

Recommendation models are evaluated in many of the same ways as topic models, discussed in Section 2.1. The RMSE on a test set is used to evaluate how well the model makes predictions globally. For a more effective evaluation, an application-specific approach like MR is used.

Extensions

There are also other use formulations of matrix factorization. Variants of matrix factorization related to latent Dirichlet allocation (LDA) [26] have been used for various social recommendation tasks [45, 93]. This approach lends a natural topical interpretation to the latent profiles, but has not been extended to leverage the multiple complementary relationships that are the study of Chapter 6.

Some aspects of recommendation have been investigated in the context of social networks. Studies have found that recommendation can improve the outcome of participants [88, 177, 132, 114, 68] and enhance the quality and sustainability of the groups themselves [169]. Considerable recommendation work has been done in the context of question and answer sites [8, 190]. Chen *et al.* showed that topic-based models for group recommendation were dramatically superior to rule-based models [45], an approach that was later refined [93]. Most research into recommendation

in a medical context has leveraged electronic medical records [13, 111, 76], which is complementary to the behavior-driven collaborative filtering methods we investigate. The two approaches could be used in concert to generate the best recommendations.

In Chapter 6, we address the issue of the impact on the groups that we recommend. In a first attempt to apply recommendation that is equally valuable to individual consumers and the community, Akoglu and Faloutsos demonstrated a system that recommends so as to optimize global properties of the network [11]. Other research has also considered the case of recommendations to groups of consumers [96, 95] using simple techniques, but it has not been integrated with recommendation of groups or applied for understanding the impact recommendation has on groups.

Patient similarity has been explored using electronic medical records by [13, 111, 76]. Studies have investigated the benefits of careful matching between patients and clinicians [68] and support groups[88, 177, 132, 114], but we are not aware of any research on automated personalized support group ranking.

CHAPTER III

LANGUAGE GAPS IN HEALTH INFORMATION RETRIEVAL

3.1 Introduction

Jamal's hand blistered up when he burned it last week. Now the skin has peeled off and the scab looks yellow. The skin around the burn still looks red or orange. It is a weekend, so he will go to the emergency room if he has to, but wants to find out first if it can wait until Monday. He tried to search the Web using the query "burn blister orange red." Most of the result pages are totally worthless, because they interpret "burn" as a sensation instead of the cause of the blister. However, one result looks especially promising because it talks about the skin turning yellow after a burn. According to the Website (http://www.pnphpbb.com/qna/What_happens_when_your_skin_turns_yellow_after_a_burn-qna107592.html), the skin turns yellow after a burn because the liver is failing, which will be fatal. Jamal started looking for information about an infected burn area, and now he wonders if he should call an ambulance to get him to the hospital before his liver fails completely. Jamal is needlessly worried because of a language gap between him and the technical medical resources he is looking for. He does not know how to describe his situation with enough precision to find relevant documents.

Language gaps are very important for health information retrieval. Despite substantial effort to understand the translation between technical medical language and consumer language, users still have difficulty finding relevant documents [108]. Because of this, current research continues to search for a way to surmount the language gap. In addition, researchers believe that personalization is key to improving health

literacy at a large scale [99], and we claim that the language gap contains valuable information needed for personalization. Language plays a central role in human culture, and many cultural differences that affect the acceptance of health information is reflected in differences in word choice. This aspect of the language gap is especially valuable in social recommendation, where cultural differences influence engagement with a group or user.

We directly model the dialect of users and documents, which tackles both aspects simultaneously. When can compare queries with documents in a different dialect using a shared topic space. At the same time, we can estimate the user’s health literacy profile: the level of technicality on each health topic. In this chapter, we develop our language gap models and evaluate them for document retrieval.

3.2 Topic Models

Documents are written on different topics, such as dieting, measuring blood glucose levels or chemotherapy. Some words, like “POD” (a variety of insulin pump) are more likely to occur in documents about measuring blood glucose than about chemotherapy. A topic model is a way to represent this relationship mathematically. It is common to assume that the probability of a word is fixed, conditioned on the topic.

Table 2 shows an example of three topic models learned by LDA. LDA is a generative model: the model describes a hypothetical process by which each document is produced. First, an author decides what kind of document to write, by selecting a mixture of topics. Then, as she writes each word she first choose a topic for that word according to the mixture, then she choose a word according to the word distribution for the topic.

It might be mysterious how an algorithm can discover topics without knowing what they are. Consider the case of just two words, say “lose” and “weight” or “lose” and “radical.” The training algorithm will pass repeatedly through the documents,

Table 2: Example topic models: probabilities of the top ten words in several topics.

Diet		Blood Tests		Physiopathology	
foodborn	10.5%	bloody	15.1%	plaque	12.6%
eaten	8.6	titer	11.6	effector	9.2
dietary	4.5	boner	4.8	radical	5.5
healthier	4.2	arthropod	4.4	physiopathology	4.9
fatal	3.1	surgical	3.7	metabolite	4.7
anion	2.9	urinary	3.2	development	3.8
daytime	2.8	abort	3.2	immunoprecipitate	3.8
calves	2.0	arthritis	2.6	chemo	3.1
frustration	1.7	blade	2.4	ultraviolet	2.9
bodily	1.7	joke	2.1	classify	2.7

making the model fit the patterns that it finds. The probability of a word given a topic will be proportional to the number of times that word occurred in a document with that topic. Because “lose” and “weight” appear in the same document frequently, the same topic will have high probabilities for both of the terms. On the other hand, “lose” and “radical” occur together many fewer times, so they will tend to go in different topics. The same process is at work with larger sets of words: if several words from the set appear together frequently in documents, LDA will learn that they all belong in the same topic.

This approach breaks down when there is a language gap. Consider the extreme case where the gap is between English and German, with no training documents containing both languages. “Glukosemeßinstrument” is the German word for “glucose meter,” but it never appears in the same document with “glucose” or “meter,” so LDA cannot learn the relationship. Instead, LDA will learn some topics that are completely English and others that are completely German. In this topic space, two documents talking about glucose meters in different languages will look completely unrelated. We could get around this problem by adding the words from a translation to every document, so that every word that had contained “Glukosemeßinstrument” also contains “glucose” and “meter.” With this modification, LDA will be able to

leverage the patterns present in both English and German to arrive at meaningful topics that work similarly for either language.

Between the two extremes (isolated extremes and fully translated), there is a spectrum of different amounts of intermixing. At some threshold, LDA will have enough evidence to merge “Glukosemeßinstrument” and “glucose” into the same topic. It is important to understand where this threshold comes from. Many different and conflicting patterns are present in the training data. Each pattern has some amount of support, defined loosely as the number of occurrences of the pattern. The supports of patterns follow a power-law distribution, with most patterns having very little support. LDA chooses a threshold of support that tries to separate the signal (high-support patterns) from the noise (low-support patterns).

The same is true for dialects: if we naïvely supply LDA with Jamal’s question about his burn and technical documents discussing diagnosis of secondary infections, the cross-dialect patterns will have support far below the threshold needed for LDA to learn the relationship. In general, the cross-dialect patterns will have much lower support than in-dialect patterns, so that most will be pruned with the noise. In the next section, we will see a model that is able to learn the cross-dialect relationships by separating the in-dialect and cross-dialect thresholds.

3.3 Dialect Topic Models

In this section, we introduce Dialectical Topic Models (diaTM) that support joint inference over topics and dialects. DiaTM learns the probability distribution for words, conditioned on both the topic and the dialects. For each topic, diaTM learns a different model for each dialect. Two factors work together to make diaTM effective at modeling topical similarity when some documents are more technical than others. First, the training data contains a variety of documents that have more or less technical document in order to maximize the occurrences of patterns that combine topically

related non-technical and technical words. Also, the model is able to separate the task of splitting topics within a dialect from the task of aligning topics. In particular, the threshold of support needed to align two topics from different dialects may be lower than the support needed within a topic-dialect topic model. By boosting the level of the cross-dialect signal and lowering the detection threshold, we are able to force alignment between topics in the different dialects.

In Section 3.5.2, we see that the forced alignment is sometimes strained but that the model is nevertheless very useful for improving cross-dialect retrieval. We compared the performance of diaTM to LDA in several key areas. Using perplexity, KL-divergence and normalized Discounted Cumulative Gain (nDCG), we found that diaTM does 63% better than LDA at modeling the collection and 98% better at identifying similar topics across dialects.

DiaTM is a generative model very much like LDA, depicted in Figure 4a. In detail, each document d in collection c is generated according to:

Choose the topics. The author selects the mixture of topics θ_d from the symmetric Dirichlet distribution $\mathcal{D}(\alpha)$.

Choose the dialects. He also selects a dialect which is a mixture π_d of extreme dialects from $\mathcal{D}(\lambda_c)$. The Dirichlet parameters reflect differences in the predominate dialects of the collection.

Choose the document length. He select the length of the document N_d from Poisson $\mathcal{P}(\epsilon)$.

Choose each word. (1) He selects the topic z_{dn} from the Multinomial distribution $\mathcal{M}(\theta_d)$. (2) He selects the dialect t_{dn} from $\mathcal{M}(\pi_d)$. (3) He selects the word w_{dn} from $\mathcal{M}(\beta_{uz})$. (4) He selects the label \mathbf{y}_{dn} from $\mathbf{T}\mathbf{u}_{dn} + \mathcal{N}(0, \sigma_y^2)$. \mathbf{y}_{dn} is a vector of observed features for each word which signals which dialect is in use. These features could be derived from collection statistics (as in our experiments) or they could be more complicated, like the local grammatical complexity.

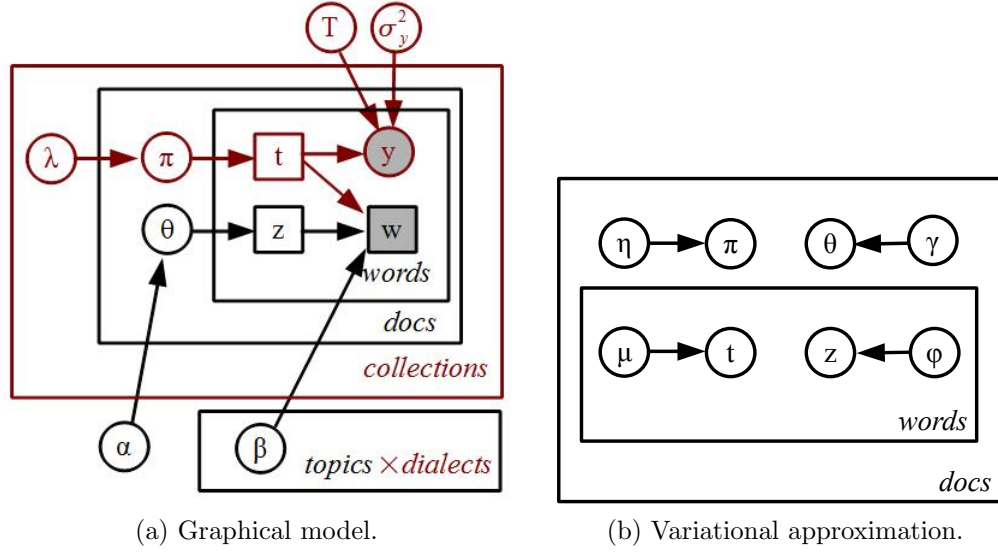


Figure 4: Diagram of the diaTM graphical and variational models.

3.3.1 Algorithm

We want to find the instance Θ^* of the diaTM model that maximizes the likelihood of the training data. However, computing the exact likelihood is intractable, so we use a variational approximation instead. First, we define the variational class by relaxing the dependencies that make inference intractable. Each variable becomes dependent only on a document- or word-specific conjugate prior, as shown in Figure 4b. We can then estimate the likelihood for a model Θ by using the most similar instance of the variational class by KL-divergence [101]. We use the variational expectation-maximization algorithm (VarEM) Algorithm 1, which starts with an initial estimate of diaTM and iteratively refines it by finding the corresponding variational model \mathbf{V} and then maximizing the likelihood with respect to the diaTM model while holding \mathbf{V} fixed.

Variational Approximation

One way to solve intractable inference problems is variational inference. This method is similar to gradient descent, in that we iteratively approximate the current solution

```

Input:  $N$ 
Output:  $\Theta$ 
Initialize  $\Theta$  ;
repeat
  foreach  $d$  in  $N$  do
     $\mathbf{V}_d \leftarrow \text{varInf}(\Theta, d)$ ;
  end
   $\Theta \leftarrow \text{maxModel}(\mathbf{V})$ ;
until convergence;
return  $\Theta$ ;

```

Algorithm 1: VarEM: Variational Expectation Maximization algorithm.

with a simpler probabilistic model that we use to find a better solution. The family of simpler models that we use is depicted in Figure 4b. In the approximation, the interactions between the four hidden variables are neglected and we model them by simple distributions from their conjugate priors. We maximize the likelihood of the training document to find its associated variational model [26], but even finding the best variational model is iterative because of interdependencies between the variational parameters, Eqn. (4). To do this, we use Algorithm 2. The indices are collection c , document d of M , word in document n of N_d , topic i of k , dialect h of H and term j . w_{dnj} indicates whether word n in document d is term j .

$$\begin{aligned}
\gamma_{di} &= \alpha + \sum_n \phi_{dni}, & \eta_{dh} &= \lambda_{ch} + \sum_n \mu_{dnh} \\
\phi_{dni} &\propto \prod_{h,j} \beta_{hij}^{w_{dnj} \mu_{dnh}} \exp \left(\Psi(\gamma_{di}) - \Psi \left(\sum_{i'} \gamma_{di'} \right) \right) \\
\mu_{dnh} &\propto \prod_{h,j} \beta_{hij}^{w_{dnj} \mu_{dnh}} \exp \left(\Psi(\eta_{dh}) - \Psi \left(\sum_{h'} \eta_{dh'} \right) + \frac{\mathbf{T}_h^{-1} \mathbf{y}_{dn} - \frac{1}{2}}{\sigma_y^2} \right)
\end{aligned} \tag{4}$$

Parameter Estimation

Given the variational approximation, it is straight-forward to determine the model that maximizes the likelihood by taking the appropriate partial derivatives, as shown in the following equations. α and λ_c are found using the Newton-Raphson method

```

Input:  $\Theta, d$ 
Output:  $V$ 
foreach  $i$  do
  | Initialize  $\gamma_i \leftarrow \alpha + N/k$ ;
end
foreach  $i, n$  do
  | Initialize  $\phi_{ni} \leftarrow 1/k$ ;
end
foreach  $h$  do
  | Initialize  $\eta_h \leftarrow \lambda_c + N/H$ ;
end
foreach  $h, n$  do
  | Initialize  $\mu_{nh} \leftarrow 1/H$ ;
end
repeat
  | foreach  $n, i$  do
  | |  $\phi'_{ni} \leftarrow \text{computePhi}$ ;
  | end
  | foreach  $n, h$  do
  | |  $\mu'_{nh} \leftarrow \text{computeOmega}$ ;
  | end
  |  $\phi_n \leftarrow \text{normalize}(\phi'_n)$ ;
  |  $\mu_n \leftarrow \text{normalize}(\mu'_n)$ ;
  | foreach  $i$  do
  | |  $\gamma_i \leftarrow \alpha + \sum_n \phi_{ni}$ ;
  | end
  | foreach  $h$  do
  | |  $\eta_h = \lambda_{ch} + \sum_n \mu_{nh}$ ;
  | end
until convergence;
return  $(\gamma, \phi, \eta, \mu)$ ;

```

Algorithm 2: *varInf()*: Variational approximation step.

```

Input:  $V$ 
Output:  $\Omega$ 
 $\alpha \leftarrow \text{findRoot}(\gamma);$ 
foreach  $c$  do
   $\lambda_c \leftarrow \text{findRoot}(\eta_c);$ 
end
 $\mathbf{T}^{-1} \leftarrow \text{regress}(\boldsymbol{\mu}, \mathbf{y});$ 
 $\sigma_y^2 \leftarrow \text{computeSigma}(\mathbf{w}, \mathbf{T}^{-1}, \mathbf{y}, \boldsymbol{\mu});$ 
return  $(\alpha, \lambda, \mathbf{T}^{-1}, \sigma_y^2);$ 

```

Algorithm 3: *maxModel()*: Model maximization step.

as for LDA [26]. \mathbf{T}^{-1} is found using linear regression. (Other regressors are also suitable.) The algorithm is presented in Algorithm 3.

$$\begin{aligned}
M(\Psi(k\alpha) - \Psi(\alpha)) + \sum_d \left(\Psi\left(\sum_{i'} \gamma_{di'}\right) - \Psi(\gamma_{di}) \right) &= 0 \\
M_c \left(\Psi\left(\sum_{h'} \lambda_{ch'}\right) - \Psi(\lambda_{ch}) \right) + \sum_{d \in c} \left(\Psi\left(\sum_{h'} \eta_{dh'}\right) - \Psi(\eta_{dh}) \right) &= 0 \quad (5) \\
\boldsymbol{\mu} = \mathbf{T}^{-1} \mathbf{y}, \quad \beta_{hij} \propto \sum_{d,n} w_{dnj} \phi_{dni} \mu_{dnt} \\
\sigma_y^2 = \frac{1}{\sum_d N_d} \sum_{d,n,h,j} w_{dnj} [\mathbf{T}_h^{-1} \mathbf{y}_{dn} (\mathbf{T}_h^{-1} \mathbf{y}_{dn} - 2 * \mu_{dnh}) + \mu_{dnh}]
\end{aligned}$$

3.3.2 Implementation

We implemented diaTM by modifying *LDA-C* (<http://www.cs.princeton.edu/~blei/lda-c>). Eqn. (5) are not computed directly: instead sufficient statistics are collected as each document is processed and the model parameters that maximize the variational estimate of the likelihood are computed using the sufficient statistics.

Computing λ requires implementing the Newton-Raphson algorithm as defined by Blei [26] but not provided in his implementation. Also, the computations for λ required higher quality libraries for digamma and trigamma, for which we used Cephes (<http://www.netlib.org/cephes>). In addition, for all calculations it was necessary

to restrict the range of numbers to avoid numerical problems. In most cases, we used the range $[10^{-100}, 1 - 10^{-100}]$.

3.4 *Topic-Adapted Latent Dirichlet Allocation*

In Jamal’s situation, it made perfect sense to model his technicality as a single variable that influenced all topics in the same ways. The reason is that in a search context we generally do not have enough information to model anything more accurate with any accuracy. However, as we work to extend these results into a social setting, a richer health literacy profile becomes both possible and important. For example, a person with diabetes would be familiar with very different language depending on the type of diabetes, her treatment plan and which books she had read. In this case, her level of health literacy cannot be expressed by a simple number: we need different estimates for different topics. We developed this idea in τ LDA.

This model shows good promise for applications in modeling health literacy. It had good performance using common performance metrics including perplexity and nDCG. More importantly, the predictive ability to identify the technicality of a document was more than 20% better than competing algorithms, which should correlate with good performance identifying health literacy.

τ LDA has two differences relative to the generative model of diaTM (Section 3.3). The graphical model is represented in Figure 5. (1) A different document technicality $\boldsymbol{\pi}_i$ is selected from $\text{Dir}(\lambda)$ for each topic i . Thus, $\boldsymbol{\pi}$ is a matrix instead of a vector. (2) The overall technicality τ_d is observed once per document instead of once per word.

We investigated three approaches for generating a single aggregate observation from the word technicalities t , each through the empirical average $\bar{y} = 1/N \sum_n t_n$. For clarity, the models are presented with two dialects (such as non-technical and technical), though extensions to higher dimensions are not difficult. Cosine regression (CR), which has a desirable interpretability property [183], is modeled by $p(\tau|\omega^T \bar{y}) =$

for LAD,

$$b_i = \text{sign}(\tau - \frac{1}{N} \sum_{ni} \omega_i \mu_n \phi_{ni}) \frac{\omega_i}{N\delta}.$$

The model parameters (α, λ, β) are computed as for diaTM, Eqn. (5). Naturally, the coupling parameters for the dialect observations depend on the regression model: for CR, $\hat{\omega} = \bar{\mathbf{h}} / \|\bar{\mathbf{h}}\|_A$, where $\bar{\mathbf{h}} = \frac{1}{M} \sum_m \tau_m \mathbb{E}_q[\bar{\mathbf{y}}_m]$, $\mathbb{E}_q[\bar{\mathbf{y}}] = \frac{1}{N} \sum_n \mu_n \phi_{nk}$, and $\mathbf{A} = \mathbb{E}_q[\bar{\mathbf{y}}_{1:M}] \mathbb{E}_q[\bar{\mathbf{y}}_{1:M}]^\top / \|\boldsymbol{\tau}_{1:M}\|^2$, $\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^\top \mathbf{A} \mathbf{x}}$ denotes the \mathbf{A} -weighted ℓ_2 -norm; for LR, linear regression; for LAD, and iterative reweighted least squares algorithm, which alternates between $\mathbf{\Lambda}^{\text{new}} = \text{diag}(\hat{\omega}^{\text{old}\top} \mathbf{Y})$ and $\hat{\omega}^{\text{new}} = (\mathbf{Y}^\top \mathbf{\Lambda}^{\text{new}} \mathbf{Y})^{-1} \mathbf{Y}^\top \mathbf{\Lambda}^{\text{new}} \mathbf{T}$.

3.5 Experiments

3.5.1 Data Sets

We collected documents from six different collections for our experiments. To represent the kinds of information lay people are interested in, we collected questions and answers from the Yahoo! Answers health category (<http://answers.yahoo.com/rss/catq?sid=396545018> accessed August 2009 through January 2010). To represent technical medical content, we used full-text journal articles from the PubMed Central Open Access Subset (<ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/articles.tar.gz> accessed September 2009) and descriptors from Medical Subject Headings (MeSH®), (<http://www.nlm.nih.gov/mesh/introduction.html> 2010 version). We also included neutral documents from the WebMD® (<http://www.webmd.com> accessed December 2009) and Centers for Disease Control and Prevention (<http://www.cdc.gov> accessed December 2009) Websites to help bridge between the slang and technical topics.

We cleaned the collections by removing any document that was not in English (using Mguesser, <http://freshmeat.net/projects/mguesser>) or had no useful content. We stemmed with the SnowBall Stemmer (<http://snowball.tartarus.org>). After cleaning, the smallest collection was PubMed containing 16,696 documents.

3.5.2 Dialect Topic Models

Vocabulary Selection

We based our vocabulary selection on data we had collected through November 2009. All of the collections except CDC were represented. For each term j in each collection c , we computed the normalized frequency f_{cj} , by dividing the frequency of the term in the collection by the frequency in the Web 1T 5-gram collection [31] with appropriate scaling to make the different collections comparable. Term j is selected for the vocabulary if $\sum_c f_{cj}$ exceeds a threshold. The threshold that we selected resulted in 20,688 terms in the vocabulary. This results in an average of 61 unique terms per document, but more than half the documents have four or fewer unique terms.

Evaluation Details

We used three dialects for the experiments, expecting them to correspond to informal (slang), formal (common) and technical dialects. We used two different kinds of features for our experiments. For some experiments we used three features closely correlated with the dialects, Eqn. (6). The slang and common features were then mapped to $[-1, 1]$ using a sigmoid function and the tech features were mapped to $[-1, 1]$ using a linear function. Because we fix T as the identity matrix when using these features, we call these the *bound* features.

$$\begin{aligned}
 S &= f_{YahooQ,j} + f_{YahooA,j} + 0.5 & y_{slang,j} &= S/\sqrt{CT} \\
 C &= f_{YahooA,j} + f_{WebMD,j} + 0.5 & y_{common,j} &= C/\sqrt{ST} \\
 T &= f_{MeSH,j} + f_{PubMed,j} + 0.5 & y_{tech,j} &= T/\sqrt{SC}
 \end{aligned} \tag{6}$$

For other experiments we used thirteen features: for each collection (counting Yahoo! questions and answers as separate collections) and each term, we computed

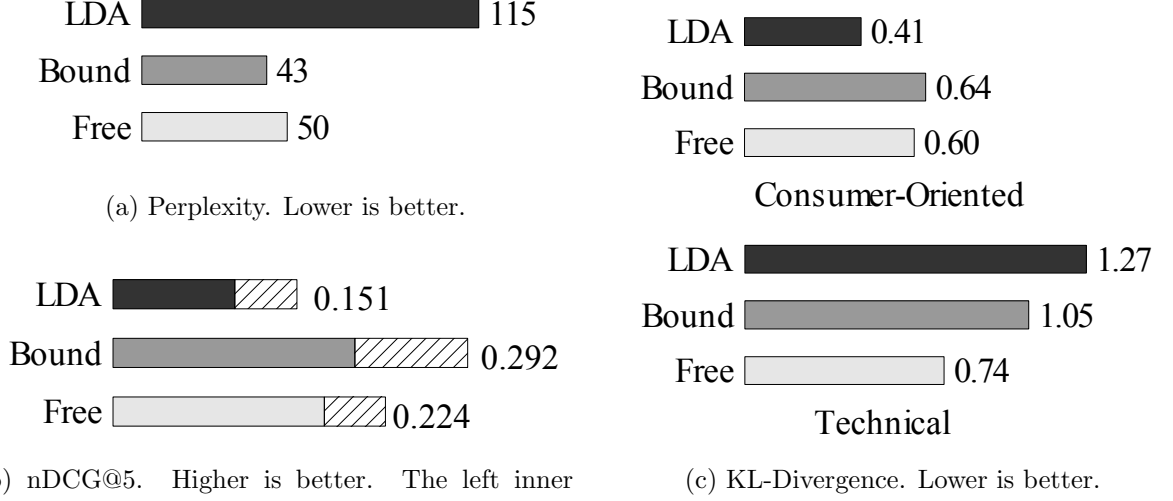


Figure 6: DiaTM significantly outperforms the LDA baseline using three important metrics.

the term frequency, $\sum_{d \in c, n} w_{dnj}$, and document frequency, the number of documents in the collection containing the term. These features are invariant across documents. The thirteenth feature is computed separately for each document as the fraction of occurrences within this document that are in the question portion of the document, $\sum_{n \in \text{question}} w_{dnj} / \sum_n w_{dnj}$. For all collections except Yahoo!, this is naturally 0. Because T is free to be learned when using these features, we call these the *free* features.

Evaluation Metrics

Perplexity. Perplexity measures how well a collection of documents is explained by a statistical model. It corresponds to the number of words a purely random model would choose from to result in the same probabilities as the model assigns the collection.

KL-Divergence. One important task of these models is to use the topics to compare documents of different technicality. For that to work, it is necessary to have good topical overlap between documents of the different dialects. In the extreme

case, if topics were partitioned so that each represented only a single dialect, then similarities across dialects would always be zero.

For a given collection c and diaTM model, it is straight-forward to compute the conditional probability distribution over topics z :

$$P(z|c) = \frac{1}{D_c} \sum_{d \in c} \frac{1}{N_d} \sum_n \sum_j \phi_{dnj} \quad (7)$$

We use KL-divergence to measure the overlap between topic distributions. KL-divergence is a metric that is commonly used to compare probability distributions: lower divergence of a distribution for a technical collection c from that of a non-technical distribution c^* indicates that there is better sharing of topics between the two collections.

$$KL_{c^*}(c) = \sum_z P(z|c^*) \ln \left(\frac{P(z|c^*)}{P(z|c)} \right) \quad (8)$$

nDCG. One valuable use for diaTM is in information retrieval: users search for documents that are more technical than the language of the query. To measure this, we simulated an information retrieval setting finding the most similar technical documents for test consumer questions. The similarity was measured using cosine similarity between the variational estimates of the topic distributions of the documents. We then computed the nDCG@5 for LDA and each variant of diaTM. nDCG@5 is a commonly used metric for information retrieval. It measures the expected usefulness of the results, discounting the intrinsic usefulness of each result more when it appears lower in the list. The measure is normalized by dividing by the best possible value given all rated results.

For ratings, we collected the top five results from several different models (approximately 30 results per query) and 200 queries and submitted the query-result pairs to Mechanical Turk for rating. The judges were asked to use the scale: 0=not relevant; 1=slightly related; 2=similar topic; 4=same topic; 5=addresses specific question. For

highly-rated pairs, we also asked the judge to provide evidence so that we could verify the reasonableness. We first submitted a random sample of 100 pairs to multiple judges. We used this data to check judge agreement. We found that it was more common for a judge to overlook the value of a document rather than to exaggerate its importance. Consequently, for pairs with multiple judgments we used the highest rating. The remainder of the pairs were each judged by a single judge, selected from the judges who had provided high quality judgments in the first round. In all, we obtained 5852 judgments on 4982 pairs. The judgments, queries and pointers to the documents are available (<https://research.cc.gatech.edu/dmirlab/node/2>).

Comparative Performance of diaTM and LDA

The results of the comparison of diaTM and LDA is shown in Figure 6. Both variants of diaTM performed well more than twice as well as LDA on perplexity. This indicates that systematic language variations play an important role in this set of documents.

The KL-divergence between topics used by consumers and those used in technical documents indicated more overlap for diaTM than for LDA. Not surprisingly, the *Free* variant, which gives the model extra flexibility in how it identifies dialects, had the best overlap. Surprisingly, LDA had better overlap between topics used by consumers and consumer-oriented content. This probably indicates a trade-off between having good topical overlap with the two different sets of documents. Still, this is surprising because we intuitively expect the consumer oriented documents to be somehow *between* the consumer and the technical documents.

For nDCG, which measures usefulness of these models for finding related documents, the *Bound* variant of diaTM is best by far. We believe that the *Free* variant has too much flexibility and overfits the perplexity at the expense of retrieval performance. Although this is a dramatic improvement over LDA, we note that we have had difficulty translating this experimental result into the real world of queries.

Table 3: Example topics found by τ LDA: the top-ten words for several topics in both the lay (β^0) and expert domains (β^1). The top row shows the technicality of each topic.

#1: $\pi = 0.06$		disease #2: $\pi = 0.15$		heredity #3: $\pi = 0.18$		medical records #4: $\pi = 0.19$		weight loss #5: $\pi = 0.26$		diagnosis #6: $\pi = 0.54$	
β^0	β^1	β^0	β^1	β^0	β^1	β^0	β^1	β^0	β^1	β^0	β^1
who	protein	problem	activ	better	nucleic	how	relat	treatment	gene	recommen	data
ask	associ	risk	chemic	below	structur	think	program	weight	analysi	medicin	method
much	immunolog	you	process	she	same	you	report	your	determin	not	import
bodi	psycholog	your	therapi	children	inhibitor	someth	previous	food	blood	tell	deriv
you	purif	fill	substanc	farther	genom	femal	web	profession	model	littl	depart
not	virolog	thought	poison	abl	possibl	googl	various	fda	amino	past	diagnost
eat	enzymolog	skin	conserv	print	pcr	mmwr	file	health	enzym	test	measur
period	induc	anyth	organ	transmiss	express	histori	databas	diet	biosynthesi	quit	complet
sometim	parasitolog	face	virus	season	chromosom	partner	establish	fat	signal	social	design
agre	patholog	regular	cell	treatment	dna	websit	analys	dose	yeast	progress	generat

3.5.3 Topic-Adapted Latent Dirichlet Allocation

Vocabulary Selection

The language gap leads to a substantial discrepancy of word usages between different domains, making it difficult to maintain a global vocabulary that is effectively balanced across domains. (Otherwise, the vocabulary could be extremely skewed such that the majority of words come from lay domains.) To this end, we first select terms (after stemming and stop-word removal) locally from each domain based on DF (document frequency) scores, and then interleave the sub-selection round-robin to form the global vocabulary (over 10K words).

Evaluation Details

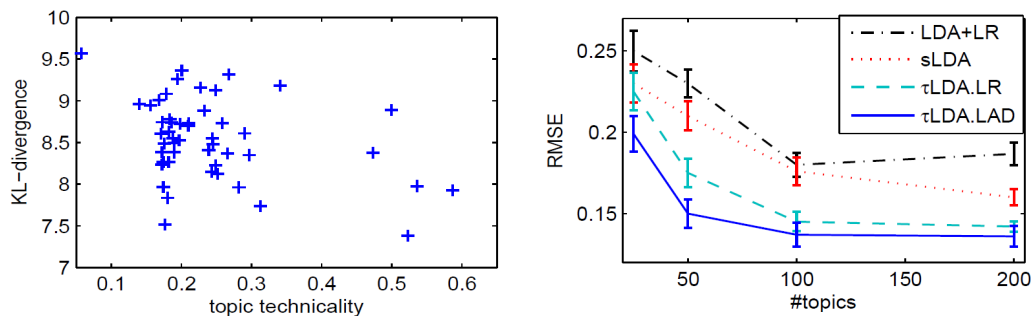
We limited our study to the case of two dialects, an informal dialect and a technical dialect. We based the observation τ_d of each document on the source of the document: Yahoo!, 0.0; WebMD, 0.3; CDC, 0.7; PubMed or MeSH, 1.0. We used a 60-40% training-testing random split and report results averaged over five trials.

Topic Quality

Table 3 shows six example topics found by τ LDA. For each topic, we show the top ten words by frequency in the informal and technical extremes. In most cases, we are able to discern the significance of a topic (shown at the top of the table) by inspecting the most common words. There is a reasonable alignment between the informal and technical versions of each topic. Moreover, the alignment is even stronger when we account for the global frequency of each word. This alignment is essential for the model to be useful for cross-dialect applications, including enabling search or recommendation of health information resources.

Contrasting the different versions of each topic, we see that the informal version uses much simpler words that are generally familiar to consumers, including “agree,” “treatment” and “progress.” On the other hand, the technical versions of the topics use much less common words (“induce,” “DNA” and “generate”) and also some unfamiliar words like “pathology,” “genome” and “PCR” (polymerase chain reaction). This validates the expectation that the model adapts the topics according to the technicality, despite being only weakly supervised.

We also calculated the technicality π of each topic. This value is also shown at the top of Table 3. This is the fraction of times the technical version of the topic was used to generate a word. A value near 0 or 1 indicates that the topic is not shared across dialects and contributes little to cross-dialect applications. Interestingly, the values ranged from 0 to 0.6, which indicates that some topics are exclusive to non-technical contexts but no topics are exclusive to technical use. In Figure 7a, we plot the KL-divergence between the informal and technical variants of each topic against the technicality of the topic. This figure is intended to show any correlation between stronger sharing of topics (technicality near 0.5) and similarity of the two versions of the topic (low KL-divergence). We did find such a correlation, but the data size was too small to ensure statistical significance.



(a) Topic variation vs. topic-technicality. (b) Domain identification accuracy.

Figure 7: τ LDA evaluation results.

Information Retrieval

We used a set of 25 queries with 100 labeled documents per query to test retrieval performance. Each query-document pair was judged by one person on a five-point scale from 0 (irrelevant) to 4 (relevant). τ LDA achieved an nDCG of 0.51 (for a loose comparison, diaTM achieves only 0.29 on a different dataset, Figure 6b), which is very good for an unsupervised model. However, this dataset has not been used to evaluate any baseline algorithm, so the number is suggestive but not definitive.

Dialect Identification

A significant advantage of τ LDA is its ability to identify the technicality at the granularity of topics. As an initial evaluation we tested how well the model could predict the aggregate dialect observation of documents. We compared the ability to predict τ using the three variants (CR, LR and LAD) against two baseline models: LDA and supervised LDA ([25], sLDA). We show the root mean squared error (RMSE) for various numbers of topics in Figure 7b, excluding CR because its performance was much worse ($\text{RMSE} > 0.3$) than any other model. It is not surprising that CR performs so poorly: it is very sensitive to noise and the “ground truth” observations we used were noisy (fixed per document source).

CHAPTER IV

QUALITY OF MEDICAL CONTENT IN SOCIAL MEDIA

4.1 *Introduction*

With the growing popularity of seeking health information online [46, 66, 74, 128, 138], the quality of the information users are finding is critical. It can be difficult for users to find quality information because much online health information is misleading or even dangerous [32, 67, 77, 125, 158, 178]. Increasingly, health information is disseminated through social media, including YouTube (<http://www.youtube.com>) and Flickr (<http://www.flickr.com>). This provides a valuable channel for health professionals [20, 168] but can equally be used by anyone else, whether qualified or not. Often patients assume that pages that have been viewed many times must be legitimate [89], but unfortunately thousands of popular videos promote anorexia [131], promote unproven, invasive medical procedures [42] or claim vaccinations are dangerous [6].

Many teenagers suffer from dissatisfaction with their weight even though they have a healthy weight or are underweight. For example, Ricciardelli and McCabe found that 60% of girls and 30% of boys desired to change their body shape [142]. *Anorexia nervosa* is an eating disorder associated with lack of appetite, fear of gaining weight and refusal to maintain a healthy weight [167]. Over a five year period, the incidence rate of anorexia among girls aged 12–24 increased from 4–5% to 6% while the online content promoting anorexia and similar disorders increased 470% to 500,000 pages [164].

In this chapter, we study the anorexia-related content on several social media sites. In Section 4.2, we examine the anorexia content on YouTube. We divide the content

into pro-anorexia (proana) content and pro-recovery content. We also examine how people interact with the content of the two types. In Section 4.3 we look at two interacting communities that promote anorexia or recovery respectively. We study the interactions between the two communities and their attempts to influence each other. Our goal is to better understand the issues surrounding the quality of health information so as to inform future interventions.

4.2 *Anorexia Content on YouTube*

We measured the prevalence of pro-anorexia and informative anorexia content on YouTube. We also examined how the consumers who viewed this information interacted with it.

4.2.1 Methods

As shown in the Figure 8, we used the YouTube application programming interface (API) to search for videos using each of the queries “anorexia”, “anorexia nervosa”, “proana” and “thinspo”. The API provides an *easy, standardized* way to get the content, while using a web browser is more time consuming and also may face challenges related to personalization. We retrieved up to 4,000 results for each query and sorting criteria (relevance, date uploaded, number of views, rating). In total, 16,000 search results were retrieved containing 7,583 videos uploaded by 3,968 users.

The top 30 most viewed videos for each of the above mentioned four queries and a subset of 30 random videos with at least 5,000 views were selected for classification by human experts. We only analyzed 140 of the 150 selected videos because eight videos were retrieved in several of the queries and two videos were newdeleted from YouTube during the reviewing process. Twenty one videos were in European languages, but experts with knowledge in the video’s language were contacted for clarification. The combined duration of the 140 videos was 11 hours.

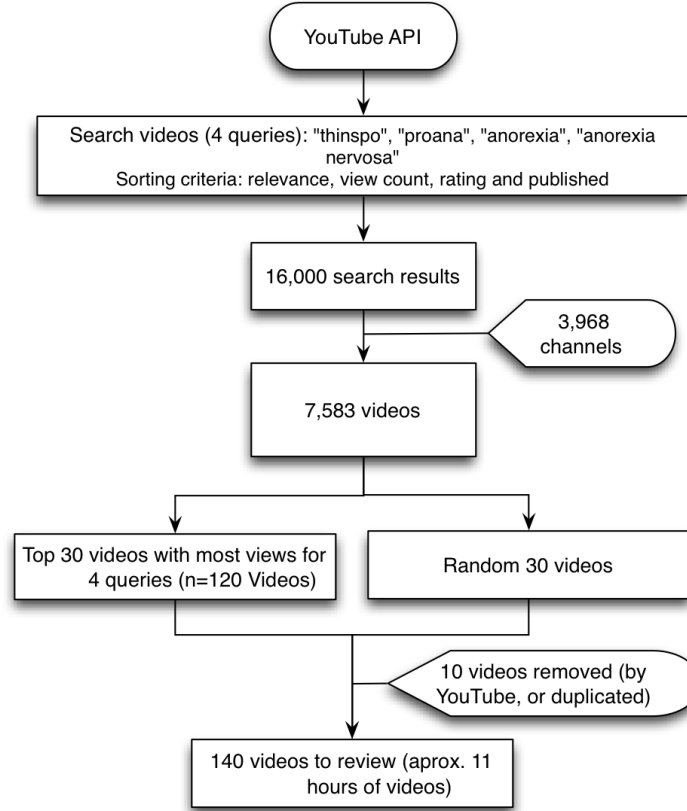


Figure 8: YouTube anorexia videos data extraction.

Three independent physicians (R, Z, K) reviewed 140 videos, and classified videos as informative (describing anorexia as an eating disorder), pro-anorexia (describing anorexia as a healthy lifestyle, promoting misleading information) or others (not related with anorexia/eating behavior). The inter-rater agreement was estimated by Fleiss Kappa [69]. For the ten videos that were classified differently by different reviewers, three additional reviewers (LFL, SAS, SV) re-classified them and reached consensus while watching the videos together.

We collected the number of times each video was favorited or commented on. We compared the rates of favoriting and like/dislike responding for videos of each class using one way analysis of variance (ANOVA). We also analyzed the content of the videos and the characteristics of the viewership for videos with demographic information available.

Table 4: Classification results for YouTube anorexia videos, including the top 30 videos per query by views and a random selection of videos with at least 5000 views.

	Total	Informative	Pro-Anorexia	Others
Top 30 videos per query	110	55% ($n = 61$)	29% ($n = 32$)	16% ($n = 17$)
Random videos	30	57% ($n = 17$)	33% ($n = 10$)	10% ($n = 3$)
Total reviewed videos	140	56% ($n = 78$)	29% ($n = 41$)	15% ($n = 21$)

Table 5: Engagement with the top 20 most viewed anorexia-related videos on YouTube, showing the number of times the videos were viewed, favorited, commented upon, liked or disliked and the percentage of each activity relative to the number of views.

		Informative	Pro-Anorexia		
Total Views	51,620,000	100.00%	9,510,000	100.00%	$P < .001$
Favorites	39,424	0.08	24,462	0.26	$P = .104$
Responses	45,486	0.08	15,209	0.15	$P < .001$
Likes	40,332	0.07	12,560	0.13	$P < .001$
Dislikes	5,154	0.01	2,649	0.02	$P = .006$

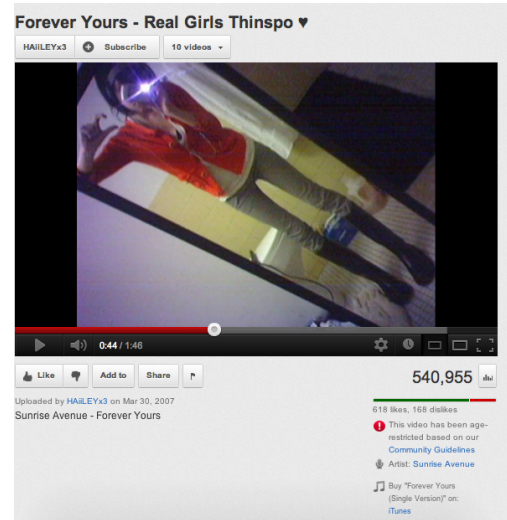
4.2.2 Results

Out of 140 videos analyzed, 29% were rated as pro-anorexia, 56% as informative and 15% as others (see Table 4). The inter-rater agreement of their classification was moderate, Fleiss Kappa = 0.494. The random selection of 30 videos with at least 5,000 views had similar percentages: 33% pro-anorexia videos, 57% informative and 10% others. Based on these results, we infer that there are about 2,000 pro-anorexia videos in the set of 7,583 videos we downloaded.

We also examined user engagement using the 20 most viewed anorexia videos (Table 5). Pro-anorexia videos are favorited 3 times more often than the informative videos. The response rate was estimated from the number of the viewers who clicked on the “Like/Dislike” icon relative to total views. Pro-anorexia video viewers responded twice more often than those of informative videos.



(a) Alternative food pyramid.



(b) Very thin models.

Figure 9: Typical pro-anorexia videos.

In most of the cases, the pro-anorexia videos featured photos of extremely thin models (Figure 9b). Most videos featured female models, but we identified a few with very thin male models. These videos were explicitly used to inspire people to become very thin. Many videos are classified as “thinspo” (inspiration to become thin). It was also common to include tips and advice for extreme weight loss. For example, Figure 9a shows a screenshot of a video in Spanish with a “thinspo” nutritional pyramid with advice such as “Smoke as much as necessary, or eat sugar-free chewing gun ” and “Drugs for losing weight such as Xenadime, Reductil.”

The informative videos were provided by a wide range of users: individuals recovering from the diseases, health organizations, news media and students. The most popular videos were produced by news agencies.

In the “others” categories there were some videos tagged with some of the keywords without any clear explanation. In some other cases the videos were from a music band named Anorexia.

In order to understand the demographic characteristics of the pro-anorexia community, we analyzed the demographic information of the viewers (Figure 10). A total

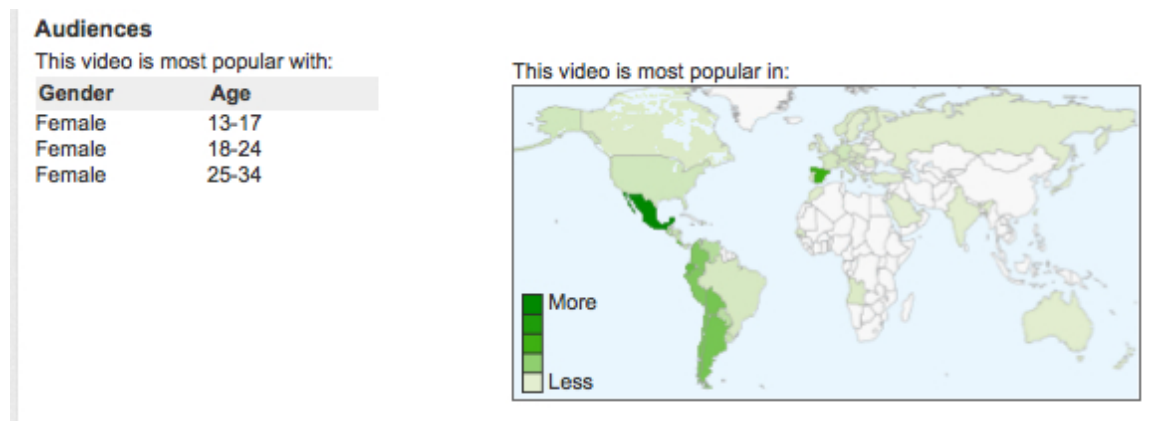


Figure 10: Demographics of the video “Princesas de Porcelana”, <http://www.youtube.com/watch?v=k8u6orW3ShQ>.

of 15 pro-anorexia videos have demographic information available. Of those, 80% had minors (13-18) as one of the top three viewer age groups and one third of the videos were not restricted by age. That implies that some videos were very popular among minors before being flagged as inappropriate for minors.

4.3 *Anorexia Communities at War*

We have seen that pro-anorexia content poses a substantial public health risk. However, previous studies were limited in the volume of data they examined and focused solely on the pro-anorexia community. In reality, the pro-anorexia community is strongly shaped by its interactions with the pro-recovery community [118]. Following Wilson et al. [176], we define pro-anorexia sites as those encouraging disordered eating. Pro-recovery sites are those which express a recovery-oriented perspective. Both kinds of sites, of which the pro-anorexia are more numerous [176], provide a platform for individual expression and community tools like discussion boards.

In this section, we examine a large corpus of pro-anorexia and pro-recovery data from the image-sharing site Flickr (<http://www.flickr.com>), to investigate the interactions between the opposing communities. We also explore the motivations of users who post pro-recovery content and its impact on the pro-anorexia community.

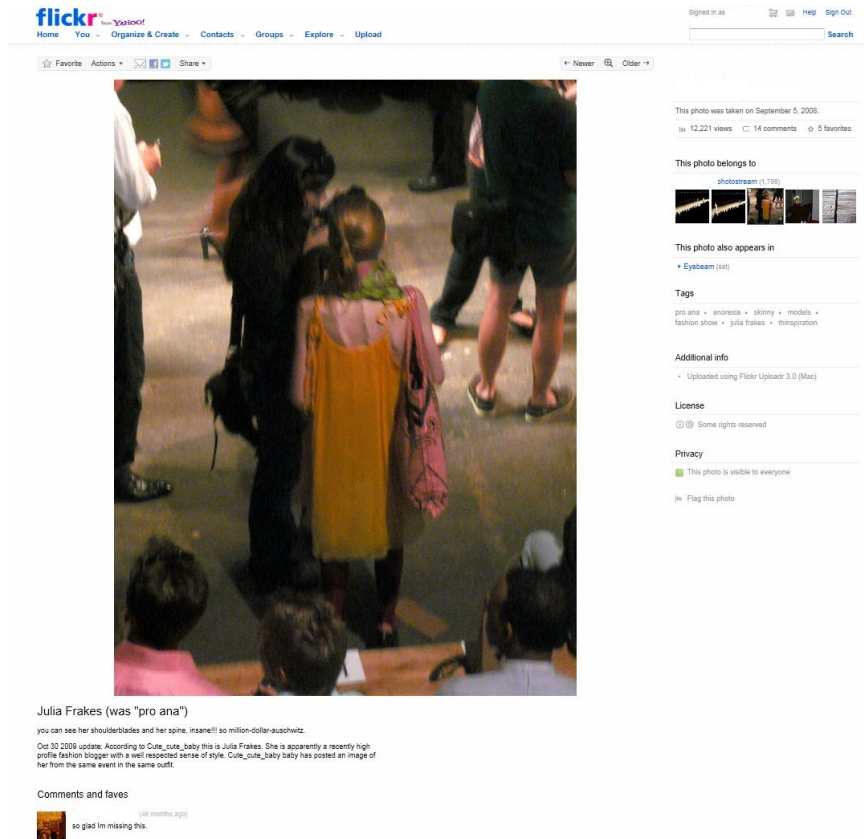


Figure 11: An example Flickr page. The image is in the top left with a textual description is below it. Comments and favoriting by other users are shown below it and the tags that categorize the image are on the right.

4.3.1 Methods

We collected four kinds of content from Flickr: (1) *contacts* represent the social network; (2) users mark pictures that they like as a *favorite*, a process that is called *favoriting*; (3) users post *comments* to pictures, with the sequence of comments often being conversational; (4) users provide *tags* that help to organize the pictures in a variety of ways. Figure 11 provides an example of a Flickr image with these annotations. This image has a creative commons license, but we chose to obscure the username to protect privacy. The research described in this sections was approved by the Yahoo IRB, and was comprised solely of observational data. We did not extract identifying information (like names or emails) except the usernames.

Flickr users can make their data public or private, but we collected only data that was public during February 2012. Data was obtained through the Flickr API (<http://www.flickr.com/services/api/>) except for favorites which were crawled from image comment pages.

We identified a set of candidate pro-anorexia and pro-recovery users using four methods.

1. Find all users who uploaded at least two images with title, description or tags containing “thinspo”, “thinspiration” or “pro-ana”. There were 162 such users.
2. Find all users who uploaded at least two photographs to the anorexia-related Flickr groups “Eating Disorders Art” (<http://www.flickr.com/groups/399147@N23>), “Anorexia Nervosa” (<http://www.flickr.com/groups/1193591@N22>), “Anorexia Help” (<http://www.flickr.com/groups/809716@N21>) or “ED Healing” (<http://www.flickr.com/groups/343610@N25>). There were 71 of such users.
3. Find all users who commented on at least two photos that had one of the aforementioned tags or that was in one of the aforementioned groups. There were 669 such users.
4. Find all users who had favorited at least two photos that were tagged as both “thinspo” and “skinny” as well as one of “pretty”, “cute” or “beautiful”. There were 14 such users.

For all four methods, users were not included if their profile had been deleted. For each of these users we obtained the information for their activity on the Flickr site:

1. Image meta-information: For the 5,000 most recent photos posted by each user, we obtained the title, tags, description, date posted, number of times that the image was viewed and the geographic location information, where available. In total, we obtained information for 543,891 photos.

2. Image comments: For the 500 most recent photos of each user, we obtained the comment text, an identifier of who left the comment and the time stamp of when it was left. In total, we gathered 2,229,489 comments on 106,877 images uploaded by 739 users.
3. Image favoriting: We extracted the list of users who favorited each of the 500 most recent photos of each user. In total, there were 642,317 favoritings pertaining to 88,337 photos uploaded by 753 users.
4. Public contacts: The list of contacts of each user was obtained. In total the (directed) contact graph contained 237,165 outgoing edges for 721 seed users. Of these, 2,821 edges were between two seed users, pointing to 543 distinct users.

Five researchers independently labeled the users according to their support for pro-anorexia or pro recovery content using a Likert scale. The data provided by the reviewers was normalized and used to categorize the users into classes. Kappa agreement in labeling was 0.51 ($P < .001$) [48], reflecting good agreement regarding the class of users. The labeling identified 172 pro-recovery users and 319 pro-anorexia users. Anecdotally, many pro-anorexia users identify themselves as having an eating disorder but pro-recovery users rarely self-identify. Of the pro-recovery users who indicate their relationship to the disease, about 20% indicate that they formerly suffered from an eating disorder.

We identified tags related to anorexia content (both pro-anorexia and pro-recovery) by representing the tags using a vector space model and selecting all tags which were at least 10 times more likely compared to images with a neutral label. A total of 25,689 *highly relevant* images contained at least one of these tags.

Table 6: Distinguishing tags for pro-recovery and pro-anorexia users on Flickr.

Pro-Recovery	home, sign, self-portrait, glass, cars, plants, building, mother, sunshine, bird, plant, autumn, garden, female, fence, dog, warm, architecture, stone, birds
Pro-Anorexia	thinspiration, doll, thinspo, skinny, thin, cigarette, sexy, landscape, legs, abstract, long, day, street, body, blonde, sister, Nikon, up, life, model

4.3.2 Results

Posting volume

We looked at the volume of posting highly relevant pro-anorexia and pro-recovery content over time. Both types of content are similar in volume, and have grown rapidly since 2009. The Spearman correlation between the two time series is 0.82 ($P < .001$), demonstrating an extremely high correlation. Pro-recovery users are, in general, more active, posting a median of 196 photos, compared to 105 photographs by pro-anorexia users ($P < .001$ by ranksum).

We also examined tags that identify the photo as dealing with the photographer himself. These included the tags “self”, “self-portrait”, and “me”. Pro-anorexia users are responsible for 24% of photographs with these tags, which is a low proportion considering that 40% of the photographs are posted by these users. However, these tags appear in 42% of the highly relevant photos (where 40% of the photographs are posted by pro-anorexia users). Therefore, pro-recovery users tend, in general, to post more photographs of themselves. When dealing with anorexia-related issues, both pro-anorexia and pro-recovery users are similarly interested in their own images.

Most Indicative Tags

To identify distinguishing tags, we computed the ratio of the probability of selecting each tag by one class of users relative to the other class. Table 6 shows the most distinguishing tags for each class. One could infer from these lists that pro-recovery

users post pictures on a wide variety of topics but pro-anorexia users are more focused on images related to body image. Also noteworthy is the use of the tag “cigarette”, which is frequently cited as a way to decrease hunger in the pro-anorexia community.

Inter- and Intra-Community Connectivity

Contacts are more likely in class: 72% of contacts by pro-recovery users were to users of the same class, while 59% of the contacts by pro-anorexia users were to users of the same class ($P < .001$ by χ^2). Similarly, comments are more likely in-class, with 83% of the comments by pro-recovery users and 74% of comments by pro-anorexia users being made to users of the same class ($P < .001$ by χ^2).

Pro-recovery users are approximately as likely to favorite a photo regardless of the posting user’s stance (56% vs. 44%), but pro-anorexia users are 8.4 times more likely to favorite a photo posted by a pro-anorexia user than by an pro-recovery user (89% vs. 11%, $P < .001$ by χ^2).

We compared the tags used by users to describe their photos cosine similarity and a BOW model. The average similarity of tags between photographs made by pro-anorexia users was 0.259, between pro-recovery users 0.202, and between the tags of pro-anorexia and pro-recovery users 0.225 ($P < .001$ by ranksum). Therefore, the similarity between pro-recovery and pro-anorexia users is greater than within pro-recovery users. This is partly because pro-recovery users have a broader range of interests (Most Indicative Tags, above), but also because pro-recovery users often choose tags associated with the pro-anorexia camp. For example, the tag “thinspiration” and its variations are used by 36.8% of pro-anorexia users and by 6.6% of the pro-recovery users. Even more striking is that the tag “pro-anorexia” (and its variations) are used by 1.7% of pro-anorexia users, but 2.4% of pro-recovery users. Overall, the Spearman correlation between tag frequencies in both communities is 0.67 ($P < .001$).

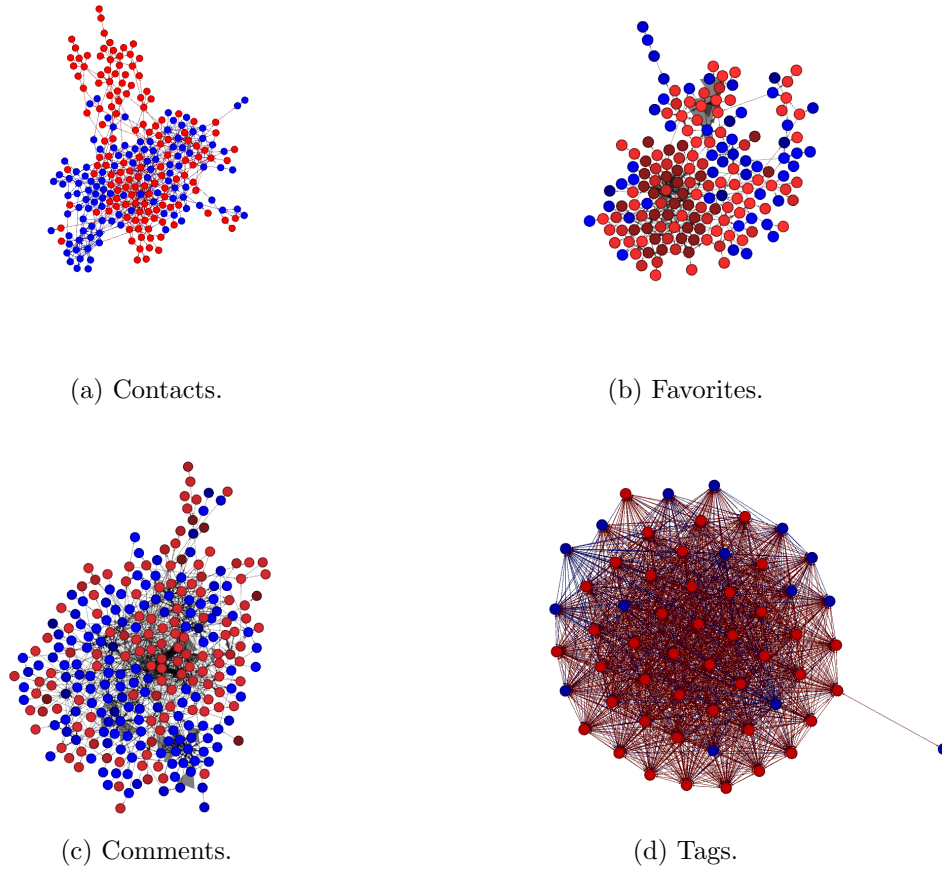


Figure 12: Network graphs according to four connection types. Each node is a user. Blue represents a pro-recovery user, red represents a pro-anorexia user. Only the main connected component is shown in each graph.

Figure 12 shows the networks laid out by Gephi (<http://gephi.org>) based on different types of connections between users. For tags, we considered two users to be similar in their tags if the cosine similarity of their tags was greater than 99%, a threshold that was selected to obtain a sparsity level similar to that of the three other networks. As evident from the graphs, the two classes are intermingled, but are most highly so when observing tags. In order to estimate the separation between classes according to the different networks, we labeled each user according to the difference between the numbers of his neighbors of each class. The predictive ability of each network was estimated by the area under the receiver operator curve (AROC). The AROC using the comments or contacts was 0.74, whereas the area using favorites was

Table 7: Rates of stopping to post highly relevant photos and average days to stopping by pro-anorexia (PA) and pro-recovery (PR) users following comments by other users.

	Rate	Days
PA comment on PA image	61%	225
PR comment on PA image	46	329
PA comment on PR image	61	366
PR comment on PR image	71	533

0.53 and 0.52 using the tags network. Thus, comments from people of a given class and the class of one’s contacts are the best predictor of one’s class.

Inter-Community Posting as an Intervention

In Section 4.3.3, we will discuss the possibility that pro-recovery users may be trying to influence pro-anorexia users by commenting on pro-anorexia images. With that as motivation, we measured whether comments by the other camp would influence future posting behavior. We identified users who stopped posting highly relevant (either pro-recovery or pro-anorexia) content, but continued to post other photographs. We defined posting cessation as stopping to post highly relevant pictures for at least 3 months, while continuing to post other pictures. This definition was used in order to ascertain that the user did not abandon the site, which could happen for several reasons. For each comment on an image, we tested for posting cessation by the user who received the comment.

The cessation rates and the average days to cessation are shown in Table 7. As the table shows, comments by pro-anorexia users have the same effect on both types of users. However, comments by pro-recovery users decrease cessation in pro-anorexia users and increase it in pro-recovery users. This is also evident in the average days to cessation, which are higher for pro-anorexia users when comments are from the opposite camp, but lower for the pro-recovery camp.

Table 8: Hazard model relative risk (RR) for posting users from the pro-recovery (PR) and pro-anorexia (PA) camps. Numbers denoted by a star are statistically significant at $P < .05$. Coefficients are calculated after normalizing the data to zero mean and unit variance.

Covariate	PR RR	PA RR
Activity prior to 30 days		
Number of photos	-28.8%*	-20.2%*
Number of highly relevant photos	1.3	-20.0*
Number of views	-6.9*	23.6*
Number of views of highly relevant photos	2.3	17.8*
Number of comments from same-class users	5.9*	-5.5
Number of comments from other-class users	-21.9*	-11.0*
Fraction of comments from same-class users	-2.7	-23.5*
Activity in past 30 days		
Number of photos	18.2*	23.7*
Number of highly relevant photos	-9.0*	-2.2
Number of views	17.7*	2.9
Number of views of highly relevant photos	22.0*	-0.7
Number of comments from same-class users	2.4	-6.6*
Number of comments from other-class users	-0.2	5.9*
Fraction of comments from same-class users	18.8*	6.3*

Another way to quantify these effects is through the use of a Cox hazard regression model [50]. For each of the classes we built a separate model, using the covariates described in Table 8. We transformed each covariate by taking the logarithm and then normalizing to zero mean and unit variance. We computed these features at a resolution of 10 days, and attempted to predict the hazard of posting a highly relevant image in the following days. The table also shows the relative risk for each covariate from the regression models.

The results of this model show that pro-anorexia users are encouraged to post additional photographs when many people view the photos they have posted and, in the short term, by comments of the pro-recovery group. Pro-recovery users are encouraged by viewings and by comments from their own group, but discouraged by comments from pro-anorexia users.

4.3.3 Discussion

This study found that about 30% of the videos on YouTube related to anorexia were misleading. Interestingly, the 30% holds both for the top search results and for randomly selected videos, from which one might conclude that it is an accurate depiction of the ratio more generally. Furthermore, these findings are similar with the studies that analyzed YouTube videos for other diseases.

Although YouTube video content has been evaluated as a source of information for several diseases, none of these studies assess viewer behavior. When we analyzed the behavior of the viewers in this study, we found that the pro-anorexia videos are favorited three times as often as and responded to twice as often as the informative videos.

We also investigated the interactions between the pro-recovery and pro-anorexia communities. We found that comments and contacts are more likely in-community than between communities. Favorites by pro-recovery users are equally likely to members of the two communities, but pro-anorexia users are 8.4 times more likely to favorite pro-anorexia photos. Taken together, these show two active communities of users, who mostly interact in their own community. The main divergence from this behavior is in marking favorite photographs. This is most likely because the receiving user can delete comments and contacts, but not favorite markings. Therefore, it is likely that the figures for comments and contacts would be different before this filtering.

Our results show that the two communities coexist separately according to the (moderated) comments and contacts, but interestingly, they are intermingled according to the tags and (unmoderated) favorite links. A plausible explanation for this is that the pro-recovery group is trying to expose itself to pro-anorexia users through the use of similar tags, thus causing images posted by them to surface in searches of pro-anorexia users. Furthermore, by marking as favorites pro-anorexia images they

are causing the users who posted them to be exposed (indirectly) to pro-recovery content.

However, by modeling the likelihood of discontinuing to post photos, we found that an intervention of comment posting by pro-recovery users is counter-productive, causing pro-anorexia users to continue posting for longer and, if they cease posting, to do so later. Previous studies have found that pro-anorexia users perceive themselves as isolated in the physical world [72]. It may be that pro-recovery comments reinforce this feeling, entrenching users in their behavior.

Thus, pro-recovery users undertake two kinds of interventions. First, they expose pro-anorexia users who search for pro-anorexia content to pro-recovery content. Second, they post comments to pro-anorexia content. The latter, at least, is detrimental, as measured by the cessation of posting.

Some aspects of our study cannot be generalized to other pro-anorexia websites. Flickr is a general-purpose photo-sharing platform where users can create their own groups and freely intermingle with multiple communities. In this context, pro-anorexia and pro-recovery communities have to co-exist in the platform and therefore interactions are likely to happen. Our study findings are most likely to generalize to other general social networks where pro-recovery and pro-anorexia users can easily interact.

Clinical Relevance

The understanding of these communities is of key interest to public health officials aiming to prevent anorexia. Burton highlighted the importance of public health community mining for public health professionals [36]. In fact, to understand the dynamics of health communities, it is of vital importance to design online social methods for health promotion [75]. Our results show that the pro-anorexia and pro-recovery communities in social media have grown in their volume in a similar manner over

time. Consequently, it is important to study the two communities and the interactions between them. That information can be extracted automatically as we did in this study in order to enact surveillance of the prevalence of the community online. As shown in this study, there are cases of pro-anorexia users actively trying to persuade members of pro-recovery communities. Public health interventions on content-based social platforms, such as Flickr and YouTube, must be aware of the possibility that social network dynamics may undermine the effects of their intervention.

Automatic analysis of the pro-anorexia communities can be used to improve online interventions, such as warning messages that have been already piloted with promising results [112]. The suggested intervention by Lewis et al. [105] to filter access to pro-anorexia websites from the search results can take advantage of our findings about the different tagging patterns used by the pro-anorexia community. An online intervention aimed at people with anorexia can be displayed when users are searching for pro-anorexia related terms. In addition, the network dynamics described in this paper can guide public health officials disseminating content in social networks so that they can gain more visibility and improve their reputation within the different online communities.

Limitations

Our study has several limitations: First, it is unknown how to generalize our results beyond YouTube and Flickr. Second, while our manual labeling is highly likely to result in correct class assignments (high precision), our collection method does not ensure high recall. As a result, we are likely missing relevant videos and users who were not identified by our collection method. Third, as our data collection relied on public APIs, we have very little interaction data such as specific viewing behavior. For example, users who only browse content but never post any photo, comment, or favorite link will be missing from our data, as will the effect of such viewing on

individual behavior. Finally, our data does not contain any clinical indication of a user’s actual state. This is especially evident in using posting cessation as a measure of engagement, where actual information on recovery or otherwise would have been of immense value. However, privacy concerns make it unlikely that such ground-truth labels can be obtained.

4.4 Conclusion

We urge scientific communities to disseminate their research results in the social media. Flagging videos by the community can prevent some of the harm, but such algorithms takes time to detect hazardous videos and therefore do not protect minors from watching them. Robust search and filtering algorithms are essential to facilitate the search of informative videos while filtering out the misleading content.

Our investigation of photo sharing behavior by the pro-anorexia community and the pro-recovery community has uncovered significant differences between the two. Better understanding of the pro-anorexia community can guide public health officials designing online interventions for people at risk of eating disorders, or to mitigate the effects of such communities on individuals. In addition, our study is a first step towards the design of advanced filtering tools that will prevent pro-anorexia content from reaching vulnerable individuals.

CHAPTER V

ONLINE HEALTH DISCUSSION GROUPS

5.1 Introduction

Tanya was a generally healthy young woman who liked to run, and was happy about her recent weight loss. When she went to her doctor with chronic fatigue, she was shocked to discover she had diabetes. The news left her feeling confused, worried and very alone, even though she was one of over 7,000 people with the same news that day.¹ Many people with chronic diseases seek support from online communities like TuDiabetes (<http://www.tudiabetes.com>), where they can learn from other patients and not feel so alone.

TuDiabetes is a diabetes community with 22,000 members, operated by the Diabetes Hands Foundation. It provides forums for discussions with broad appeal and more focused groups where people with specific interests can carve out a niche. The forums are very popular, but they provide limited organization so that valuable older discussions are buried by recent chatter. For example, the forum “New to Diabetes?” currently has 85 pages of discussions. One of its discussions alone has 105 pages of posts.² In fact, Manny Hernandez, the president of the Diabetes Hands Foundation, stated that searching for older forum content is what frustrates their users the most. They also have little capacity for developing a shared sense of identity around a specific topic. The groups are intended to organize people and discussions around specific interests, potentially providing more organized access to relevant information and valuable social connections. However, only a small number of groups are active

¹1.9 million Americans were diagnosed with diabetes in 2010 [127].

²Accessed August 20, 2012.

enough to regularly benefit their members. For example, only twelve groups averaged 3 or more posts per week over the past three months. We surmise that a group below this threshold does not have significant conversation between members.

The vitality of a group is complex and difficult to define analytically. However, the energy that consumers put into groups when they join or participate in discussions provides a window to some important aspects of vitality. The Diabetes Hands Foundation is actively seeking ways to increase the number of active groups that enrich the community. For example, they regularly discuss some of the groups in the newsletter they produce. They also recently (December 2011) introduced a page on the site that organizes some of the groups by category, in the hope that consumers would have an easier time finding these groups.

In this chapter, we seek to identify why some groups in health communities are more successful than others at attracting new users and discussion activity. To accomplish this, we analyze the correlations between group features and the rates of social events in the groups of TuDiabetes. We use event history analysis [5] to identify the correlations, incorporating novel methods to handle censoring for privacy and significant correlations between group features.

Although our techniques cannot establish the causes of differences in group vitality, they identify important relationships that warrant close attention, for example by user studies. We hypothesize that these patterns would be typical of health communities and other communities where the participants have an extrinsic motivation to participate. First, TuDiabetes, like many online communities, mainly advertises the popular groups. For example, the list of groups is sorted by the amount or recency of activity. We find that community members who have been members of TuDiabetes for fewer than nine months are attracted to these more visible groups *en masse*, so that these groups become increasingly similar in terms of who is participating. This process results in a small set of densely connected groups that reduces valuable depth

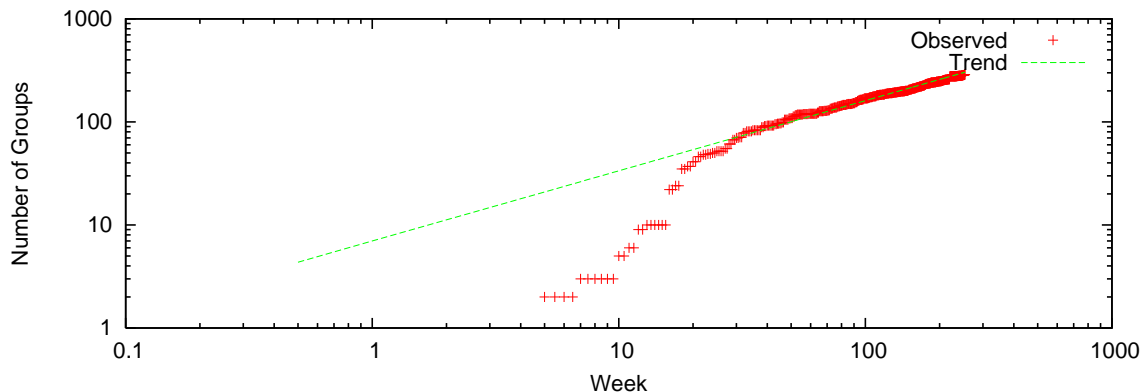


Figure 13: Number of groups in TuDiabetes by week.

and diversity [73]. Community managers can reverse this trend by increasing the diversity of groups that are visible, by for example, personalized recommendation that selectively steers users into different groups. The Diabetes Hands Foundation has responded to these results by highlighting some of the less popular groups in newsletters to increase their visibility. They are also eager to provide personalized recommendations to their users.

Another important insight is that the majority of active long-term community members mainly interacted with other long-term users, thus limiting their contribution to the community. On the other hand, some long-term members make a conscious effort to invest into the community by becoming active in groups that need reviving. The differentiation between these two behaviors seems to happen for most users by the time they have been members of the community for nine months, offering a critical time for community managers to encourage more long-term members to invest in the community.

Section 5.2 will analyze the social interactions that happen in the TuDiabetes network. Section 5.3 will explain how event history analysis can be used to identify group characteristics that are predictive of different levels of activity in the group. Our experiments will be described in Section 5.4 with a careful analysis in Section 5.5 before we conclude in Section 5.6.

5.2 Activity in Online Health Communities

TuDiabetes provides a rich community experience for people with interest in diabetes, including a social network, personal pages and blogs, 26 discussion forums and the ability to forge group identities around specific topics. Our operationalized goal is to understand what group characteristics drive people to join and participate in discussions, so that we can identify ways to increase the population of healthy groups. To this end, we repeatedly crawled TuDiabetes for event history data (IRBprotocol H11049). In order to preserve privacy, user identifiers were replaced with a cryptographic hash, data were only recorded for groups with at least 25 members and times were recorded at the granularity of one week. From this data, we extracted events corresponding to three types of activities: joins, discussion starts and discussion posts.

At the end of our study, the community had 22,000 members and 570 groups with an average of 72 members per group. Figure 13 shows how the number of groups has been increasing over time, estimated using the date of the earliest discussion in each group. After censoring, we were able to observe the membership of 266 groups accounting for 84% of all memberships. The memberships were collected at 25 different times (initially irregular and eventually weekly) between April 2011 and January 2012 in order to detect the approximate time each new member joined. 9,769 users were members of at least one group at some time during the study, on average 4 groups each. For each user who participated in a discussion, we counted the number of replies by that user during each week (44,922 (*discussion, user, week*)-triples). We observed all of the discussions that happened in the forums (25,162) and in the 266 observed groups (7,914) from the inception of TuDiabetes. 253 groups had at least one discussion, averaging 31 discussions per group. On average, each discussion had 9 replies. We saw some evidence that users had left groups (12% of user-group memberships) or that discussions had been deleted (1% of discussions). For example, some users who had posted in groups were no longer members of the group during

our study. Whereas these deletion events are a significant fraction of the interaction events in TuDiabetes, they could warrant additional study.

The groups have varying character. The largest group is the group for patients who use Minimed Paradigm insulin pumps. 34 of the 1400 members have “liked” this group in TuDiabetes and 12 have “liked” it in Facebook. The members start approximately a dozen discussions each month, each consisting of a dozen posts on average. For lighter-weight communication, they also make heavy use of the “wall.” Discussion ranges from seeking help with troubleshooting the devices to general advice about managing diabetes. The diversity of discussion topics, including topics with no obvious relationship to insulin pumps, highlights the role this group plays for its members as a miniature subcommunity.

For contrast, we examined the group on traveling with diabetes. We observed very little conversation happening in the discussions—most posts on discussions were by a user who was new to the discussion. Social conversation is rare and only takes place on the “wall.” Also, few members posted to this group over an extended period of time. Only one member has ever “liked” this group. Each of the few discussion threads is focused on the topic and has a very long lifetime as someone new finds it relevant and adds a reply years after the discussion started. Because the interactions with the group appear utilitarian, we surmise that it fulfills no significant social role for its members.

User activity levels follow a few patterns. Meredith (not her real name) first joined TuDiabetes a year ago. She posted in a forum her first week, but did not join any groups for two months, at which time she joined the appropriate regional group, like nearly everyone else. She also joined the Minimed group and a group for people who had had diabetes a long time. The regional group finally had its first discussion a half-year later, shortly before she joined two additional groups. One might imagine that she had forgotten all about the regional group until she read that discussion.

Meredith is like most community members in that her activity is fairly constant after an initial burst when she first joined and another burst when she discovered groups.

Another user, Fred, joined TuDiabetes in October, 2009. He started by joining an insulin pump group and posted 21 messages in 8 forums his first month. Since then, he has joined a new group every month or two; he is in many of the same groups as Meredith. He now normally posts around a hundred replies each month, mainly in the forums. In contrast to Meredith, Fred’s activity level has increased over time. While members like Fred are not common, in aggregate they contribute a fifth of the activity in TuDiabetes.

For the purposes of event analysis, we define the *age* of a user as the number of weeks since the user joined the TuDiabetes community. Figure 14 shows the average number of events of each type per user as a function of user age. For this calculation, we only include users who have some activity (approximately half of the members of TuDiabetes never join any group or participate in any discussion) and only consider join events that we observed during the study. New users join several groups during the first week, dropping rapidly according to a power law thereafter. There is a marked change in trajectory at about 9 months that is likely caused by users like Fred who contribute a disproportionate amount of activity.

Participation in discussions reveals a similar pattern. The number of discussions and their replies both also drop off according to a power law. Unlike group joining behavior, discussion participation does not begin to rise for older members. They do however, in aggregate, continue to post the same number of replies to fewer discussions, suggesting that they choose to participate in discussions with a more interactive nature.

Mirroring the changing trends that we see in Figure 14, we split the users into three groups according to their age at the time of an event. We call a user “young” for the first four weeks after he joins the community. At a given time, there are typically

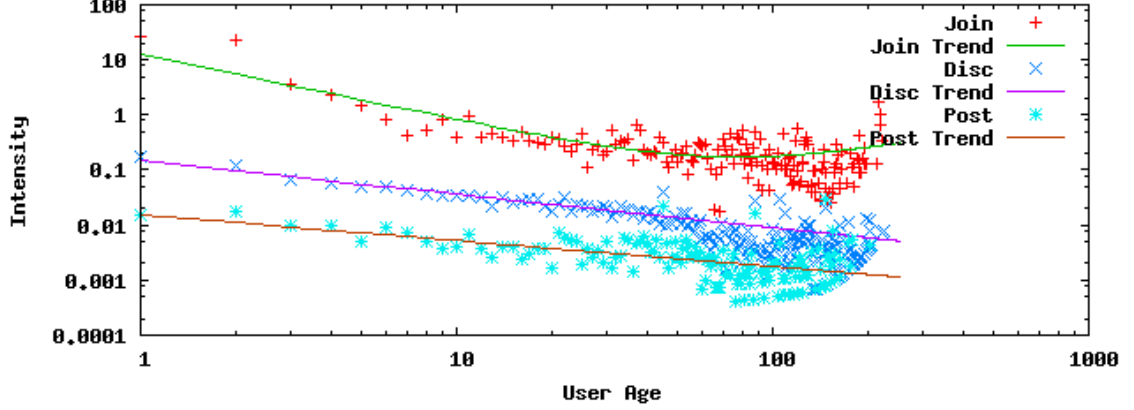


Figure 14: Group event intensity by user age.

200–300 young users providing a fifth of the activity. Similarly, middle users have been in the community for 5–39 weeks and old users 40 or more weeks. There can be as many as 1,500 old users active in a month, although in most months only a few hundred are active, accounting for a fifth of the activity. In the next section, we show how event history analysis can identify the factors that influence group activity by users in different age categories.

5.3 *Event History Analysis for Social Networks*

In a community, events happen at various times for different reasons, most of which are not directly related to the community. However, if there is a group characteristic that influences the timing of events (for example, starting a fascinating discussion may increase the rate of posts in the group), the relationship can be quantified using event history analysis [5]. Event history analysis is a statistical technique that identifies when specific features are correlated with differences in the rates that events happen, even when most of driving forces behind individual events are hidden.

As is typical with event history analysis, we model the event times with Poisson processes and Cox intensity functions [50],

$$\lambda_{eg}(t) = \alpha_e(t) \exp(\beta_e^T \phi_g(t)).$$

The intensity function $\lambda_{eg}(t)$ gives the average rate of events of type e in group g at time t . The baseline intensity function $\alpha_e(t)$ provides the average rate of events of type e for the community as whole. $\phi_g(t)$ is a vector of features for group g that are hypothesized to effect the rate of events. β_e is a vector of parameters to be learned that controls how much each feature influences the rate of events. This model allows different types of events (for example a new member joining or someone posting in a discussion) to have different rates that are influenced by the features in different ways.

We estimate β using relative risk assessment through the partial likelihood function [51]. For a Poisson process, the probability of event e occurring in group g at time t is proportional its intensity function. The partial likelihood of the event is its probability conditioned on an event occurring at that time,

$$\frac{\lambda_{eg}(t)}{\lambda(t)} = \frac{\alpha_e(t) \exp(\beta_e^T \phi_g(t))}{\sum_{fh} \alpha_f(t) \exp(\beta_f^T \phi_h(t))}.$$

Let $\delta(t)$ and $\delta_{eg}(t)$ denote the number of events time t , possibly restricted to a specific type e and group g . Then the probability of this particular set of events without replacement is

$$\frac{\delta(t)!}{\prod_{eg} \delta_{eg}(t)!} \prod_{eg} \left(\frac{\alpha_e(t) \exp(\beta_e^T \phi_g(t))}{\sum_{fh} \alpha_f(t) \exp(\beta_f^T \phi_h(t))} \right)^{\delta_{eg}(t)}.$$

Combining the partial likelihoods for all events, we get the objective function

$$\begin{aligned} \text{minimize } \mathbb{L} = & \sum_t \delta(t) \ln \left(\sum_{eg} \alpha_e(t) \exp(\beta_e^T \phi_g(t)) \right) \\ & - \sum_{teg} \delta_{eg}(t) \beta_e^T \phi_g(t). \end{aligned} \quad (9)$$

Given a set of actual events $(e_i, g_i, t_i)_i$ and estimates of the features for all groups at those times $(\phi_h(t_i))_{hi}$, the optimal β is found using stochastic gradient descent

[191]; the contribution to the gradient from each event is

$$\gamma_i = \left(\frac{\alpha_{e_i}(t_i) \exp(\beta_{e_i}^T \phi_{g_i}(t_i))}{\sum_{fh} \alpha_f(t_i) \exp(\beta_f^T \phi_h(t_i))} - 1 \right) \phi_{g_i}(t_i).$$

With learning rate η and N_e total events of type e , the update for each event becomes

$$\beta_e \leftarrow \beta_e - \frac{\eta}{N_e} \gamma_i \quad (10)$$

We iterate numerous times over the randomly permuted events, decaying the learning rate on each iteration.

5.3.1 Censoring

Our data exhibit a number of different kinds of censoring that makes the true values of the dynamic features at the time of an event unknown. For example, we may know that a user joined a group sometime between when the user joined the community and when she first posted in the group. Thus, if $\phi_g(t)$ contains features counting the members in the group, the partial likelihood function Eqn. (9) is an unobserved random variable. To properly handle the censoring, we optimize with respect to the expected partial likelihood,

$$\begin{aligned} \mathbb{E} \mathbb{L}_i &= \mathbb{E} \ln \left(\sum_{eg} \alpha_e(t_i) \exp(\beta_e^T \phi_g(t_i)) \right) - \mathbb{E}(\beta_{e_i}^T \phi_{g_i}(t_i)) \\ &\approx \ln \left(\sum_{eg} \alpha_e(\mathbb{E} t_i) \exp(\beta_e^T \mathbb{E} \phi_g(t_i)) \right) - \beta_{e_i}^T \mathbb{E} \phi_{g_i}(t_i), \end{aligned}$$

where the expectations are conditioned on all observed events, regardless of the event time. Note that the features are functions of the history up to the time of the event but the expected values of the features are functions of observations at all times. For counts of events that are constrained to intervals, we prorate each event linearly over the interval and then aggregate the prorated values.

5.3.2 Interpretation

Relative risk, the impact of a unit change in a feature on the rate of events, is commonly used to interpret intensity models [5]. For the Cox model, the relative risk of feature i is simply $\exp(\beta_i)$. To make the relative risk meaningful in our application, we transform the features to remove trends in mean and variance so that the relative risk consistently reflects the effect of a change in feature value of one unit of standard deviation.

Many of the natural features that might be predictive of event intensities are highly correlated. For example, it is obvious that a group with more members is likely to also have more discussions. This makes interpretation impossible because the model could attribute the relative risk of the shared information to either feature. To make the model identifiable, we use partial singular value decomposition (SVD) to transform these features into a basis that explicitly models the sharing.

The SVD also helps with interpretation in another way. A known problem with interpretation of regression-based intensity models (including the Cox model) is that the full impact to the risk from knowing a single feature is not present in the model [5]. For example, knowing the value of one feature also gives a good estimate for correlated features. As a result, the relative risk of a feature underestimates the impact of the feature. When we train the model directly from the original feature space, there is no way to recover these interactions. In contrast, when the model is trained in the reduced basis, we can define the full relative risk for a single feature as

$$RR_i = \exp \left(\frac{\mathbf{U}_i \boldsymbol{\Sigma}^{-1} \beta_i}{\|\mathbf{U}_i \boldsymbol{\Sigma}^{-1}\|} \left(\sum_j \mathbf{U}_j \right) \boldsymbol{\Sigma}^{-2} \mathbf{U}_i^T \right), \quad (11)$$

where partial SVD approximates the row vector of feature i for each group and time in some fixed order with $\mathbf{U}_i \boldsymbol{\Sigma} \mathbf{V}^T$. We might extend this definition to the full relative risk $RR_{\mathcal{J}}$ of a set \mathcal{J} of observed features and further define the conditional relative risk $RR_{i|\mathcal{J}}$ as the additional risk from knowing feature i when features \mathcal{J} are already

known as $CRR_{i|\mathcal{J}} = RR_{\{i\} \cup \mathcal{J}} / RR_{\mathcal{J}}$. Then SVD enables us to compute the CRR by replacing U_i in Eqn. 11 with its component orthogonal to U_j for every $j \in \mathcal{J}$.

5.4 Experiments

For each group ($n = 569$) and time step ($n = 251$), we estimated the features: (1) Number of times an event (join (J), new discussion (D) or post on a discussion (P)) occurred in the group. For J and P events, the counts were aggregated by the age of the user at the time of the event (young Y, middle M, old O or unknown U), resulting in nine types of event count features. (2) For each of these feature types, we also computed the time-weighted counts using four half-lives, one week (Wk), four weeks (Mo), sixteen weeks (Qt) and 64 weeks (Yr), yielding 36 additional features. (3) We identified the activeness weight of each user in a group based on the length of time since the most recent post in the group, using a half-life of 26 weeks. We tallied the number of active users, the fraction of active users with unknown current age and the fraction of known-age active users who are young or old for four more features. (4) For each previously mentioned feature, we also computed a collaborative filtering (CF) equivalent. To compute feature \hat{f}_g derived from feature f_g for group g at time step t , we used the weighted average of the value for other groups h , $\sum_h \langle g, h \rangle f_h$. $\langle g, h \rangle$ is the Pearson correlation between the members of groups g and h at time t , so that the computed feature is most similar to the feature in groups with similar members. The name is inspired by item-based collaborative filtering [150].

We randomly split all groups between training (80%), validation (10%) and testing (10%). We use regression (linear or log-linear) to remove the trend in both the mean and variance of each feature, based on the training subset. We then apply partial SVD to reduce the features to twenty dimensions, which is adequate for selecting the top ten features and accounts for 71% of the sum of the singular values. We select the best parameters for stochastic gradient descent with a parameter sweep.

Table 9: Most significant shared features for group event dynamics. *Statistically significant $P < .002$, likelihood ratio.

Covariate	Post Young		Post Middle		Post Old		Disc	
	RR	CRR	RR	CRR	RR	CRR	RR	CRR
CF J U Yr	-2.5%	-2.5%*	-2.5%	-2.5%*	-2.5%	-2.5%*	0.7%	0.7%
P O	-2.4	-2.2	-2.4	-2.2	0.8	0.8	-5.2	-4.6
J U Qt	-1.9	-1.4 *	-1.9	-1.4 *	0.7	0.6 *	-4.5	-3.3
P U	-1.7	1.2	-1.7	1.2	0.5	-0.5	-3.6	3.2
CF P U Qt	-1.9	0.8 *	-1.9	0.8 *	0.6	-0.1 *	-4.7	1.9
CF J Y Yr	1.4	0.8 *	1.4	0.8 *	-0.5	-0.3 *	3.5	1.9
CF J O Mo	1.0	-0.4	0.9	-0.4 *	-0.5	-0.1 *	2.4	-0.6
CF J O	0.7	0.5	0.7	0.4	-0.4	0.0	2.3	2.1

Covariate	Join Young		Join Middle		Join Old	
	RR	CRR	RR	CRR	RR	CRR
CF J U Yr	-4.3%	-4.3%*	-4.3%	-4.3%*	0.8%	0.8%
P O	-4.3	-3.9	-4.3	-3.9	1.6	1.5
J U Qt	-3.0	-2.2 *	-3.0	-2.2 *	1.3	1.2 *
P U	-3.1	1.9	-3.1	1.9	1.1	-0.7
CF P U Qt	-3.3	1.5 *	-3.3	1.4 *	1.1	0.6
CF J Y Yr	2.5	1.3 *	2.5	1.3 *	-0.9	-0.6
CF J O Mo	1.5	-1.0	1.5	-1.0	-1.5	-1.0 *
CF J O	1.0	0.5	1.0	0.6	-1.6	-0.2 *

5.4.1 Results

We first consider the effect of features that are selected uniformly for all event types. Table 9 shows for each of the selected features and for each subtype of event the full relative risk of the feature by itself and the conditional relative risk given the previously selected features. Recall that the relative risks reported here are the relative change in event rates given a one-standard-deviation change in a feature value.

CF J U Yr, J U Qt. At these scales, events by users of unknown age are essentially proxy for all events. If many users have joined the group during the past quarter or many users have joined similar groups in the past year, it is likely that only old users will be active in this group. Activity by other users is suppressed by about 6.0% while activity by old members will increase by about 1.7%.

P O: Posts by old users similarly predict increased activity by old users by about 1.1% while decreasing activity by other users by 3.4%.

P U, CF P U Qt. Total posts and posts in similar groups during the past quarter, when considered in isolation, also suppress activity by most users. Interestingly, together with the preceding features they predict greater activity for young and middle users by about 3.2%. Total posts make very little difference for old members.

CF J Y Yr, CF J O Mo, CF J O. Young members joining similar groups over the past year predicts greater activity by young and middle-age users (1.5%), while again slightly suppressing interest by old members (0.5%). Because the event counts are transformed by logarithm, the opposition between CF J O Mo and CF J O is equivalent to the ratio feature $CF J O Mo / CF J O$. An unusual increase in the number of old members joining the group in the previous month is correlated with reduced activity by other members (1.5%). These factors are not predictive of long-time member activity.

The most significant features (Table 10) for each type of event individually gives additional insight into event dynamics. J U is in opposition to numerous features involving join events in similar groups or in recent time, so an unusually high number of total members is predictive of more activity by 9.7%. More previous posts in the group, especially in comparison against posts in similar groups, is also predictive of more activity by 2.1%. New discussions started in similar groups during the prior week is predictive of more posts in this group by 1.8%. Discussion dynamics are not differentiated by user age, but not surprisingly are quite similar to the dynamics for posts by young users. The main additions are a strong correlation with the number of posts (3.2%) and the number of joins by young users in similar groups during the past year (2.0%).

Table 10: Most significant features for group event dynamics by event type. *Statistically significant $P < .002$, likelihood ratio.

Post Young			Post Middle			Post Old		
Covariate	RR	CRR	Covariate	RR	CRR	Covariate	RR	CRR
CF J U Yr	-2.5%	-2.5%*	CF J U Yr	-2.5%	-2.5%*	P O	0.8%	0.8%
P O	-2.4	-2.2	P O	-2.4	-2.2	J U Qt	0.7	0.6
J U Qt	-1.9	-1.4 *	J U Qt	-1.9	-1.4 *	P O Yr	0.7	0.4
J U	1.3	1.2	J U	1.3	1.2	CF J U Yr	0.7	0.4
P Y	1.0	0.7 *	P Y	1.0	0.7 *	CF P U Qt	0.6	-0.3
CF P U Qt	-1.9	-0.9 *	CF P U Qt	-1.9	-1.0 *	CF J Y	-0.4	-0.2
CF D Wk	1.9	1.8	CF D Wk	1.9	1.8	CF P U Wk	-0.4	-0.4
CF J U	-2.1	-1.9	CF J U	-2.1	-1.8	P O Wk	0.2	-0.2

Join Young			Join Middle			Join Old		
Covariate	RR	CRR	Covariate	RR	CRR	Covariate	RR	CRR
P O	-4.3%	-4.3%*	P O	-4.3%	-4.3%*	P O Yr	2.0%	2.0%
CF J U Yr	-4.3	-3.6	CF J U Yr	-4.3	-3.6	CF J O Yr	-1.8	-1.3
J U	2.3	2.2 *	J U	2.3	2.2 *	J U Qt	1.3	1.2 *
CF J O Yr	1.6	-1.4	CF J O Yr	1.6	-1.4	CF P O Mo	-1.1	-0.8
CF J U	-3.5	-3.0 *	CF J U	-3.5	-3.0 *	P O	1.6	1.0
P Y	1.8	1.2 *	CF P U Wk	2.4	1.2 *	CF J O Qt	-1.8	1.8
P O Wk	1.6	1.4	CF J U Wk	1.4	1.3	P U	1.1	0.6 *
J U Mo	-1.4	-1.2	CF P U Mo	-1.9	-1.0	J U Mo	1.4	0.6 *

For long-term users, group features have much less effect on behavior rates, which could be because of less predictability or less variability. Increased activity is correlated with the number of posts by long term members and the number of joins in this and similar groups over the previous year. A spike in joining by old members in the previous quarter is correlated with additional joining (3.1%), but a spike in posts in similar groups by old members in the previous month is correlated with a decrease in joining (1.4%). Posts by these users are decreased when there have been many recent posts by these users in this or similar groups (0.9%).

We trained our model with 180 parameters using 86,101 events, giving nearly 500 events per parameter, well above the threshold of ten required to avoid underfitting

and overfitting. We tested the significance of CRR for each feature as we selected it. Each model was compared to the best-performing previous model using the likelihood ratio test. The values that are significant ($P \leq .002$) are marked in Tables 9 and 10. Coincidentally, unmarked parameters have $P > .1$. It often happened that we would select several features that individually made the performance worse but as a set significantly improved performance. Our algorithm selects the parameters for all of the features in concert, so it is reasonable to suppose that the selected features are important to the accuracy of the model even though they do not individually test as significant, but there is no evidence to support this possibility. Our results show that greedily selecting the feature that yields the highest increase in CRR is not the same as greedily selecting the feature that yields the most significant improvement, but it is not clear which strategy performs better in the end.

5.5 Discussion

Our results reveal many interesting insights, although our methods can not identify motivations like a user study could. As expected, we found a marked difference between the dynamics of young and old users. Across all significant features, the polarity of the impact was opposite between young and old users, suggesting that homophily is likely at work in these communities. The groups that are attractive to old users are less attractive to young members and *vice versa*. In particular, groups with many old members joining and posting were likely to attract new old-time members but not many younger members. It may be that the long-term members have different needs, but even so this reduces the amount they invest in new members.

In contrast, we also expected to see different dynamics governing young and middle age behavior because of the tremendous difference in event rates (Section 5.2). Instead, the dynamics are identical, implying that the only difference between a young and a middle user is waning engagement.

Young members were much more active in groups that were visible, that is groups that had many members or recent activity. This results in the popular groups becoming increasingly popular: we identified a correlation between young user activity and activity in similar groups which is most likely because popular groups become more similar as the same users join them. This trend may improve the social ties between new members, as they will repeatedly encounter the same people in different groups, strengthening the relationships. For example, contacts outside of a group have a significant impact on social roles within the group [92]. On the other hand, it means that there is little diversity in the membership of the most popular groups. We confirmed this by examining the eigenvalues of the membership and post matrices, which showed that a few coherent sets of users have been increasingly dominating the discussions. Moreover, peaks in discussions across groups frequented by young members are highly correlated, suggesting that there is not much to distinguish the groups. A lack of diversity is troubling, because Goel et al. showed that a strong set of less popular products (here groups) is crucial to satisfy an increasing population of users [73]. Another issue is that this situation limits the ability of less active groups to attract young members. In fact, the groups that are popular for young members are remarkably stable over time, as shown by the correlation between current activity by young members and the join rate of young members a year prior, who are now old members.

On the other hand, behavior by old users appears to be in part shaped by the desire to invest in the community. We saw that old members are much more likely to join groups that have had little activity and to start new discussions in these groups. The analysis shows a common pattern of old users joining a lethargic group and investing heavily over the course of a month. Usually the number of old members contributing to a group is quite small: the total number of posts by old members actually decreases as the number of old members in the group increases. This makes

sense if old members are intent on investing in the community, because there is less benefit to investing in a group that already has leadership. In contrast to young members, they do not flock to similar groups: the existence of groups with similar members actually implies that this group will be less active.

Naturally, our analysis is limited to the dynamics of users who participate in events. Many users presumably persist in following the same dynamics with an ever-falling baseline intensity function. By the time the user becomes an old user, the user is no longer participating much in the community and so is not reflected in the analysis. As such, the analysis is dominated by the relatively rare active old users. Because two very different motivations seem to underlie the old user dynamics, we hypothesize three different types of old user. Naturally, real old users are likely to be a mixture of these idealized types. Most have very little activity in the community; some form cliques with other old users and do not interact with newer members; others take ownership of the community and seek to invest in making it a profitable place for a wide variety of users. Multiple types of old users is also consistent with the greater difficulty modeling the old user behavior, as evidenced by the poor significance test results. The first type dominates social dynamics for the first nine months of a user, at which time the other two types dominate. We believe that if users can be kept involved in the community until this critical time, it should be possible to increase the number of users who take an active role in developing the community.

The lack of population diversity among popular groups is also concerning. It appears to arise in large part from the way that groups are advertised according to popularity. Many groups are close to having the critical mass required to be successful, but will only slowly attract new members under the current policies. There is great potential for identifying these groups and doing targeted advertising to attract members in keeping with the distinctive character of a group. Additionally, recruiting someone who is at the threshold of becoming an old member to take ownership

of the group would simultaneously cement this user as an active contributor to the community and improve the group’s prospects. These observations suggest that personalized recommendation of groups could contribute to the robustness of a health community.

5.6 Conclusion

Health forums provide a valuable social resource to people who are facing dramatic changes in lifestyle because of a disease. However, there is significant variation in user participation in different groups, which reduces the effectiveness of the community as a whole in meeting user needs. In this chapter, we have used event history analysis to understand the factors that affect the activity level in TuDiabetes groups. In doing so, we have also contributed a new way to interpret event history models with strongly correlated dynamic features. We find that new community members tend to densely populate the groups that are most visible, raising concerns about the amount of diversity between these popular groups. We also identified a critical time period in which community administrators can cultivate users who can provide active leadership in the community. These insights are valuable for improving the vitality of groups in online health communities and other communities with a strong incentive for user participation.

CHAPTER VI

SOCIAL RECOMMENDATION

6.1 Introduction

On-line health communities such as TuDiabetes [62] play an important role in helping patients and their caregivers cope with disease treatment and management outside the context of hospitals and clinics. The members of health forums are diverse: some consumers have been diagnosed with a life-changing disease like diabetes, others are concerned about mysterious symptoms and still others just want to lose some weight and get into better shape. They come from cultures with different perspectives on health that strongly influence their receptivity to health information. Some seldom think about health, while others are conversant in the latest research related to their disease. With all of this variety come significant cognitive gaps that impair effective communication of health information and the participation of new members in the health communities. A consumer who joins an inappropriate group will not make strong social contacts, may reject the discussion as culturally irrelevant and generally will not benefit to the same extent that she would in a more appropriate group. On the other hand, communities provide a wide variety of member-created groups: some are active with good quality interaction, but many have few members and little interaction. In between are groups that have the potential to go either way, depending on whether they can attract the interest of the right kinds of members and can facilitate valuable interaction among their members. If groups could have help attracting a critical mass of appropriate members, they would be much more likely to develop sustainable group dynamics. Similarly, if the members had assistance finding relevant information to discuss and filling in coverage gaps, the increased diversity

would make the discussion more useful, would more fully engage the members and would reduce the chance that biased information would result in the group taking an extreme view.

Chronic disease management requires patients to learn about the disease and make substantial lifestyle changes to reduce the risk of complications. In this context, community involvement can be invaluable to promote patient understanding and adherence [152, 103]. Taking diabetes as an example, support groups supplement classes offered by clinicians by bringing patients and caregivers together to help each other learn about diabetes and encourage each other to work hard to manage the disease. In addition, communities offer social support from people who have experienced similar emotions caused by coping with the disease. In many cases, people are turning to online support groups because they can be accessed when they are needed instead of at scheduled times, offer perceived anonymity [103] and access to a group of people that is not geographically constrained.

There are many health support groups available, but choosing a group carefully can improve the patient's outcome [88]. Currently, online support groups are found either by word of mouth, static lists of good groups or dynamic lists that identify the most popular groups. Word of mouth is ideal but is only available to a small fraction of patients. Static lists are not easily kept up-to-date and provide no personalization. Unpersonalized dynamic lists contribute to an influx of new members into groups that are already popular, often to the extent that the popular groups become increasingly similar.

Personalized recommendation can play many valuable roles in online communities, helping users find interesting groups to join, discussions to read and people who might make good friends. It can also be used from the other direction to find resources that will benefit groups or discussions. Recommendation in communities alters the mixture of participants in groups and discussions, and so may impact the character

of the recommended resources. This is in stark contrast to recommendation of books or movies, which causes very little impact on the recommended item. On the one hand, this can benefit the community by increasing the diversity of active groups and improving access to a wealth of older discussions. It can also be a useful tool for community administrators as they seek to recruit active volunteers and cultivate group identities. On the other hand, careless recommendation can damage a group, for example by bringing in too many new members in a small period of time.

The unique characteristics of health communities and their participants influence recommendation. Participants often have a greater incentive to invest time in the community, owing to the toll of the disease on their life. As a result, more users are highly active so that the power-law distributions of activity counts are more extreme. In addition, users are more likely to follow up on recommendations, leaving a stronger impact on the recommended resources. We use symmetric recommendation tasks in the evaluation to better understand how these issues interact with our recommendation models. Privacy is another aspect in which health communities are distinct. A person’s health is often a sensitive topic that can influence a person’s self respect, attitudes of other people and may have implications for employment. In addition, users may share information in public groups that would ordinarily have legal protection. Although the information is technically public, recommendation algorithms subject it to high levels of analysis that may feel especially invasive to users in a health context. The techniques that we present in this paper could naturally incorporate additional information like the content of discussions or friendship networks, but we selected a subset of information that demonstrates the value of our methods while limiting the sensitivity of the information we collected.

State-of-the-art recommendation is based on matrix factorization [19, 147]. However, two major challenges arise when applying widely-used matrix factorization models. (1) Online health communities inherently have multiple types of entities (like patients, groups and discussions) and relations (membership in a group, participation in a discussion). Therefore, it helps to use many relations to improve the accuracy of recommendations. (2) The recommendation is based on social behaviors of users in online health communities that are power-law distributed, which is an important factor influencing recommendation accuracy. In this paper, we demonstrate how current state-of-the-art recommender system algorithms can be adapted to comparing entities in online health discussion communities by overcoming the above two challenges (Section 6.2). In particular, we address the problem of recommendation using multiple relations with wide-ranging implicit ratings that are derived from social behavior. For multiple typical social recommendation tasks, we evaluate a class of multi-task matrix factorization models that has the flexibility to share the same latent profiles completely or partially between tasks or to use independent profiles for each task. Moreover, in order to better model the social behavior data, we propose two new techniques, Gaussian transform encoding and power-law link function encoding, to transform between the distribution produced by latent factor models and power-law-distributed social behaviors.

In Section 6.3, we evaluate the proposed methods in order to identify the strategy that provides the best basis for making social recommendations, using data from the TuDiabetes community (<http://www.tudiabetes.org>). Section 6.4 addresses how well these approaches are applicable to real-world social recommendation. Each of the various encodings is best for some recommendation tasks. In general, our approach works well for recommending any kind of social resource to users, but is less effective for recommendation to groups, like linking discussions to groups. We will conclude in Section 6.5.

6.2 *Social Recommendation*

There are several ways in which recommendation can increase the value of online health communities to their participants. As communities grow, recommendation can unearth valuable resources that would otherwise be difficult to find, resources ranging from groups to people to discussions with long-term value. A healthy variety of groups increases the population of users that the community can serve (compare [73]) and groups serve as a topical index to discussions. However, in communities that organize groups by popularity, most groups will lack the visibility they need to attract new members. Group recommendation is one way in which a community can support a larger number of active groups, by advertising each group to just the users who are most likely to be interested. Recommendation can also play a role in protecting and fostering group identity. People are naturally the most valuable resource in a social environment, fueling groups and discussion. People recommendation can identify possible leaders for a group, experts to address questions in a discussion or interested people to help stimulate a discussion. Discussions, on the other hand, record the collective inspirations, struggles and knowledge of the community. Many discussions are of value long after they were started and long after they can be easily found. About 75% of the discussions in TuDiabetes, for example, take place in busy open forums where older discussions might be on the hundredth page of discussions. Recommendation can surface the old discussions that are still useful.

6.2.1 Encoding Social Events

For social health recommendation, we observe the number of posts a user invested in a discussion instead of direct ratings. These counts can be aggregated into implicit ratings for other types of relationships; for example, the affinity of a user-group pair is one more (for joining the group) than the number of posts by the user in the group. In practice, event counts follow a power-law distribution which is very different from the

approximately Gaussian distribution produced by PMF. Indeed, a naïve application of PMF to power-law data is very unstable, because many of the observations are in the very improbable region where the condition number is numerically infinity.

Binary Encoding

One simple approach is to encode the interactions with binary indicators. The interaction counts are replaced with the value 1, regardless how many interactions there were. Although this seems to discard most of the information, it retains the network structure and information like node degrees that are highly correlated with the interaction counts.

Gaussian Transform Encoding

In Gaussian transform encoding, the interaction counts are transformed so that they fit a standard Gaussian distribution. After this, the model can be trained using standard matrix factorization techniques. We observed that the probability of small interaction counts is orders of magnitude less than would be predicted by a simple power-law distribution, so we fit a truncated power law distribution to the probability of the interactions between entities u and g , $\mathbb{P}(A_{ug}|u, g) \approx \omega_u \omega_g b \min(A_{ug}, \tau)^m$ for $A_{ug} > 0$. m and b are the parameters for the power-law distribution and ω_u and ω_g are scaling factors for user u and group g respectively. τ is a threshold that controls truncation for small interaction counts. We can transform the random variable A_{ug} into a new variable R_{ug} that has Gaussian distribution by equating the complementary cumulative density functions. That is, R_{ug} is chosen so that the probability of the interval (R_{ug}, ∞) under a standard Gaussian distribution is the same as the probability of (A_{ug}, ∞) under the power-law distribution,

$$\frac{1}{2} \operatorname{erfc}(R_{ug}/\sqrt{2}) = \omega_u \omega_g b \sum_{i=A_{ug}+1}^{\infty} \exp(m \min(i, \tau)),$$

where erfc is the complementary error function. For $A_{ug} \geq \tau$,

$$R_{ug} = \sqrt{2} \text{erfc}^{-1} \left(-2 \frac{\omega_u \omega_g b}{m+1} A_{ug}^{m+1} \right).$$

Power-Law Link Function Encoding

Power-law link function encoding is essentially the reverse of Gaussian transform encoding: the model predictions are transformed into the power law space.

$$x_t(d) = \frac{1}{2} + \left(- \left(\frac{1}{2} + \frac{1}{2} \text{erfc} \left(\frac{d}{\sqrt{2}} + \epsilon \right) \right)^{(m+1) \exp(-b)} \right)^{\frac{1}{m+1}},$$

where ϵ is a small value that improves numerical stability. Numerical stability is an issue because without ϵ the condition number is large for some reasonable values of d . ϵ should be chosen so that $x_t(0)$ is about twice the maximal value of A_{ug} . In this space, the squared error loss function (Eqn. (3)) is not appropriate because substantial differences in scale. Instead, we use squared relative error,

$$\frac{1}{2} \sum_{ug} \left(1 - \frac{x_t(U_u T_{ug} G_g^T)}{A_{ug}} \right)^2.$$

6.3 Experiments

We collected interactions in the TuDiabetes online diabetes community through March 7, 2012 (IRB protocol H11049). At that time, TuDiabetes had about 23,000 members (U). Since its inception, the community has offered 26 forums (F) containing threaded discussion (D) in which users discuss a variety of issues related to life with diabetes. It also has 569 groups (G) that users can join in order to be a part of a smaller community with more focused discussion. We recorded the user-group membership for all groups with at least 25 members. We also recorded the number of times each user participated in each discussion in the forums or groups. We aggregated these counts to give us the amount of activity for each pair of entities: number of posts per user-discussion (UD); number of posts per user-forum (UF); number of posts plus

Table 11: Group event data by event type.

Reln	Training	Validation		Testing	
	Pos	Pos	Neg	Pos	Neg
UD	234,162	26,228	26,208	26,728	26,237
UF	28,922	5,801	1,151	5,850	1,141
UG	29,905	6,191	3,523	6,240	3,518
DF	22,071	2,807	1,350	2,819	1,358
DG	6,466	1,008	2,409	1,010	2,411
Total	321,526	42,035	34,641	42,647	34,665

1 for being a member per user-group (UG); number of posts per discussion for the forum (DF) or group (DG) in which the discussion occurred.

We experimented with four different encodings for the interaction data. The natural encoding (N) was the number of interactions involving a pair of entities, rescaled so that the values for each kind of interaction were comparable. The power-law link function encoding (L) used the natural encoding, but incorporated a link function that transformed the values produced by the model to fall on the same power-law distribution as the observed interaction counts. The Gaussian transform encoding (T) transformed the interaction counts to the tail of a standard Gaussian distribution. The binary encoding (B) used 1 to indicate observed interactions regardless of the count. The encodings are described in more detail in Section 6.2.1.

We split the interaction data into training (80%), validation (10%) and testing (10%) sets. The sampling process was stratified by degree distributions to ensure good coverage for testing. If a user-group interaction was included in the validation or testing set, then all user-discussion interactions for this user and group were promoted to same set. To measure overfitting with RMSE, for each observed relationship in the validation and test sets, we added two negative observations consisting of one of the entities from the observed relationship and a randomly chosen other entity of the correct type.¹ The amount of data is detailed in Table 11.

¹Selected examples that matched a real observation were rejected.

Tasks.

Group recommendation ($G \rightarrow U$). Personalized recommendation helps users find relevant groups to join, considering only the user’s point of view.

Symmetric group recommendation ($U \leftrightarrow G$). Symmetric recommendation simultaneously recommends groups to users and users to groups. This more difficult task attempts to balance the needs of the user for an interesting group and the needs of the group for sustainable growth.

Discussion recommendation ($D \rightarrow U$). Often, older discussions are still of interest to new users, but it can be nearly impossible to locate them. Thus, we have the task of recommending discussions to users.

People recommendation ($U \rightarrow D$). Some discussions would benefit from the input of an experienced user with relevant expertise. The purpose of this task is to identify people who are relevant to a discussion, offering lively interaction or expertise.

Discussion linkage ($D \leftrightarrow G$). One reason a discussion may be hard to locate is that most of the discussions take place in the forums, where there is very little structure. If discussions were cross-linked to relevant groups, the groups would provide additional indexing to help locate the discussion. It is important to check both that the group is relevant to the discussion and that the discussion is relevant to the group.

Evaluation Metrics.

Root Mean Squared Error (RMSE). RMSE measures how faithful the model is to recover actual activity levels. It is closely related to standard deviation and provides an intuitive measure of the uncertainty in the model. We report RMSE relative to the transformed activity levels, because it is not appropriate for raw activity levels.

$$RMSE = \sqrt{\frac{\sum_{ij} (R_{ij} - R_{ij}^*)^2}{n}},$$

where the sums are taken over n test data points and R^* is the model prediction.

Mean Rank of k (MRk). Model predictions are used to order prospective recommendations. Here, we want to measure how a test datum point will be ordered relative to k random unobserved entities. MRk reports the average location of the test entity in the list, from 1 (first in the list) to $k + 1$ (last in the list) [97]. A good model should rank the test entity towards the front of the list, resulting in a low score. For a random predictor, $MRk = k/2 + 1$. To compute the MRk for a specific recommendation task, for each test instance we count the number of entities that have a lesser (L), equal (S) or greater (H , $T = L + S + H$) prediction according to the model. The expected value and variance of the rank r can be computed by averaging the means and variances of the hypergeometric distributions with between H and $H + S$ positive entities.

$$\mathbb{E}(r) = 1 + \frac{k(H + S/2)}{T}$$

$$MRk = \frac{\sum_i \mathbb{E}(r_i)}{n}.$$

For symmetric recommendation tasks, we compute the rank statistics for each entity from a test pair. The rank for the whole pair is the maximum of either rank.

$$\mathbb{E}(r_{ij}) = \max(\mathbb{E}(r_i), \mathbb{E}(r_j)).$$

Implementation. We based our implementation on a Matlab implementation of probabilistic matrix factorization (PMF) and Bayesian PMF (BPMF) [148, 147]. We modified the code to support collective matrix factorization with partially shared latent profiles. We also changed the code to discount the learning rate on each epoch i by a factor of \sqrt{i} [18]. We tuned the parameters based on validation RMSE after 20 epochs.

6.3.1 Results

For this evaluation, we trained a large number of different models, covering each algorithm PMF, collective PMF (CPMF), PCPMF and their Bayesian analogs BPMF,

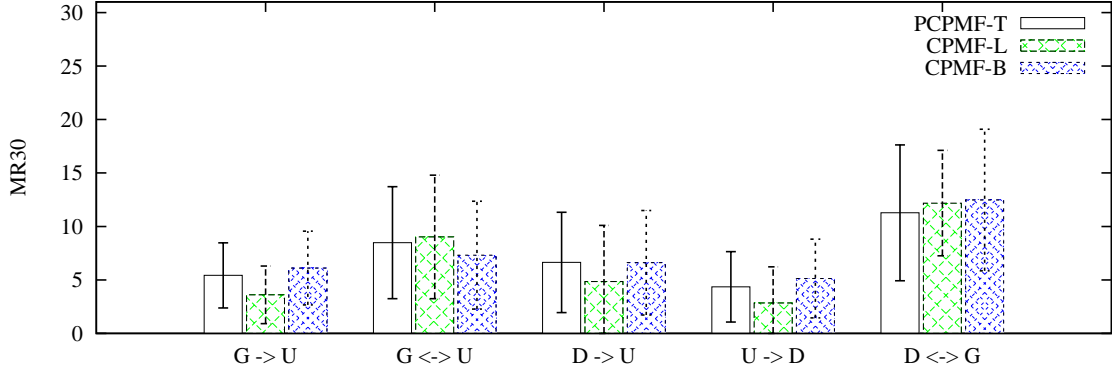


Figure 15: Comparison of performance with different encodings.

BCPMF and BPCPMF in each encoding N, T, L, B. We do not report further on the natural encoding N because all algorithms except PMF and BPMF were highly unstable. We also do not further report on the Bayesian models because they consistently performed worse than random on task-specific performance. Bayesian models do not perform well in terms of MR because they are reluctant to commit to the order of test instances.

Encoding

Figure 15 shows the performance in terms of MR30 for the best algorithms on each of the core tasks. Power-law link function encoding (L) is optimal for simple recommendation tasks.² For symmetric group recommendation, the binary encoding was a better choice. Discussion linkage has the characteristic that each discussion interacts with exactly one group so that any discussions used for training have no discussion-group interactions in the training data. Thus, incorporating information on other kinds of interactions involving the discussion and group is the only way this task can be accomplished. On this most difficult task, the transform encoding (T) performed best.

²All observations are significant at $P < .001$ unless specified otherwise.

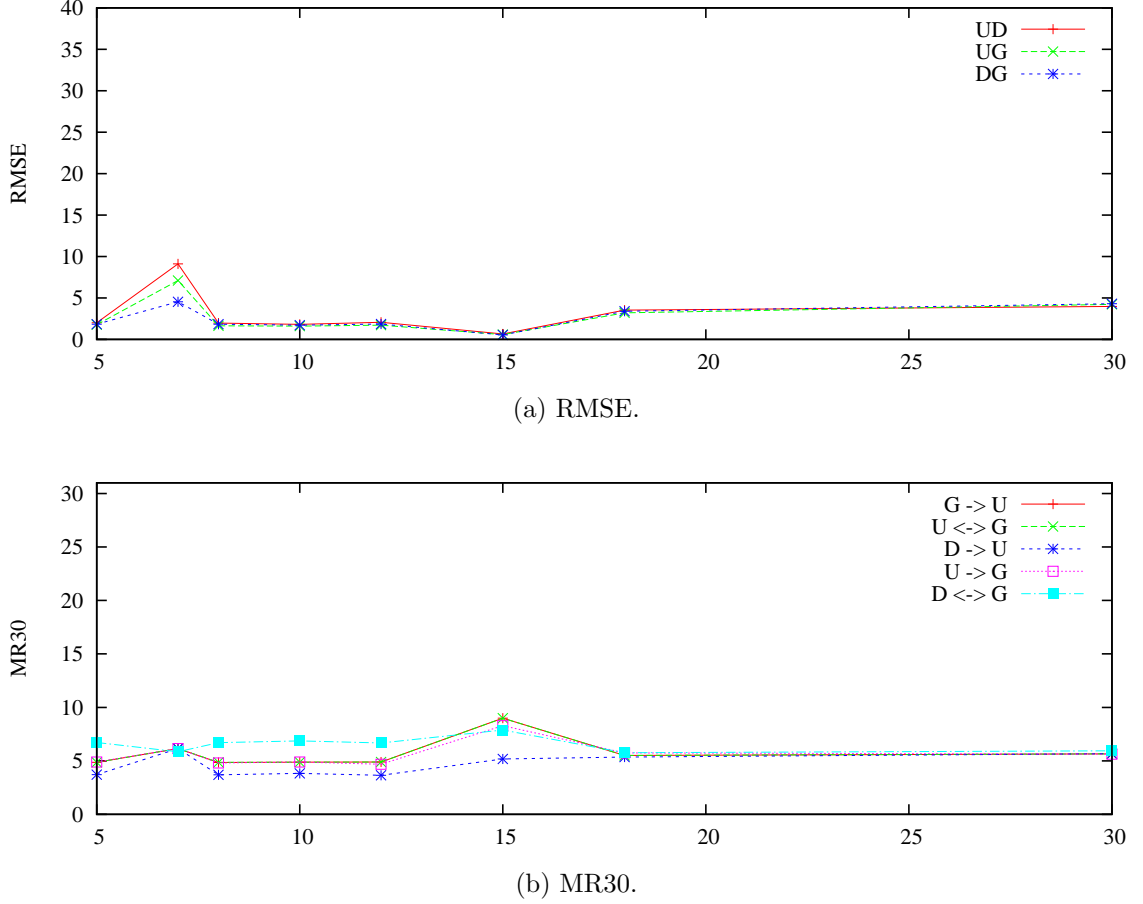


Figure 16: Comparison of performance of CPMF-T as a function of the number of dimensions.

Number of Dimensions

Figure 16 shows the effect of the number of dimensions on the performance. The RMSE appears best at 15 dimensions, but the task-specific performance is better with around 10 dimensions, with the exception of discussion linkage, which benefits from 18 dimensions ($P < .09$).

Parameter Sharing

In order to compare the effect of sharing dimensions across tasks, we evaluated the performance of PCPMF-T as we varied the number of shared dimensions (Figure 17). PMF-T (no sharing) has the best performance by RMSE but the poorest MR30

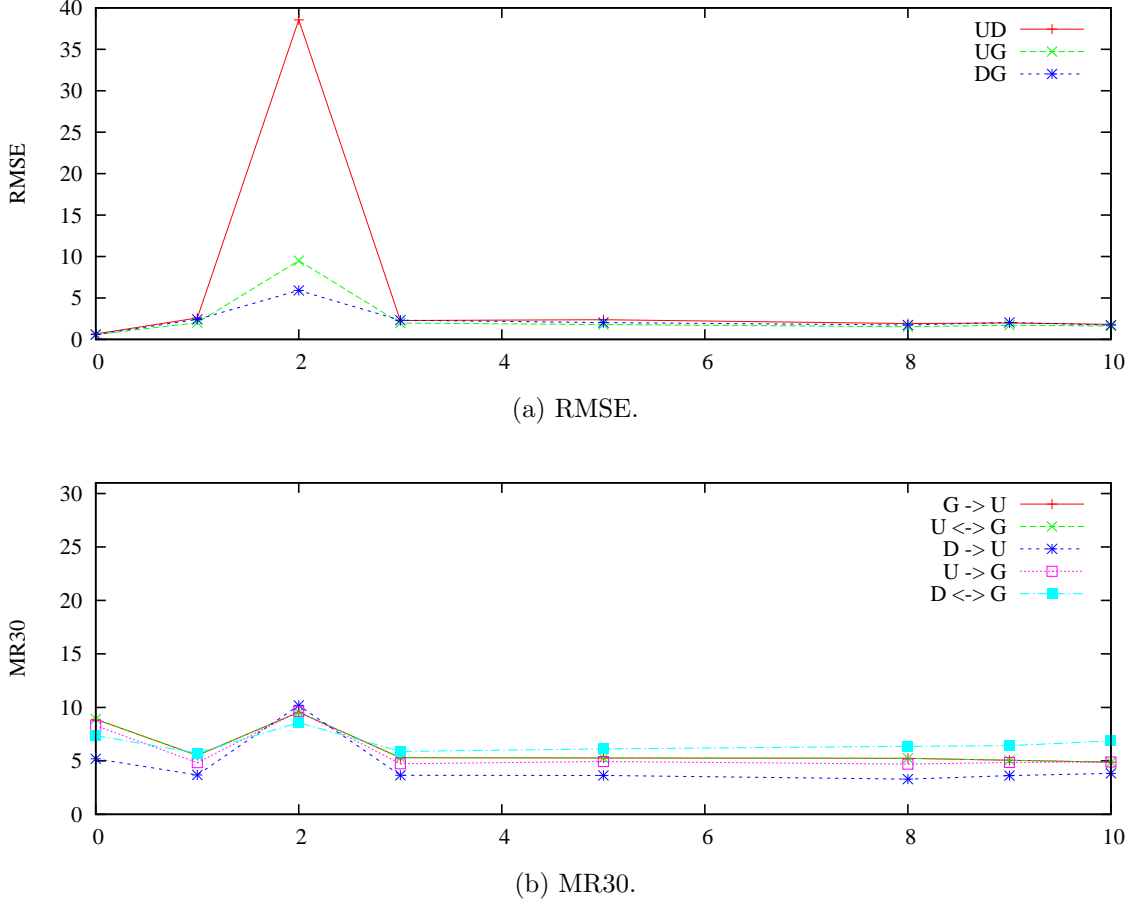


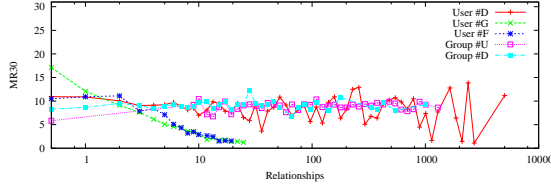
Figure 17: Comparison of performance of PCPMF-T with 10 dimensions as a function of the number of shared dimensions.

performance on the tasks. Depending on the task, the optimal sharing is 80–100% of the dimensions.

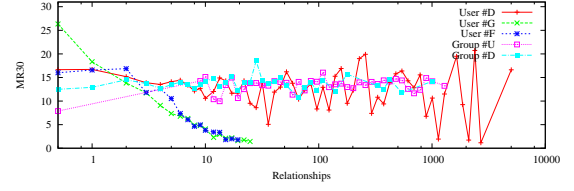
Effect of Amount of Training Data

In a social network, there is a substantial disparity in the amount of training data for different users or items. We generally expect the performance to be substantially better for users with substantially more data. We selected the CPMF and PCPMF models with the best MR30 for each task and show the results in Figure 18.

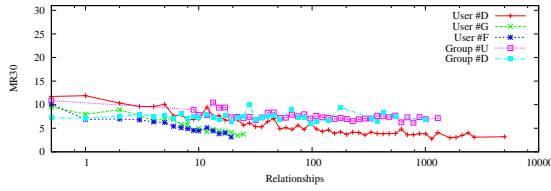
Group recommendation: The performance is remarkably stable for groups regardless of the number of users or discussions. The apparent improvement in performance for groups with small numbers of users is an illusion, coming from a few groups that



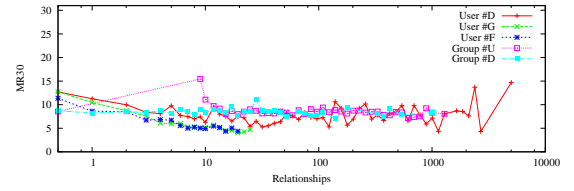
(a) G->U CPMF-L



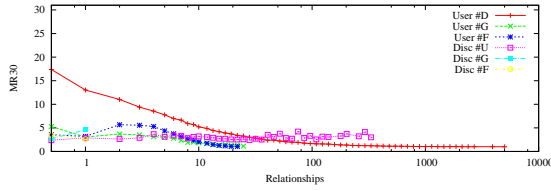
(b) G->U PCPMF-L



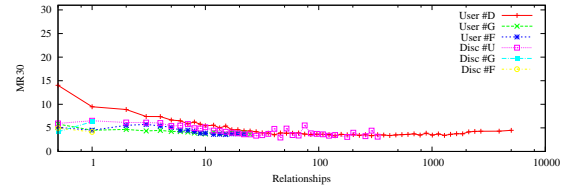
(c) G<->U CPMF-B



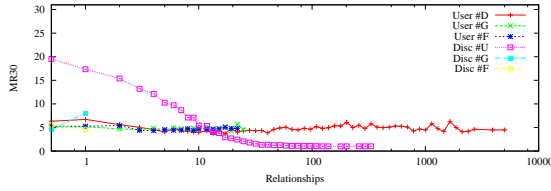
(d) G<->U PCPMF-T



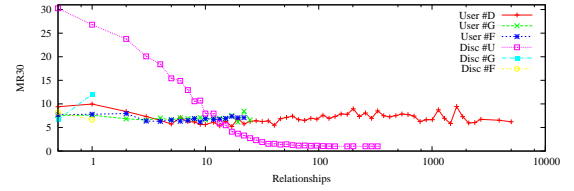
(e) D->U CPMF-L



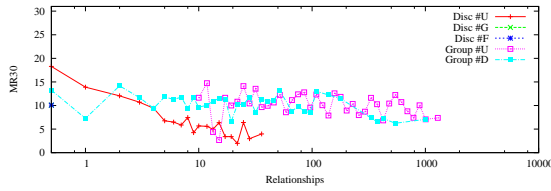
(f) D->U PCPMF-T



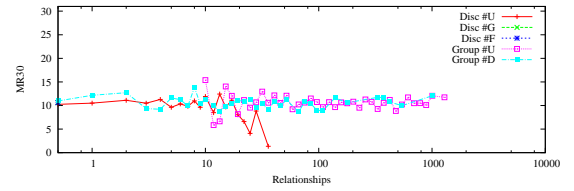
(g) U->D CPMF-L



(h) U->D PCPMF-L



(i) D<->G CPMF-T



(j) D<->G PCPMF-T 15 dimensions

Figure 18: Impact of available training data on MR30 for best CPMF and PCPMF models.

have more than 25 members but appear to have fewer due to censoring. Likewise the number of discussions in which the user participates makes little difference. However, the performance improves quickly for users who participate in more groups or more forums. Note that users who are generally more active will be participating in more groups and more forums. As a result, our analysis cannot distinguish whether either in isolation would have the same impact on performance.

Symmetric group recommendation: In contrast to group recommendation, the symmetric version slowly improves in performance with additional information of any kind.

Discussion recommendation: Discussion recommendation performance does not depend much on the number of participants. However, the performance is much higher for discussions that happen in forums (Disc #G= 0) than in groups (Disc #G= 1). The dependence on the number of discussions in which the user has participated is very strong. Although the PCPMF algorithm is generally weaker than CPMF, it has significantly better performance for users who have participated in fewer than 10 discussions. The dependence on the number of group memberships for the user has an unexpected shape—performance actually deteriorates when the user joins her second group, finally returning to the initial level of performance for people who are members of about 7 groups.

People recommendation: This task measures our ability to identify users who have posted on a given discussion. The performance does not depend much on the characteristics of the users, but only on the number of users in the discussion.

Discussion linkage: Performance on this task depends only on the number of users who participate in the discussion. As for discussion recommendation, the PCPMF algorithm offers an improvement for discussions with fewer than 5 participants.

6.4 *Analysis*

Our experiments yield insight into the appropriate technology choices for social recommendation. Some of the algorithms performed more poorly in this domain than would be expected. The Bayesian models did not work well for recommendation and are also very slow. As a result, they are best avoided for social recommendation. Likewise, unshared dimensions cause performance problems: in this application, there is a great deal of useful information in the indirectly-related relationships. The shared dimensions mediate the sharing of this information between tasks. On the other hand, models with a few dimensions left unshared improve performance in cases when an entity has participated in few discussions. For best overall performance, both kinds of models should be combined in a mixture of experts that transitions smoothly between the models depending on the available training data.

We saw repeatedly in our results that there is little correlation between RMSE and MR. This suggests that RMSE is a poor surrogate for performance in social recommendation applications. Our implementation, like nearly all MF-based algorithms, used RMSE as the loss function for training because it is tractable mathematically and ensures that the predictions of the model are near the observed values. This all suggests that training directly with a variant of MR as a loss function could yield much better performance than we report.

We found that the characteristics of the task influence the appropriate encoding. This can be addressed by using different encodings for each type of relationship [189], by using a mixture of experts approach or by learning a separate model for each task. In the latter case, the models are all trained on the full variety of data so that they can leverage other types of information, but each model specializes in a specific task. For group recommendation, a binary encoding is ideal. This may imply that social connections are more significant than the levels of interaction when a user is deciding whether to join a group, or it may simply be an artifact of the binary MR_k evaluation.

It is less clear what distinguishes the Gaussian transform and power-law link function encodings. The transform that is involved is the same, just taking a different direction. The distinctions arise from numerical stability, efficiency and subtle differences in the loss function. In most application, the Gaussian transform encoding should be preferred because it is easily implemented and requires less tuning. However, the power-law link function encoding yielded significantly better performance on more tasks.

The results also show that it is fairly easy to recommend resources to users. The tasks for these recommendations have MR30 of around 5, meaning that users are likely to find useful recommendations when presented with a list of manageable size (5–7 recommendations). However, the results for people recommendation make it clear that this approach is not at all selective of the quality of the recommended users, so that this approach would not work well for finding experts. What the approach can do is predict users who are likely to post to a discussion, which can be combined with other methods that directly tackle matching user expertise to discussions. The combined system would be able to identify users who have relevant expertise and motivation to participate. We also found that discussion linkage is very difficult, mainly owing to the fact that each discussion is only in one forum or group, so that every problem is a cold-start problem. Collective factorization offers some ability to leverage common users, but content-based approaches are expected to provide much better performance for this task.

6.5 Conclusion

Recommendation plays an important role in online health communities. We have analyzed several algorithms for recommendation in online health communities. Collective probabilistic matrix factorization is very effective when any of the encodings of social interactions between entities that we examined. This method achieves an

MR30 of near 5 for recommendations to users, which is adequate for this class of recommendations. Other kinds of recommendation are more challenging, especially recommendations made to groups.

CHAPTER VII

FUTURE WORK

Language Gap. We developed algorithms for topic models that bridge the language gap by incorporating differences coming from different levels of technicality. There are a number of significant directions to extend this research.

Information Retrieval. We have been working with Google to attempt to make this technology work in a Web-scale search context. We have encountered two main problems: real queries are too short to analyze topically; the topics are too coarse, so that words within the same cluster may be essentially unrelated. These problems affect topic models in general, and solving them would have wide-ranging benefits. We will seek to learn the fine-grained relationships of words within topics, so that topic models can be used without decimating precision.

Social Dialects. Public health surveillance tracks the behavior of populations that are at risk for certain diseases in order to focus education efforts and prepare for epidemics. Social media analysis is important in this vein because it allows public health professionals to model the social connections between different geographical regions. However, many populations are located in the same geographical area but are cut off by cultural differences that are often signaled by differences in dialect. We propose to extend diaTM so public health professionals can analyze social movement trends at the level of local populations.

Health Literacy Profiles. By leveraging models that reflect topics at multiple levels of technicality, we hope to be able to estimate how literate an individual consumer is with respect to different health topics. This information can then improve our ability to recommend understandable documents and valuable mentoring relationships. It

can further be extended to measure the impact of health literacy campaigns on the audience’s understanding.

Quality of Medical Content in Social Media. We found that proana content promoting dangerous eating habits is abundant in social media and is not controlled by existing technologies. Much of the difficulty comes from the strong similarity in language used by the proana and pro-recovery communities. We propose to extend the τ LDA model, which works well for separating documents of different technicality, as the basis of a classifier for filtering proana content. This is especially important when combined with our work in recommendation, because we do not want to recommend dangerous medical content from social media.

Event History Analysis. We developed a technique using SVD to untangle the influences of many highly-correlated features in an event history model. Our interpretation was based on greedily selecting the most influential features. However, it would be much more informative to select groups of correlated features and use a visualization to represent the significant sets of features in a meaningful way.

Online Health Discussion Groups. We have identified some important features that appear to influence the vitality of discussion groups, including the composition of members, distinctiveness from other groups and the effort required to find the group. We plan a user study to validate and refine our discoveries. We also intend to develop methods to help improve the vitality of discussion groups so that they can more effectively server their members. This integrates with our work in social recommendation in that the members we recommend will impact the character of the group. We also propose to develop systems that community administrators can use to identify possible leadership for a group and a dashboard that shows the strengths and weaknesses of a group, which can guide attempts to curate the group.

Social Recommendation. We are pursuing three different directions for social recommendation.

Group Recommendation Prototype. We have identified algorithms that are suitable for group recommendation. We are in the process of constructing a website that allows us to offer recommendations to users of TuDiabetes and collect feedback on its usefulness. This prototype will provide us with a valuable platform for conducting user studies of various kinds with the TuDiabetes members.

Recommendation and Diversity. We have hypothesized that personalized recommendation will result in more groups attracting a critical mass of users and more diversity between groups. We will test this hypothesis using simulation and user studies, comparing personalized recommendation to popular group recommendation and random group recommendation. We will also investigate whether it is beneficial to explicitly model diversity in the recommendation system.

Event History Matrix Factorization: The technique that we develop for social recommendation in this thesis is a direct adaptation of conventional matrix factorization. We hypothesize that combining event history analysis with matrix factorization will result in a much more natural model for social recommendation that also results in improved recommendations and improved user models. We will develop this approach and test it with data from TuDiabetes.

CHAPTER VIII

CONCLUSION

In this thesis, we have developed the idea of helping consumers access valuable medical content. We have made contributions in four distinct areas.

Language Gap. We wanted to facilitate finding relevant technical information for a consumer who was not familiar with the relevant medical terminology. We developed two variants of topic modeling that allow the direct modeling of variations in how words are selected. Using these models, we were able to achieve about a 240% improvement over LDA on the task of finding technical medical content that is on the same topic as a consumer question. The models are also able to characterize the technicality of a document at the granularity of topics, opening fascinating possibilities for assessing consumer health literacy and learning.

Quality of Online Content. We analyzed a YouTube videos promoting a dangerous eating disorder and found that these videos were twice as engaging for their viewers as the informative videos. We also found that existing community-based safeguards were not responding quickly enough to prevent minors from watching the videos in large numbers. We then studied the interactions of two communities in conflict, one promoting the eating disorder and the other promoting recovery. We found evidence that the pro-recovery community was attempting to engage and persuade the pro-anorexia community, but their efforts were counterproductive. These findings have important implications for public health professionals who want to reach marginalized groups of consumers and for any system that facilitates finding reliable social health content.

Health Discussion Groups. We studied the factors that make some health discussion groups more vitally active than others. To enable the analysis, we developed a significant new algorithm to perform event history analysis reliably with correlated features. Our analysis was able to expose differences in motivation and behavior for members of the diabetes forum. We discovered that new users are not very selective in the groups that they join, which is causing long-term sustainability problems for communities that advertise groups uniformly to all members. We also found an opportunity for community administrators to encourage long-term members to take a leadership role in some less popular groups. Incidentally, these findings have been of great value to the community that we studied, because the administrators have been able to make changes in how they advertise groups and how they reach out to long-term members to promote leadership.

Social Recommendation. We sought to identify algorithms that are appropriate for recommendation in a social context. We introduced two new variations on matrix factorization that greatly improve its stability for problems with power-law-distributed ratings, which are typical of social recommendation. We also developed techniques for performing symmetric recommendation from latent profile models, so that the recommendation will account for the impact on the recommended resource as well as the user. Finally, we identified specific combinations of matrix factorization that are most effective at several key recommendation tasks. The performance on recommendations to users was generally quite good, with a mean rank of about 5 out of 30. On the other hand, we found that recommending connections between discussions and groups was very difficult and we found not satisfactory algorithm. These results provide essential groundwork to anyone implementing social recommendation systems.

Publications

- CRAIN, S. P., HUANG, J., and ZHA, H., “A scalable assistant librarian: hierarchical classification of books,” in *SIGIR*, 2008.
- CRAIN, S. P. and JIAO, Y., “Its time you drove: Deep retrieval with ontological visualization and exploration,” tech. rep., Oak Ridge National Laboratory, 2009.
- CRAIN, S. P., YANG, S.-H., JIAO, Y., and ZHA, H., “Dialect topic modeling for improved consumer medical search,” in *AMIA Ann. Sym.*, (Washington, D.C.), American Medical Informatics Association, 2010.
- CRAIN, S. P., YANG, S.-H., and ZHA, H., “Understanding group dynamics in health forums,” in *ASONAM*, 2012.
- CRAIN, S. P., YANG, S. H., ZHOU, K., and ZHA, H., *Mining text data*, ch. Dimensionality reduction and topic modeling: from latent semantic indexing to latent Dirichlet allocation and beyond. Kluwer, 2012.
- CRAIN, S. P. and ZHA, H., “Consumer medical information retrieval relevance judgments,” 2010.
- CRAIN, S. P., ZHOU, K., and ZHA, H., “Recommendation in online health communities,” in *HI-BI-BI*, 2012.
- SYED-ABDUL, S., FERNANDEZ-LUQUE, L., CRAIN, S. P., HSU, M.-H., LI, Y.-C., JIAN, W.-S., WANG, Y.-C., DORJSUREN, K., CHULUUNBAATAR, Z., and NGUYEN, A., “Health related misinformation promoted through social media: The YouTube case,” In Submission.
- YANG, S. H., CRAIN, S. P., and ZHA, H., “Bridging the language gap: topic-level adaptation for cross-domain knowledge transfer,” in *AI Stat*, 2011.
- YOM-TOV, E., FERNANDEZ-LUQUE, L., WEBER, I., and CRAIN, S. P., “Pro-anorexia and pro-recovery photo sharing: A tale of two warring tribes,” In submission.

REFERENCES

- [1] “Italy pact to stop skinny models.” <http://news.bbc.co.uk/2/hi/europe/6204865.stm>, Dec. 22 2006. Accessed: 2012-08-06. (Archived by WebCite® at <http://www.webcitation.org/69iemB3eT>).
- [2] “Bmi related disease.” <http://www.bmicalculator.org/bmi-related-disease/>, 2009. Accessed: 2012-08-03. (Archived by WebCite® at <http://www.webcitation.org/69dLDJf2E>).
- [3] “O negative blood type.” <http://onegativebloodtype.com/o-negative-blood-type.html>, 2010. Accessed: 2012-08-17. (Archived by WebCite® at <http://www.webcitation.org/6A02J19Nb>).
- [4] “Israel passes law banning use of underweight models.” <http://www.bbc.co.uk/news/world-middle-east-17450275>, Mar. 20 2012. Accessed: 2012-06-25. (Archived by WebCite® at <http://www.webcitation.org/68gylnuDp>).
- [5] AALEN, O. O., BORGAN, O., and GJESSING, H. K., *Survival and event history analysis: a process point of view*. Springer, 2008.
- [6] ACHE, K. A. and WALLACE, L. S., “Human papillomavirus vaccination coverage on YouTube,” *Am J Prev Med*, vol. 35, no. 4, pp. 389–392, 2008.
- [7] AGARWAL, D., CHEN, B., and ELANGO, P., “Fast online learning through offline initialization for time-sensitive recommendation,” in *KDD*, pp. 703–712, 2010.
- [8] AGICHTEN, E., CASTILLO, C., DONATO, D., GIONIS, A., and MISHNE, G., “Finding high-quality content in social media,” in *WSDM*, pp. 183–194, 2008.
- [9] AHMED, A. and XING, E. P., “Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective,” 2010.
- [10] AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E., and XING, E. P., “Mixed membership stochastic blockmodels,” *Journal of Machine Learning Research*, vol. 9, pp. 1981–2014, June 2008.
- [11] AKOGLU, L. and FALOUTSOS, C., “Valuepick: Towards a value-oriented dual-goal recommender system,” in *ICDMW*, pp. 1151–1158, 2010.
- [12] ARCELUS, J., MITCHELL, A. J., WALES, J., and NIELSEN, S., “Mortality rates in patients with anorexia nervosa and other eating disorders. a meta-analysis of 36 studies,” *Arch Gen Psychiatry*, vol. 68, no. 7, pp. 724–731, 2011.

- [13] ARNOLD, C. W., EL-SADEN, S. M., BUI, A. A. T., and TAIRA, R., “Clinical case-based retrieval using latent topic analysis,” in *AMIA Ann. Sym.*, 2010.
- [14] ASUNCION, A., WELLING, M., SMYTH, P., and TEH, Y. W., “On smoothing and inference for topic models,” in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI ’09, (Arlington, Virginia, United States), pp. 27–34, AUAI Press, 2009.
- [15] BACKSTROM, L., HUTTENLOCHER, D., KLEINBERG, J., and LAN, X., “Group formation in large social networks: membership, growth, and evolution,” in *KDD*, pp. 44–54, 2006.
- [16] BAHL, L., BAKER, J., JELINEK, E., and MERCER, R., “Perplexity—a measure of the difficulty of speech recognition tasks,” in *Program, 94th Meeting of the Acoustical Society of America*, vol. 62, p. S63, 1977.
- [17] BARDONE-CONE, A. M. and CASS, K. M., “What does viewing a pro-anorexia website do? an experimental examination of website exposure and moderating effects,” *Int J Eat Disord*, vol. 40, no. 6, pp. 537–548, 2007.
- [18] BARTLETT, P. L., HAZAN, E., and RAKHLIN, A., “Adaptive online gradient descent,” in *NIPS*, 2007.
- [19] BELL, R., KOREN, Y., and VOLINSKY, C., “The BellKor 2008 solution to the Netflix prize,” Tech. Rep. 12, 2008.
- [20] BENNETT, E., “Found in cache: social media resources for health care professionals.” <http://ebennett.org/hsn1/>, 2011. Accessed: 2012-07-30. (Archived by WebCite® at url <http://www.webcitation.org/69YNu4tLR>).
- [21] BICKMORE, T. W., PFEIFER, L. M., and PAASCHE-ORLOW, M. K., “Using computer agents to explain medical documents to patients with low health literacy,” *Patient Educ. Couns.*, vol. 75, no. 3, pp. 315–20, 2009.
- [22] BICKMORE, T. W., PFEIFER, L. M., and JACK, B. W., “Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents,” in *CHI ’09: Proceedings of the 27th international conference on Human factors in computing systems*, (New York, NY, USA), pp. 1265–1274, ACM, 2009.
- [23] BLEI, D. M., GRIFFITHS, T., JORDAN, M., and TENENBAUM, J., “Hierarchical topic models and the nested chinese restaurant process,” in *NIPS*, 2003.
- [24] BLEI, D. M. and LAFFERTY, J. D., “A correlated topic model of science,” *AAS*, vol. 1, no. 1, pp. 17–35, 2007.
- [25] BLEI, D. M. and MCAULIFFE, J. D., “Supervised topic models,” in *Advances in Neural Information Processing Systems*, vol. 21, pp. 121–128, 2008.

- [26] BLEI, D. M., NG, A. Y., and JORDAN, M. I., “Latent Dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
- [27] BORZEKOWSKI, D. L. G., SCHENK, S., WILSON, J. L., and PEEBLES, R., “e-Ana and e-Mia: A content analysis of pro-eating disorder web sites,” *American Journal of Public Health*, vol. 100, no. 8, pp. 1526–1534, 2010.
- [28] BOYD-GRABER, J. and BLEI, D. M., “Multilingual topic models for unaligned text,” in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI ’09, (Arlington, Virginia, United States), pp. 75–82, AUAI Press, 2009.
- [29] BOYD-GRABER, J. L. and BLEI, D., “Syntactic topic models,” in *Advances in Neural Information Processing Systems 21* (KOLLER, D., SCHUURMANS, D., BENGIO, Y., and BOTTOU, L., eds.), pp. 185–192, 2009.
- [30] BRANDES, U., LERNER, J., and SNIJDERS, T. A. B., “Networks evolving step by step: statistical analysis of dyadic event data,” *ASONAM*, pp. 200–205, 2009.
- [31] BRANTS, T. and FRANZ, A., “Web 1T 5-gram version 1.” Linguistic Data Consortium, Philadelphia, PA, 2006.
- [32] BRIONES, R., NAN, X., MADDEN, K., and WAKS, L., “When vaccines go viral: An analysis of HPV vaccine coverage on YouTube,” *Health Commun.*, vol. 27, no. 5, pp. 478–485, 2012.
- [33] BRODER, A., CICCULO, P., GABRILOVICH, E., JOSIFOVSKI, V., METZLER, D., RIEDEL, L., and YUAN, J., “Online expansion of rare queries for sponsored search,” in *WWW ’09: Proceedings of the 18th international conference on World wide web*, (New York, NY, USA), pp. 511–520, ACM, 2009.
- [34] BRODY, S. and ELHADAD, N., “Detecting salient aspects in online reviews of health providers,” in *Proceedings of the AMIA Annual Symposium 2010*, (Washington, D.C.), American Medical Informatics Association, 2010.
- [35] BUNTINE, W. and JAKULIN, A., “Discrete component analysis,” in *Subspace, Latent Structure and Feature Selection* (SAUNDERS, C., GROBELNIK, M., GUNN, S., and SHAW-TAYLOR, J., eds.), vol. 3940 of *Lecture Notes in Computer Science*, pp. 1–33, Springer Berlin / Heidelberg, 2006.
- [36] BURTON, S., MORRIS, R., DIMOND, M., HANSEN, J., GIRAUD-CARRIER, C., WEST, J., HANSON, C., and BARNES, M., “Public health community mining in YouTube,” in *SIGHIT IHI*, pp. 81–90, 2012.
- [37] BUTLER, B., “Membership size, communication activity and sustainability: a resource-based model of online social structures,” *Info. Sys. Res.*, vol. 12, no. 4, pp. 346–362, 2001.

- [38] BUTTS, C. T., “A relational event framework for social action,” *Sociological Methodology*, vol. 38, no. 1, pp. 155–200, 2008.
- [39] CAN, A. B. and BAYKAL, N., “Medicoport: A medical search engine for all,” *Computer Methods and Programs in Biomedicine*, vol. 86, no. 1, pp. 73 – 86, 2007.
- [40] CAWSEY, A., GRASSO, F., and PARIS, C., *The adaptive Web.*, ch. Adaptive information for consumers of healthcare. Springer Berlin / Heidelberg, 2007.
- [41] CENTOLA, D., GONZÁLEZ-AVELLA, J. C., EGUÍLUZ, V. M., and MIGUEL, M. S., “Homophily, cultural drift, and the co-evolution of cultural groups,” *Journal of Conflict Resolution*, vol. 51, no. 6, pp. 905–929, 2007.
- [42] CHAFE, R., BORN, K. B., SLUTSKY, A. S., and LAUPACIS, A., “The rise of people power,” *Nature*, vol. 472, pp. 410–411, 2011.
- [43] CHAN, C. V., MATTHEWS, L. A., and KAUFMAN, D. R., “A taxonomy characterizing complexity of consumer eHealth literacy,” in *AMIA Annual Symposium Proceedings*, vol. 2009, pp. 86–90, 2009.
- [44] CHANG, J., BOYD-GRABER, J., GERRISH, S., WANG, C., and BLEI, D., “Reading tea leaves: How humans interpret topic models,” in *Advances in Neural Information Processing Systems 22* (BENGIO, Y., SCHUURMANS, D., LAFFERTY, J., WILLIAMS, C. K. I., and CULOTTA, A., eds.), pp. 288–296, 2009.
- [45] CHEN, W.-Y., CHU, J.-C., LUAN, J., BAI, H., WANG, Y., and CHANG, E. Y., “Collaborative filtering for Orkut communities: discovery of user latent behavior,” in *WWW*, pp. 681–690, 2009.
- [46] CHOU, W. Y., HUNT, Y., FOLKERS, A., and AUGUSTSON, E., “Cancer survivorship in the age of YouTube and social media: a narrative analysis,” *J Med Internet Res*, vol. 13, no. 1, p. e7, 2011.
- [47] CHURCH, K. W. and GALE, W. A., “Poisson mixtures,” *Natural Language Engineering*, vol. 1, pp. 163–190, 1995.
- [48] COHEN, J., “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [49] CONCATO, J., FEINSTEIN, A. R., and HOLFORD, T. R., “The risk of determining risk with multivariable models,” *Annals Of Internal Medicine*, vol. 118, no. 3, pp. 201–210, 1993.
- [50] COX, D. R., “Regression models and life tables (with discussion),” *J of the Roy Stat Soc: Ser B (Stat Meth)*, vol. 34, pp. 187–220, 1972.
- [51] COX, D. R., “Partial likelihood,” *Biometrika*, vol. 62, pp. 269–276, 1975.

- [52] CRAIN, S. P., HUANG, J., and ZHA, H., “A scalable assistant librarian: hierarchical classification of books,” in *SIGIR*, 2008.
- [53] CRAIN, S. P. and JIAO, Y., “Its time you DROVE: Deep retrieval with ontological visualization and exploration,” tech. rep., Oak Ridge National Laboratory, 2009.
- [54] CRAIN, S. P., YANG, S.-H., JIAO, Y., and ZHA, H., “Dialect topic modeling for improved consumer medical search,” in *AMIA Ann. Sym.*, (Washington, D.C.), American Medical Informatics Association, 2010.
- [55] CRAIN, S. P., YANG, S.-H., and ZHA, H., “Understanding group dynamics in health forums,” in *ASONAM*, 2012.
- [56] CRAIN, S. P., YANG, S. H., ZHOU, K., and ZHA, H., *Mining text data*, ch. Dimensionality reduction and topic modeling: from latent semantic indexing to latent Dirichlet allocation and beyond. Kluwer, 2012.
- [57] CRAIN, S. P. and ZHA, H., “Consumer medical information retrieval relevance judgments,” 2010.
- [58] CRAIN, S. P., ZHOU, K., and ZHA, H., “Recommendation in online health communities,” in *HI-BI-BI*, 2012.
- [59] CUSTERS, K. and VAN DEN BULCK, J., “Viewership of pro-anorexia websites in seventh, ninth and eleventh graders,” *European Eating Disorders Review*, vol. 17, no. 3, pp. 214–219, 2009.
- [60] DECOSTER, V. A., “The emotions of adults with diabetes: a comparison across race,” *Soc Work In Health Care*, vol. 36, no. 4, pp. 79 – 99, 2003.
- [61] DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., and HARSHMAN, R., “Indexing by latent semantic analysis,” *Journal of the American Society for Information Science*, vol. 41, pp. 391–407, Sept. 1990.
- [62] DIABETES HANDS FOUNDATION, “TuDiabetes.org: A community of people touched by diabetes.” <http://www.tudiabetes.org>, 2010. Accessed 12/16/2010.
- [63] DOUPI, P. and VAN DER LEI, J., “Towards personalized Internet health information: the STEPPS architecture,” *Medical informatics and the Internet in Medicine*, vol. 27, pp. 139–151, September 2002.
- [64] DOYLE, G. and ELKAN, C., “Accounting for burstiness in topic models,” in *ICML*, 2009.
- [65] ELHADAD, N., *User-sensitive text summarization: application to the medical domain*. PhD thesis, Columbia University, New York, NY, USA, 2006. Adviser-Mckeown, Kathleen.

- [66] FERNANDEZ-LUQUE, L., ELAHI, N., and GRAJALES, F. J., "An analysis of personal medical information disclosed in YouTube videos created by patients with multiple sclerosis," *Stud Health Technol Inform*, vol. 150, pp. 292–296, 2009.
- [67] FERNANDEZ-LUQUE, L., KARLSEN, R., and MELTON, G. B., "HealthTrust: A social network approach for retrieving online health videos," *J Med Internet Res*, vol. 14, no. 1, p. e22, 2012.
- [68] FISHER, T. L., BURNET, D. L., HUANG, E. S., CHIN, M. H., and KATHLEEN A, C., "Cultural leverage: Interventions using culture to narrow racial disparities in health care," *Medical Care Research and Review*, vol. 64, pp. suppl 243S–282S, Oct 2007.
- [69] FLEISS, J. L., "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.
- [70] FOX, S. and JONES, S., "The social life of health information." http://www.pewinternet.org/~media/Files/Reports/2009/PIP_Health_2009.pdf, 2009. Accessed: 2009-12-17. (Archived by WebCite® at <http://www.webcitation.org/5m5IixBMx>).
- [71] GAVARD, J. A., LUSTMAN, P. J., and CLOUSE, R. E., "Prevalence of depression in adults with diabetes: An epidemiological evaluation," *Diabetes Care*, vol. 16, no. 8, pp. 1167–1178, 1993.
- [72] GAVIN, J., RODHAM, K., and POYER, H., "The presentation of "pro-anorexia" in online group interactions," *Qualitative Health Research*, vol. 18, no. 3, pp. 325–33, 2008.
- [73] GOEL, S., BRODER, A., GABRILOVICH, E., and PANG, B., "Anatomy of the long tail: ordinary people with extraordinary tastes," in *WSDM*, pp. 201–210, 2010.
- [74] GÓMEZ-ZÚÑIGA, B., FERNANDEZ-LUQUE, L., POUSADA, M., HERNÁNDEZ-ENCUENTRA, E., and ARMAYONES, M., "ePatients on youtube: Analysis of four experiences from the patients' perspective," *Medicine 2.0*, vol. 1, 2012.
- [75] GOSSELIN, P. and POITRAS, P., "Use of an internet "viral" marketing software platform in health promotion," *J Med Internet Res*, vol. 10, no. 4, p. e47, 2008.
- [76] GRANNIS, S. J., OVERHAGE, J. M., and McDONALD, C., "Real world performance of approximate string comparators for use in patient matching.," *Stud Health Technol Inform*, vol. 107, pp. 43–47, 2004.
- [77] GREENBERG, L., D'ANDREA, G., and LORENCE, D., "Setting the public agenda for online health search: a white paper and action agenda," *J Med Internet Res*, vol. 6, no. 2, p. e18, 2004.

- [78] GRIFFITHS, T. L. and STEYVERS, M., “Finding scientific topics,” in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, (Irvine, CA), pp. 5228–5235, National Academies of Sciences and Engineering, April 2004.
- [79] GRIFFITHS, T. L., STEYVERS, M., BLEI, D. M., and TENENBAUM, J. B., “Integrating topics and syntax,” in *Advances in Neural Information Processing Systems 17* (SAUL, L. K., WEISS, Y., and BOTTOU, L., eds.), pp. 537–544, Cambridge, MA: MIT Press, 2005.
- [80] GUPTA, S., PHUNG, D., ADAMS, B., and VENKATESH, S., “A matrix factorization framework for jointly analyzing multiple nonnegative data sources,” in *Proceedings of the SIAM Workshop on Text Mining*, 2011.
- [81] GUYOT, J., RADHOUANI, S., and FALQUET, G., “Ontology-based multilingual information retrieval,” in *In CLEF Workshop, Working Notes Multilingual Track*, pp. 21–23, 2005.
- [82] HARPER, K., SPERRY, S., and THOMPSON, J. K., “Viewership of pro-eating disorder websites: Association with body image and eating disturbances,” *International Journal of Eating Disorders*, vol. 41, no. 1, pp. 92–95, 2008.
- [83] HARTLINE, C., “Dying to fit in—literally! learning to love our bodies and ourselves.” http://www.edreferral.com/body_image.htm/#Dying%20to%20Fit%20In, 2004. Accessed: 2012-08-03. (Archived by WebCite® at <http://www.webcitation.org/69dL42Ehp>).
- [84] HEATH, M. T., *Scientific computing: an introductory survey*. McGraw Hill, 2002.
- [85] HEISE, D. R., “Modeling event structures,” *Journal of Mathematical Society*, vol. 14, pp. 139–169, 1989.
- [86] HOEK, H. W. and VAN HOEKEN, D., “Review of the prevalence and incidence of eating disorders,” *International Journal of Eating Disorders*, vol. 34, no. 4, pp. 383–96, 2003.
- [87] HOFMANN, T., “Probabilistic latent semantic analysis,” in *Proc. of Uncertainty in Artificial Intelligence, UAI99*, p. 21, Citeseer, 1999.
- [88] HUMPHREYS, K., “Clinicians’ referral and matching of substance abuse patients to self-help groups after treatment,” *Psychiatr Serv*, vol. 48, pp. 1445–1449, Nov 1997.
- [89] INTERNET ACCURACY PROJECT, “Internet accuracy project.” <http://accuracyproject.org>, 2012. Accessed: 2012-07-30. (Archived by WebCite® at <http://www.webcitation.org/69YXuoJri>).

- [90] IVERSON, S. A., HOWARD, K. B., and PENNEY, B. K., “Impact of Internet use on health-related behaviors and the patient-physician relationship: a survey-based study and review,” *Journal of the American Osteopathic Association*, vol. 108, pp. 699–711, December 2008.
- [91] JONES, Q., RAVID, G., and RAFAELI, S., “Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration,” *Info. Sys. Res.*, vol. 15, no. 2, pp. 194–210, 2004.
- [92] KAMEDA, T., OHTSUBO, Y., and TAKEZAWA, M., “Centrality in sociocognitive networks and social influence: an illustration in a group decision-making context,” *J of Personality and Soc Psych*, vol. 73, no. 2, pp. 296–309, 1997.
- [93] KANG, Y. and YU, N., “Soft-constraint based online lda for community recommendation,” in *Advances in Multimedia Information Processing - PCM 2010*, pp. 494–505, Springer, 2011.
- [94] KIM, H.-S. and SUNDAR, S. S., “Using interface cues in online health community boards to change impressions and encourage user contribution,” in *CHI*, pp. 599–608, 2011.
- [95] KIM, H. K., OH, H. Y., GU, J. C., and KIM, J. K., “Commenders: A recommendation procedure for online book communities,” *Elect. Comm. Res. and App.*, vol. 10, no. 5, pp. 501 – 509, 2011.
- [96] KIM, J. K., KIM, H. K., OH, H. Y., and RYU, Y. U., “A group recommendation system for online communities,” *International Journal of Information Management*, vol. 30, no. 3, pp. 212 – 219, 2010.
- [97] KOREN, Y., “Factorization meets the neighborhood: a multifaceted collaborative filtering model,” in *KDD*, pp. 426–434, 2008.
- [98] KOREN, Y., “Factor in the neighbors: Scalable and accurate collaborative filtering,” *TKDD*, vol. 4, no. 1, pp. 1–24, 2010.
- [99] KREPS, G. L. and NEUHAUSER, L., “New directions in ehealth communication: Opportunities and challenges,” *Patient Educ. Couns.*, vol. 78, pp. 329–336, 2010.
- [100] KRIPALANI, S., JACOBSON, K. L., BROWN, S., MANNING, K., RASK, K. J., and JACOBSON, T. A., “Development and implementation of a health literacy training program for medical residents,” *Medical Education Online*, vol. 11, 2006.
- [101] KULLBACK, S. and LEIBLER, R. A., “On information and sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, March 1951.

- [102] KUMMERVOLD, P. E., CHRONAKI, C. E., LAUSEN, B., PROKOSCH, H. U., RASMUSSEN, J., SANTANA, S., STANISZEWSKI, A., and WANGBERG, S. C., “eHealth trends in Europe 2005–2007: A population-based survey,” *J Med Internet Res*, vol. 10, no. 4, p. e42, 2008.
- [103] KUMMERVOLD, P. E., GAMMON, D., BERGVIK, S., JOHNSEN, J. A., HASVOLD, T., and ROSENVINGE, J. H., “Social support in a wired world: use of online mental health forums in Norway,” *Nord. J. Psychiatry*, vol. 56, no. 1, pp. 59–65, 2002.
- [104] KUNST, H., GROOT, D., LATTHE, P. M., LATTHE, M., and KHAN, K. S., “Accuracy of information on apparently credible websites: survey of five common health topics,” *BMJ*, vol. 324, no. 7337, pp. 581–582, 2002.
- [105] LEWIS, S. P. and ARBUTHNOTT, A. E., “Searching for thinspiration: the nature of internet searches for pro-eating disorder websites,” *Cyberpsychol Behav Soc Netw*, vol. 15, no. 4, pp. 200–4, 2012.
- [106] LI, W., BLEI, D., and MCCALLUM, A., “Nonparametric Bayes Pachinko allocation,” in *UAI 07*, 2007.
- [107] LINKLETTER, M., GORDON, K., and DOOLEY, J., “The choking game and YouTube: a dangerous combination,” *Clin Pediatr (Phila)*, vol. 49, no. 3, pp. 274–279, 2010.
- [108] LUO, G., TANG, C., YANG, H., and WEI, X., “MedSearch: a specialized search engine for medical information retrieval,” in *CIKM*, pp. 143–152, 2008.
- [109] LUSTMAN, P. J., GRIFFITH, L. S., FREEDLAND, K. E., and CLOUSE, R. E., “The course of major depression in diabetics,” *Gen Hosp Psychiatry*, vol. 19, no. 2, pp. 138–143, 1997.
- [110] MAIL FOREIGN SERVICE, “Second Ralph Lauren model in Photoshop row as she’s airbrushed to become impossibly skinny,” *MailOnline*, 2009. Accessed: 2012-08-03. (Archived by WebCite® at <http://www.webcitation.org/69dLeCHyN>).
- [111] MALIN, B. and AIROLDI, E., “Confidentiality preserving audits of electronic medical record access,” *Studies in Health Technology and Informatics*, vol. 129, no. Pt 1, pp. 320–324, 2007.
- [112] MARTIJN, C., SMEETS, E., JANSEN, A., HOEYMANS, N., and SCHOEMAKER, C., “Don’t get the message: The effect of a warning text before visiting a proanorexia website,” *International Journal of Eating Disorders*, vol. 42, no. 2, pp. 139–145, 2009.
- [113] MCGEE, J., “Unrealistic body image in the fashion industry.” <http://voices.yahoo.com/unrealistic-body-image-fashion-industry-5845700.html>,

2010. Accessed: 2012-08-06. (Archived by WebCite® at <http://www.webcitation.org/69i2E3r33>).
- [114] McLELLAN, A. T. and ALTERMAN, A. I., “Patient treatment matching: a conceptual and methodological review with suggestions for future research,” *NIDA Res Monogr*, vol. 106, pp. 114–135, 1991.
 - [115] MCPHERSON, M., LOVIN, L. S., and COOK, J. M., “Birds of a feather: Homophily in social networks,” *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.
 - [116] MEI, Q., LING, X., WONDRA, M., SU, H., and ZHAI, C., “Topic sentiment mixture: modeling facets and opinions in weblogs,” in *WWW ’07: Proceedings of the 16th international conference on World Wide Web*, (New York, NY, USA), pp. 171–180, ACM Press, 2007.
 - [117] MEYER, P., “Decomposition of supermartingales: the uniqueness theorem,” *Illinois Journal of Mathematics*, vol. 7, pp. 1–17, 1963.
 - [118] MIAH, A. and RICH, E., *The Medicalization of Cyberspace*, ch. The bioethics of cybermedicalization, pp. 107–116. Routledge, 2008.
 - [119] MIMNO, D., WALLACH, H. M., NARADOWSKY, J., SMITH, D. A., and MCCALLUM, A., “Polylingual topic models,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 880–889, August 2009.
 - [120] MIMNO, D. M. and MCCALLUM, A., “Topic models conditioned on arbitrary features with Dirichlet-multinomial regression,” in *UAI*, pp. 411–418, 2008.
 - [121] MINSKY, M., *The psychology of computer vision*, ch. A framework for representing knowledge. McGraw-Hill, 1975.
 - [122] MITCHELL, K. J. and YBARRA, M. L., “Online behavior of youth who engage in self-harm provides clues for preventive intervention,” *Prev Med*, vol. 45, no. 5, pp. 392–396, 2007.
 - [123] MOOLGAVKAR, S. H. and VENZON, D. J., “General relative risk regression models for epidemiologic studies,” *Am J Epidemiol*, vol. 126, pp. 949–961, 1987.
 - [124] MOYER, E., “Instagram follows Tumblr, Pinterest; bans self-harm posts.” http://news.cnet.com/8301-1023_3-57418395-93, April 21 2012. Accessed: 2012-08-06. (Archived by WebCite® at <http://www.webcitation.org/69iiDIL4v>).
 - [125] MUKEWAR, S., MANI, P., LOPEZ, R., and SHEN, B., “YouTube: a friend or foe when you are taking care of IBD patients?,” *The American Journal of Gastroenterology*, 2011.

- [126] MULVEEN, R. and HEPWORTH, J., “An interpretative phenomenological analysis of participation in a pro-anorexia internet site and its relationship with disordered eating,” *Journal of health psychology*, vol. 11, no. 2, pp. 283–96, 2006.
- [127] NATIONAL CENTER FOR CHRONIC DISEASE PREVENTION AND HEALTH PROMOTION, DIVISION OF DIABETES TRANSLATION, “National diabetes fact sheet,” 2011.
- [128] NATIONAL RESEARCH CORPORATION, “1 in 5 americans use social media for health care information,” *Press Release*, Feb 28 2011.
- [129] NEUHAUSER, L. and KREPS, G. L., “Online cancer communication: Meeting the literacy, cultural and linguistic needs of diverse audiences,” *Patient Educ. Couns.*, vol. 71, pp. 365–377, June 2008.
- [130] NEZLEK, J. B., HAMPTON, C. P., and SHEAN, G. D., “Clinical depression and day-to-day social interaction in a community sample,” *Journal of Abnormal Psychology*, vol. 109, pp. 11–19, Feb. 2000.
- [131] NORRIS, M. L., BOYDELL, K. M., PINHAS, L., and KATZMAN, D. K., “Anorexia and the Internet: a review of pro-anorexia websites,” *Int J Eat Disord*, vol. 39, no. 6, pp. 443–447, 2006.
- [132] OUIMETTE, P. C., FINNEY, J. W., GIMA, K., and MOOS, R. H., “A comparative evaluation of substance abuse treatment III. Examining mechanisms underlying patient-treatment matching hypotheses for 12-step and cognitive-behavioral treatments for substance abuse,” *Alcohol Clin Exp Res*, vol. 23, pp. 545–551, Mar 1999.
- [133] PANDEY, A., PATNI, N., SINGH, M., SOOD, A., and SINGH, G., “YouTube as a source of information on the H1N1 influenza pandemic,” *Am J Prev Med*, vol. 38, no. 3, pp. e1–3, 2010.
- [134] PANT, S., DESHMUKH, A., MURUGIAH, K., KUMAR, G., SACHDEVA, R., and MEHTA, J. L., “Assessing the credibility of the “YouTube approach” to health information on acute myocardial infarction,” *Clin Cardiol*, 2012.
- [135] PATEREK, A., “Improving regularized singular value decomposition for collaborative filtering,” in *Proc. KDD Cup and Workshop*, vol. 2007, pp. 5–8, 2007.
- [136] PAUL, M. and GIRJU, R., “Cross-cultural analysis of blogs and forums with mixed-collection topic models,” *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1408–1417, Aug 2009.
- [137] PEDUZZI, P., CONCATO, J., FEINSTEIN, A. R., and HOLFORD, T. R., “Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates,” *J Clin Epidemiol*, vol. 48, no. 12, pp. 1503–1510, 1995.

- [138] PEW RESEARCH CENTER, “Fall tracking survey 2008—health,” in *Pew Internet & American life project*, December 2008.
- [139] ÅHLFELDT, H., BORIN, L., DAUMKE, P., GRABAR, N., HALLETT, C., HARDCASTLE, D., KOKKINAKIS, D., MANCINI, C., MARKÓ, K., MERKEL, M., PIETSCH, C., POWER, R., SCOTT, D., SILVERVAR, A., GRONOSTAJ, M. T., WILLIAMS, S., and WILLIS, A., “Literature review on patient-friendly documentation systems,” tech. rep., The Open University, UK, May 2006.
- [140] REISINGER, J., WATERS, A., SILVERTHORN, B., and MOONEY, R. J., “Spherical topic models,” in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (FÜRNKRANZ, J. and JOACHIMS, T., eds.), (Haifa, Israel), pp. 903–910, Omnipress, June 2010.
- [141] RENDLE, S., FREUDENTHALER, C., and SCHMIDT-THIEME, L., “Factorizing personalized Markov chains for next-basket recommendation,” *WWW*, p. 811, 2010.
- [142] RICCIARDELLI, L. and MCCABE, M., “Children’s body image concerns and eating disturbance: A review of the literature,” *Clin Psychol Rev*, vol. 21, pp. 325–344, 2001.
- [143] RIDINGS, C. and WASKO, M., “Online discussion group sustainability: Investigating the interplay between structural dynamics and social dynamics over time,” *JAIS*, vol. 11, pp. 95–121, Feb 2010.
- [144] ROULEAU, C. R. and VON RANSON, K. M., “Potential risks of pro-eating disorder websites,” *Clin. Psychol. Rev.*, vol. 4, pp. 525–31, 2011.
- [145] RUDD, R. E., MOEYKENS, B. A., and COLTON, T., *The annual review of adult learning and literacy*, vol. 1 of *Jossey-Bass higher and adult education series*, ch. Health and literacy: a review of medical and public health literature, pp. 158–199. San Francisco: University of California Press, 1999.
- [146] SAHAY, S. and RAM, A., “Socia-semantic health information access,” in *AI and Health Communication. AAAI Spring Symposium 2011*, 2011.
- [147] SALAKHUTDINOV, R., “Bayesian probabilistic matrix factorization.” Source code, 2011.
- [148] SALAKHUTDINOV, R. and MNIH, A., “Bayesian probabilistic matrix factorization using Markov chain Monte Carlo,” *ICML*, pp. 880–887, 2008.
- [149] SALAKHUTDINOV, R. and MNIH, A., “Probabilistic matrix factorization,” *NIPS*, vol. 20, pp. 1257–1264, 2008.
- [150] SARWAR, B., KARYPIS, G., KONSTAN, J., and REIDL, J., “Item-based collaborative filtering recommendation algorithms,” *WWW*, pp. 285–295, 2001.

- [151] SCHWARTZBERG, J. G., VANGHEEST, J. B., and WANG, C. C., eds., *Understanding health literacy: implications for medicine and public health*. American Medical Association, 2005.
- [152] SHAW, S. J., HUEBNER, C., ARMIN, J., ORZEC, K., and VIVIAN, J., “The role of culture in health literacy and chronic disease screening and management,” *Journal of Immigrant and Minority Health*, vol. 11, no. 6, pp. 460–467, 2008.
- [153] SHEN, C. and MONGE, P., “Who connects with whom? a social network analysis of an online open source software community,” *First Monday*, vol. 16, June 6 2011.
- [154] SINGH, A. G., SINGH, S., and SINGH, P. P., “YouTube for information on rheumatoid arthritis – a wakeup call?,” *J Rheumatol*, vol. 39, no. 5, pp. 899–903, 2012.
- [155] SMITH, B. and FELLBAUM, C., “Medical WordNet: a new methodology for the construction and validation of information resources for consumer health,” in *COLING*, (Morristown, NJ, USA), p. 371, Association for Computational Linguistics, 2004.
- [156] SNIJDERS, T. A. B., “Statistical methods for network dynamics,” in *Proc of the Scientific Meeting of the Italian Statistical Society*, 2006.
- [157] SOOD, A., SARANGI, S., PANDEY, A., and MURUGIAH, K., “YouTube as a source of information on kidney stone disease,” *Urology*, vol. 77, no. 3, pp. 558–562, 2011.
- [158] STEINBERG, P. L., WASON, S., STERN, J. M., DETERS, L., KOWAL, B., and SEIGNE, J., “YouTube as source of prostate cancer information,” *Urology*, vol. 75, no. 3, pp. 619–622, 2010.
- [159] STEWART, G. W., “On the early history of the singular value decomposition,” *SIAM Review*, vol. 35, no. 4, pp. 551–566, 1993.
- [160] SYED-ABDUL, S., FERNANDEZ-LUQUE, L., CRAIN, S. P., HSU, M.-H., LI, Y.-C., JIAN, W.-S., WANG, Y.-C., DORJSUREN, K., CHULUUNBAATAR, Z., and NGUYEN, A., “Health related misinformation promoted through social media: The YouTube case,” In Submission.
- [161] TEH, Y. W., JORDAN, M. I., BEAL, M. J., and BLEI, D. M., “Hierarchical Dirichlet processes,” *JASA*, vol. 101, 2006.
- [162] TELFORD, K., KRALIK, D., and KOCH, T., “Acceptance and denial: implications for people adapting to chronic illness: literature review,” *Journal of Advanced Nursing*, vol. 55, no. 4, pp. 457–464, 2006.

- [163] TITOV, I. and McDONALD, R., “Modeling online reviews with multi-grain topic models,” in *Proceeding of the 17th international conference on World Wide Web*, WWW '08, (New York, NY, USA), pp. 111–120, ACM, 2008.
- [164] UNIDAD EDITORIAL INTERNET, S.L., “Aumentan un 470% las páginas web que fomentan la anorexia,” *El Mundo*, 15/02/2011 2011.
- [165] U.S. NATIONAL LIBRARY OF MEDICINE, “MedlinePlus Connect: linking electronic health records (EHRs) to MedlinePlus health information.” <http://www.nlm.nih.gov/medlineplus/connect>, Nov 2010.
- [166] U.S. NATIONAL LIBRARY OF MEDICINE, “MedlinePlus®: Trusted health information for you.” <http://www.nlm.nih.gov/medlineplus>, 2010. Accessed 12/16/2010.
- [167] U.S. NATIONAL LIBRARY OF MEDICINE, “Medical subject headings descriptors,” 2012.
- [168] VAN DE BELT, T. H., BERBEN, S. A., SAMSOM, M., ENGELEN, L. J., and SCHOONHOVEN, L., “Use of social media by Western European hospitals: longitudinal study,” *J Med Internet Res*, vol. 14, no. 3, p. e61, 2012.
- [169] VAN KNIPPENBERG, D. and SCHIPPERS, M. C., “Work group diversity,” *Annual Review of Psychology*, vol. 58, pp. 514–541, 2007.
- [170] VU, D. Q., ASUNCION, A. U., HUNTER, D. R., and SMYTH, P., “Dynamic egocentric models for citation networks,” in *ICML*, 2011.
- [171] VUCKOVIC, N. H., McMULLEN, C., and SCHNEIDER, J., “Clinician awareness of low health literacy,” in *12th Annual Conference of the HMO Research Network*, 2006.
- [172] WALLACH, H., MIMNO, D., and MCCALLUM, A., “Rethinking lda: Why priors matter,” in *Advances in Neural Information Processing Systems 22* (BENGIO, Y., SCHUURMANS, D., LAFFERTY, J., WILLIAMS, C. K. I., and CULOTTA, A., eds.), pp. 1973–1981, 2009.
- [173] WALLACH, H. M., “Topic modeling: beyond bag-of-words,” in *ICML*, 2006.
- [174] WALLACH, H. M., MURRAY, I., SALAKHUTDINOV, R., and MIMNO, D., “Evaluation methods for topic models,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, (New York, NY, USA), pp. 1105–1112, ACM, 2009.
- [175] WEI, X. and CROFT, W. B., “LDA-based document models for ad-hoc retrieval,” in *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, (New York, NY, USA), pp. 178–185, ACM, 2006.

- [176] WILSON, J. L., PEEBLES, R., HARDY, K. K., and LITT, I. F., “Surfing for thinness: A pilot study of pro-eating disorder web site usage in adolescents with eating disorders,” *Pediatrics*, vol. 118, no. 6, pp. e1635–e1643, 2006.
- [177] WINZELBERG, A. and HUMPHREYS, K., “Should patients’ religiosity influence clinicians’ referral to 12-step self-help groups? evidence from a study of 3,018 male substance abuse patients,” *J Consult Clin Psychol*, vol. 67, pp. 790–794, Oct 1999.
- [178] WOLFSEN, C., BURLING, M. G., TOMCZYK, P., and WOLFSEN, H., “Social media for survivors of esophageal cancer,” *The American Journal of Gastroenterology*, 2011.
- [179] WOOD, R. T. A. and WOOD, S. A., “An evaluation of two United Kingdom online support forums designed to help people with gambling issues,” *Journal of Gambling Issues*, pp. 5–30, June 2009.
- [180] WU, F. and HUBERMAN, B. A., “How public opinion forms,” in *WINE*, pp. 334–341, 2008.
- [181] YANG, S.-H., BIAN, J., and ZHA, H., “Hybrid generative/discriminative learning for automatic image annotation,” in *UAI’10: Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 2010.
- [182] YANG, S. H., CRAIN, S. P., and ZHA, H., “Bridging the language gap: topic-level adaptation for cross-domain knowledge transfer,” in *AI Stat*, 2011.
- [183] YANG, S.-H. and ZHA, H., “Language pyramid and multi-scale text analysis,” in *Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM ’10*, (New York, NY, USA), pp. 639–648, ACM, 2010.
- [184] YANG, S.-H., ZHA, H., and HU, B.-G., “Dirichlet-Bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora,” in *NIPS*, 2009.
- [185] YI-FRAZIER, J. P., SMITH, R. E., VITALIANO, P. P., YI, J. C., MAI, S., HILLMAN, M., and WEINGER, K., “A person-focused analysis of resilience resources and coping in patients with diabetes,” *Stress and Health*, vol. 26, no. 1, pp. 51–60, 2010.
- [186] YOM-TOV, E., FERNANDEZ-LUQUE, L., WEBER, I., and CRAIN, S. P., “Pro-anorexia and pro-recovery photo sharing: A tale of two warring tribes,” In submission.
- [187] ZENG, Q., KOGAN, S., ASH, N., GREENES, R. A., and BOXWALA, A. A., “Characteristics of consumer terminology for health information retrieval,” *Methods Inf Med*, vol. 41, pp. 289–298, 2002.

- [188] ZENG, Q. T., CROWELL, J., PLOVNICK, R. M., KIM, E., NGO, L., and DIBBLE, E., “Assisting consumer health information retrieval with query recommendations,” *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 80 – 90, 2006.
- [189] ZHOU, D., ZHU, S., YU, K., SONG, X., TSENG, B. L., ZHA, H., and GILES, C. L., “Learning multiple graphs for document recommendations,” *WWW*, p. 141, 2008.
- [190] ZHOU, Y., CONG, G., CUI, B., JENSEN, C., and YAO, J., “Routing questions to the right users in online communities,” in *ICDE*, pp. 700 –711, 2009.
- [191] ZINKEVICH, M., SMOLA, A. J., WEIMER, M., and LI, L., “Parallelized stochastic gradient descent,” in *NIPS*, 2010.

INDEX

A

analysis of variance 62
ANOVA.....*see* analysis of variance
AROC.....*see* receiver operator curve,
area under

B

bag of words 25
BMI.....*see* body mass index
body mass index 29
BOW.....*see* bag of words

C

consumer.....8

D

DCG... *see* discounted cumulative gain
diabetes.....3

E

eating disorder 3
anorexia
pro-recovery 6
proana 6, 115
ERGM *see* graph model, random,
exponential
error
root mean squared 39, 59, 103–105,
107, 108, 111
event.....10
event history analysis 31

F

frames 10
frequency
document, inverse 24

G

graph model, random, exponential..31
group
vitality.....11

H

health information resource.....8
health information resources
consumer-targeted.....4
technical 5

I

inference 31
Institute Review Board 66, 81, 102
interaction 10
interactions, social 4
IRB.....*see* Institute Review Board

K

KL-divergence ... *see* Kullback-Leibler
divergence
Kullback-Leibler divergence . 22, 45, 54

L

language gap 3, 8
latent Dirichlet allocation.....*see also*
topic model, 4–6, 14, 15, 17–20,
22, 25, 26, 39, 42–45, 49, 55, 56,
59, 117
hierarchical 27
polylingual 5
topic adapted 6, 28, 50, 51, 58, 59,
115
latent semantic indexing.....14, 19
probabilistic.....14
LDA.....*see* latent Dirichlet allocation
LSI *see* latent semantic indexing

M

matrix factorization
Bayesian 105, 106
collective 105–110
partial.....105–110
probabilistic 101, 105–110
matrix factorization, Bayesian.....106
maximum likelihood estimation 31

N	
NLP .. <i>see</i> natural language processing	
O	
observable	10
P	
parameter learning	31
participant	10
polysemy	9
privacy	11
profile, latent	11
Q	
quality	3
quality, content	10
R	
reciever operator curve, area under .	71
recommendation	4
risk	
relative	
conditional	88, 89, 91, 92

S	
singular value decomposition	7, 14, 35, 87, 88, 115
social media	3
SVD . <i>see</i> singular value decomposition	
synonymy	9
T	
topic	9
topic model	<i>see</i>
<i>also</i> latent Dirichlet allocation;	
latent semantic indexing,	4
dialect .	6, 28, 44–46, 49–52, 55, 56, 59, 114
V	
vitality, discussion group	3
W	
weight	
body	29

VITA

Steven P. Crain received a Certificate of Completion from Howell Public Schools in June 1988. He then attended Michigan State University, which granted a degree of Bachelor of Science in Mathematics with high honor in June 1992. He has since done graduate coursework at Gordon-Conwell Theological Seminary, the Assemblies of God Theological Seminary and Massachusetts Institute of Technology. He worked from February 1996 until August 2007 in numerous technical capacities (including system and database administration, system programming, software engineering and system security) for Eco Software and Primus Telecommunications. He has been studying for the Degree of Doctor of Philosophy at Georgia Institute of Technology since August 2007, working as a Graduate Research Assistant for Dr. Hongyuan Zha, briefly for Dr. Haessun Park and Instructor of Record for a course in numerical analysis. In July 2012 he joined the faculty of Oberlin College as visiting instructor in the Department of Computer Science.