STRUCTURED TOPIC MODELS: JOINTLY MODELING WORDS AND THEIR ACCOMPANYING MODALITIES

A Dissertation Presented

by

XUERUI WANG

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2009

Computer Science

STRUCTURED TOPIC MODELS: JOINTLY MODELING WORDS AND THEIR ACCOMPANYING MODALITIES

A Dissertation Presented

by

XUERUI WANG

Approved as to style and content by:

Andrew McCallum, Chair

James Allan, Member

David Jensen, Member

Padhraic Smyth, Member

John Staudenmayer, Member

Andrew G. Barto, Department Chair
Computer Science

ACKNOWLEDGMENTS

Writing the acknowledgments is the most enjoyable part of this thesis. Many people helped me during my graduate school career, and it is very exciting to have this opportunity to acknowledge them. Much of the research in this dissertation would not have occurred without the hard work and great insight of my colleagues.

My first thanks must go to Andrew McCallum. He has been such a wonderful advisor, and every aspect of this thesis has benefitted from his invaluable guidance and support throughout my graduate studies. At times I was a difficult, even timid graduate student, Andrew had tremendous patience to explain ideas within and beyond the scope of our research. What I will miss most, you bet, is the quick, at-will discussion either in my cube or in his office, coming away with new insights on almost everything from coding tricks to the state-of-the-art machine learning topics.

In addition to Andrew, my thesis has been shaped by a great committee. Having James Allan as a committee member has provided me an information retrieval resource with which I could apply topic models to many real-world IR problems. I have benefited much from David Jensen's suggestions and his pedagogy, which helped make graduate school a very enjoyable experience. Much of the work described are based on work by Padhraic Smyth's research group, and I learned a great deal from them. John Staudenmayer brought a much-needed statistics perspective to a thesis on statistical topic models. For their guidance and support I am also grateful to my master advisers Tom Mitchell and Wenhuang Liu. Their clever ideas initiated the stage for my subsequent progress.

I have been fortunate to have great collaborators in the past few years: Andrei Broder, Andres Corrada-Emmanuel, Greg Druck, Henry Foley, Marcus Fontoura, Ev-

geniy Gabrilovich, C. Lee Giles, Rebecca Hutchinson, Vanja Josifovski, Marcel Just, Michael Kelm, David Kulp, Wei Li, Natasha Mohanty, Sharlene Newman, Radu Niculescu, Chris Pal, Bo Pang, Francisco Pereira, Jimeng Sun, Xing Wei, John Yen, and Haizheng Zhang. I have also benefitted from interactions with many other members of IESL. For this, I thank Ron Bekkerman, Kedar Belare, Gaurav Chandalia, Aron Culotta, Alex Dingle, Rob Hall, Gary Huang, Pallika Kanani, Gideon Mann, David Mimno, Jason Naradowsky, Khashayar Rohanemanesh, Adam Saunders, Karl Schultz, Charles Sutton, Andrew Tolopko, Hanna Wallach, Michael Wick, Rachel Shorey, and Limin Yao.

Last, but definitely not least, I am endlessly grateful to my dear parents Yulan and Chunxiang, my beautiful wife Xing, and my naughty son Richard. The love and support from my family have made my graduate school career, and all of my work, worthwhile. I dedicate this thesis to them.

ABSTRACT

STRUCTURED TOPIC MODELS: JOINTLY MODELING WORDS AND THEIR ACCOMPANYING MODALITIES

MAY 2009

XUERUI WANG

B.E., TSINGHUA UNIVERSITY

M.E., TSINGHUA UNIVERSITY

M.S., CARNEGIE MELLON UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Andrew McCallum

The abundance of data in the information age poses an immense challenge for us: how to perform large-scale inference to understand and utilize this overwhelming amount of information. Such techniques are of tremendous intellectual significance and practical impact. As part of this grand challenge, the goal of my Ph.D. thesis is to develop effective and efficient statistical topic models for massive text collections by incorporating extra information from other modalities in addition to the text itself.

Text documents are not just text, and different kinds of additional information are naturally interleaved with text. Most previous work, however, pays attention to only one modality at a time, and ignore the others. In my thesis, I will present a series of probabilistic topic models to show how we can bridge multiple modalities of information, in a united fashion, for various tasks. Interestingly, joint inference over

multiple modalities leads to many findings that can not be discovered from just one modality alone, as briefly illustrated below:

Email is pervasive nowadays. Much previous work in natural language processing modeled text using latent topics ignoring the social networks. On the other hand, social network research mainly dealt with the existence of links between entities without taking into consideration the language content or topics on those links. The author-recipient-topic (ART) model, by contrast, steers the discovery of topics according to the relationships between people, and learns topic distributions based on the direction-sensitive messages sent between entities.

However, the ART model does not explicitly identify groups formed by entities in the network. Previous work in social network analysis ignores the fact that different groupings arise for different topics. The group-topic (GT) model, a probabilistic generative model of entity relationships and textual attributes, simultaneously discovers groups among the entities and topics among the corresponding text.

Many of the large datasets do not have static latent structures; they are instead dynamic. The topics over time (TOT) model explicitly models time as an observed continuous variable. This allows TOT to see long-range dependencies in time and also helps avoid a Markov model's risk of inappropriately dividing a topic in two when there is a brief gap in its appearance. By treating time as a continuous variable, we also avoid the difficulties of discretization.

Most topic models, including all of the above, rely on the bag of words assumption. However, word order and phrases are often critical to capturing the meaning of text. The topical n-grams (TNG) model discovers topics as well as meaningful, topical phrases simultaneously.

In summary, we believe that these models are clear evidence that we can better understand and utilize massive text collections when additional modalities are considered and modeled jointly with text.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS . . . . . **iv**

ABSTRACT . . . . . **vi**

LIST OF TABLES . . . . . **xi**

LIST OF FIGURES . . . . . **xv**

CHAPTER

1. AN INTRODUCTION TO STATISTICAL TOPIC MODELS **1**

1.1 Latent Dirichlet Allocation 1
1.2 Motivation 4
1.3 Related Work 8
1.4 Evaluating Topic Models 9
1.5 Outline 9

2. TOPIC AND ROLE DISCOVERY IN SOCIAL NETWORKS **12**

2.1 Author-Recipient-Topic Models 15

2.1.1 Inference by Collapsed Gibbs Sampling 18

2.2 Experimental Results 19

2.2.1 Topics and Prominent Relations from ART 22
2.2.2 Stochastic Blockstructures and Roles 24
2.2.3 Perplexity Comparison between AT and ART 30

2.3 Role-Author-Recipient-Topic Models 32
2.4 Experimental Results with RART 36
2.5 Summary 37

3. JOINT GROUP AND TOPIC DISCOVERY FROM RELATIONS AND TEXT 38

3.1 Group-Topic Model 41
3.2 Experimental Results 45

3.2.1 The US Senate Dataset 47
3.2.2 The United Nations Dataset 50

3.2.2.1 Overlapping Time Intervals 53

3.3 Summary 55

4. TOPICS OVER TIME: A NON-MARKOV CONTINUOUS-TIME MODEL OF TOPICAL TRENDS 57

4.1 Topics over Time 60
4.2 Datasets 66
4.3 Experimental Results 68

4.3.1 Topics Discovered for Addresses 68
4.3.2 Topics Discovered for Email 71
4.3.3 Topics Discovered for NIPS 73
4.3.4 Time Prediction 74
4.3.5 Topic Distribution Profile over Time 75
4.3.6 Topic Co-occurrences over Time 75

4.4 Summary 76

5. PHRASE AND TOPIC DISCOVERY WITH APPLICATION TO INFORMATION RETRIEVAL 78

5.1 N-gram based Topic Models 80

5.1.1 Bigram Topic Model (BTM) 80
5.1.2 LDA Collocation Model (LDACOL) 81
5.1.3 Topical N-gram Model (TNG) 82

5.2 Experimental Results 85

5.2.1 Ad-hoc Retrieval 88
5.2.2 Difference between Topical N-grams and LDA in IR Applications 89
5.2.3 Comparison of BTM, LDACOL and TNG on TREC Ad-hoc Retrieval 92

5.3 Summary 93

6. CONCLUSIONS 95

APPENDICES

A. COLLAPSED GIBBS SAMPLING DERIVATION FOR ART 101
B. COLLAPSED GIBBS SAMPLING DERIVATION FOR GT 103
C. COLLAPSED GIBBS SAMPLING DERIVATION FOR TOT 105
D. COLLAPSED GIBBS SAMPLING DERIVATION FOR TNG 107
E. ALL 50 ART TOPICS FOR ENRON DATASET 109

BIBLIOGRAPHY 122

LIST OF TABLES

Table **Page**

1.1 Notation used in this manuscript 3

2.1 An illustration of several topics from a 50-topic run for the Enron email dataset. Each topic is shown with the top 10 words and their corresponding conditional probabilities. The quoted titles are our own summary for the topics. Below are prominent author-recipient pairs for each topic. For example, Mary Hain was an in-house lawyer at Enron; Eric Bass was the coordinator of a fantasy football league within Enron. See all 50 topics in Appendix E. 23

2.2 The four topics most prominent in McCallum's email exchange with Padhraic Smyth, from a 50-topic run of ART on 9 months of McCallum's email. The topics provide an extremely salient summary of McCallum and Smyth's relationship during this time period: they wrote a grant proposal together; they set up many meetings; they discussed machine learning models; they were friendly with each other. Each topic is shown with the 10 highest-probability words and their corresponding conditional probabilities. The quoted titles are our own summary for the topics. Below are prominent author-recipient pairs for each topic. The people other than smyth also appear in very sensible associations: stowell is McCallum's proposal budget administrator; McCallum also wrote a proposal with John Lafferty and Fernando Pereira; McCallum also sets up meetings, discusses machine learning and has friendly discourse with his graduate student advisees: ronb, wellner, casutton, and culotta; he does not, however, discuss the details of proposal-writing with them. 24

2.3 Pairs considered most alike by ART and SNA on McCallum email. All pairs produced by the ART model are accurately quite similar. This is not so for the top SNA pairs. Many users are considered similar by SNA merely because they appear in the corpus mostly sending email only to McCallum. However, this causes people with very different roles to be incorrectly declared similar—such as McCallum's spouse and the JMLR editor. 29

2.4 Pairs with the highest rank difference between ART and SNA on McCallum email. The traditional SNA metric indicates that these pairs of people are different, while ART indicates that they are similar. There are strong relations between all pairs. 30

2.5 An illustration of two roles from a 50-topic, 15-role run for the McCallum email dataset. Each role is shown with the most prominent users (their short descriptions in parenthesis) and the corresponding conditional probabilities. The quoted titles are our own summary for the roles. For example, in Role 3, the users are all employees (or mailing lists) of the IT support staff at UMass CS, except for *allan*, who, however, was the professor chairing the department's computing committee. 35

2.6 An illustration of the role distribution of two users from a 50-topic, 15-role run for the McCallum email dataset. Each user is shown with his most prominent roles (their short descriptions in parenthesis) and the corresponding conditional probabilities. For example, considering user *pereira* (Fernando Pereira), his top five role assignments are all appropriate, as viewed through McCallum's email. 36

3.1 Additional notation (to Table 1.1) used in this chapter 42

3.2 Average AI for different models for both Senate and UN datasets. The group cohesion in (joint) GT is significantly better than in (serial) baseline, as well as the blockstructures model that does not use text at all. 46

3.3 Top words for topics generated with the mixture of unigrams model on the Senate dataset. The headers are our own summary of the topics. 47

3.4 Top words for topics generated with the GT model on the Senate dataset. The topics are influenced by both the words and votes on the bills. 48

3.5 Senators in the four groups corresponding to Topic Education + Domestic in Table 3.4. 49

3.6 Senators that switch groups the most across topics for the 101st-109th Senates 50

3.7 Top words for topics generated from mixture of unigrams model with the UN dataset (1990-2003). Only text information is utilized to form the topics, as opposed to Table 3.8 where our GT model takes advantage of both text and voting information. 51

3.8 Top words for topics generated from the GT model with the UN dataset (1990-2003) as well as the corresponding groups for each topic (column). The countries listed for each group are ordered by their 2005 GDP (PPP) and only the top 5 countries are shown in groups that have more than 5 members. 52

3.9 Results for 15-year-span slices of the UN dataset (1960-2000). The top probable words are listed for all topics, but only the groups corresponding the most dominant topic are shown (Topic 3). We list the countries for each group ordered by their 2005 GDP (PPP)and only show the top 5 countries in groups that have more than 5 members. We do not repeat the results in Table 3.8 for the most recent window (1990-2003). 54

4.1 Average KL divergence between topics for TOT vs. LDA on three datasets. TOT finds more distinct topics. 74

4.2 Predicting the decade, in the Address dataset. L1 Error is the difference between predicted and true decade. In the Accuracy column, we see that TOT predicts exactly the correct decade nearly twice as often as LDA. 74

5.1 The four topics from a 50-topic run of TNG on 13 years of NIPS research papers with their closest counterparts from LDA. The **Title** above the word lists of each topic is our own summary of the topic. To better illustrate the difference between TNG and LDA, we list the n-grams ($n > 1$) and unigrams separately for TNG. Each topic is shown with the 20 sorted highest-probability words. The TNG model produces clearer word list for each topic by associating many generic words (such as "set", "field", "function", etc.) with other words to form n-gram phrases. 86

5.2 Comparison of LDA and TNG on TREC retrieval performance (average precision) of eight queries on the SJMN dataset. The top four queries obviously contain phrase(s), and thus TNG achieves much better performance. On the other hand, the bottom four queries do not contain common phrase(s) after preprocessing (stopword and punctuation removal). Surprisingly, TNG still outperforms LDA on some of these queries. 91

5.3 Comparison of the bigram topic model ($\lambda = 0.7$), LDA collocation model ($\lambda = 0.9$) and the topical n-gram Model ($\lambda = 0.8$) on TREC retrieval performance (average precision) on the SJMN dataset. * indicates statistically significant differences in performance with 95% confidence according to the Wilcoxon test. TNG performs significantly better than other two models overall. 92

5.4 Comparison of the bigram topic model ($\lambda = 0.7$), LDA collocation model ($\lambda = 0.9$) and the topical n-gram Model ($\lambda = 0.8$) on TREC retrieval performance (average precision). The values of λ were tuned on the SJMN dataset. * indicates statistically significant differences in performance with 95% confidence according to the Wilcoxon test. TNG performs significantly better than other two models on both datasets. 93

LIST OF FIGURES

Figure **Page**

1.1 The graphical model representation of latent Dirichlet allocation (plate notation). 2

1.2 The abstract representation of all possible directed topic models of two conditional dependencies, with two observed modalities. In all figures, z is a latent topic, and w and v are observed information from two modalities. For simplicity, w and v are treated symmetrically, and they are just two different modalities, that is, we do not distinguish $w \longrightarrow v$ and $v \longrightarrow w$. 6

2.1 Three related models, and the ART model. In all models, each observed word, w, is generated from a multinomial word distribution, ϕ_z, specific to a particular topic/author, z, however topics are selected differently in each of the models. In LDA, the topic is sampled from a per-document topic distribution, θ, which in turn is sampled from a Dirichlet over topics. In the Author Model, there is one topic associated with each author (or category), and authors are sampled uniformly. In the Author-Topic model, the topic is sampled from a per-author multinomial distribution, θ, and authors are sampled uniformly from the observed list of the document's authors. In the Author-Recipient-Topic model, there is a separate topic-distribution for each author-recipient pair, and the selection of topic-distribution is determined from the observed author, and by uniformly sampling a recipient from the set of recipients for the document. 15

2.2 Perplexity on McCallum dataset for different values of the hyperparameters α and β. Perplexity and the corresponding experimental setting are discussed in detail in Section 2.2.3. From the perplexity plot, the ART model is not very sensitive to the hyperparameter values. 19

2.3 Power-law relationship between the frequency of occurrence of of an author (or an author-recipient pair) and the rank determined by the above frequency of occurrence. In the author plots, we treat both the sender and the recipients as authors. 21

2.4 **Left:** SNA Inverse JS Network. **Middle:** ART Inverse JS Network. **Right:** AT Inverse JS Network. Darker shades indicate higher similarity. 26

2.5 SNA Inverse JS Network for a 10 topic run on McCallum Email Data. Darker shades indicate higher similarity. Graph partitioning was calculated with the 128 authors that had ten or more emails in McCallum's Email Data. The block from 0 to 30 are people in and related to McCallum's research group at UMass. The block from 30 to 50 includes other researchers around the world. 28

2.6 Perplexity comparison of AT and ART on two datasets. We plot the information rate (logarithm of perplexity) here. The difference between AT and ART is significant under one-tailed t-test (Enron dataset: p-value < 0.01 except for 10 topics with p-value $= 0.018$; McCallum dataset: p-value $< 1e-5$). 31

2.7 Three possible variants for the Role-Author-Recipient-Topic (RART) model. 33

3.1 The Group-Topic model 41

4.1 Three topic models: LDA and two perspectives on TOT 61

4.2 Four topics discovered by TOT (above) and LDA (bottom) for the Address dataset. The titles are our own interpretation of the topics. Histograms show how the topics are distributed over time; the fitted beta PDFs are shown also. (For LDA, beta distributions are fit in a post-hoc fashion). The top words with their probability in each topic are shown below the histograms. The TOT topics are better localized in time, and TOT discovers more event-specific topical words. 67

4.3 Four topics discovered by TOT (above) and LDA (bottom) for the McCallum dataset, showing improved results with TOT. For example, the Faculty Recruiting topic is correctly identified in the spring in the TOT model, but LDA confuses it with other interactions among faculty. 70

4.4 Two topics discovered by TOT (above) and LDA (bottom) for the NIPS dataset. For example, on the left, two major approaches to dynamic system modeling are mixed together by LDA, but TOT more clearly identifies waning interest in Recurrent Neural Networks, with a separate topic (not shown) for rising interest in Markov models. 72

4.5 The distribution over topics given time in the NIPS data set. Note the rich collection of shapes that emerge from the Bayesian inversion of the collection of per-topic Beta distributions over time. 76

4.6 Eight topics co-occurring strongly with the "classification" topic in the NIPS dataset. Other co-occurring topics are labeled as a combined background topic. Classification with neural networks declined, while co-occurrence with SVMs, boosting and NLP are on the rise. The x-axis is the proceeding number, e.g., 1 corresponding to NIPS 1987 and 17 corresponding to NIPS 2003. 77

5.1 Three n-gram based topic models 81

6.1 The autocorrelation plot of perplexity in the Gibbs Chain for iterations 10001-10500 of the ART Model on the McCallum dataset. Y-axis is the autocorrelations for perplexity at varying time lags. The randomness of Gibbs samples is ascertained by near-zero autocorrelations for any and all time-lag separations. 96

CHAPTER 1

AN INTRODUCTION TO STATISTICAL TOPIC MODELS

The main idea behind statistical topic models is the assumption that documents are mixtures of topics, where a topic is a probability distribution over words. The discovery of topics is driven by the word co-occurrence patterns in a text collection. The majority of topic models are statistical generative models in which documents arise from a generative process. A primary goal of topic models is to invert the generative process through various standard statistical techniques and to infer the latent topics from which a collection of text documents were generated. Once the latent topics are discovered, it becomes much easier to understand these massive text collections, and they can be used as a succinct representation of documents for various tasks.

Research in statistical models of co-occurrence has led to the development of a variety of useful mechanisms for discovering low-dimensional, multi-faceted summaries of documents. In this chapter we review the most recently popular topic model, latent Dirichlet allocation (LDA) in detail [8, 20]. The topic models in subsequent chapters, all motivated by incorporating accompanying modalities of text, can be explained as extensions of LDA, and will be presented in similar ways.

1.1 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) and many other topic models share fundamentally the same idea: each document is a mixture of topics and each topic in turn is a

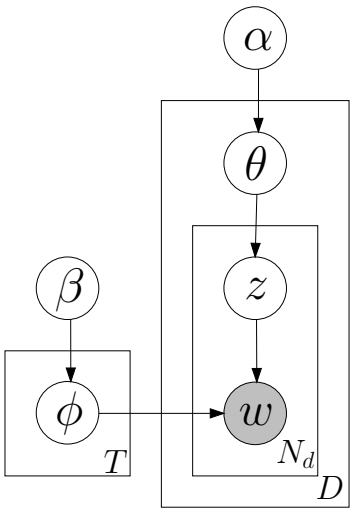


Figure 1.1. The graphical model representation of latent Dirichlet allocation (plate notation).

mixture of words. In LDA, this mixture behavior is captured through the following generative process:

1. draw discrete distributions ϕ_z from a Dirichlet prior β for each topic z;

2. for each document d, draw a discrete distribution θ_d from a Dirichlet prior α; then for each word w in document d:

   (a) draw z from discrete θ_d; and

   (b) draw w from discrete ϕ_z.

The generative process described here does not make any assumptions about the order of words as they appear in documents. This is known as the bag-of-words assumption that is very common in many language processing tasks.

Machine learning researchers usually convert this kind of generative process into a graphical model representation, to convey the idea more succinctly . The corresponding graphical model representation is shown in Figure 1.1. The repeated choices of topics and words can be conveniently illustrated using plate notation that repre-

SYMBOL	DESCRIPTION
T	number of topics
D	number of documents
V	number of unique words (vocabulary size)
N_d	number of word tokens in document d

Table 1.1. Notation used in this manuscript

sent replicates with the number in the lower right corner referring to the number of samples. Shaded and unshaded variables indicate observed and latent (i.e., hidden, or unobserved) variables, respectively. Arrows indicate conditional dependencies between variables that provide great convenience for inferring the latent variables. Some common notation used throughout this thesis is presented in Table 1.1.

When fitting a LDA model, the goal is to find the best set of latent variables that can explain the observed words in documents, assuming that the model actually generated the text collection. This involves inferring the probability distribution over words ϕ associated with each topic, the distribution over topics θ for each document, and the topic responsible for generating each word. The hyperparameters α and β are used as a prior to smooth the distribution over topics θ and the distribution over words ϕ, respectively. These hyperparameters can be inferred from the observed data, but that is often not necessary in practice, as demonstrated by the sensitivity analysis of the models presented in this thesis.

Posterior inference can be conducted via standard statistical techniques such as Gibbs sampling [4], variational methods [27] and expectation-propagation [42]. Throughout this thesis, we focus on Gibbs sampling since it is easy to understand and to implement. Note that direct Gibbs sampling would sample all latent random variables including θ and ϕ as well and not surprisingly the chain would converge (mix) very slowly. Instead, we use *collapsed* Gibbs sampling by integrating out θ and

ϕ mathematically, which converges much faster due to simpler and smaller sample space.

The posterior distribution of ϕ can help understand the underlying semantic topics discussed in the text collection by looking at the top probable words, and the posterior distribution of θ provides a semantically meaningful low-dimensional representation of documents, which can be subsequently used for various tasks such as document classification and information retrieval.

1.2 Motivation

LDA captures the semantic topics by only looking at the co-occurrence of words in text documents. However, text documents are not just text. Intuitively, if we can use additional information, we might be able to discover more useful topics in many different applications. Given a task, with multiple modality information, how shall we incorporate everything together to discover topics?

Assume that we have w and v, random variables standing for two observed modalities (for example, words and timestamps), and a latent, low-dimensional topic random variable z. The abstract representation of all possible directed topic models of two conditional dependencies[1] are shown in Figure 1.2. For simplicity, w and v are treated symmetrically, and they are just two different modalities, that is, we do not distinguish $w \longrightarrow v$ and $v \longrightarrow w$. Otherwise, more variations can be formed by switching w and v.

First of all, the goal of topic models is to discover a low-dimensional latent representation, thus in Figure 1.2, (D), (E), and (F) are not good choices conceptually, in which topics are generated from, for example, words. In (D), (E) and (G), the

[1]Three or more dependencies are possible and meaningful for certain applications, but in general with complexity more difficult to manage. Discussion in this regard is beyond the scope of this thesis.

dependency between w and v has no influence on the latent topic z, thus we can exclude them as well.

Among the three remaining configurations (A), (B) and (C), the choice of configuration could depend on the application. Given a problem, we consider three major factors in deciding the best configuration:

- **Domain Knowledge**. Background information or accumulated knowledge about the relationship between modality could be used as a basis to design topic models for multiple-modality data. For instance, an author decides what he/she wants to write (topics).

- **Parameterization Difficulty**. Based on the intrinsic properties of the modalities (such as continuity vs discreteness), it is very common that one direction of dependency is much easier to parameterize than the other direction, which more often results in close-form mathematical derivations.

- **Inference Efficiency**. The goal of the topic models is to discover interpretable representations of textual data in lower dimensional space, and we do not want to map some modality into a sparser representation, which usually leads to huge inference burden. Thus in topic models, we want words to be generated from topics but not vice versa.

For example, a large email archive not only contains the text message, but also represents a complex social network of senders and recipients. The social network is not independent from the body messages: e.g., a professor may talk about different things to his students than to his secretary, and thus use different languages. Ignoring the social network, a plain LDA will not capture these kinds of subtle difference. On the other hand, two secretaries may never communicate with each other, but perform almost identical roles, and we would expect their messages contain requests for photocopying, travel bookings, and meeting room arrangements (i.e., they use

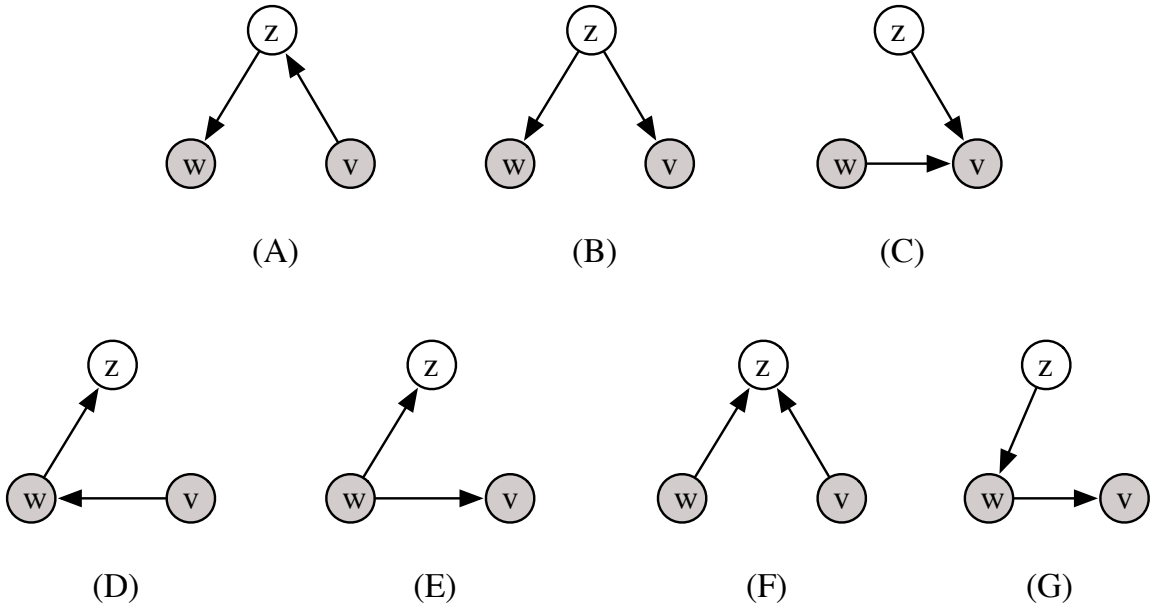


Figure 1.2. The abstract representation of all possible directed topic models of two conditional dependencies, with two observed modalities. In all figures, z is a latent topic, and w and v are observed information from two modalities. For simplicity, w and v are treated symmetrically, and they are just two different modalities, that is, we do not distinguish $w \longrightarrow v$ and $v \longrightarrow w$.

similar languages). Studying the social network alone would not help us identify such roles. We need a new model that can discover topics that is influenced by the accompanying modality—the social structure in which messages are sent and received. In this case, (A) in Figure 1.2 is more preferable since when a person writes to another, by common sense, the content of the message (topics and words) is decided by the author-recipient pair. With this configuration, the role discovery in an email network is feasible by conditioning topics on the social links as shown in Chapter 2.

We give another example showing that heeding only one modality is less desirable. Consider a legislative body and imagine its members forging alliances (forming groups), and voting accordingly. However, different alliances arise depending on the topic of the resolution up for a vote. For example, one grouping of the legislators may arise on the issue of taxation, while a quite different grouping may occur for votes on

foreign trade. Similar patterns of topic-based affiliations would arise in other types of entities as well, e.g., research paper co-authorship relations between people and citation relations between papers. Without the text, all these different groupings would look like the same. Correspondingly, grouping patterns can help distinguish topics described in very similar languages, such as "nuclear arsenal" and "nuclear arms race". The topic-specific grouping needs a model like (B) in Figure 1.2 since the topics would be the cause why a particular grouping arises.

Another modality that is often ignored, but exists everywhere, is time. Many of the large datasets to which topic models are applied do not have *static* co-occurrence patterns; they are instead *dynamic*. The data are often collected over time, and generally patterns present in the early part of the collection are not in effect later. Topics rise and fall in prominence; they split apart; they merge to form new topics; words change their correlations. Not modeling time can confound co-occurrence patterns and result in unclear, sub-optimal topic discovery. For example, in topic analysis of U.S. Presidential State-of-the-Union addresses, LDA confounds Mexican-American War (1846-1848) with some aspects of World War I (1914-1918), because LDA is unaware of the 70-year separation between the two events. A continuous modality (such as timestamp) is easier to generate from a discrete modality in a relatively small dimensionality (such as topics). From timestamps to topics, it would obviously need much more effort to describe. That is, the time sensitive topics would be best captured by a model like (B) in Figure 1.2 .

The bag-of-words assumption is prevalent in topic models as shown above, but word order, which we consider as an accompanying modality of text as well, is not only important for syntax, but also important for lexical meaning. For example, the phrase "white house" carries a special meaning beyond the appearance of its individual words, whereas "yellow house" does not. Note, however, that whether or not a phrase is a collocation may depend on the topic context. In the context

of a document about real estate, "white house" may not be a collocation. Phrases often have specialized meaning, but not always. For instance, "neural networks" is considered a phrase because of its frequent use as a fixed expression. However, it specifies two distinct concepts: biological neural networks in neuroscience and artificial neural networks in modern usage. Without consulting the context in which the term is located, it is hard to determine its actual meaning. Adding phrases increases the model's complexity, but it could be useful in certain contexts. The generation of an observed word is conditioned on the topic and another word, which is a perfect match for (C) in Figure 1.2 where the generation of words depends on both global context (topics) and local context (preceding word).

1.3 Related Work

Statistical topic models haven been actively studied in recent years, and many of them are dealing with multi-modality information as well and could be roughly mapped to one of the above configurations:

- *Configuration (A):* Author Model and Author-Topic Model of words and their authors [38, 50, 53], Topic Models with Meta Features of words and their arbitrary meta features [41], etc.
- *Configuration (B):* Citation-Topic Model [15] of words and research paper citations, Supervised LDA Model [7] of words and their class label, etc.
- *Configuration (C):* Syntax Topic Models [21], Bigram Topic Model [59], LDA-Collocation Model [22] of word sequences with Markov dependencies, etc.

Most of the above models will be discussed in detail with comparison to our own models in the following chapters.

1.4 Evaluating Topic Models

There are not yet gold standard metrics for evaluating topic models commonly accepted by the topic models community. In general, like clustering, the probability of a new document given by a vector of word indices are in terms of latent variables (the topic assignments of the words in the document), with intractable integration. In models of Configuration (A), since the topics are generated from an observed modality, such probability could be relatively easily calculated, and also termed as *perplexity* after some transformation. The use of perplexity originated from text compression, and in my opinion, it is not a good metric for evaluating topic model, because in most cases, we are more interested in the interpretability of text instead of compressibility.

When perplexity is not easy to calculate such as in LDA, Configuration (B) and (C), one can also use Chib's method [12] to approximately estimate marginal likelihood from samples of the topic assignments to words for the new document with the harmonic mean approximation. However, Chibs method is known to be very unstable (high to infinite variance).

As an alternative, instead of evaluating the topic models directly, many researchers evaluate the use of topic models is improving some supervised tasks such as text classification and information retrieval.

I do not have a strong preference on any of the above metrics. In this thesis, depending on the application, different metrics are explored.

1.5 Outline

The remainder of this thesis is laid out as follows:

In Chapter 2, I present the *Author-Recipient-Topic* (ART) model, a directed graphical model of words in a message generated given their author and a set of recipients. The model is similar to the Author-Topic (AT) model [50, 53], but with the crucial enhancement that it conditions the per-message topic distribution jointly

on both the author and individual recipients, rather than on individual authors. Thus the discovery of topics in the ART model is influenced by the social structure in which messages are sent and received. Each topic consists of a multinomial distribution over words. Each author-recipient pair has a multinomial distribution over topics. We can also easily calculate marginal distributions over topics conditioned solely on an author, or solely on a recipient, in order to find the topics on which each person is most likely to send or receive. Most importantly, we can also effectively use these person-conditioned topic distributions to measure similarity between people, and thus discover people's roles by clustering using this similarity.

In Chapter 3, the *Group-Topic* (GT) model is presented to consider not only the relations between objects but also the attributes of the relations (for example, the text associated with the relations) when assigning group membership. The GT model can be viewed as an extension of the stochastic blockstructures model [28, 46] with the key addition that group membership is conditioned on a latent variable associated with the attributes of the relation. Thus the discovery of groups is guided by the emerging topics, and the discovery of topics is guided by emerging groups. Resolutions that would have been assigned the same topic in a model using words alone may be assigned to different topics if they exhibit distinct voting patterns. Distinct word-based topics may be merged if the entities vote very similarly on them.

In Chapter 4, I present *Topics over Time* (TOT), a topic model that explicitly models time jointly with word co-occurrence patterns. Significantly, and unlike some recent work with similar goals, our model does not discretize time, and does not make Markov assumptions over state transitions in time. Rather, TOT parameterizes a continuous distribution over time associated with each topic, and topics are responsible for generating both observed timestamps as well as words. Parameter estimation is thus driven to discover topics that simultaneously capture word co-occurrences *and* locality of those patterns in time. When a strong word co-occurrence pattern appears

for a brief moment in time then disappears, TOT will create a topic with a narrow time distribution. (Given enough evidence, arbitrarily small spans can be represented, unlike schemes based on discretizing time.) When a pattern of word co-occurrence remains consistent across a long time span, TOT will create a topic with a broad time distribution.

In Chapter 5, we propose a new *topical n-gram* (TNG) model that automatically determines unigram words and phrases based on context and assign mixture of topics to both individual words and n-gram phrases. The ability to form phrases only where appropriate is unique to our model, distinguishing it from the traditional collocation discovery methods, where a *discovered* phrase is always treated as a *collocation* regardless of the context (which would possibly make us incorrectly conclude that "white house" remains a phrase in a document about real estate). Thus, TNG is not only a topic model that uses phrases, but also help linguists discover meaningful phrases in right context, in a completely probabilistic manner.

In Chapter 6, we summarize the ideas of the thesis and point to directions of future work.

CHAPTER 2

TOPIC AND ROLE DISCOVERY IN SOCIAL NETWORKS

Social network analysis (SNA) is the study of mathematical models for interactions among people, organizations and groups. With the recent availability of large datasets of human interactions [51, 71], the popularity of services like Facebook and LinkedIn, and the salience of the connections among the 9/11 hijackers, there has been growing interest in social network analysis.

Historically, research in the field has been led by social scientists and physicists [3, 33, 65, 66], and previous work has emphasized binary interaction data, with directed and/or weighted edges. There has not, however, previously been significant work by researchers with backgrounds in statistical natural language processing, nor analysis that captures the richness of the *language contents* of the interactions—the words, the topics, and other high-dimensional specifics of the interactions between people.

Using pure network connectivity properties, SNA often aims to discover various categories of nodes in a network. For example, in addition to determining that a node-degree distribution is heavy-tailed, we can also find those particular nodes with an inordinately high number of connections, or with connections to a particularly well-connected subset (group or block) of the network [1, 29, 28, 30, 32, 46]. Furthermore, using these properties we can assign "roles" to certain nodes [33, 69]. However, it is clear that network properties are not enough to discover all the roles in a social network. Consider email messages in a corporate setting, and imagine a situation in which a tightly knit group of users trade email messages with each other in a

roughly symmetric fashion. Thus, at the network level they appear to fulfill the same role. But perhaps, one of the users is in fact a manager for the whole group—a role that becomes obvious only when one accounts for the language content of the email messages.

Outside of the social network analysis literature, similarly, statistical topic models are also developed to discover the low-dimensional latent structures that are responsible to form documents in a corpus. For example, Probabilistic Latent Semantic Indexing [25] and Latent Dirichlet Allocation [8] robustly discover multinomial word distributions of these topics. Hierarchical Dirichlet Processes [57] can determine an appropriate number of topics for a corpus. The Author-Topic Model [53] learns topics conditioned on the mixture of authors that composed a document. However, none of these models are appropriate for SNA, in which we aim to capture the directed interactions and relationships between people.

This chapter presents the *Author-Recipient-Topic* (ART) model [40], a directed graphical model of words in a message generated given their author and a set of recipients. The model is similar to the Author-Topic (AT) model, but with the crucial enhancement that it conditions the per-message topic distribution jointly on both the author and individual recipients, rather than on individual authors. Thus the discovery of topics in the ART model is influenced by the social structure in which messages are sent and received. Each topic consists of a multinomial distribution over words. Each author-recipient pair has a multinomial distribution over topics. We can also easily calculate marginal distributions over topics conditioned solely on an author, or solely on a recipient, in order to find the topics on which each person is most likely to send or receive.

Most importantly, we can also effectively use these person-conditioned topic distributions to measure similarity between people, and thus discover people's roles by

clustering using this similarity.[1] For example, people who receive messages containing requests for photocopying, travel bookings, and meeting room arrangements can all be said to have the role "administrative assistant," and can be discovered as such because in the ART model they will all have these topics with high probability in their receiving distribution. Note that we can discover that two people have similar roles even if in the graph they are connected to very different sets of people.

We demonstrate this model on the Enron email corpus comprising 147 people and 23k messages, and also on about 9 months of incoming and outgoing mail of Andrew McCallum, comprising 825 people and 14k messages. We show not only that ART discovers extremely salient topics, but also gives evidence that ART predicts people's roles better than AT and SNA. Also, we show that the similarity matrix produced by ART is different from both the SNA matrix and the AT matrix in several appropriate ways. Furthermore, we find that the ART model gives a significantly lower perplexity on previously unseen messages than AT, which shows that ART is a better topic model for email messages.

We also describe an extension of the ART model that explicitly captures *roles* of people, by generating role associations for the author and recipient(s) of a message, and conditioning the topic distributions on the role assignments. The model, which we term *Role-Author-Recipient-Topic* (RART), naturally represents that one person can have more than one role. We describe several possible RART variants, and describe experiments with one of these variants.

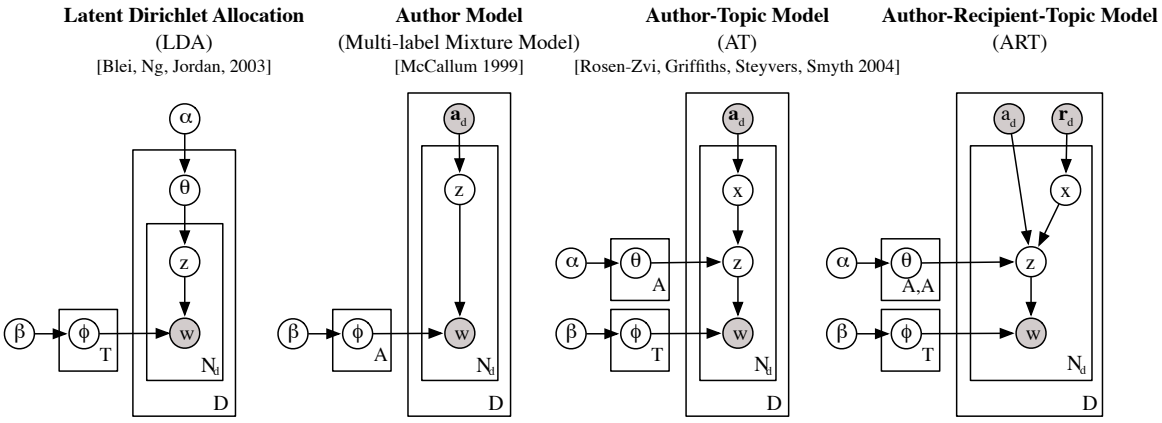


Figure 2.1. Three related models, and the ART model. In all models, each observed word, w, is generated from a multinomial word distribution, ϕ_z, specific to a particular topic/author, z, however topics are selected differently in each of the models. In LDA, the topic is sampled from a per-document topic distribution, θ, which in turn is sampled from a Dirichlet over topics. In the Author Model, there is one topic associated with each author (or category), and authors are sampled uniformly. In the Author-Topic model, the topic is sampled from a per-author multinomial distribution, θ, and authors are sampled uniformly from the observed list of the document's authors. In the Author-Recipient-Topic model, there is a separate topic-distribution for each author-recipient pair, and the selection of topic-distribution is determined from the observed author, and by uniformly sampling a recipient from the set of recipients for the document.

2.1 Author-Recipient-Topic Models

Before describing the ART model, we first describe two related models both of which are tightly related to the Latent Dirichlet Allocation model we described in Chapter 1. The graphical model representations for all models are shown in Figure 2.1 in which we list the LDA model as well for comparison. In addition to the notation in Table 1.1, here A represents the number of email accounts (senders and recipients).

The Author model, also termed a Multi-label Mixture Model [38], is a Bayesian network that simultaneously models document content and its authors' interests with

[1]The clustering may be either external to the model by simple greedy-agglomerative clustering, or internal to the model by introducing latent variables for the sender's and recipient's roles, as described in the Role-Author-Recipient-Topic (RART) model toward the end of this chapter.

a 1-1 correspondence between topics and authors. For each document d, a set of authors $\mathbf{a}_d$ is observed. To generate each word, an author, z, is sampled uniformly from the set, and then a word, w, is generated by sampling from an author-specific multinomial distribution ϕ_z. The Author-Topic (AT) model is a similar Bayesian network, in which each author's interests are modeled with a *mixture* of topics [53]. In its generative process for each document d, a set of authors, $\mathbf{a}_d$, is observed. To generate each word, an author x is chosen uniformly from this set, then a topic z is selected from a topic distribution θ_x that is specific to the author, and then a word w is generated from a topic-specific multinomial distribution ϕ_z. However, as described previously, none of these models is suitable for modeling message data.

An email message has one sender and in general more than one recipients. We could treat both the sender and the recipients as "authors" of the message, and then employ the AT model, but this does not distinguish the author and the recipients of the message, which is undesirable in many real-world situations. A manager may send email to a secretary and vice versa, but the nature of the requests and language used may be quite different. Even more dramatically, consider the large quantity of junk email that we receive; modeling the topics of these messages as undistinguished from the topics we write about as authors would be extremely confounding and undesirable since they do not reflect our expertise or roles.

Alternatively we could still employ the AT model by ignoring the recipient information of email and treating each email document as if it only has one author. However, in this case (which is similar to the LDA model) we are losing all information about the recipients, and the connections between people implied by the sender-recipient relationships.

Thus, we propose an Author-Recipient-Topic (ART) model for email messages. The ART model captures topics and the directed social network of senders and recipients by conditioning the multinomial distribution over topics distinctly on both

the author and one recipient of a message. Unlike AT, the ART model takes into consideration both author and recipients, in addition to modeling the email content as a mixture of topics.

The ART model is a Bayesian network that simultaneously models message content, as well as the directed social network in which the messages are sent. In its generative process, for each message d, an author, a_d, and a set of recipients, $\mathbf{r}_d$, are observed. To generate each word, a recipient, x, is chosen uniformly from $\mathbf{r}_d$, and then a topic z is chosen from a multinomial topic distribution $\theta_{a_d x}$, where the distribution is specific to the author-recipient pair (a_d, x). This distribution over topics could also be smoothed against a distribution conditioned on the author only, although we did not find that to be necessary in our experiments. Finally, the word w is generated by sampling from a topic-specific multinomial distribution ϕ_z. The result is that the discovery of topics is guided by the social network in which the collection of message text was generated.

In the ART model, given the hyperparameters α and β, an author a_d, and a set of recipients $\mathbf{r_d}$ for each message d, the joint distribution of the topic mixture θ_{ij} for each author-recipient pair (i, j), the word mixture ϕ_t for each topic t, a set of recipients $\mathbf{x}$, a set of topics $\mathbf{z}$ and a set of words $\mathbf{w}$ in the corpus is given by:

$$P(\Theta, \Phi, \mathbf{x}, \mathbf{z}, \mathbf{w}|\alpha, \beta, \mathbf{a}, \mathbf{r}) = \prod_{i=1}^{A}\prod_{j=1}^{A} p(\theta_{ij}|\alpha) \prod_{t=1}^{T} p(\phi_t|\beta) \prod_{d=1}^{D}\prod_{i=1}^{N_d} (P(x_{di}|\mathbf{r}_d)P(z_{di}|\theta_{a_d x_{di}})P(w_{di}|\phi_{z_{di}}))$$

Integrating over Θ and Φ, and summing over $\mathbf{x}$ and $\mathbf{z}$, we get the marginal distribution of a corpus:

$$P(\mathbf{w}|\alpha, \beta, \mathbf{a}, \mathbf{r})$$
$$= \iint \prod_{i=1}^{A}\prod_{j=1}^{A} p(\theta_{ij}|\alpha) \prod_{t=1}^{T} p(\phi_t|\beta) \prod_{d=1}^{D}\prod_{i=1}^{N_d} \sum_{x_{di}=1}^{A} (P(x_{di}|\mathbf{r}_d) \sum_{z_{di}=1}^{T} (P(z_{di}|\theta_{a_d x_{di}})P(w_{di}|\phi_{z_{di}})))\mathrm{d}\Phi\mathrm{d}\Theta$$

2.1.1 Inference by Collapsed Gibbs Sampling

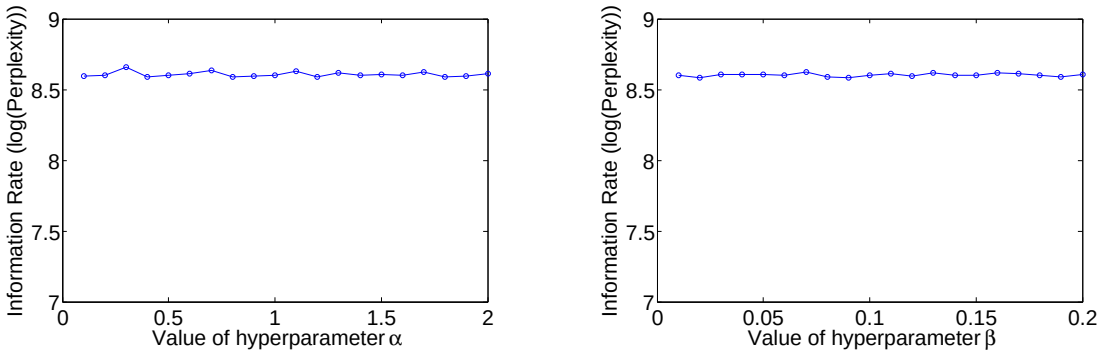
Inference on models in the LDA family cannot be performed exactly. Three standard approximate inference methods have been used to obtain practical results: variational methods [8], Gibbs sampling [20, 50, 53], and expectation propagation [20, 42]. We choose collapsed Gibbs sampling for its ease of implementation. Note that we adopt conjugate priors (Dirichlet) for the multinomial distributions, and thus we can easily integrate out θ and ϕ, analytically capturing the uncertainty associated with them. In this way we facilitate the sampling—that is, we need not sample θ and ϕ at all. One could estimate the values of the hyperparameters of the ART model, α and β, from data using a Gibbs EM algorithm [4]. In the particular applications discussed in this chapter, after trying out many different hyperparameter settings, we find that the sensitivity to hyperparameters is not very strong, as shown in Figure 2.2. Thus, again for simplicity, we use fixed symmetric Dirichlet distributions ($\alpha = 50/T$ and $\beta = 0.1$) in all our experiments.

We need to derive $P(x_{di}, z_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r})$, the conditional distribution of a topic and recipient for the word w_{di} given all other words' topic and recipient assignments, $\mathbf{x}_{-di}$ and $\mathbf{z}_{-di}$, to carry out the collapsed Gibbs sampling procedure for ART. We begin with the joint probability of the whole dataset, and by the chain rule, the above conditional probability can be obtained with ease:

$$P(x_{di}, z_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r}) \propto \frac{\alpha_{z_{di}} + n'_{a_d x_{di} z_{di}}}{\sum_{t=1}^{T}(\alpha_t + n'_{a_d x_{di} t})} \frac{\beta_{w_{di}} + m'_{z_{di} w_{di}}}{\sum_{v=1}^{V}(\beta_v + m'_{z_{di} v})}$$

where n'_{ijt} is the number of tokens (excluding w_{di}) assigned to topic t and the author-recipient pair (i, j), and m'_{tv} represent the number of tokens (excluding w_{di}) of word v assigned to topic t (n_{ijt} and m_{tv} are the corresponding counts including w_{di} used below).

The posterior estimates of θ and ϕ given the training set can be calculated by



(a) Perplexity vs. α ($\beta = 0.1$ and $T = 50$) (b) Perplexity vs. β ($\alpha = 1$ and $T = 50$)

Figure 2.2. Perplexity on McCallum dataset for different values of the hyperparameters α and β. Perplexity and the corresponding experimental setting are discussed in detail in Section 2.2.3. From the perplexity plot, the ART model is not very sensitive to the hyperparameter values.

$$\hat{\theta}_{ijz} = \frac{\alpha_z + n_{ijz}}{\sum_{t=1}^{T}(\alpha_t + n_{ijt})}, \hat{\phi}_{tw} = \frac{\beta_w + m_{tw}}{\sum_{v=1}^{V}(\beta_v + m_{tv})} \tag{2.1}$$

Detailed derivation of collapsed Gibbs sampling for ART is provided in Appendix A. An overview of the collapsed Gibbs sampling procedure we use is shown in Algorithm 1.

2.2 Experimental Results

We present results with the Enron email corpus and the personal email of a researcher (Andrew McCallum). The Enron email corpus, is a large body of email messages subpoenaed as part of the investigation by the Federal Energy Regulatory Commission (FERC), and then placed in the public record. The original dataset con-

Algorithm 1 Inference and Parameter Estimation in ART

initialize the author and topic assignments randomly for all tokens
repeat
  for $d = 1$ to D **do**
    for $i = 1$ to N_d **do**
      draw x_{di} and z_{di} from $P(x_{di}, z_{di} | \mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r})$
      update $n_{a_d x_{di} z_{di}}$ and $m_{z_{di} w_{di}}$ (or equivalently, $n'_{a_d x_{di} z_{di}}$ and $m'_{z_{di} w_{di}}$)
    end for
  end for
until the Markov chain reaches its equilibrium
compute the posterior estimates of θ and ϕ

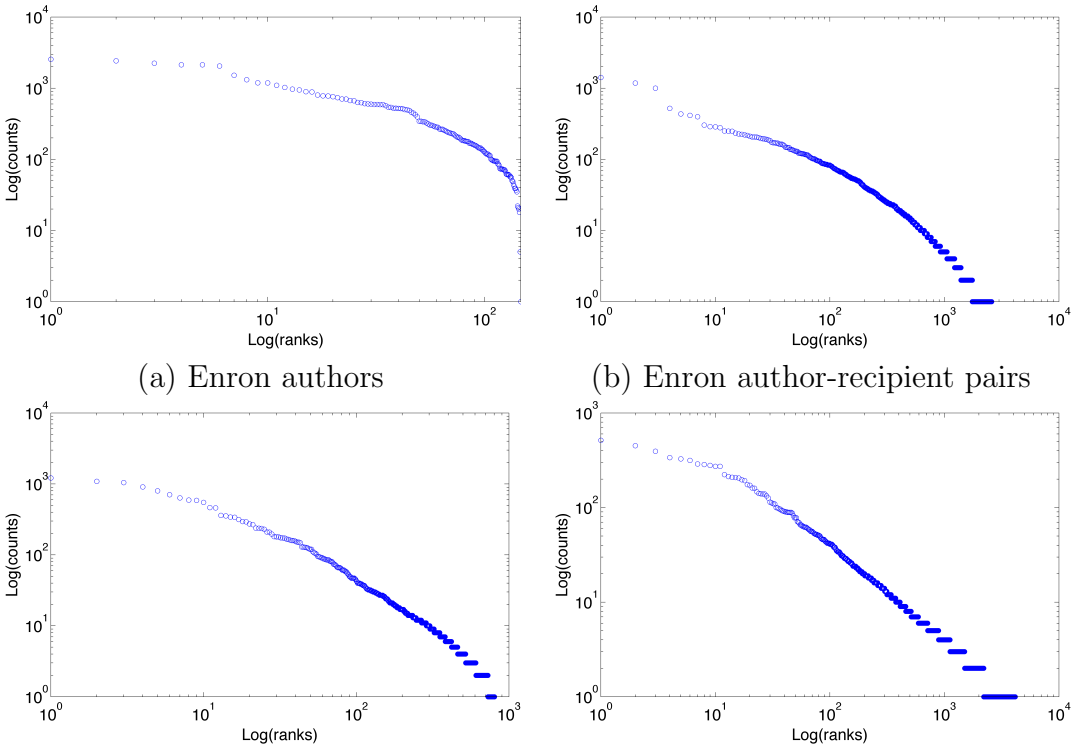
tains 517,431 messages, however MD5 hashes on contents, authors and dates show only 250,484 of these to be unique.

Although the Enron email dataset contains the email folders of 150 people, two people appear twice with different usernames, and we remove one person who only sent automated calendar reminders, resulting in 147 people for our experiments. We hand-corrected variants of the email addresses for these 147 users to capture the connectivity of as much of these users' emails as possible. The total number of email messages traded among these users is 23,488. We did not model email messages that were not received by at least one of the 147 users.

In order to capture only the new text entered by the author of a message, it is necessary to remove "quoted original messages" in replies. We eliminate this extraneous text by a simple heuristic: all text in a message below a "forwarded message" line or timestamp is removed. This heuristic certainly incorrectly loses words that are interspersed with quoted email text. Only words formed as sequences of alphabetic characters are kept, which results in a vocabulary of 22,901 unique words. To remove sensitivity to capitalization, all text is downcased.

Our second dataset consists of the personal email sent and received by McCallum between January and September 2004. It consists of 13,633 unique messages written by 825 authors. In a typical power-law behavior, most of these authors wrote only

a few messages, while 128 wrote ten or more emails. After applying the same text normalization filter (lowercasing, removal of quoted email text, etc.) that was used for the Enron dataset, we obtained a text corpus containing 457,057 word tokens, and a vocabulary of 22,901 unique words.



(a) Enron authors (b) Enron author-recipient pairs

(c) McCallum authors (d) McCallum author-recipient pairs

Figure 2.3. Power-law relationship between the frequency of occurrence of of an author (or an author-recipient pair) and the rank determined by the above frequency of occurrence. In the author plots, we treat both the sender and the recipients as authors.

By conditioning topic distributions on author-recipient pairs instead of authors, the data we have may look sparser considering that we have substantially more author-recipient pairs than authors. However, as shown in Figure 2.3, we can find that the number of emails of an author-recipient pair and its rank determined by the count still follow a power-law behavior, as for authors. For example, in the McCallum dataset, 500 of possible 680,625 author-recipient pairs are responsible for 70% of the email exchange. That is, even though the data are sparser for the ART model, the

power-law behavior makes it still possible to obtain a good estimate of the topic distributions for prominent author-recipient pairs.

We initialize the Gibbs chains on both datasets randomly, and find that the results are very robust to different initializations. By checking the perplexity, we find that usually the Gibbs chain converges after a few hundred iterations, and we run 10,000 iterations anyway to make sure it converges.

2.2.1 Topics and Prominent Relations from ART

Table 2.2 shows the highest probability words from eight topics in an ART model trained on the 147 Enron users with 50 topics. The quoted titles are our own interpretation of a summary for the topics. The clarity and specificity of these topics are typical of the topics discovered by the model. For example, Topic 17 (Document Review) comes from the messages discussing review and comments on documents; Topic 27 (Time Scheduling) comes from the messages negotiating meeting times.

Beneath the word distribution for each topic are the three author-recipient pairs with highest probability of discussing that topic—each pair separated by a horizontal line, with the author above the recipient. For example, Hain, the top author of messages in the "Legal Contracts" topic, was an in-house lawyer at Enron. By inspection of messages related to "Sports Pool", Eric Bass seems to have been the coordinator for a fantasy football league among Enron employees. In the "Operations" topic, it is satisfying to see Beck, who was the Chief Operating Officer at Enron; Kitchen was President of Enron Online; and Lavorato was CEO of Enron America. In the "Government Relations" topic, we see Dasovich, who was a Government Relation Executive, Shapiro, who was Vice President of Regulatory Affairs, Steffes, who was Vice President of Government Affairs, and Sanders, who was Vice President of WholeSale Services. In "Wireless" we see that Hayslett, who was Chief Financial Officer and

Topic 5 "Legal Contracts"		Topic 17 "Document Review"		Topic 27 "Time Scheduling"		Topic 45 "Sports Pool"	
section	0.0299	attached	0.0742	day	0.0419	game	0.0170
party	0.0265	agreement	0.0493	friday	0.0418	draft	0.0156
language	0.0226	review	0.0340	morning	0.0369	week	0.0135
contract	0.0203	questions	0.0257	monday	0.0282	team	0.0135
date	0.0155	draft	0.0245	office	0.0282	eric	0.0130
enron	0.0151	letter	0.0239	wednesday	0.0267	make	0.0125
parties	0.0149	comments	0.0207	tuesday	0.0261	free	0.0107
notice	0.0126	copy	0.0165	time	0.0218	year	0.0106
days	0.0112	revised	0.0161	good	0.0214	pick	0.0097
include	0.0111	document	0.0156	thursday	0.0191	phillip	0.0095
M.Hain	0.0549	G.Nemec	0.0737	J.Dasovich	0.0340	E.Bass	0.3050
J.Steffes		B.Tycholiz		R.Shapiro		M.Lenhart	
J.Dasovich	0.0377	G.Nemec	0.0551	J.Dasovich	0.0289	E.Bass	0.0780
R.Shapiro		M.Whitt		J.Steffes		P.Love	
D.Hyvl	0.0362	B.Tycholiz	0.0325	C.Clair	0.0175	M.Motley	0.0522
K.Ward		G.Nemec		M.Taylor		M.Grigsby	

Topic 34 "Operations"		Topic 37 "Power Market"		Topic 41 "Gov. Relations"		Topic 42 "Wireless"	
operations	0.0321	market	0.0567	state	0.0404	blackberry	0.0726
team	0.0234	power	0.0563	california	0.0367	net	0.0557
office	0.0173	price	0.0280	power	0.0337	www	0.0409
list	0.0144	system	0.0206	energy	0.0239	website	0.0375
bob	0.0129	prices	0.0182	electricity	0.0203	report	0.0373
open	0.0126	high	0.0124	davis	0.0183	wireless	0.0364
meeting	0.0107	based	0.0120	utilities	0.0158	handheld	0.0362
gas	0.0107	buy	0.0117	commission	0.0136	stan	0.0282
business	0.0106	customers	0.0110	governor	0.0132	fyi	0.0271
houston	0.0099	costs	0.0106	prices	0.0089	named	0.0260
S.Beck	0.2158	J.Dasovich	0.1231	J.Dasovich	0.3338	R.Hayslett	0.1432
L.Kitchen		J.Steffes		R.Shapiro		T.Geaccone	
S.Beck	0.0826	J.Dasovich	0.1133	J.Dasovich	0.2440	T.Geaccone	0.0737
J.Lavorato		R.Shapiro		J.Steffes		R.Hayslett	
S.Beck	0.0530	M.Taylor	0.0218	J.Dasovich	0.1394	R.Hayslett	0.0420
S.White		E.Sager		R.Sanders		D.Fossum	

Table 2.1. An illustration of several topics from a 50-topic run for the Enron email dataset. Each topic is shown with the top 10 words and their corresponding conditional probabilities. The quoted titles are our own summary for the topics. Below are prominent author-recipient pairs for each topic. For example, Mary Hain was an in-house lawyer at Enron; Eric Bass was the coordinator of a fantasy football league within Enron. See all 50 topics in Appendix E.

Treasurer, was an avid user of the Blackberry brand wireless, portable email system.

Results on the McCallum email dataset are reported in Table 2.2.

Topic 5 "Grant Proposals"		Topic 31 "Meeting Setup"		Topic 38 "ML Models"		Topic 41 "Friendly Discourse"	
proposal	0.0397	today	0.0512	model	0.0479	great	0.0516
data	0.0310	tomorrow	0.0454	models	0.0444	good	0.0393
budget	0.0289	time	0.0413	inference	0.0191	don	0.0223
work	0.0245	ll	0.0391	conditional	0.0181	sounds	0.0219
year	0.0238	meeting	0.0339	methods	0.0144	work	0.0196
glenn	0.0225	week	0.0255	number	0.0136	wishes	0.0182
nsf	0.0209	talk	0.0246	sequence	0.0126	talk	0.0175
project	0.0188	meet	0.0233	learning	0.0126	interesting	0.0168
sets	0.0157	morning	0.0228	graphical	0.0121	time	0.0162
support	0.0156	monday	0.0208	random	0.0121	hear	0.0132
smyth mccallum	0.1290	ronb mccallum	0.0339	casutton mccallum	0.0498	mccallum culotta	0.0558
mccallum stowell	0.0746	wellner mccallum	0.0314	icml04-web icml04-chairs	0.0366	mccallum casutton	0.0530
mccallum lafferty	0.0739	casutton mccallum	0.0217	mccallum casutton	0.0343	mccallum ronb	0.0274
mccallum smyth	0.0532	mccallum casutton	0.0200	nips04work mccallum	0.0322	mccallum saunders	0.0255
pereira lafferty	0.0339	mccallum wellner	0.0200	weinman mccallum	0.0250	mccallum pereira	0.0181

Table 2.2. The four topics most prominent in McCallum's email exchange with Padhraic Smyth, from a 50-topic run of ART on 9 months of McCallum's email. The topics provide an extremely salient summary of McCallum and Smyth's relationship during this time period: they wrote a grant proposal together; they set up many meetings; they discussed machine learning models; they were friendly with each other. Each topic is shown with the 10 highest-probability words and their corresponding conditional probabilities. The quoted titles are our own summary for the topics. Below are prominent author-recipient pairs for each topic. The people other than smyth also appear in very sensible associations: stowell is McCallum's proposal budget administrator; McCallum also wrote a proposal with John Lafferty and Fernando Pereira; McCallum also sets up meetings, discusses machine learning and has friendly discourse with his graduate student advisees: ronb, wellner, casutton, and culotta; he does not, however, discuss the details of proposal-writing with them.

2.2.2 Stochastic Blockstructures and Roles

The stochastic equivalence hypothesis from SNA states that nodes in a network that behave stochastically equivalently must have similar roles. In the case of an email network consisting of message counts, a natural way to measure equivalence is to examine the probability that a node communicated with other nodes. If two nodes have similar probability distribution over their communication partners, we should

consider them role-equivalent. Lacking a true distance measure between probability distributions, we can use some symmetric measure, such as the Jensen-Shannon (JS) divergence, to obtain a symmetric matrix relating the nodes in the network. Since we want to consider nodes/users that have a small JS divergence as equivalent, we can use the inverse of the divergence to construct a symmetric matrix in which larger numbers indicate higher similarity between users.

Standard recursive graph-cutting algorithms on this matrix can be used to cluster users, rearranging the rows/columns to form approximately block-diagonal structures. This is the familiar process of 'blockstructuring' used in SNA. We perform such an analysis on two datasets: a small subset of the Enron users consisting mostly of people associated with the Transwestern Pipeline Division within Enron, and the entirety of McCallum's email.

We begin with the Enron TransWestern Pipeline Division. Our analysis here employed a "closed-universe" assumption—only those messages traded among considered authors in the dataset were used.

The traditional SNA similarity measure (in this case JS divergence of distributions on recipients from each person) is shown in the left matrix in Figure 2.4. Darker shading indicates that two users are considered more similar. A related matrix resulting from our ART model (JS divergence of recipient-marginalized topic distributions for each email author) appears in the middle of Figure 2.4. Finally, the results of the same analysis using topics from the AT model rather than our ART model can be seen on the right. The three matrices are similar, but have interesting differences.

Consider Enron employee Geaccone (user 9 in all the matrices in Figure 2.4). According to the traditional SNA role measurement, Geaccone and McCarty (user 8) have very similar roles, however, both the AT and ART models indicate no special similarity. Inspection of the email messages for both users reveals that Geaconne was an Executive Assistant, while McCarty was a Vice-President—rather different

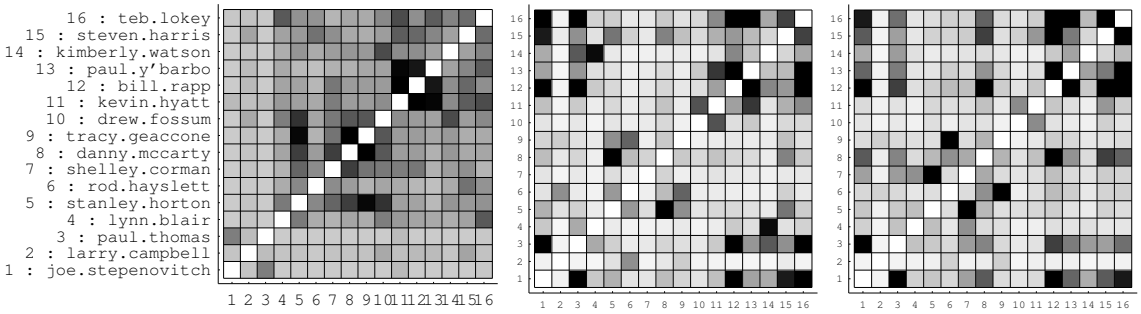


Figure 2.4. Left: SNA Inverse JS Network. **Middle:** ART Inverse JS Network. **Right:** AT Inverse JS Network. Darker shades indicate higher similarity.

roles—and, thus the output of ART and AT is more appropriate. We can interpret these results as follows: SNA analysis shows that they wrote email to similar sets of people, but the ART analysis illustrates that they used very different language when they wrote to these people.

Comparing ART against AT, both models provide similar role distance for Geaccone versus McCarty, but ART and AT show their differences elsewhere. For example, AT indicates a very strong role similarity between Geaconne and Hayslett (user 6), who was her boss (and CFO & Vice President in the Division); on the other hand, ART more correctly designates a low role similarity for this pair—in fact, ART assigns low similarity between Geaconne and all others in the matrix, which is appropriate because she is the only executive assistant in this small sample of Enron employees.

Another interesting pair of people is Blair (user 4) and Watson (user 14). ART predicts them to be role-similar, while the SNA and AT models do not. ART's prediction seems more appropriate since Blair worked on "gas pipeline logistics" and Watson worked on "pipeline facility planning", two very similar jobs.

McCarty, a Vice-President and CTO in the Division, also highlights differences between the models. The ART model puts him closest to Horton (user 5), who was President of the Division. AT predicts that he is closest to Rapp (user 12), who was

merely a lawyer that reviewed business agreements, and also close to Harris (user 15), who was only a mid-level manager.

Using ART in this way emphasizes role similarity, but not group membership. This can be seen by considering Thomas (user 3, an energy futures trader), and his relation to both Rapp (user 12, the lawyer mentioned above), and Lokey (user 16, a regulatory affairs manager). These three people work in related areas, and both ART and AT fittingly indicate a role similarity between them, (ART marginally more so than AT). On the other hand, traditional SNA results (Figure 2.4 left) emphasizes *group memberships* rather than role similarity by placing users 1 through 3 in a rather distinct blockstructure; they are the only three people in this matrix who were not members of the Enron Transwestern Division group, and these three exchanged more email with each other than with the people of the Transwestern Division.

Based on the above examples, and other similar examples, we posit that the ART model is more appropriate than SNA and AT in predicting role similarity. We thus would claim that the ART model yields more appropriate results than the SNA model in predicting role-equivalence between users, and somewhat better than the AT model in this capacity.

We also carried out this analysis with the personal email for McCallum to further validate the difference between the ART and SNA predictions. There are 825 users in this email corpus, while only 128 wrote ten or more emails. We perform the blockstructure analysis with these 128 users, shown in Figure 2.5. The blocks discovered are quite meaningful, e.g., the block from 0 to 30 are people in and related to McCallum's research group at UMass, and the block from 30 to 50 includes other researchers around the world.

Table 2.3 shows the closest pairs in terms of JS divergence, as calculated by the ART model and the SNA model. The difference in quality between the ART and SNA halves of the table is striking.

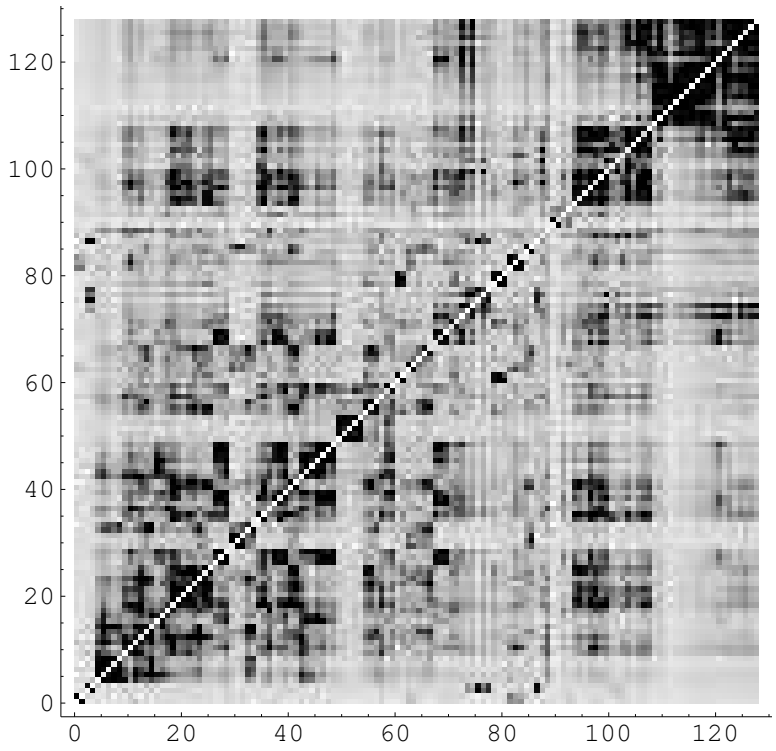


Figure 2.5. SNA Inverse JS Network for a 10 topic run on McCallum Email Data. Darker shades indicate higher similarity. Graph partitioning was calculated with the 128 authors that had ten or more emails in McCallum's Email Data. The block from 0 to 30 are people in and related to McCallum's research group at UMass. The block from 30 to 50 includes other researchers around the world.

Almost all the pairs predicted by the ART model look reasonable while many of those predicted by SNA are the opposite. For example, ART matches editor and reviews, two email addresses that send messages managing journal reviews. User mike and mikem are actually two different email addresses for the same person. Most other coreferent email addresses were pre-collapsed by hand during preprocessing; here ART has pointed out a mistaken omission, indicating the potential for ART to be used as a helpful component of an automated coreference system. Users aepshtey and smucker were students in a class taught by McCallum. Users coe, laurie and kate are all UMass CS Department administrative assistants; they rarely send email to each other, but they write about similar things. User ang is Andrew Ng from Stanford; joshuago is

Pairs considered most alike by ART	
User Pair	*Description*
editor reviews	Both journal review management
mike mikem	Same person! (manual coreference error)
aepshtey smucker	Both students in McCallum's class
coe laurie	Both UMass admin assistants
mcollins tom.mitchell	Both ML researchers on SRI project
mcollins gervasio	Both ML researchers on SRI project
davitz freeman	Both ML researchers on SRI project
mahadeva pal	Both ML researchers, discussing hiring
kate laurie	Both UMass admin assistants
ang joshuago	Both on organizing committee for a conference
Pairs considered most alike by SNA	
User Pair	*Description*
aepshtey rasmith	Both students in McCallum's class
donna editor	Spouse is unrelated to journal editor
donna krishna	Spouse is unrelated to conference organizer
donna ramshaw	Spouse is unrelated to researcher at BBN
donna reviews	Spouse is unrelated to journal editor
donna stromsten	Spouse is unrelated to visiting researcher
donna yugu	Spouse is unrelated to grad student
aepshtey smucker	Both students in McCallum's class
rasmith smucker	Both students in McCallum's class
editor elm	Journal editor and its Production Editor

Table 2.3. Pairs considered most alike by ART and SNA on McCallum email. All pairs produced by the ART model are accurately quite similar. This is not so for the top SNA pairs. Many users are considered similar by SNA merely because they appear in the corpus mostly sending email only to McCallum. However, this causes people with very different roles to be incorrectly declared similar—such as McCallum's spouse and the JMLR editor.

Joshua Goodman of Microsoft Research; they are both on the organizing committee of a new conference along with McCallum.

On the other hand, the pairs declared most similar by the SNA model are mostly extremely poor. Most of the pairs include donna, and indicate pairs of people who are similar only because in this corpus they appeared mostly sending email only to McCallum, and not others. User donna is McCallum's spouse. Other pairs are more sensible. For example, aepshtey, smucker and rasmith were all students in McCallum's

User Pair	*Description*
editor reviews	Both journal editors
jordan mccallum	Both ML researchers
mccallum vanessa	A grad student working in IR
croft mccallum	Both UMass faculty, working in IR
mccallum stromsten	Both ML researchers
koller mccallum	Both ML researchers
dkulp mccallum	Both UMass faculty
blei mccallum	Both ML researchers
mccallum pereira	Both ML researchers
davitz mccallum	Both working on an SRI project

Table 2.4. Pairs with the highest rank difference between ART and SNA on McCallum email. The traditional SNA metric indicates that these pairs of people are different, while ART indicates that they are similar. There are strong relations between all pairs.

class. User elm is Erik Learned-Miller who is correctly indicated as similar to editor since he was the Production Editor for the Journal of Machine Learning Research.

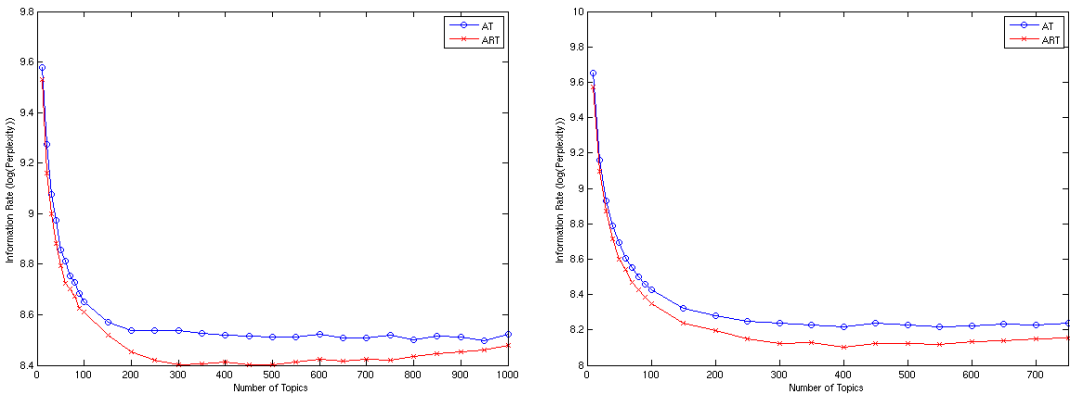
To highlight the difference between the SNA and ART predictions, we present Table 2.4, which was obtained by using both ART and SNA to rank the pairs of people by similarity, and then listing the pairs with the highest rank *differences* between the two models. These are pairs that SNA indicated were different, but ART indicated were similar. In every case, there are role similarities between the pairs.

2.2.3 Perplexity Comparison between AT and ART

Models for natural languages are often evaluated by perplexity as a measure of the goodness of fit of models. The lower perplexity a language model has, the better it predicts the unseen words given the words we previously saw.

The perplexity of a previously unseen message d consisting of words $\mathbf{w}_d$ can be defined as follows, when the author a_d and the recipient(s) $\mathbf{r}_d$ are given:

$$\text{Perplexity}(\mathbf{w}_d) = \exp\left(-\frac{\log(p(\mathbf{w}_d|a_d, \mathbf{r}_d))}{N_d}\right),$$



(a) Enron dataset (b) McCallum dataset

Figure 2.6. Perplexity comparison of AT and ART on two datasets. We plot the information rate (logarithm of perplexity) here. The difference between AT and ART is significant under one-tailed t-test (Enron dataset: p-value < 0.01 except for 10 topics with p-value $= 0.018$; McCallum dataset: p-value $< 1e-5$).

where ($\hat{\theta}$ and $\hat{\phi}$ defined in Equation 2.1)

$$p(\mathbf{w}_d|a_d, \mathbf{r}_d) = \prod_{i=1}^{N_d} \left(\frac{1}{|\mathbf{r}_d|} \sum_{r \in \mathbf{r}_d} \sum_{t=1}^{T} \hat{\theta}_{a_d r t} \hat{\psi}_{t w_{di}} \right).$$

We randomly split our datasets into a training set (9/10) and a test set (the remaining 1/10). In the test sets, 92.37% (Enron) and 84.51% (McCallum) of the author-recipient pairs also appear in the training sets. Ten Markov chains are run with different initializations, and the samples at the 2000th iteration are used to estimate $\hat{\theta}$ and $\hat{\phi}$ by Equation 2.1. We report the average information rate (logarithm of perplexity) with different number of topics on two datasets in Figure 2.6.

As clearly shown in the figure, ART has significantly better predictive power than AT over a large number of randomly selected test documents on both datasets under one-tailed t-test. Particularly on the Enron dataset, ART uses much fewer number of topics to achieve the best predictive performance. We can also find that the lowest perplexity obtained by ART is not achievable by AT with any parameter setting on both datasets. Both these results provide evidence that ART discovers meaningful

topics in the context of a social network and is indeed more appropriate to message data than AT.

Here we do not compare perplexity between ART and LDA, however AT (which ART dominates in perplexity) has already been shown to have better perplexity than LDA [49]. Due to the much simpler model structure, the author model [38] has much worse perplexity. Measured on both data sets, the information rates (log perplexity) are larger than 10, whereas ART's information rates are mostly between 8 and 9.

2.3 Role-Author-Recipient-Topic Models

To better explore the roles of authors, an additional level of latent variables can be introduced to explicitly model roles. Of particular interest is capturing the notion that a person can have multiple *roles* simultaneously—for example, a person can be both a professor and a mountain climber. Each role is associated with a set of topics, and these topics may overlap. For example, professors' topics may prominently feature research, meeting times, grant proposals, and friendly relations; climbers' topics may prominently feature mountains, climbing equipment, and also meeting times and friendly relations.

We incorporate into the ART model a new set of variables that take on values indicating role, and we term this augmented model the *Role-Author-Recipient-Topic* (RART) model. In RART, authors, roles and message-contents are modeled simultaneously. Each author has a multinomial distribution over roles. Authors and recipients are mapped to some role assignments, and a topic is selected based on these roles. Thus we have a clustering model, in which appearances of topics are the underlying data, and sets of correlated topics gather together clusters that indicate roles. Each sender-role and recipient-role pair has a multinomial distribution over topics, and each topic has a multinomial distribution over words.

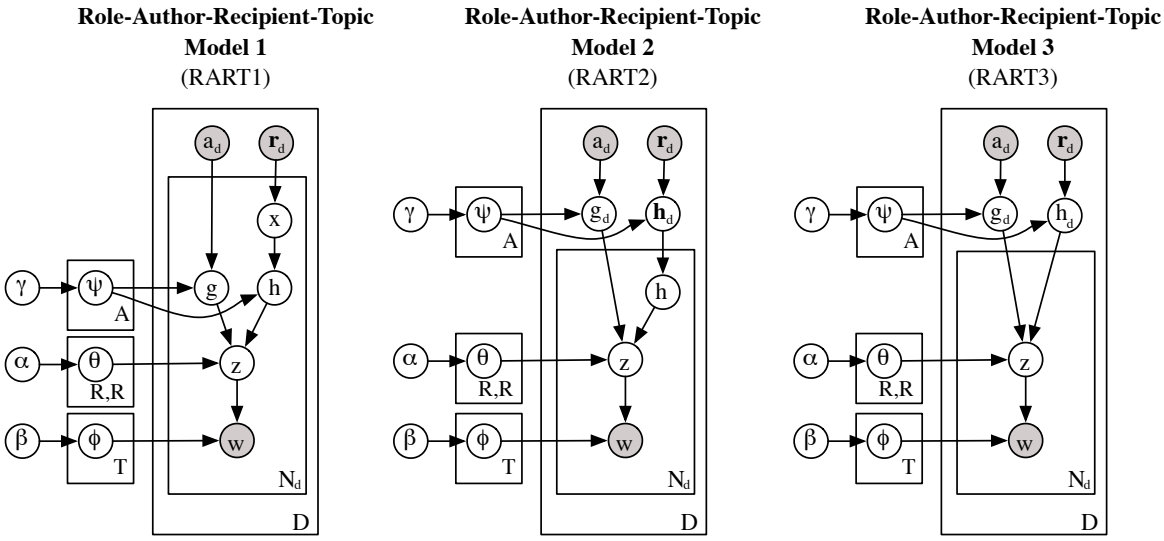


Figure 2.7. Three possible variants for the Role-Author-Recipient-Topic (RART) model.

As shown in Figure 2.7, different strategies can be employed to incorporate the "role" latent variables. First in RART1, role assignments can be made separately for each word in a document. This model represents that a person can change role during the course of the email message. In RART2, on the other hand, a person chooses one role for the duration of the message. Here each recipient of the message selects a role assignment, and then for each word, a recipient (with corresponding role) is selected on which to condition the selection of topic. In RART3, the recipients together result in the selection of a common, shared role, which is used to condition the selection of every word in the message. This last model may help capture the fact that a person's role may depend on the other recipients of the message, but also restricts all recipients to a single role.

We describe the generative process of RART1 in this chapter in detail, and leave the other two for exploration elsewhere. In its generative process for each message, an author, a_d, and a set of recipients, $\mathbf{r}_d$, are observed. To generate each word, a recipient, x, is chosen at uniform from $\mathbf{r}_d$, and then a role g for the author, and a

role h for the recipient x are chosen from two multinomial role distributions ψ_{a_d} and ψ_x, respectively. Next, a topic z is chosen from a multinomial topic distribution θ_{gh}, where the distribution is specific to the author-role recipient-role pair (g, h). Finally, the word w is generated by sampling from a topic-specific multinomial distribution ϕ_z.

In the RART1 model, given the hyperparameters α, β and γ, an author a_d, and a set of recipients $\mathbf{r}_d$ for each message d, the joint distribution of the topic mixture θ_{ij} for each author-role recipient-role pair (i, j), the role mixture ψ_k for each author k, the word mixture ϕ_t for each topic t, a set of recipients $\mathbf{x}$, a set of sender roles $\mathbf{g}$, a set of recipient roles $\mathbf{h}$, a set of topics $\mathbf{z}$ and a set of words $\mathbf{w}$ is given by (we define R as the number of roles):

$$
\begin{aligned}
& P(\Theta, \Phi, \Psi, \mathbf{x}, \mathbf{g}, \mathbf{h}, \mathbf{z}, \mathbf{w} | \alpha, \beta, \gamma, \mathbf{a}, \mathbf{r}) \\
= & \prod_{i=1}^{R} \prod_{j=1}^{R} p(\theta_{ij}|\alpha) \prod_{t=1}^{T} p(\phi_t|\beta) \prod_{k=1}^{A} p(\psi_k|\gamma) \prod_{d=1}^{D} \prod_{i=1}^{N_d} P(x_{di}|\mathbf{r}_d) P(g_{di}|a_d) P(h_{di}|x_{di}) P(z_{di}|\theta_{g_{di}h_{di}}) P(w_{di}|\phi_{z_{di}})
\end{aligned}
$$

Integrating over Ψ, Θ and Φ, and summing over $\mathbf{x}$, $\mathbf{g}$, $\mathbf{h}$ and $\mathbf{z}$, we get the marginal distribution of a corpus, similar to what we showed for ART.

To perform inference on RART models, the collapsed Gibbs sampling formulae can be derived in a similar way as in Appendix A, but in a more complex form.

Extensive experiments have been conducted with the RART1 model. Because we introduce two sets of additional latent variables (author role and recipient role), the sampling procedure at each iteration is significantly more complex. To make inference more efficient, we can instead perform it in two distinct parts. One strategy we have found useful is to first train an ART model, and use a sample to obtain topic assignments and recipient assignments for each word token. Then, in the next stage, we treat topics and recipients as observed (locked). Although such a strategy may not be recommended for arbitrary graphical models, we feel this is reasonable here because we find that a single sample from collapsed Gibbs sampling on the ART

Role 3 "IT Support at UMass CS"		Role 4 "Working on the SRI CALO Project"	
olc (lead Linux sysadmin)	0.2730	pereira (prof. at UPenn)	0.1876
gauthier (sysadmin for CIIR group)	0.1132	claire (UMass CS business manager)	0.1622
irsystem (mailing list CIIR sysadmins)	0.0916	israel (lead system integrator at SRI)	0.1140
system (mailing list for dept. sysadmins)	0.0584	moll (prof. at UMass)	0.0431
allan (prof., chair of computing committee)	0.0515	mgervasio (computer scientist at SRI)	0.0407
valerie (second Linux sysadmin)	0.0385	melinda.gervasio (same person as above)	0.0324
tech (mailing list for dept. hardware)	0.0360	majordomo (SRI CALO mailing list)	0.0210
steve (head of dept. of IT support)	0.0342	collin.evans (computer scientist at SRI)	0.0205

Table 2.5. An illustration of two roles from a 50-topic, 15-role run for the McCallum email dataset. Each role is shown with the most prominent users (their short descriptions in parenthesis) and the corresponding conditional probabilities. The quoted titles are our own summary for the roles. For example, in Role 3, the users are all employees (or mailing lists) of the IT support staff at UMass CS, except for *allan*, who, however, was the professor chairing the department's computing committee.

model yields good assignments. The following results are based on a 15-role, 50-topic run of RART1 on McCallum email dataset.

Our results show that the RART model does indeed automatically discover meaningful person-role information by its explicit inclusion of a role variable. We show the most prominent users in two roles in Table 2.5. For instance, the users most prominent in Role 3 are all employees (or mailing lists) of the IT support staff at UMass CS, except for *allan*, who, however, was the professor chairing the department's computing committee. Role 4 seems to represent "working on the SRI CALO project." Most of its top prominent members are researchers working on CALO project, many of them at SRI. The sender *majordomo* sends messages from an SRI CALO mailing list. Users *claire* and *moll* were, however, unrelated with the project, and we do not know the reason they appear in this role. The users *mgervasio* and *melinda.gervasio* are actually the same person; satisfyingly RART found that they have very similar role distributions.

allan (James Allan)		pereira (Fernando Pereira)	
Role 10 (grant issues)	0.4538	Role 2 (natural language researcher)	0.5749
Role 13 (UMass CIIR group)	0.2813	Role 4 (working on SRI CALO Project)	0.1519
Role 2 (natural language researcher)	0.0768	Role 6 (proposal writing)	0.0649
Role 3 (IT Support at UMass CS)	0.0326	Role 10 (grant issues)	0.0444
Role 4 (working on SRI CALO Project)	0.0306	Role 8 (guests at McCallum's house)	0.0408

Table 2.6. An illustration of the role distribution of two users from a 50-topic, 15-role run for the McCallum email dataset. Each user is shown with his most prominent roles (their short descriptions in parenthesis) and the corresponding conditional probabilities. For example, considering user *pereira* (Fernando Pereira), his top five role assignments are all appropriate, as viewed through McCallum's email.

2.4 Experimental Results with RART

One objective of the RART model is to capture the multiple roles that a person has. The role distribution of two users are shown in Table 2.6. For example, user *allan* (James Allan) mentioned above has a role in "IT support," but also has a role as a "member of the Center for Intelligent Information Retrieval," as a "grant proposal writer," and as a "natural language researcher." Although not a member of the "SRI CALO Project," *allan*'s research is related to CALO, and perhaps this is the reason that CALO appears (weakly) among his roles. Consider also user *pereira* (Fernando Pereira); his top five role assignments are all exactly appropriate, as viewed through McCallum's email.

As expected, one can observe interesting differences in the sender versus recipient topic distributions associated with each role. For instance, in Role 4 "SRI CALO," the top three topics for a sender role are Topic 27 "CALO information," Topic 11 "mail accounts," and Topic 36 "program meetings," but for its recipient roles, most prominent are Topic 48 "task assignments," Topic 46 "a CALO-related research paper," and Topic 40 "java code".

2.5 Summary

We have presented the Author-Recipient-Topic model, a Bayesian network for social network analysis that discovers discussion topics conditioned on the sender-recipient relationships in a corpus of messages. To the best of our knowledge, this model combines for the first time the directionalized connectivity graph from social network analysis with the clustering of words to form topics from probabilistic language modeling.

The model can be applied to discovering topics conditioned on message sending relationships, clustering to find social roles, and summarizing and analyzing large bodies of message data. The model would form a useful component in systems for routing requests, expert-finding, message recommendation and prioritization, and understanding the interactions in an organization in order to make recommendations about improving organizational efficiency.

CHAPTER 3

JOINT GROUP AND TOPIC DISCOVERY FROM RELATIONS AND TEXT

Research in the field of social network analysis (SNA) has led to the development of mathematical models that discover patterns in interaction between entities [65]. Besides role discovery we discussed in the previous chapter, another of the objectives of SNA is to detect salient groups of entities. Group discovery has many applications, such as understanding the social structure of organizations [11] or native tribes [70], uncovering criminal organizations [52], and modeling large-scale social networks in Internet services such as Friendster.com or LinkedIn.com.

Social scientists have conducted extensive research on group detection, especially in fields such as anthropology [70] and political science [14, 24]. Recently, statisticians and computer scientists have begun to develop models that specifically discover group memberships [6, 28, 30, 46]. One such model is the stochastic blockstructures model [46], which discovers the latent structure, groups or classes based on pair-wise relation data. A particular relation holds between a pair of entities (people, countries, organizations, etc.) with some probability that depends only on the class (group) assignments of the entities. The relations between all the entities can be represented with a directed or undirected graph. The class assignments can be inferred from a graph of observed relations or link data using Gibbs sampling [46]. This model is extended in [28] to automatically select an arbitrary number of groups by using a Chinese Restaurant Process prior.

The aforementioned models discover latent groups only by examining whether one or more relations exist between a pair of entities. The Group-Topic (GT) model pre-

sented in this chapter [63], on the other hand, considers not only the relations between objects but also the attributes of the relations (for example, the text associated with the relations) when assigning group membership.

The GT model can be viewed as an extension of the stochastic blockstructures model [28, 46] with the key addition that group membership is conditioned on a latent variable associated with the attributes of the relation. In our experiments, the attributes of relations are words, and the latent variable represents the topic responsible for generating those words. Unlike previous methods, our model captures the *(language) attributes* associated with interactions between entities, and uses distinctions based on these attributes to better assign group memberships.

Consider a legislative body and imagine its members forging alliances (forming groups), and voting accordingly. However, different alliances arise depending on the topic of the resolution up for a vote. For example, one grouping of the legislators may arise on the issue of taxation, while a quite different grouping may occur for votes on foreign trade. Similar patterns of topic-based affiliations would arise in other types of entities as well, e.g., research paper co-authorship relations between people and citation relations between papers, with words as attributes on these relations.

In the GT model, the discovery of groups is guided by the emerging topics, and the discovery of topics is guided by emerging groups. Both modalities are driven by the common goal of increasing data likelihood. Consider the voting example again; resolutions that would have been assigned the same topic in a model using words alone may be assigned to different topics if they exhibit distinct voting patterns. Distinct word-based topics may be merged if the entities vote very similarly on them. Likewise, multiple different divisions of entities into groups are made possible by conditioning them on the topics.

The GT model simultaneously clusters entities to groups and clusters words into topics, unlike models that generate topics solely based on word distributions such as

Latent Dirichlet Allocation [8]. In this way the GT model discovers salient topics relevant to relationships between entities in the social network—topics which the models that only examine words are unable to detect. Erosheva et al. [15] provide a general formulation for mixed membership, of which LDA is a special case, and they apply it to soft clustering of papers by topics using words from the text and references. In work parallel to ours but different from GT, Airoldi et al. [2] extend the general mixed membership model to also incorporate stochastic blockstructures models of the form arising in the network literature. Their application is to protein-protein interactions.

Exploring the notion that entities in the same group have the same behavior and that the behavior of an entity can be explained by its (hidden) group membership, Jakulin and Buntine [26] develop a discrete PCA model for discovering groups in the 108 US Senate. In the model each entity can belong to each of the k groups with a certain probability, and each group has its own specific pattern of behaviors. Therefore, an entity's behavior depends on the probability of belonging to a group and the probability that the group has that behavior. They apply this model to voting data in the 108th US Senate where the behavior of an entity is its vote on a resolution. A similar model is developed in [47] that examines group cohesion and voting similarity in the Finnish Parliament. We apply our GT model also to voting data. However, unlike [26, 47], our model considers the relation between a pair of voting entities and does not try to predict the *actual* vote of an entity on a resolution. Since our goal is to cluster entities based on the similarity of their voting patterns, we are only interested in whether a pair of entities voted the same or differently, not their actual yes/no votes. The complete negation of a resolution might share the same topic (e.g., increasing vs. decreasing budget) with the original resolution, and not surprisingly, the actual votes on them would be opposite; however, pairs of entities would tend to vote same on both resolutions. To capture this, we model

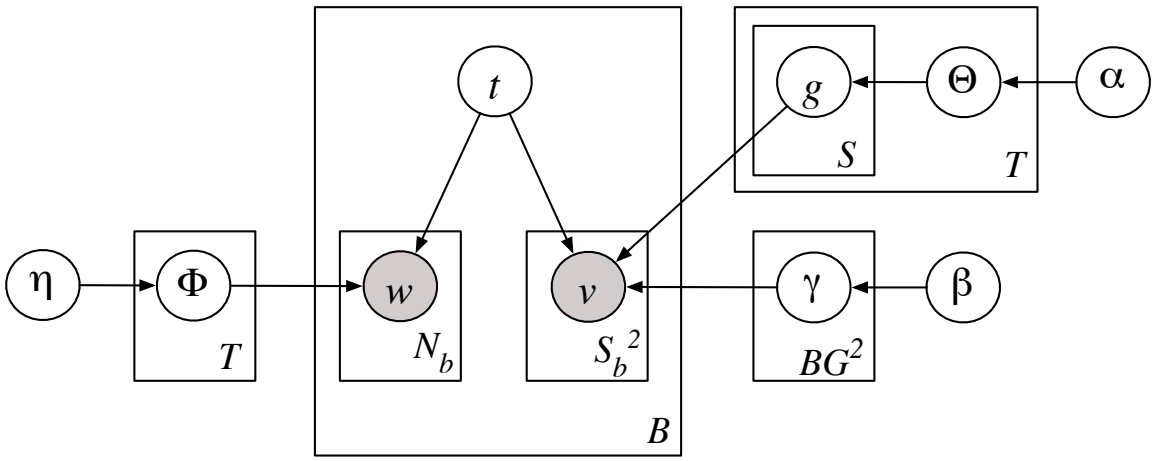


Figure 3.1. The Group-Topic model

relations as *agreement* between entities, not the yes/no vote itself. This kind of "content-ignorant" feature is similarly found in some work on web log clustering [5].

We demonstrate the capabilities of the GT model by applying it to two large sets of voting data: one from US Senate and the other from the General Assembly of the UN. The model clusters voting entities into coalitions and simultaneously discovers topics for word attributes describing the relations (bills or resolutions) between entities. We find that the groups obtained from the GT model are significantly more cohesive (p-value $< .01$) than those obtained from the blockstructures model. The GT model also discovers new and more salient topics in both the Senate and UN datasets—in comparison with topics discovered by only examining the words of the resolutions, the GT topics are either split or joined together as influenced by the voters' patterns of behavior.

3.1 Group-Topic Model

The Group-Topic Model is a directed graphical model that clusters entities with relations between them, as well as attributes of those relations. The relations may be

SYMBOL	DESCRIPTION
g_{it}	entity i's group assignment in topic t
t_b	topic of an event b
$w_k^{(b)}$	the kth token in the event b
$V_{ij}^{(b)}$	entity i and j's groups behaved same (1) or differently (2) on the event b
S	number of entities
G	number of groups
B	number of events
N_b	number of word tokens in the event b
S_b	number of entities who participated in the event b

Table 3.1. Additional notation (to Table 1.1) used in this chapter

either directed or undirected and have multiple attributes. In this chapter, we focus on undirected relations and have words as the attributes on relations.

In the generative process for each event (an interaction between entities), the model first picks the topic t of the event and then generates all the words describing the event where each word is generated independently according to a multinomial (discrete) distribution ϕ_t, specific to the topic t. To generate the relational structure of the network, first the group assignment, g_{st} for each entity s is chosen conditionally from a particular multinomial (discrete) distribution θ_t over groups for each topic t. Given the group assignments on an event b, the matrix $V^{(b)}$ is generated where each cell $V_{ij}^{(b)}$ represents if the groups of two entities (i and j) behaved the same or not during the event b, (e.g., voted the same or not on a bill). Each element of V is sampled from a binomial (Bernoulli) distribution $\gamma_{g_i g_j}^{(b)}$. In addition to the notation in Table 1.1, the TG specific notation is summarized in Table 3.1, and the graphical model representation of the model is shown in Figure 3.1.

Without considering the topic of an event, or by treating all events in a corpus as reflecting a single topic, the simplified model (only the right part of Figure 3.1) becomes equivalent to the stochastic blockstructures model [46]. To match the block-

structures model, each event defines a relationship, *e.g.*, whether in the event two entities' groups behave the same or not. On the other hand, in our model a relation may have multiple attributes (which in our experiments are the words describing the event, generated by a per-topic multinomial (discrete) distribution).

When we consider the complete model, the dataset is dynamically divided into T sub-blocks each of which corresponds to a topic. The complete GT model is as follows:

$$\begin{aligned}
t_b &\sim \text{Uniform}(\frac{1}{T}) \\
w_{it}|\phi_t &\sim \text{Multinomial}(\phi_t) \\
\phi_t|\eta &\sim \text{Dirichlet}(\eta) \\
g_{it}|\theta_t &\sim \text{Multinomial}(\theta_t) \\
\theta_t|\alpha &\sim \text{Dirichlet}(\alpha) \\
V_{ij}^{(b)}|\gamma_{g_i g_j}^{(b)} &\sim \text{Binomial}(\gamma_{g_i g_j}^{(b)}) \\
\gamma_{gh}^{(b)}|\beta &\sim \text{Beta}(\beta).
\end{aligned}$$

We want to perform joint inference on (text) attributes and relations to obtain topic-wise group memberships. Since inference can not be done exactly on such complicated probabilistic graphical models, we employ collapsed Gibbs sampling to conduct inference. Note that we adopt conjugate priors in our setting, and thus we can easily integrate out θ, ϕ and γ to decrease the uncertainty associated with them. This simplifies the sampling since we do not need to sample θ, ϕ and γ at all, unlike in [46]. In our case we need to compute the conditional distribution $P(g_{st}|\mathbf{w}, \mathbf{V}, \mathbf{g}_{-st}, \mathbf{t}, \alpha, \beta, \eta)$ and $P(t_b|\mathbf{w}, \mathbf{V}, \mathbf{g}, \mathbf{t}_{-b}, \alpha, \beta, \eta)$, where $\mathbf{g}_{-st}$ denotes the group assignments for all entities except entity s in topic t, and $\mathbf{t}_{-b}$ represents the topic assignments for all events except event b. Beginning with the joint probability of a dataset, and using the chain rule, we can obtain the conditional probabilities conveniently. In our setting, the

relationship we are investigating is always symmetric, so we do not distinguish R_{ij} and R_{ji} in our derivations (only $R_{ij}(i \leq j)$ remain). Thus

$$
\begin{aligned}
& P(g_{st}|\mathbf{V}, \mathbf{g}_{-st}, \mathbf{w}, \mathbf{t}, \alpha, \beta, \eta) \\
\propto \; & \frac{\alpha_{g_{st}} + n_{tg_{st}} - 1}{\sum_{g=1}^{G}(\alpha_g + n_{tg}) - 1} \prod_{b=1}^{B} \left(I(t_b = t) \prod_{h=1}^{G} \frac{\prod_{k=1}^{2} \prod_{x=1}^{d_{g_{st}hk}^{(b)}} \left(\beta_k + m_{g_{st}hk}^{(b)} - x\right)}{\prod_{x=1}^{\sum_{k=1}^{2} d_{g_{st}hk}^{(b)}} \left((\sum_{k=1}^{2}(\beta_k + m_{g_{st}hk}^{(b)}) - x \right)} \right),
\end{aligned}
$$

where n_{tg} represents how many entities are assigned into group g in topic t, c_{tv} represents how many tokens of word v are assigned to topic t, $m_{ghk}^{(b)}$ represents how many times group g and h vote same ($k = 1$) and differently ($k = 2$) on event b, $I(t_b = t)$ is an indicator function, and $d_{g_{st}hk}^{(b)}$ is the increase in $m_{g_{st}hk}^{(b)}$ if entity s were assigned to group g_{st} than without considering s at all (if $I(t_b = t) = 0$, we ignore the increase in event b). Furthermore,

$$
\begin{aligned}
& P(t_b|\mathbf{V}, \mathbf{g}, \mathbf{w}, \mathbf{t}_{-b}, \alpha, \beta, \eta) \\
\propto \; & \left(\frac{\prod_{v=1}^{V} \prod_{x=1}^{e_v^{(b)}} (\eta_v + c_{t_b v} - x)}{\prod_{x=1}^{\sum_{v=1}^{V} e_v^{(b)}} \left(\sum_{v=1}^{V} (\eta_v + c_{t_b v}) - x \right)} \right)^{\lambda} \prod_{g=1}^{G} \prod_{h=g}^{G} \frac{\prod_{k=1}^{2} \Gamma(\beta_k + m_{ghk}^{(b)})}{\Gamma(\sum_{k=1}^{2}(\beta_k + m_{ghk}^{(b)}))},
\end{aligned}
$$

where $e_v^{(b)}$ is the number of tokens of word v in event b. Note that $m_{ghk}^{(b)}$ is not a constant and changes with the assignment of t_b since it influences the group assignments of all entities that vote on event b. We use a weighting parameter λ to rescale the likelihoods from different modalities, as is also common in speech recognition when the acoustic and language models are combined. The GT model uses information from two different modalities. In general, the likelihood of the two modalities is not directly comparable, since the number of occurrences of each type may vary greatly (e.g., there may be far more pairs of voting entities than word occurrences).

3.2 Experimental Results

We present experiments applying the GT model to the voting records of members of two legislative bodies: the US Senate and the UN General Assembly. To make sure of convergence, we run the Markov chains for 10,000 iterations, (which by inspection are stable after a few hundred iterations), and use the topic and group assignments in the last Gibbs sample. According to some analysis similar to Figure 2.2 (but we use the Agreement Index defined below instead of perplexity), the GT model is not sensitive to the value of hyperparameters[1]. For simplicity, we set $\alpha = 1$, $\beta = 5$, and $\eta = 1$ in all experiments.

For comparison, we present the results of a baseline method that first uses a mixture of unigrams to discover topics and associate a topic with each resolution, and then runs the blockstructures model [46] separately on the resolutions assigned to each topic. This baseline approach is similar to the GT model in that it discovers both groups and topics, and has different group assignments on different topics. However, whereas the GT model performs joint inference simultaneously, the baseline performs inference serially. Note that our baseline is still more powerful than the blockstructures models, since it models the topic associated with each event, and allows the creation of distinct groupings dependent on different topics.

In this chapter, we are interested in the quality of both the groups and the topics. In the political science literature, group cohesion is quantified by the *Agreement Index (AI)* [26, 47], which measures the similarity of votes cast by members of a group during a particular roll call. The AI for a particular group on a given roll call i is based on the number of group members that vote Yea(y_i), Nay(n_i) or Abstain(a_i) in the roll call i. Higher AI index means better cohesion.

[1] λ is relatively more sensitive, and we set it by trial and error for each dataset

Datasets	Avg. AI for GT	Avg. AI for Baseline	p-value	Blockstructures
Senate	0.8294	0.8198	$< .01$	0.7850
UN	0.8664	0.8548	$< .01$	0.7934

Table 3.2. Average AI for different models for both Senate and UN datasets. The group cohesion in (joint) GT is significantly better than in (serial) baseline, as well as the blockstructures model that does not use text at all.

$$AI_i = \frac{\max\{y_i, n_i, a_i\} - \frac{y_i+n_i+a_i-\max\{y_i,n_i,a_i\}}{2}}{y_i + n_i + a_i}$$

The blockstructures model assumes that members of a legislative body have the same group affiliations irrespective of the topic of the resolution on vote. However, it is likely that members form their groups based on the topic of the resolution being voted on. We quantify the extent to which a member s switches groups with a *Group Switch Index* (GSI).

$$GSI_s = \sum_{i,j}^{T} \frac{\text{abs}(\vec{s_i} - \vec{s_j})}{|G(s,i)| - 1 + |G(s,j)| - 1}$$

where $\vec{s_i}$ and $\vec{s_j}$ are bit vectors of the length of the size of the legislative body. The k_{th} bit of $\vec{s_i}$ is set if k is in the same group as s on topic i and similarly $\vec{s_j}$ corresponds to topic j. $G(s,i)$ is the group of s on topic i which has a size of $|G(s,i)|$ and $G(s,j)$ is the group of s on topic j. We present entities that frequently change their group alliance according to the topics of resolutions.

Group cohesion from the GT model is found to be significantly greater than the baseline group cohesion under a pairwise t-test, as shown in Table 3.2, which indicates that the GT's joint inference is better able to discover cohesive groups. We find that nearly every document has a higher Agreement Index across groups using the GT model as compared to the baseline. As expected, stochastic blockstructures without text [46] is even worse than our baseline.

Economic	Education	Military Misc.	Energy
federal	education	government	energy
labor	school	military	power
insurance	aid	foreign	water
aid	children	tax	nuclear
tax	drug	congress	gas
business	students	aid	petrol
employee	elementary	law	research
care	prevention	policy	pollution

Table 3.3. Top words for topics generated with the mixture of unigrams model on the Senate dataset. The headers are our own summary of the topics.

3.2.1 The US Senate Dataset

Our Senate dataset consists of the voting records of Senators in the 101st-109th US Senate (1989-2005) obtained from the Library of Congress THOMAS database. During a roll call for a particular bill, a Senator may respond *Yea* or *Nay* to the question that has been put to vote, else the vote will be recorded as *Not Voting*. We do not consider *Not Voting* as a unique vote since most of the time it is a result of a Senator being absent from the session of the US Senate. The text associated with each resolution is composed of its index terms provided in the database. There are 3423 resolutions in our experiments (we excluded roll calls that were not associated with resolutions). Each bill may come up for vote many times in the U.S. Senate, each time with an attached amendment, and thus many relations may have the same attributes (index terms). Since there are far fewer words than pairs of votes, we adjust the text likelihood to the 5th power (weighting factor 5) in the experiments with this dataset so as to balance its influence during inference.

We cluster the data into 4 topics and 4 groups (cluster sizes are suggested by a political science professor) and compare the results of GT with the baseline. The most likely words for each topic from the traditional mixture of unigrams model is shown in Table 3.3, whereas the topics obtained using GT are shown in Table 3.4.

Economic	Education + Domestic	Foreign	Social Security + Medicare
labor	education	foreign	social
insurance	school	trade	security
tax	federal	chemicals	insurance
congress	aid	tariff	medical
income	government	congress	care
minimum	tax	drugs	medicare
wage	energy	communicable	disability
business	research	diseases	assistance

Table 3.4. Top words for topics generated with the GT model on the Senate dataset. The topics are influenced by both the words and votes on the bills.

The GT model collapses the topics Education and Energy together into Education and Domestic, since the voting patterns on those topics are quite similar. The new topic Social Security + Medicare did not have strong enough word coherence to appear in the baseline model, but it has a very distinct voting pattern, and thus is clearly found by the GT model. Thus GT discovers topics that are salient in that they correlate with people's behavior and relations, not simply word co-occurrences.

Examining the group distribution across topics in the GT model, we find that on the topic Economic the Republicans form a single group whereas the Democrats split into 3 groups indicating that Democrats have been somewhat divided on this topic. With regard to Education + Domestic and Social Security + Medicare, Democrats are more unified whereas the Republicans split into 3 groups. The group membership of Senators on Education + Domestic issues is shown in Table 3.5. We see that the first group of Republicans include a Democratic Senator from Texas, a state that usually votes Republican. Group 2 (majority Democrats) includes Sen. Chafee who is known to be pro-environment and is involved in initiatives to improve education, as well as Sen. Jeffords who left the Republican Party to become an Independent and has championed legislation to strengthen education and environmental protection.

Group 1	Group 3	Group 4
73 Republicans	Cohen(R-ME)	Armstrong(R-CO)
Krueger(D-TX)	Danforth(R-MO)	Garn(R-UT)
Group 2	Durenberger(R-MN)	Humphrey(R-NH)
90 Democrats	Hatfield(R-OR)	McCain(R-AZ)
Chafee(R-RI)	Heinz(R-PA)	McClure(R-ID)
Jeffords(I-VT)	Kassebaum(R-KS)	Roth(R-DE)
	Packwood(R-OR)	Symms(R-ID)
	Specter(R-PA)	Wallop(R-WY)
	Snowe(R-ME)	Brown(R-CO)
	Collins(R-ME)	DeWine(R-OH)
		Thompson(R-TN)
		Fitzgerald(R-IL)
		Voinovich(R-OH)
		Miller(D-GA)
		Coleman(R-MN)

Table 3.5. Senators in the four groups corresponding to Topic Education + Domestic in Table 3.4.

Nearly all the Senators in Group 4 (in Table 3.5) are advocates for education and many of them have been awarded for their efforts (e.g., Sen. Fitzgerald has been honored by the NACCP for his active role in Early Care and Education, and Sen. McCain has been added to the ASEE list as a *True Hero* in American Education). Sen. Armstrong was a member of the Education committee; Sen. Voinovich and Sen. Symms are strong supporters of early education and vocational education, respectively; and Sen. Roth has constantly voted for tax deductions for education. It is also interesting to see that Sen. Miller (D-GA) appears in a Republican group; although he is in favor of educational reforms, he is a conservative Democrat and frequently criticizes his own party—even backing Republican George W. Bush over Democrat John Kerry in the 2004 Presidential election.

Many of the Senators in Group 3 have also focused on education and other domestic issues such as energy, however, they often have a more liberal stance than those in Group 4, and come from states that are historically less conservative. Senators

Senator	Group Switch Index
Shelby(D-AL)	0.6182
Heflin(D-AL)	0.6049
Voinovich(R-OH)	0.6012
Johnston(D-LA)	0.5878
Armstrong(R-CO)	0.5747

Table 3.6. Senators that switch groups the most across topics for the 101st-109th Senates

Hatfield, Heinz, Snowe, Collins, Cohen and others have constantly promoted pro-environment energy options with a focus on renewable energy, while Sen. Danforth has presented bills for a more fair distribution of energy resources. Sen. Kassebaum is known to be uncomfortable with many Republican views on domestic issues such as education, and has voted against voluntary prayer in school. Thus, both Groups 3 and 4 differ from the Republican core (Group 2) on domestic issues, and also differ from each other.

The Senators that switch groups the most across topics in the GT model are shown in Table 3.6 based on their GSIs. Sen. Shelby(D-AL) votes with the Republicans on Economic, with the Democrats on Education + Domestic and with a small group of maverick Republicans on Foreign and Social Security + Medicare. Both Sen. Shelby and Sen. Heflin are Democrats from a fairly conservative state (Alabama) and are found to side with the Republicans on many issues.

3.2.2 The United Nations Dataset

The second dataset involves the voting record of the UN General Assembly [58]. We focus first on the resolutions discussed from 1990-2003, which contain votes of 192 countries on 931 resolutions. If a country is present during the roll call, it may choose to vote *Yes*, *No* or *Abstain*. Unlike the Senate dataset, a country's vote can have one of three possible values instead of two. Because we parameterize agreement

Everything Nuclear	Human Rights	Security in Middle East
nuclear	rights	occupied
weapons	human	israel
use	palestine	syria
implementation	situation	security
countries	israel	calls

Table 3.7. Top words for topics generated from mixture of unigrams model with the UN dataset (1990-2003). Only text information is utilized to form the topics, as opposed to Table 3.8 where our GT model takes advantage of both text and voting information.

and not the votes themselves, this 3-value setting does not require any change to our model. In experiments with this dataset, we use a weighting factor 500 for text (adjusting the likelihood of text by a power of 500 so as to make it comparable with the likelihood of pairs of votes for each resolution). We cluster this dataset into 3 topics and 5 groups (again, numbers are suggested by a political science professor).

The most probable words in each topic from the mixture of unigrams model is shown in Table 3.7. For example, Everything Nuclear constitutes all resolutions that have anything to do with the use of nuclear technology, including nuclear weapons. Comparing these with topics generated from the GT model shown in Table 3.8, we see that the GT model splits the discussion about nuclear technology into two separate topics, Nuclear Arsenal which is generally about countries obtaining nuclear weapons and management of nuclear waste, and Nuclear Arms Race which focuses on the arms race between Russia and the US and preventing a nuclear arms race in outer space. These two issues had drastically different voting patterns in the U.N., as can be seen in the contrasting group structure for those topics in Table 3.8. The countries in Table

GROUP ↓	Nuclear Arsenal	Human Rights	Nuclear Arms Race
	nuclear	rights	nuclear
	states	human	arms
	united	palestine	prevention
	weapons	occupied	race
	nations	israel	space
1	Brazil	Brazil	UK
	Columbia	Mexico	France
	Chile	Columbia	Spain
	Peru	Chile	Monaco
	Venezuela	Peru	East-Timor
2	USA	Nicaragua	India
	Japan	Papua	Russia
	Germany	Rwanda	Micronesia
	UK...	Swaziland	
	Russia	Fiji	
3	China	USA	Japan
	India	Japan	Germany
	Mexico	Germany	Italy...
	Iran	UK...	Poland
	Pakistan	Russia	Hungary
4	Kazakhstan	China	China
	Belarus	India	Brazil
	Yugoslavia	Indonesia	Mexico
	Azerbaijan	Thailand	Indonesia
	Cyprus	Philippines	Iran
5	Thailand	Belarus	USA
	Philippines	Turkmenistan	Israel
	Malaysia	Azerbaijan	Palau
	Nigeria	Uruguay	
	Tunisia	Kyrgyzstan	

Table 3.8. Top words for topics generated from the GT model with the UN dataset (1990-2003) as well as the corresponding groups for each topic (column). The countries listed for each group are ordered by their 2005 GDP (PPP) and only the top 5 countries are shown in groups that have more than 5 members.

3.8 are ranked by their GDP in 2005.[2] Thus, again the GT model is able to discover

[2]http://en.wikipedia.org/wiki/List_of_countries_by_GDP_%28PPP%29. In Table 3.8, we omit some countries (represented by ...) in order to incorporate other interesting but relatively low ranked countries (for example, Russia) in the GDP list.

salient topics—topics that reflect the voting patterns and coalitions, not simply word co-occurrence alone.

As seen in Table 3.8, groups formed in Nuclear Arms Race are unlike the groups formed in the remaining topics. These groups map well to the global political situation of that time when, despite the end of the Cold War, there was mutual distrust between Russia and the US with regard to the continued manufacture of nuclear weapons. For missions to outer space and nuclear arms, India was a staunch ally of Russia, while Israel was an ally of the US.

3.2.2.1 Overlapping Time Intervals

In order to understand changes and trends in topics and groups over time, we run the GT model on resolutions that were discussed during overlapping time windows of 15 years, from 1960-2000, each shifted by a period of 5 years. We consider 3823 unique resolutions in this way. The topics as well as the group distribution for the most dominant topic during each time period are shown in Table 3.9.

Over the years there is a shift in the topics discussed in the UN, which corresponds well to the events and issues in history. During 1960-1975 the resolutions focused on countries having the right to self-determination, especially countries in Africa which started to gain their freedom during this time. Although this topic continued to be discussed in the subsequent time period, the focus of the resolutions shifted to the role of the UN in controlling nuclear weapons as the Cold War conflict gained momentum in the late 70s. While there were few resolutions condemning the racist regime in South Africa between 1965-1980, this was the topic of many resolutions during 1970-1985—culminating in the UN censure of South Africa for its discriminatory practices.

Other topics discussed during the 70s and early 80s were Israel's occupation of neighboring countries and nuclear issues. The reduction of arms was primarily discussed during 1975-1990, the time period during which the US and Soviet Union had

Time Period	Topic 1	Topic 2	Topic 3	Group distributions for Topic 3				
				Group 1	Group2	Group3	Group4	Group5
60-75	Nuclear	Procedure	Africa Indep.	India	USA	Argentina	USSR	Turkey
	operative	committee	calling	Indonesia	Japan	Colombia	Poland	
	general	amendment	right	Iran	UK	Chile	Hungary	
	nuclear	assembly	africa	Thailand	France	Venezuela	Bulgaria	
	power	deciding	self	Philippines	Italy	Dominican	Belarus	
65-80	Independence	Finance	Weapons	Cuba	India	Algeria	USSR	USA
	territories	budget	nuclear	Albania	Indonesia	Iraq	Poland	Japan
	independence	appropriation	UN		Pakistan	Syria	Hungary	UK
	self	contribution	international		Saudi	Libya	Bulgaria	France
	colonial	income	weapons		Egypt	Afghanistan	Belarus	Italy
70-85	N. Weapons	Israel	Rights	Mexico	China	USA	Brazil	India
	nuclear	israel	africa	Indonesia		Japan	Turkey	USSR
	international	measures	territories	Iran		UK	Argentina	Poland
	UN	hebron	south	Thailand		France	Colombia	Vietnam
	human	expelling	right	Philippines		Italy	Chile	Hungary
75-90	Rights	Israel/Pal.	Disarmament	Mexico	USA	Algeria	China	India
	south	israel	UN	Indonesia	Japan	Vietnam	Brazil	
	africa	arab	international	Iran	UK	Iraq	Argentina	
	israel	occupied	nuclear	Thailand	France	Syria	Colombia	
	rights	palestine	disarmament	Philippines	USSR	Libya	Chile	
80-95	Disarmament	Conflict	Pal. Rights	USA	China	Japan	Guatemala	Malawi
	nuclear	need	rights	Israel	India	UK	St Vincent	
	US	israel	palestine		Russia	France	Dominican	
	disarmament	palestine	israel		Spain	Italy		
	international	secretary	occupied		Hungary	Canada		
85-00	Weapons	Rights	Israel/Pal.	Poland	China	USA	Russia	Cameroon
	nuclear	rights	israeli	Czech R.	India	Japan	Argentina	Congo
	weapons	human	palestine	Hungary	Brazil	UK	Ukraine	Ivory C.
	use	fundamental	occupied	Bulgaria	Mexico	France	Belarus	Liberia
	international	freedoms	disarmament	Albania	Indonesia	Italy	Malta	

Table 3.9. Results for 15-year-span slices of the UN dataset (1960-2000). The top probable words are listed for all topics, but only the groups corresponding the most dominant topic are shown (Topic 3). We list the countries for each group ordered by their 2005 GDP (PPP)and only show the top 5 countries in groups that have more than 5 members. We do not repeat the results in Table 3.8 for the most recent window (1990-2003).

talks about disarmament. During 1980-1995 the central topic of discussion was the Israeli-Palestinian conflict; this time period includes the beginning of the *Intifada* revolt in Palestine and the Gulf War. This topic continued to be important in the next time period (1985-2000), but in the most recent slice (1990-2003, Table 3.8) it has become a part of a broader topic on human rights by combining other human rights related resolutions that appear as a separate topic during 1985-2000. The human rights issue continues to be the primary topic of discussion during 1990-2003.

Throughout the history of the UN, the US is usually in the same group as Europe and Japan. However, as we can see in Table 3.9 during 1985-2000, when the Israeli-Palestinian conflict was the most dominant topic, US and Israel form a group of their own separating themselves from Europe. In other topics discussed during 1985-2000, US and Israel are found to be in the same group as Europe and Japan.

Another interesting result of considering the groups formed over the years is that, except for the last time period (1990-2003), countries in eastern Europe such as Poland, Hungary, Bulgaria, etc., form a group along with USSR (Russia). However, in the last time window on most topics they become a part of the group that consists of the western Europe, Japan and the US. This shift corresponds to the end of the communist regimes in these countries that were supported by the Soviet Union. It is also worth mentioning that before 1990, our model assigned East Germany to the same group as other eastern European countries and USSR (Russia), while it assigned West Germany to the same group as western European countries.[3]

3.3 Summary

We present the Group-Topic model that jointly discovers latent groups in a network as well as clusters of attributes (or topics) of events that influence the interaction between entities in the network. The model extends prior work on latent group discovery by capturing not only pair-wise relations between entities but also multiple attributes of the relations (in particular, the model considers words describing the relations). In this way the GT model obtains more cohesive groups as well as fresh topics that influence the interaction between groups. The model could be applied to variables of other data types in addition to voting data. We are now using the model to analyze the citations in academic papers to capture the topics of research papers and discover research groups. It would also apply to a much larger network of entities (people, organizations, etc.) that frequently appear in newswire articles.

The model can be altered suitably to consider other attributes characterizing relations between entities in a network. In ongoing work we are extending the Group-

[3]This is not shown in Table 3.9 because they are missing from the 2005 GDP data.

Topic model to capture a richer notion of topic, where the attributes describing the relations between entities are represented by a mixture of topics.

CHAPTER 4

TOPICS OVER TIME: A NON-MARKOV CONTINUOUS-TIME MODEL OF TOPICAL TRENDS

Many of the large datasets to which topic models are applied do not have *static* co-occurrence patterns; they are instead *dynamic*. The data are often collected over time, and generally patterns present in the early part of the collection are not in effect later. Topics rise and fall in prominence; they split apart; they merge to form new topics; words change their correlations. For example, across 17 years of the Neural Information Processing Systems (NIPS) conference, activity in "analog circuit design" has fallen off somewhat, while research in "support vector machines" has recently risen dramatically. The topic "dynamic systems" used to co-occur with "neural networks," but now co-occurs with "graphical models."

However most of the topic models are unaware of these dependencies on document timestamps. Not modeling time can confound co-occurrence patterns and result in unclear, sub-optimal topic discovery. For example, in topic analysis of U.S. Presidential State-of-the-Union addresses, LDA confounds Mexican-American War (1846-1848) with some aspects of World War I (1914-1918), because LDA is unaware of the 70-year separation between the two events. Some previous work has performed some post-hoc analysis—discovering topics without the use of timestamps and then projecting their occurrence counts into discretized time [20]—but this misses the opportunity for time to improve topic discovery.

This chapter presents *Topics over Time (TOT)* [60], a topic model that explicitly models time jointly with word co-occurrence patterns. Significantly, and unlike

some recent work with similar goals, our model does not discretize time, and does not make Markov assumptions over state transitions in time. Rather, TOT parameterizes a continuous distribution over time associated with each topic, and topics are responsible for generating both observed timestamps as well as words. Parameter estimation is thus driven to discover topics that simultaneously capture word co-occurrences *and* locality of those patterns in time.

When a strong word co-occurrence pattern appears for a brief moment in time then disappears, TOT will create a topic with a narrow time distribution. (Given enough evidence, arbitrarily small spans can be represented, unlike schemes based on discretizing time.) When a pattern of word co-occurrence remains consistent across a long time span, TOT will create a topic with a broad time distribution. In current experiments, we use a Beta distribution over a (normalized) time span covering all the data, and thus we can also flexibly represent various skewed shapes of rising and falling topic prominence.

The model's generative storyline can be understood in two different ways. We fit the model parameters according to a generative model in which a per-document multinomial distribution over topics is sampled from a Dirichlet, then for each word occurrence we sample a topic; next a per-topic multinomial generates the word, *and* a per-topic Beta distribution generates the document's timestamp. Here the timestamp (which in practice is always observed and constant across the document) is associated with each word in the document. We can also imagine an alternative, corresponding generative model in which the timestamp is generated once per document, conditioned directly on the per-document mixture over topics. In both cases, the likelihood contribution from the words and the contribution from the timestamps may need to be weighted by some factor, as in the balancing of acoustic models and language models in speech recognition. The later generative storyline more directly corresponds to common datasets (with one timestamp per document); the former is easier to fit,

and can also allow some flexibility in which different parts of the document may be discussing different time periods.

Note that, in contrast to other work that models trajectories of individual topics over time, TOT topics and their meaning are modeled as constant over time. TOT captures changes in the occurrence (and co-occurrence conditioned on time) of the topics themselves, not changes in the word distribution of each topic. The classical view of splitting and merging of topics is thus reflected as dynamic changes in the co-occurrence of *constant* topics. While choosing to model individual topics as mutable could be useful, it can also be dangerous. Imagine a subset of documents containing strong co-occurrence patterns across time: first between birds and aerodynamics, then aerodynamics and heat, then heat and quantum mechanics—this could lead to a single topic that follows this trajectory, and lead the user to inappropriately conclude that birds and quantum mechanics are time-shifted versions of the same topic. Alternatively, consider a large subject like medicine, which has changed drastically over time. In TOT we choose to model these shifts as changes in topic *co-occurrence*—a decrease in occurrence of topics about blood-letting and bile, and an increase in topics about MRI and retrovirus, while the topics about blood, limbs, and patients continue to co-occur throughout. We do not claim that this point of view is better, but the difference makes TOT much simpler to understand and implement.

In comparison to more complex alternatives such as [9, 41], the relative simplicity of TOT is a great advantage—not only for the relative ease of understanding and implementing it, but also because this approach can in the future be naturally injected into other more richly structured topic models we discussed in previous chapters, such as the Author-Recipient-Topic model to capture changes in social network roles over time [39], and the Group-Topic model to capture changes in group formation over time [62] .

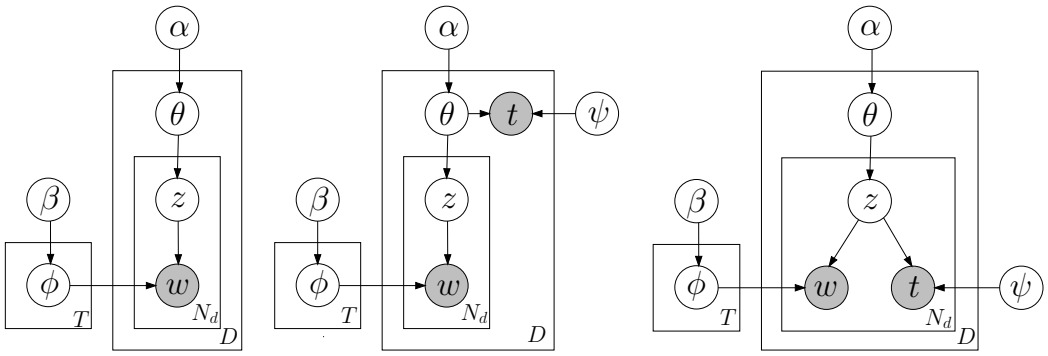
We present experimental results with three real-world datasets. On more than two centuries of U.S. Presidential State-of-the-Union addresses, we show that TOT discovers topics with both time-localization and word-clarity improvements over LDA. On the 17-year history of the NIPS conference, we show clearly interpretable topical trends, as well as a two-fold increase in the ability to predict time given a document. On nine months of a researcher's email archive (the McCallum dataset used in Chapter 2), we show another example of clearly interpretable, time-localized topics, such as springtime faculty recruiting. On all three datasets, TOT provides more distinct topics, as measured by KL divergence.

4.1 Topics over Time

Unlike in LDA, in TOT, topic discovery is influenced not only by word co-occurrences, but also temporal information. Rather than modeling a sequence of state changes with a Markov assumption on the dynamics, TOT models (normalized) absolute timestamp values. This allows TOT to see long-range dependencies in time, to predict absolute time values given an unstamped document, and to predict topic distributions given a timestamp. It also helps avoid a Markov model's risk of inappropriately dividing a topic in two when there is a brief gap in its appearance. The graphical model representations of our TOT models are shown in Figure 4.1 in which LDA is listed as well for comparison.

Time is intrinsically continuous. Discretization of time always begs the question of selecting the slice size, and the size is invariably too small for some regions and too large for others.

TOT avoids discretization by associating with each topic a continuous distribution over time. Many possible parameterized distributions are possible. Our earlier experiments were based on Gaussian. All the results in this chapter employ the Beta distribution (which can behave versatile shapes), for which the time range of



(a) LDA (b) TOT model, alternate view (c) TOT model, for collapsed Gibbs sampling

Figure 4.1. Three topic models: LDA and two perspectives on TOT

the data used for parameter estimation is normalized to a range from 0 to 1. Another possible choice of bounded distributions is the Kumaraswamy distribution [31]. Double-bounded distributions are appropriate because the training data are bounded in time. If it is necessary to ask the model to predict in a small window into the future, the bounded region can be extended, yet still estimated based on the data available up to now. Note that Beta distribution can only have a single mode (apart from the special case where two modes at both ends of the range), and this essentially rules out topics that recur. TOT would treat recurring topics as separated topics.

Topics over Time is a generative model of timestamps and the words in the timestamped documents. There are two ways of describing its generative process. The first, which corresponds to the process used in collapsed Gibbs sampling for parameter estimation, is as follows:

1. Draw T multinomials ϕ_z from a Dirichlet prior β, one for each topic z;

2. For each document d, draw a multinomial θ_d from a Dirichlet prior α; then for each word w_{di} in document d:

   (a) Draw a topic z_{di} from multinomial θ_d;

   (b) Draw a word w_{di} from multinomial $\phi_{z_{di}}$;

(c) Draw a timestamp t_{di} from Beta $\psi_{z_{di}}$.

The graphical model is shown in Figure 4.1(c). Although, in the above generative process, a timestamp is generated for each word token, all the timestamps of the words in a document are observed as the same as the timestamp of the document. One might also be interested in capturing burstiness, and some solution such as Dirichlet compound multinomial model (DCM) can be easily integrated into the TOT model [35]. In our experiments there are a fixed number of topics, T; although a non-parametric Bayes version of TOT that automatically integrates over the number of topics would certainly be possible.

As shown in the above process, the posterior distribution of topics depends on the information from two modalities—both text and time. TOT parameterization is

$$\begin{aligned}
\theta_d|\alpha &\sim \text{Dirichlet}(\alpha) \\
\phi_z|\beta &\sim \text{Dirichlet}(\beta) \\
z_{di}|\theta_d &\sim \text{Multinomial}(\theta_d) \\
w_{di}|\phi_{z_{di}} &\sim \text{Multinomial}(\phi_{z_{di}}) \\
t_{di}|\psi_{z_{di}} &\sim \text{Beta}(\psi_{z_{di}}).
\end{aligned}$$

Inference can not be done exactly in this model. We employ collapsed Gibbs sampling to perform approximate inference. Note that we adopt conjugate prior (Dirichlet) for the multinomial distributions, and thus we can easily integrate out θ and ϕ, analytically capturing the uncertainty associated with them. In this way we facilitate the sampling—that is, we need not sample θ and ϕ at all. Because we use the continuous Beta distribution rather than discretizing time, sparsity is not a big concern in fitting the temporal part of the model. For simplicity and speed we estimate these Beta distributions ψ_z by the method of moments, once per iteration of collapsed Gibbs sampling. One could estimate the values of the hyperparameters of

the TOT model, α and β, from data using a Gibbs EM algorithm [4]. In the particular applications discussed in this chapter, we find that the sensitivity to hyperparameters is not very strong after conducting a similar analysis as in Figure 2.2 of time prediction performance vs the values of hyperparameters. Thus, again for simplicity, we use fixed symmetric Dirichlet distributions ($\alpha = 50/T$ and $\beta = 0.1$) in all our experiments.

In the collapsed Gibbs sampling procedure above, we need to calculate the conditional distribution $P(z_{di}|\mathbf{w}, \mathbf{t}, \mathbf{z}_{-di}, \alpha, \beta, \Psi)$, where $\mathbf{z}_{-di}$ represents the topic assignments for all tokens except w_{di}. We begin with the joint probability of a dataset, and using the chain rule, we can obtain the conditional probability conveniently as

$$P(z_{di}|\mathbf{w}, \mathbf{t}, \mathbf{z}_{-di}, \alpha, \beta, \Psi) \propto (m_{dz_{di}} + \alpha_{z_{di}} - 1) \times \frac{n_{z_{di}w_{di}} + \beta_{w_{di}} - 1}{\sum_{v=1}^{V}(n_{z_{di}v} + \beta_v) - 1} \frac{(1-t_{di})^{\psi_{z_{di}1}-1} t_{di}^{\psi_{z_{di}2}-1}}{B(\psi_{z_{di}1}, \psi_{z_{di}2})},$$

where n_{zv} is the number of tokens of word v are assigned to topic z, m_{dz} represent the number of tokens in document d are assigned to topic z. Detailed derivation of collapsed Gibbs sampling for TOT is provided in Appendix C. An overview of the collapsed Gibbs sampling procedure we use is shown in Algorithm 2.

Algorithm 2 Inference on TOT

```
initialize topic assignment randomly for all tokens
for iter = 1 to N_iter do
    for d = 1 to D do
        for w = 1 to N_d do
            draw z_dw from P(z_dw | w, t, z_-dw, α, β, Ψ)
            update n_{z_dw w} and m_{d z_dw}
        end for
    end for
    for z = 1 to T do
        update ψ_z
    end for
end for
compute the posterior estimates of θ and φ
```

Although a document is modeled as a mixture of topics, there is typically only one timestamp associated with a document. The above generative process describes data in which there is a timestamp associated with each word. When fitting our model from typical data, each training document's timestamp is copied to all the words in the

document. However, after fitting, if actually run as a generative model, this process would generate different timestamps for the words within the same document. In this sense, thus, it is formally a deficient generative model, but still remains powerful in modeling large dynamic text collections.

An alternative generative process description of TOT (better suited to generate an unseen document), is one in which a single timestamp is associated with each document, generated by rejection or importance sampling, from a mixture of per-topic Beta distributions over time with mixtures weight as the per-document θ_d over topics. As before, this distribution over time is ultimately parameterized by the set of timestamp-generating Beta distributions, one per topic. The graphical model for this alternative generative process is shown in Figure 4.1(b).

Using this model we can predict a timestamp given the words in the document. To facilitate the comparison with LDA, we can discretize the timestamps (only for this purpose). Given a document, we predict its timestamp by choosing the discretized timestamp that maximizes the posterior which is calculated by multiplying the timestamp probability of all word tokens from their corresponding topic-wise Beta distributions over time, that is, $\arg\max_t \prod_{i=1}^{N_d} p(t|\psi_{z_i})$.

It is also interesting to consider obtaining a distribution over topics, conditioned on a timestamp. This allows us to see the topic occurrence patterns over time. By Bayes rule, $\mathrm{E}(\theta_{z_i}|t) = P(z_i|t) \propto p(t|z_i)P(z_i)$ where $P(z_i)$ can be estimated from data or simply assumed as uniform. Examples of expected topic distributions θ_d conditioned on timestamps are shown in Section 4.3.

Regarding parameter estimation, the two processes in Figure 4.1 (b) and (c) can become equivalent when we introduce a balancing hyperparameter between the likelihood from two modalities. In the second process, not surprisingly, the generation of one timestamp would be overwhelmed by the plurality of words generated under the bag of words assumption. To balance the influence from two different modalities, a

tunable hyperparameter is needed which is responsible for the relative weight of the time modality versus the text modality. Thus we use such a weighting parameter to rescale the likelihoods from different modalities, as is also common in speech recognition when the acoustic and language models are combined, and in the Group-Topic model [62] in which relational blockstructures and topic models are integrated. Here a natural setting for the weighting parameter is the inverse of the number of words N_d in the document, which is equivalent to generating N_d independent and identically distributed (i.i.d.) samples from the document-specific mixture of Beta distributions. Thus, it is probabilistically equivalent to drawing N_d samples from the individual Beta distributions according to the mixture weights θ_d, which exactly corresponds to the generative process in Figure 4.1 (c). In practice, it is also important to have such a hyperparameter when the likelihoods from discrete and continuous modalities are combined. We find that this hyperparameter is quite sensitive, and set it by trial and error.

Several previous studies have examined topics and their changes across time. Rather than jointly modeling word co-occurrence and time, many of these methods use post-hoc or pre-discretized analysis.

The first style of non-joint modeling involves fitting a time-unaware topic model, and then ordering the documents in time, slicing them into discrete subsets, and examining the topic distributions in each time-slice. One example is Griffiths and Steyvers' study of PNAS proceedings [20], in which they identified hot and cold topics based on examination of topic mixtures estimated from an LDA model.

The second style of non-joint modeling pre-divides the data into discrete time slices, and fits a separate topic model in each slice. Examples of this type include the experiments with the Group-Topic model we discussed in the previous chapter, in which several decades worth of U.N. voting records (and their accompanying text) were divided into 15-year segments; each segment was fit with the GT model, and

trends were compared. Similarly, in TDT tasks, *timelines* are constructed for a set of news stories[56, 55]. A χ^2 test is performed to identify days on which the number of occurrences of named entities or noun phrases produces a statistic above a given threshold; consecutive days under this criterion are stitched together to form an interval to be added into the timeline.

4.2 Datasets

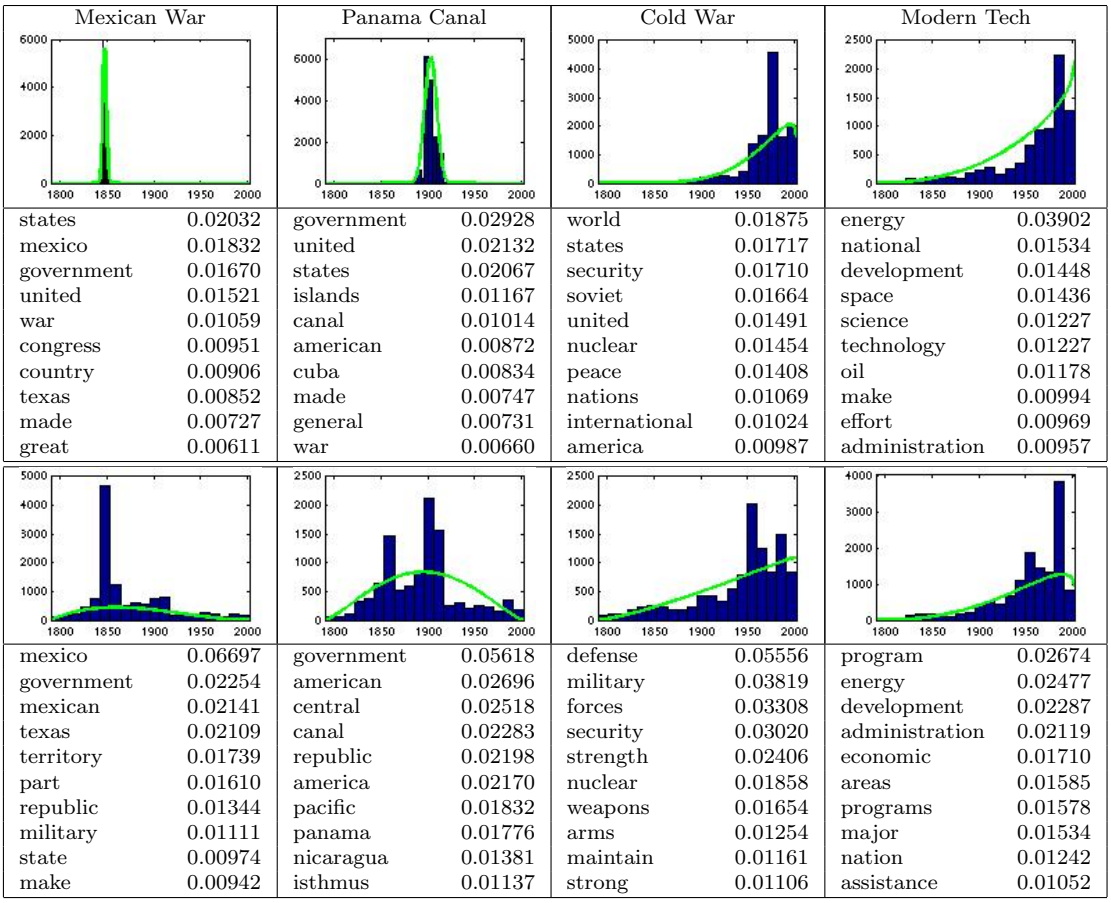
We present experiments with the TOT model on three real-world data sets: 9 months of email sent and received by Andrew McCallum (described in Chapter 2), 17 years of NIPS conference papers, and 21 decades of U.S. Presidential State-of-the-Union Addresses. In all cases, for simplicity, we fix the number of topics $T = 50$[1].

The State of the Union is an annual message presented by the President to Congress, describing the state of the country and his plan for the future. Our dataset[2] consists of the transcripts of 208 addresses during 1790-2002 (from George Washington to George W. Bush). We remove stopwords and numbers, and all text is downcased. Because the topics discussed in each address are so diverse, and in order to improve the robustness of the discovered topics, we increase the number of documents in this dataset by splitting each transcript into 3-paragraph "documents". The resulting dataset has 6,427 (3-paragraph) documents, 21,576 unique words, and 674,794 word tokens in total. Each document's timestamp is determined by the date on which the address was given.

The NIPS dataset (provided to us by Gal Chechik) consists of the full text of the 17 years of proceedings from 1987 to 2003 Neural Information Processing Systems (NIPS) Conferences. In addition to downcasing and removing stopwords and numbers, we

[1]It would be straightforward to automatically infer the number of topics using algorithms such as Hierarchical Dirichlet Process [57].

[2]http://www.gutenberg.org/dirs/etext04/suall11.txt



Mexican War		Panama Canal		Cold War		Modern Tech	
states	0.02032	government	0.02928	world	0.01875	energy	0.03902
mexico	0.01832	united	0.02132	states	0.01717	national	0.01534
government	0.01670	states	0.02067	security	0.01710	development	0.01448
united	0.01521	islands	0.01167	soviet	0.01664	space	0.01436
war	0.01059	canal	0.01014	united	0.01491	science	0.01227
congress	0.00951	american	0.00872	nuclear	0.01454	technology	0.01227
country	0.00906	cuba	0.00834	peace	0.01408	oil	0.01178
texas	0.00852	made	0.00747	nations	0.01069	make	0.00994
made	0.00727	general	0.00731	international	0.01024	effort	0.00969
great	0.00611	war	0.00660	america	0.00987	administration	0.00957
mexico	0.06697	government	0.05618	defense	0.05556	program	0.02674
government	0.02254	american	0.02696	military	0.03819	energy	0.02477
mexican	0.02141	central	0.02518	forces	0.03308	development	0.02287
texas	0.02109	canal	0.02283	security	0.03020	administration	0.02119
territory	0.01739	republic	0.02198	strength	0.02406	economic	0.01710
part	0.01610	america	0.02170	nuclear	0.01858	areas	0.01585
republic	0.01344	pacific	0.01832	weapons	0.01654	programs	0.01578
military	0.01111	panama	0.01776	arms	0.01254	major	0.01534
state	0.00974	nicaragua	0.01381	maintain	0.01161	nation	0.01242
make	0.00942	isthmus	0.01137	strong	0.01106	assistance	0.01052

Figure 4.2. Four topics discovered by TOT (above) and LDA (bottom) for the Address dataset. The titles are our own interpretation of the topics. Histograms show how the topics are distributed over time; the fitted beta PDFs are shown also. (For LDA, beta distributions are fit in a post-hoc fashion). The top words with their probability in each topic are shown below the histograms. The TOT topics are better localized in time, and TOT discovers more event-specific topical words.

also remove the words appearing less than five times in the corpus—many of them produced by OCR errors. Two letter words (primarily coming from equations), are removed, except for "ML", "AI", "KL", "BP", "EM" and "IR." The dataset contains 2,326 research papers, 24,353 unique words, and 3,303,020 word tokens in total. Each document's timestamp is determined by the year of the proceedings.

4.3 Experimental Results

In this section, we present the topics discovered by the TOT model and compare them with topics from LDA. We also demonstrate the ability of the TOT model to predict the timestamps of documents, more than doubling accuracy in comparison with LDA. We furthermore find topics discovered by TOT to be more distinct from each other than LDA topics (as measured by KL Divergence). Finally we show how TOT can be used to analyze topic co-occurrence conditioned on a timestamp. Topics presented in this section are extracted from a single sample at the 1000th iteration of the Gibbs sampler. For the address dataset, 1000 iterations of the Gibbs sampler took 3 hours on a dual-processor Opteron (Linux), 2 hours for the McCallum dataset, and 10 hours for the NIPS dataset.

4.3.1 Topics Discovered for Addresses

The State-of-the-Union addresses contain the full range of United States history. Analysis of this dataset shows strong temporal patterns. Some of them are broad historical issues, such as a clear "American Indian" topic throughout the 1800s and peaking around 1860, or the rise of "Civil Rights" across the second half of the 1900s. Other sharply localized trends are somewhat influenced by the individual president's communication style, such as Theodore Roosevelt's sharply increased use of the words "great", "men", "public", "country", and "work".

Four TOT topics, their most likely words, their Beta distributions over time, their actual histograms over time, as well as comparisons against their most similar LDA topic (by KL divergence), are shown in Figure 4.2. Immediately we see that the TOT topics are more neatly and narrowly focused in time; (time analysis for LDA is done post-hoc). An immediate and obvious effect is that this helps the reader understand more precisely when and over what length of time the topical trend was occurring. For example, in the leftmost topic, TOT clearly shows that the Mexican-American

war (1846-1848) occurred in the few years just before 1850. In LDA, on the other hand, the topic spreads throughout American history; it has its peak around 1850, but seems to be getting confused by a secondary peak around the time of World War I, (when "war" words were used again, and relations to Mexico played a small part). It is not so clear what event is being captured by LDA's topic.

The second topic, "Panama Canal," is another vivid example of how TOT can successfully localize a topic in time, and also how jointly modeling words and time can help sharpen and improve the topical word distribution. The Panama Canal (constructed during 1904-1914) is correctly localized in time, and the topic accurately describes some of the issues motivating canal construction: the sinking of the *U.S.S. Maine* in a Cuban harbor, and the long time it took U.S. warships to return to the Caribbean via Cape Horn. The LDA counterpart is not only widely spread through time, but also confounding topics such as modern trade relations with Central America and efforts to build the Panama Railroad in the 1850s.

The third topic shows the rise and fall of the Cold War, with a peak on the Reagan years, when Presidential rhetoric on the subject rose dramatically. Both TOT and LDA topics mention "nuclear," but only TOT correctly identifies "soviet". LDA confounds what is mostly a cold war topic (although it misses "soviet") with words and events from across American history, including small but noticeable bumps for World War I and the Civil War. TOT correctly has its own separate topic for World War I.

Lastly, the rightmost topics in Figure 4.2, "Modern Tech," shows a case in which the TOT topic is not necessarily better—just interestingly different than the LDA topic. The TOT topic, with mentions of *energy, space, science, and technology*, is about modern technology and energy. Its emphasis on modern times is also very distinct in its time distribution. The closest LDA topic also includes energy, but focuses on economic development and assistance to other nations. Its time distribution shows

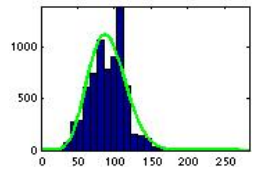
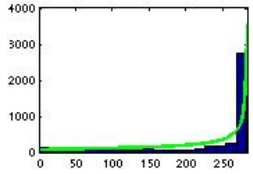
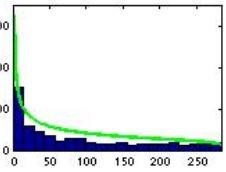
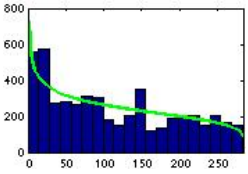
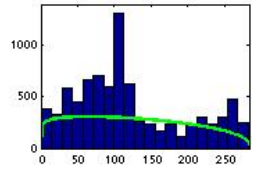
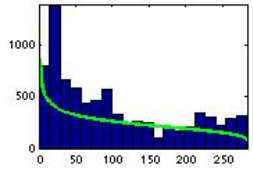
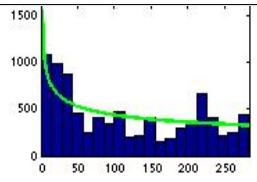
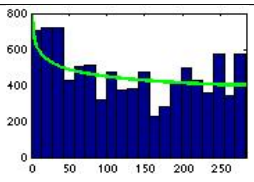
Faculty Recruiting		ART Paper		MALLET		CVS Operations	
							
cs	0.03572	xuerui	0.02113	code	0.05668	check	0.04473
april	0.02724	data	0.01814	files	0.04212	page	0.04070
faculty	0.02341	word	0.01601	mallet	0.04073	version	0.03828
david	0.02012	research	0.01408	java	0.03085	cvs	0.03587
lunch	0.01766	topic	0.01366	file	0.02947	add	0.03083
schedule	0.01656	model	0.01238	al	0.02479	update	0.02539
candidate	0.01560	andres	0.01238	directory	0.02080	latest	0.02519
talk	0.01355	sample	0.01152	version	0.01664	updated	0.02317
bruce	0.01273	enron	0.01067	pdf	0.01421	checked	0.02277
visit	0.01232	dataset	0.00960	bug	0.01352	change	0.02156
							
cs	0.05137	email	0.09991	code	0.05947	paper	0.06106
david	0.04592	ron	0.04536	mallet	0.03922	page	0.05504
bruce	0.02734	messages	0.04095	version	0.03772	web	0.04257
lunch	0.02710	data	0.03408	file	0.03702	title	0.03526
manmatha	0.02391	calo	0.03236	files	0.02534	author	0.02763
andrew	0.02332	message	0.03053	java	0.02522	papers	0.02741
faculty	0.01764	enron	0.03028	cvs	0.02511	email	0.02204
april	0.01740	project	0.02415	directory	0.01978	pages	0.02193
shlomo	0.01657	send	0.02023	add	0.01932	nips	0.01967
al	0.01621	part	0.01680	checked	0.01481	link	0.01860

Figure 4.3. Four topics discovered by TOT (above) and LDA (bottom) for the McCallum dataset, showing improved results with TOT. For example, the Faculty Recruiting topic is correctly identified in the spring in the TOT model, but LDA confuses it with other interactions among faculty.

an extra bump around the decade of the Marshal Plan (1947-1951), and a lower level during George W. Bush's presidency—both inconsistent with the time distribution learned by the TOT topic.

4.3.2 Topics Discovered for Email

In Figure 4.3 we demonstrate TOT on the McCallum dataset. Email is typically full of seasonal phenomena (such as paper deadlines, summer semester, etc.). One such seasonal example is the "Faculty Recruiting" topic, which (unlike LDA) TOT clearly identifies and localizes in the spring. The LDA counterpart is widely spread over the whole time period, and consequently, it cannot separate faculty recruiting from other types of faculty interactions and collaboration. The temporal information captured by TOT plays a very important role in forming meaningful time-sensitive topics.

The topic "ART paper" reflects a surge of effort in collaboratively writing a paper on the Author-Recipient-Topic model. Although the co-occurrence pattern of the words in this topic is strong and distinct, LDA failed to discover a corresponding topic—likely because it was a relatively short-lived phenomena. The closest LDA topic shows the general research activities, work on the DARPA CALO project, and various collaborations with SRI to prepare the Enron email dataset for public release. Not only does modeling time help TOT discover the "ART paper" task, but an alternative model that relied on coarse time discretization may miss such topics that have small time spans.

The "MALLET" topic shows that, after putting in an intense effort in writing and discussing Java programming for the MALLET toolkit, McCallum had less and less time to write code for the toolkit. In the corresponding LDA topic, MALLET development is confounded with CVS operations—which were later also used for managing collaborative writing of research papers.

TOT appropriately and clearly discovers a separate topics for "CVS operations," seen in the rightmost column. The closest LDA topic is the previously discussed one that merges MALLET and CVS. The second closest LDA topic (bottom right) discusses research paper writing, but not CVS. All these examples show that TOT's

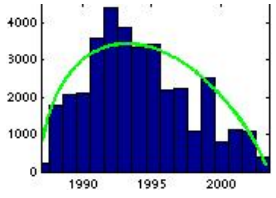
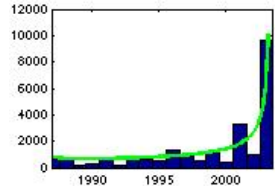
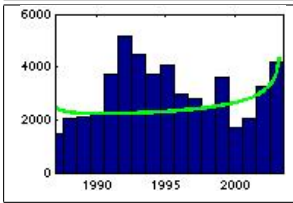
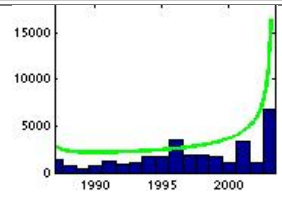
Recurrent NN		Game Theory	
			
state	0.05963	game	0.02850
recurrent	0.03765	strategy	0.02378
sequence	0.03616	play	0.01490
sequences	0.02462	games	0.01473
time	0.02402	player	0.01451
states	0.02057	agents	0.01346
transition	0.01300	expert	0.01281
finite	0.01242	strategies	0.01123
length	0.01154	opponent...	0.01088
strings	0.01013	nash	0.00848
			
state	0.05957	game	0.01784
sequence	0.03939	strategy	0.01357
sequences	0.02625	play	0.01131
time	0.02503	games	0.00940
states	0.02338	algorithm	0.00915
recurrent	0.01451	expert	0.00898
markov	0.01398	time	0.00837
transition	0.01369	player	0.00834
length	0.01164	return	0.00750
hidden	0.01072	strategies	0.00640

Figure 4.4. Two topics discovered by TOT (above) and LDA (bottom) for the NIPS dataset. For example, on the left, two major approaches to dynamic system modeling are mixed together by LDA, but TOT more clearly identifies waning interest in Recurrent Neural Networks, with a separate topic (not shown) for rising interest in Markov models.

use of time can help it pull apart distinct events, tasks and topics that may be confusingly merged by LDA.

4.3.3 Topics Discovered for NIPS

Research paper proceedings also present interesting trends for analysis. Successfully modeling trends in the research literature can help us understand how research fields evolve, and measure the impact of differently shaped profiles in time.

Figure 4.4 shows two topics discovered from the NIPS proceedings. "Recurrent Neural Networks" is clearly identified by TOT, and correctly shown to rise and fall in prominence within NIPS during the 1990s. LDA, unaware of the fact that Markov models superseded Recurrent Neural Networks for dynamic systems in the later NIPS years, and unaware of the time-profiles of both, ends up mixing the two methods together. LDA has a second topic elsewhere that also covers Markov models.

On the right, we see "Games" and game theory. This is an example in which TOT and LDA yield nearly identical results, although, if the terms beyond simply the first ten are examined, one sees that LDA is emphasizing board games, such as chess and backgammon, while TOT used its ramping-up time distribution to more clearly identify game theory as part of this topic (e.g., the word "Nash" occurs in position 12 for TOT, but not in the top 50 for LDA).

We have been discussing the salience and specificity of TOT's topics. Distances between topics can also be measured numerically. Table 4.1 shows the average distance of word distributions between all pairs of topics, as measured by KL Divergence. In all three datasets, the TOT topics are more distinct from each other. Partially because the Beta distribution is rarely multi-modal, the TOT model strives to separate events that occur during different time spans, and in real-world data, time differences are often correlated with word distribution differences that would have been more difficult to tease apart otherwise. The MALLET-CVS-paper distinction in the McCallum dataset is one example. (Events with truly multi-modal time distributions would be modeled with alternatives to the Beta distribution.)

Table 4.1. Average KL divergence between topics for TOT vs. LDA on three datasets. TOT finds more distinct topics.

	Address	Email	NIPS
TOT	0.6266	0.6416	0.5728
LDA	0.5965	0.5943	0.5421

Table 4.2. Predicting the decade, in the Address dataset. L1 Error is the difference between predicted and true decade. In the Accuracy column, we see that TOT predicts exactly the correct decade nearly twice as often as LDA.

	L1 Error	E(L1)	Accuracy
TOT	1.97	1.99	0.19
LDA	2.54	2.62	0.09

4.3.4 Time Prediction

One interesting feature of our approach (not shared by state-transition-based Markov models of topical shifts) is the capability of predicting the timestamp given the words in a document. This task also provides another opportunity to quantitatively compare TOT against LDA.

On the State-of-the-Union Address dataset, we measure the ability to predict the decade given the text of the address, as measured in accuracy, L1 error and average L1 distance to the correct decade (number of decades difference between predicted and correct decade). We randomly split our datasets into a training set (9/10) and a test set (the remaining 1/10) and do 10-fold cross-validation. As shown in Table 4.2, TOT achieves double the accuracy of LDA, and provides an L1 relative error reduction of 20%[3].

[3]When treating the timestamp as class label, a discriminative linear SVM classifier gives 72.4% accuracy under the same setting, but yields much less interpretability as expected.

4.3.5 Topic Distribution Profile over Time

It is also interesting to consider the TOT model's distribution over topics as a function of time. The time distribution of each *individual* topic is described as a Beta distribution (having flexible mean, variance and skewness), but even more rich and complex profiles emerge from the *interactions* among these Beta distributions. TOT's approach to modeling topic distributions conditioned on time stamp—based on multiple time-generating Betas, inverted with Bayes rule—has the dual advantages of a relatively simple, easy-to-fit parameterization, while also offering topic distributions with a flexibility that would be more difficult to achieve with a direct, non-inverted parameterization, (*i.e.*, one generating topic distributions directly conditioned on time, without Bayes-rule inversion).

The expected topic mixture distributions for the NIPS dataset are shown in Figure 4.5. The topics are consistently ordered in each year, and the heights of a topic's region represents the relative weight of the corresponding topic given a timestamp, calculated using the procedure described in Section 4.1. We can clearly see that topic mixtures change dramatically over time, and have interesting shapes. NIPS begins with more emphasis on neural networks, analog circuits and cells, but now emphasizes more SVMs, optimization, probability and inference.

4.3.6 Topic Co-occurrences over Time

We can also examine topic *co*-occurrences over time, which, as discussed earlier, are dynamic for many large text collections. In the following, we say two topics z_1 and z_2 (strongly) co-occur in a document d if both θ_{z_1} and θ_{z_2} are greater than some threshold h (we set $h = 2/T$); then we can count the number of documents in which certain topics (strongly) co-occur, and map out how co-occurrence patterns change over time.

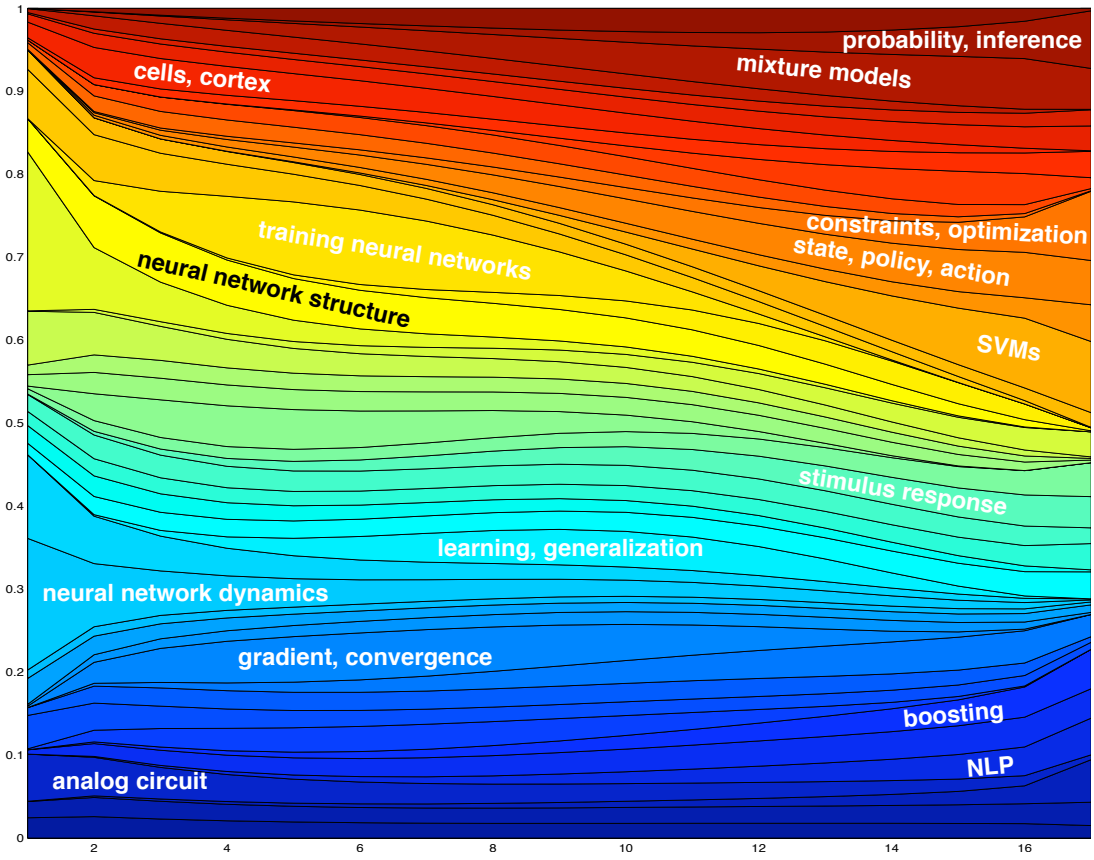


Figure 4.5. The distribution over topics given time in the NIPS data set. Note the rich collection of shapes that emerge from the Bayesian inversion of the collection of per-topic Beta distributions over time.

Figure 4.6 shows the prominence profile over time of those topics that co-occur strongly with the NIPS topic "classification." We can see that at the beginning NIPS, this problem was solved primarily with neural networks. It co-occurred with the "digit recognition" in the middle 90's. Later, probabilistic mixture models, boosting and SVM methods became popular.

4.4 Summary

This chapter has presented Topic over Time (TOT), a model that jointly models both word co-occurrences and localization in continuous time. Results on three

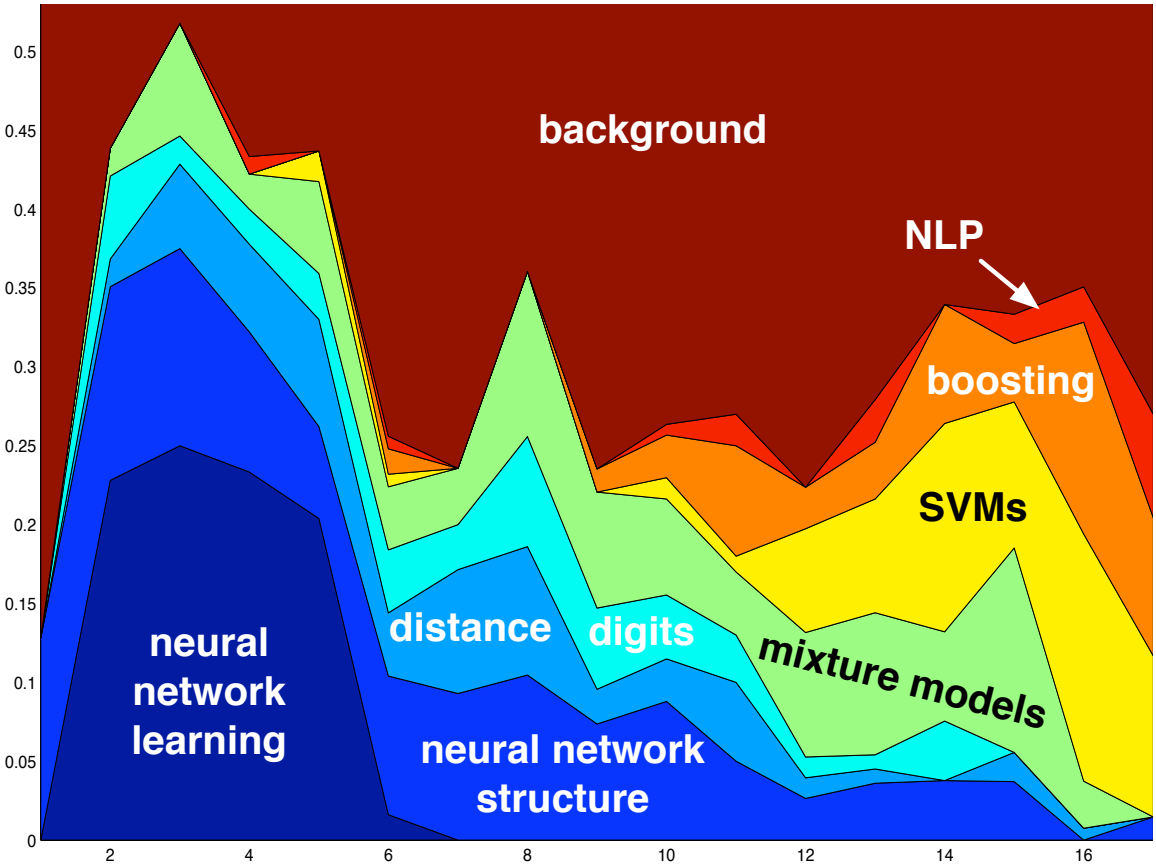


Figure 4.6. Eight topics co-occurring strongly with the "classification" topic in the NIPS dataset. Other co-occurring topics are labeled as a combined background topic. Classification with neural networks declined, while co-occurrence with SVMs, boosting and NLP are on the rise. The x-axis is the proceeding number, e.g., 1 corresponding to NIPS 1987 and 17 corresponding to NIPS 2003.

real-world datasets show the discovery of more salient topics that are associated with events, and clearly localized in time. We also show improved ability to predict time given a document. Reversing the inference by Bayes rule, yields a flexible parameterization over topics conditioned on time, as determined by the interactions among the many per-topic Beta distributions.

Unlike some related work with similar motivations, TOT does not require discretization of time or Markov assumptions on state dynamics. The relative simplicity of our approach provides advantages for injecting these ideas into other topic models.

CHAPTER 5

PHRASE AND TOPIC DISCOVERY WITH APPLICATION TO INFORMATION RETRIEVAL

Although the bag-of-words assumption is prevalent in document classification and topic models as we showed in previous chapters, the great majority of natural language processing methods represent word order, including n-gram language models for speech recognition, finite-state models for information extraction and context-free grammars for parsing. Word order is not only important for syntax, but also important for lexical meaning. A collocation is a phrase with meaning beyond the individual words.

N-gram phrases are fundamentally important in many areas of natural language processing and text mining, including parsing, machine translation and information retrieval. In general, phrases as the whole carry more information than the sum of its individual components, thus they are much more crucial in determining the topics of collections than individual words. Most topic models such as latent Dirichlet allocation (LDA) [8], however, assume that words are generated independently from each other, i.e., under the bag-of-words assumption. Adding phrases increases the model's complexity, but it could be useful in certain contexts. The possible over complicacy caused by introducing phrases makes these topic models completely ignore them. It is true that these models with the bag-of-words assumption have enjoyed a big success, and attracted a lot of interests from researchers with different backgrounds. We believe that a topic model considering phrases would be definitely more useful in certain applications.

Assume that we conduct topic analysis on a large collection of research papers. The acknowledgment sections of research papers have a distinctive vocabulary. Not surprisingly, we would end up with a particular topic on acknowledgment (or funding agencies) since many papers have an acknowledgment section that is not tightly coupled with the content of papers. One might therefore expect to find words such as "thank", "support" and "grant" in a single topic. One might be very confused, however, to find words like "health" and "science" in the same topic, unless they are presented in context: "National Institutes of Health" and "National Science Foundation".

Phrases often have specialized meaning, but not always. For instance, "neural networks" is considered a phrase because of its frequent use as a fixed expression. However, it specifies two distinct concepts: biological neural networks in neuroscience and artificial neural networks in modern usage. Without consulting the context in which the term is located, it is hard to determine its actual meaning. In many situations, topic is very useful to accurately capture the meaning. Furthermore, topic can play a role in phrase discovery. Considering learning English, a beginner usually has difficulty in telling "strong tea" from "powerful tea" [36], which are both grammatically correct. The topic associated with "tea" might help to discover the misuse of "powerful".

In this chapter, we propose a new topical n-gram (TNG) model [61] that automatically determines unigram words and phrases based on context and assign mixture of topics to both individual words and n-gram phrases. The ability to form phrases only where appropriate is unique to our model, distinguishing it from the traditional collocation discovery methods with which a *discovered* phrase is always treated as a *collocation* regardless of the context (which would possibly make us incorrectly conclude that "white house" remains a phrase in a document about real estate). Thus, TNG is not only a topic model that uses phrases, but also help linguists discover

meaningful phrases in right context, in a completely probabilistic manner. We show examples of extracted phrases and more interpretable topics on the NIPS data, and we present better information retrieval performance on an ad-hoc retrieval task over TREC collections, compared with other similar models.

5.1 N-gram based Topic Models

Before presenting our topical n-gram model, we first describe two related n-gram models, with the graphical models shown in Figure 5.1. For simplicity, all the models discussed in this section make the 1^{st} order Markov assumption, that is, they are actually bigram models. However, all the models have the ability to provide higher order n-grams ($n > 2$) by concatenating consecutive bigrams.

5.1.1 Bigram Topic Model (BTM)

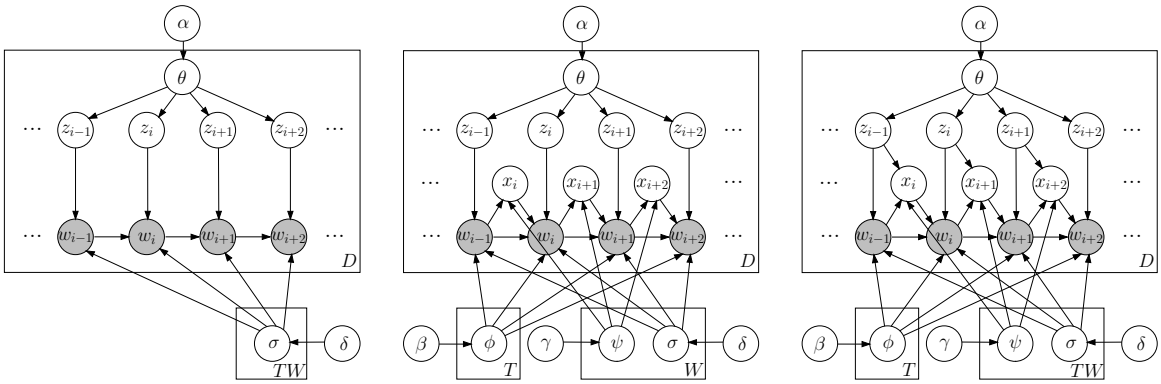
A bigram topic model was recently developed [59] on the basis of the hierarchical Dirichlet language model [34], by incorporating the concept of topic into bigram models. This model is one solution for the "neural network" example discussed earlier. We assume a dummy word w_0 existing at the beginning of each document. The graphical model presentation of this model is shown in Figure 5.1(a). The generative process of this model can be described as follows:

1. draw discrete distributions σ_{zw} from a Dirichlet prior δ for each topic z and each word w;

2. for each document d, draw a discrete distribution $\theta^{(d)}$ from a Dirichlet prior α; then for each word $w_i^{(d)}$ in document d:

   (a) draw $z_i^{(d)}$ from discrete $\theta^{(d)}$; and

   (b) draw $w_i^{(d)}$ from discrete $\sigma_{z_i^{(d)} w_{i-1}^{(d)}}$.



(a) Bigram topic model (b) LDA-Collocation model (c) Topical n-gram model

Figure 5.1. Three n-gram based topic models

Obviously, in most cases, two consecutive words do not form bigrams, and we believe that forming bigrams in appropriate context will help us understand largest text collections better.

5.1.2 LDA Collocation Model (LDACOL)

Starting from the LDA topic model, the LDA collocation model [22] introduces a new set of random variables (for bigram status) $\mathbf{x}$ ($x_i = 1$: w_{i-1} and w_i form a bigram; $x_i = 0$: they do not) that denote if a bigram can be formed with the previous token, in addition to the two sets of random variables $\mathbf{z}$ and $\mathbf{w}$ in LDA. Thus, it has the power to decide if to generate a bigram or a unigram. At this aspect, it is more realistic than the bigram topic model which always generates bigrams. After all, unigrams are the major components in a document. We assume the status variable x_1 is observed, and only a unigram is allowed at the beginning of a document. If we want to put more constraints into the model (e.g., no bigram is allowed for sentence/paragraph boundary; only a unigram can be considered for the next word after a stopword is removed; etc.), we can assume that the corresponding status variables are observed as well. This model's graphical model presentation is shown in Figure 5.1(b).

The generative process of the LDA collocation model is described as follows:

1. draw discrete distributions ϕ_z from a Dirichlet prior β for each topic z;

2. draw Bernoulli distributions ψ_w from a Beta prior γ for each word w;

3. draw discrete distributions σ_w from a Dirichlet prior δ for each word w;

4. for each document d, draw a discrete distribution $\theta^{(d)}$ from a Dirichlet prior α; then for each word $w_i^{(d)}$ in document d:

   (a) draw $x_i^{(d)}$ from Bernoulli $\psi_{w_{i-1}^{(d)}}$;

   (b) draw $z_i^{(d)}$ from discrete $\theta^{(d)}$; and

   (c) draw $w_i^{(d)}$ from discrete $\sigma_{w_{i-1}^{(d)}}$ if $x_i^{(d)} = 1$; else draw $w_i^{(d)}$ from discrete $\phi_{z_i^{(d)}}$.

Note that in the LDA Collocation model, bigrams do not have topics since the second term of a bigram is generated from a distribution σ_v conditioned on the previous word v only.

5.1.3 Topical N-gram Model (TNG)

The topical n-gram model (TNG) is not a simple combination of the bigram topic model and LDA collocation model. It can solve the problem associated with the "neural network" example as the bigram topic model, and automatically determine whether a composition of two terms is indeed a bigram as in the LDA collocation model. However, like other collocation discovery methods, a discovered bigram is always a bigram in the LDA Collocation model no matter what the context is.

One of the key contributions of our model is to make it possible to decide whether to form a bigram for the same two consecutive word tokens depending on their nearby context (i.e., co-occurrences). As in the LDA collocation model, we may assume some $\mathbf{x}$'s are observed for the same reason as we discussed in Section 5.1.2. The graphical model presentation of this model is shown in Figure 5.1(c). Its generative process can be described as follows:

1. draw discrete distributions ϕ_z from a Dirichlet prior β for each topic z;

2. draw Bernoulli distributions ψ_{zw} from a Beta prior γ for each topic z and each word w;

3. draw discrete distributions σ_{zw} from a Dirichlet prior δ for each topic z and each word w;

4. for each document d, draw a discrete distribution $\theta^{(d)}$ from a Dirichlet prior α; then for each word $w_i^{(d)}$ in document d:

   (a) draw $x_i^{(d)}$ from Bernoulli $\psi_{z_{i-1}^{(d)} w_{i-1}^{(d)}}$;

   (b) draw $z_i^{(d)}$ from discrete $\theta^{(d)}$; and

   (c) draw $w_i^{(d)}$ from discrete $\sigma_{z_i^{(d)} w_{i-1}^{(d)}}$ if $x_i^{(d)} = 1$; else draw $w_i^{(d)}$ from discrete $\phi_{z_i^{(d)}}$.

Note that our model is a generalization of BTM and of LDACOL. Both BTM (by setting all x's to 1) and LDACOL (by making σ conditioned on previous word only) are the special cases of our TNG models.

Before discussing the inference problem of our model, I will discuss the topic consistency of terms in a bigram. As shown in the above, the topic assignments for the two terms in a bigram are not required to be identical. We can take the topic of the first/last word token or the most common topic in the phrase, as the topic of the phrase. In this chapter, we will use the topic of the last term as the topic of the phrase for simplicity, since long noun phrases do truly sometimes have components indicative of different topics, and its last noun is usually the "head noun". Alternatively, we could enforce consistency in the model with ease, by simply adding two more sets of arrows ($z_{i-1} \rightarrow z_i$ and $x_i \rightarrow z_i$). Accordingly, we could substitute Step 4(b) in the above generative process with "draw $z_i^{(d)}$ from discrete $\theta^{(d)}$ if $x_i^{(d)} = 1$; else let $z_i^{(d)} = z_{i-1}^{(d)}$;" In this way, a word has the option to inherit a topic assignment from its

previous word if they form a bigram phrase. However, from our experimental results, the first choice yields visually better topics. From now on, we will focus on the model shown in Figure 5.1(c).

Finally we want to point out that the topical n-gram model is not only a new framework for distilling n-gram phrases depending on nearby context, but also a more sensible topic model than the ones using word co-occurrences alone.

In state-of-the-art hierarchical Bayesian models such as latent Dirichlet allocation, exact inference over hidden topic variables is typically intractable due to the large number of latent variables and parameters in the models. Approximate inference techniques such as variational methods [27], Gibbs sampling [4] and expectation propagation [42] have been developed to address this issue. We use collapsed Gibbs sampling again to conduct approximate inference in this chapter. To reduce the uncertainty introduced by θ, ϕ, ψ, and σ, we could integrate them out with no trouble because of the conjugate prior setting in our model. Starting from the joint distribution $P(\mathbf{w}, \mathbf{z}, \mathbf{x}|\alpha, \beta, \gamma, \delta)$, we can work out the conditional probabilities $P(z_i^{(d)}, x_i^{(d)}|\mathbf{z}_{-i}^{(d)}, \mathbf{x}_{-i}^{(d)}, \mathbf{w}, \alpha, \beta, \gamma, \delta)$ conveniently[1] using Bayes rule, where $\mathbf{z}_{-i}^{(d)}$ denotes the topic assignments for all word tokens except word $w_i^{(d)}$, and $\mathbf{x}_{-i}^{(d)}$ represents the bigram status for all tokens except word $w_i^{(d)}$. During collapsed Gibbs sampling, we draw the topic assignment $z_i^{(d)}$ and the bigram status $x_i^{(d)}$ iteratively[2] for each word token $w_i^{(d)}$ according to the following conditional probability distribution:

$$
\begin{aligned}
& P(z_i^{(d)}, x_i^{(d)}|\mathbf{z}_{-i}^{(d)}, \mathbf{x}_{-i}^{(d)}, \mathbf{w}, \alpha, \beta, \gamma, \delta) \\
\propto\ & (\gamma_{x_i^{(d)}} + p_{z_{i-1}^{(d)} w_{i-1}^{(d)} x_i} - 1)(\alpha_{z_i^{(d)}} + q_{dz_i^{(d)}} - 1) \times \begin{cases} \frac{\beta_{w_i^{(d)}} + n_{z_i^{(d)} w_i^{(d)}} - 1}{\sum_{v=1}^{W} (\beta_v + n_{z_i^{(d)} v}) - 1} & \text{if } x_i^{(d)} = 0 \\ \frac{\delta_{w_i^{(d)}} + m_{z_i^{(d)} w_{i-1}^{(d)} w_i^{(d)}} - 1}{\sum_{v=1}^{W} (\delta_v + m_{z_i^{(d)} w_{i-1}^{(d)} v}) - 1} & \text{if } x_i^{(d)} = 1 \end{cases}
\end{aligned}
$$

[1] As shown in Appendix A, one could further calculate $P(z_i^{(d)}|\cdots)$ and $P(x_i^{(d)}|\cdots)$ as in a traditional Gibbs sampling procedure.

[2] For some observed $x_i^{(d)}$, only $z_i^{(d)}$ needs to be drawn.

where n_{zw} represents how many times word w is assigned into topic z as a unigram, m_{zwv} represents how many times word v is assigned to topic z as the 2^{nd} term of a bigram given the previous word w, p_{zwk} denotes how many times the status variable $x = k$ (0 or 1) given the previous word w and the previous word's topic z, and q_{dz} represents how many times a word is assigned to topic z in document d. Note all counts here do include the assignment of the token being visited. Details of the collapsed Gibbs sampling derivation are provided in Appendix D.

Simple manipulations give us the posterior estimates of θ, ϕ, ψ, and σ as follows:

$$\begin{aligned} \hat{\theta}_z^{(d)} &= \frac{\alpha_z + q_{dz}}{\sum_{t=1}^{T}(\alpha_t + q_{dt})} \qquad \hat{\phi}_{zw} = \frac{\beta_w + n_{zw}}{\sum_{v=1}^{W}(\beta_v + n_{zv})} \\ \hat{\psi}_{zwk} &= \frac{\gamma_k + p_{zwk}}{\sum_{k=0}^{1}(\gamma_k + p_{zwk})} \quad \hat{\sigma}_{zwv} = \frac{\delta_v + m_{zwv}}{\sum_{v=1}^{W}(\delta_v + m_{zwv})} \end{aligned} \tag{5.1}$$

As discussed in the bigram topic model [59], one could certainly infer the values of the hyperparameters in TNG using a Gibbs EM algorithm [4]. In the particular experiments discussed in this chapter, we find that sensitivity to hyperparameters is not a big concern after some analysis of average precision versus 5 different values for each hyperparameter. For simplicity and feasibility in our Gigabyte TREC retrieval tasks, we skip the inference of hyperparameters, and use some reported empirical values for them instead to show salient results.

5.2 Experimental Results

We apply the topical n-gram model to the NIPS proceedings dataset described in Chapter 4. Topics found from a 50-topic run on the NIPS dataset (10,000 Gibbs sampling iterations) of the topical n-gram model are shown in Table 5.1 as anecdotal evidence, with comparison to the corresponding closest (by KL divergence) topics found by LDA. We use a symmetric priors $\alpha = 1$, $\beta = 0.01$, $\gamma = 0.1$, and $\delta = 0.01$.

The "Reinforcement Learning" topic provides an extremely salient summary of the corresponding research area. The LDA topic assembles many common words used in

Reinforcement Learning			Human Receptive System		
LDA	n-gram (2+)	n-gram (1)	LDA	n-gram (2+)	n-gram (1)
state	reinforcement learning	action	motion	receptive field	motion
learning	optimal policy	policy	visual	spatial frequency	spatial
policy	dynamic programming	reinforcement	field	temporal frequency	visual
action	optimal control	states	position	visual motion	receptive
reinforcement	function approximator	actions	figure	motion energy	response
states	prioritized sweeping	function	direction	tuning curves	direction
time	finite-state controller	optimal	fields	horizontal cells	cells
optimal	learning system	learning	eye	motion detection	figure
actions	reinforcement learning rl	reward	location	preferred direction	stimulus
function	function approximators	control	retina	visual processing	velocity
algorithm	markov decision problems	agent	receptive	area mt	contrast
reward	markov decision processes	q-learning	velocity	visual cortex	tuning
step	local search	goal	vision	light intensity	moving
dynamic	state-action pair	space	moving	directional selectivity	model
control	markov decision process	step	system	high contrast	temporal
sutton	belief states	environment	flow	motion detectors	responses
rl	stochastic policy	system	edge	spatial phase	orientation
decision	action selection	problem	center	moving stimuli	light
algorithms	upright position	steps	light	decision strategy	stimuli
agent	reinforcement learning methods	transition	local	visual stimuli	cell

Speech Recognition			Support Vector Machines		
LDA	n-gram (2+)	n-gram (1)	LDA	n-gram (2+)	n-gram (1)
recognition	speech recognition	speech	kernel	support vectors	kernel
system	training data	word	linear	test error	training
word	neural network	training	vector	support vector machines	support
face	error rates	system	support	training error	margin
context	neural net	recognition	set	feature space	svm
character	hidden markov model	hmm	nonlinear	training examples	solution
hmm	feature vectors	speaker	data	decision function	kernels
based	continuous speech	performance	algorithm	cost functions	regularization
frame	training procedure	phoneme	space	test inputs	adaboost
segmentation	continuous speech recognition	acoustic	pca	kkt conditions	test
training	gamma filter	words	function	leave-one-out procedure	data
characters	hidden control	context	problem	soft margin	generalization
set	speech production	systems	margin	bayesian transduction	examples
probabilities	neural nets	frame	vectors	training patterns	cost
features	input representation	trained	solution	training points	convex
faces	output layers	sequence	training	maximum margin	algorithm
words	training algorithm	phonetic	svm	strictly convex	working
frames	test set	speakers	kernels	regularization operators	feature
database	speech frames	mlp	matrix	base classifiers	sv
mlp	speaker dependent	hybrid	machines	convex optimization	functions

Table 5.1. The four topics from a 50-topic run of TNG on 13 years of NIPS research papers with their closest counterparts from LDA. The **Title** above the word lists of each topic is our own summary of the topic. To better illustrate the difference between TNG and LDA, we list the n-grams ($n > 1$) and unigrams separately for TNG. Each topic is shown with the 20 sorted highest-probability words. The TNG model produces clearer word list for each topic by associating many generic words (such as "set", "field", "function", etc.) with other words to form n-gram phrases.

reinforcement learning, but in its word list, there are quite a few generic words (such as "function", "dynamic", "decision") that are common and highly probable in many other topics as well. In TNG, we can find that these generic words are associated with other words to form n-gram phrases (such as "markov decision process", etc.) that are only highly probable in reinforcement learning. More importantly, by forming n-gram phrases, the unigram word list produced by TNG is also cleaner. For example, because of the prevalence of generic words in LDA, highly related words (such as "q-learning" and "goal") are not ranked highly enough to be shown in the top 20

word list. On the contrary, they are ranked very high in the TNG's unigram word list.

In the other three topics (Table 5.1), we can find similar phenomena as well. For example, in "Human Receptive System", some generic words (such as "field", "receptive") are actually the components of the popular phrases in this area as shown in the TNG model. "system" is ranked high in LDA, but almost meaningless, and on the other hand, it does not appear in the top word lists of TNG. Some extremely related words (such as "spatial"), ranked very high in TNG, are absent in LDA's top word list. In "Speech Recognition", the dominating generic words (such as "context", "based", "set", "probabilities", "database") make the LDA topic less understandable than even just TNG's unigram word list.

In many situations, a crucially related word might be not mentioned enough to be clearly captured in LDA, on the other hand, it would become very salient as a phrase due to the relatively stronger co-occurrence pattern in an extremely sparse setting for phrases. The "Support Vector Machines" topic provides such an example. We can imagine that "kkt" will be mentioned no more than a few times in a typical NIPS paper, and it probably appears only as a part of the phrase "kkt conditions". TNG satisfyingly captures it successfully as a highly probable phrase in the SVM topic.

As we discussed before, higher-order n-grams ($n > 2$) can be approximately modeled by concatenating consecutive bigrams in the TNG model, as shown in Table 5.1 (such as "markov decision process", "hidden markov model" and "support vector machines", etc.).

To quantitively evaluate the topical n-gram model, we could use some standard measures such as perplexity and document classification accuracy. However, to convincingly illustrate the power of the TNG model on larger, more real scale, here we apply the TNG model to a much larger standard text mining task—we employ the

TNG model within the language modeling framework to conduct ad-hoc retrieval on Gigabyte TREC collections.

5.2.1 Ad-hoc Retrieval

Traditional information retrieval (IR) models usually represent text with a bags-of-words assumption indicating that words occur independently. However, this is not an accurate statement about natural language. To address this problem, researchers have been working on capturing word dependencies. There are mainly two types of dependencies being studied: (1) topical (semantic) dependency, which is also called long-distance dependency. Two words are considered dependent when their meanings are related and they co-occur often, such as "fruit" and "apple." Among models capturing semantic dependency, the LDA-based document models [67] is one recent example. For IR applications, a major advantage of topic models (document expansion), compared to online query expansion in pseudo relevance feedback, is that they can be trained offline, thus more efficient in handling a new query; (2) phrase dependency, also called short-distance dependency. As reported in literature, retrieval performance can be boosted if the similarity between a user query and a document is calculated by common phrases instead of common words [16, 17, 43, 54]. Most research on phrases in information retrieval has employed an independent collocation discovery module. In this way, a phrase can be indexed exactly as an ordinary word.

The topical n-gram model automatically and simultaneously takes cares of both semantic co-occurrences and phrases. Also, it does not need a separate module for phrase discovery, and everything can be seamlessly integrated into the language modeling framework, which is one of the most popular statistically principled approaches to IR. In this section, we illustrate the difference in IR experiments of the TNG and LDA models, and compare the IR performance of all three models in Figure 5.1 on TREC collections introduced below.

The SJMN dataset, taken from TREC with standard queries 51-150 that are taken from the *title* field of TREC topics, covers materials from San Jose Mercury News in 1991. For validation purpose, we also consider the WSJ dataset with standard queries 51-100 and 151-200 that are also from the TREC topics (title only). All text is downcased and only alphabetic characters are kept. Stop words in both the queries and documents are removed, according to a common stopword list in the Bow toolkit [37]. If any two consecutive tokens were originally separated by a stopword, no bigram is allowed to be formed. In total, the SJMN dataset we use contains 90,257 documents, 150,714 unique words, and 21,156,378 tokens, which is order of magnitude larger than the NIPS dataset. Relevance judgments are taken from the the judged pool of the top retrieved documents by various participating retrieval systems from previous TREC conferences.

The number of topics is set to be 100 for all models with 10,000 Gibbs sampling iterations, and the same hyperparameter setting (with symmetric priors $\alpha = 1$, $\beta = 0.01$, $\gamma = 0.1$, and $\delta = 0.01$) for the NIPS dataset are used. Here, the focus of this experiment is to compare our results to another published model [67] that also uses LDA for information retrieval, not to achieve state-of-the-art results in TREC retrieval that need significant, non-modeling effort to achieve (such as stemming).

5.2.2 Difference between Topical N-grams and LDA in IR Applications

From both of LDA and TNG, a word distribution for each document can be calculated, which thus can be viewed as a document model. With these distributions, the likelihood of generating a query can be computed to rank documents, which is the basic idea in the query likelihood (QL) model in IR. When the two models are directly applied to do ad-hoc retrieval, the TNG model performs significant better than the LDA model under the Wilcoxon test at 95% level. Among of 4881 relevant documents for all queries, LDA retrieves 2257 of them but TNG gets 2450, 8.55% more. The

average precision for TNG is 0.0709, 61.96% higher than its LDA counterpart (0.0438). Although these results are not the state-of-the-art IR performance, we claim that, if used alone, TNG represent a document better than LDA. The average precisions for both models are very low, because corpus-level topics may be too coarse to be used as the only representation in IR [13, 67]. Significant improvements in IR can be achieved through a combination with the basic query likelihood model.

In the query likelihood model, each document is scored by the likelihood of its model generating a query Q, $P_{LM}(Q|d)$. Let the query $Q = (q_1, q_2, ..., q_{L_Q})$. Under the bag-of-words assumption, $P_{LM}(Q|d) = \prod_{i=1}^{L_Q} P(q_i|d)$, which is often specified by the document model with Dirichlet smoothing [74],

$$P_{LM}(q|d) = \frac{N_d}{N_d + \mu} P_{ML}(q|d) + (1 - \frac{N_d}{N_d + \mu}) P_{ML}(q|\text{coll}),$$

where N_d is the length of document d, $P_{ML}(q|d)$ and $P_{ML}(q|\text{coll})$ are the maximum likelihood (ML) estimates of a query term q generated in document d and in the entire collection, respectively, and μ is the Dirichlet smoothing prior (in our reported experiments we used a fixed value with $\mu = 1000$ as in [67]).

To calculate the query likelihood from the TNG model within the language modeling framework, we need to sum over the topic variable and bigram status variable for each token in the query token sequence. Given the posterior estimates $\hat{\theta}$, $\hat{\phi}$, $\hat{\psi}$, and $\hat{\sigma}$ (Equation 5.1), the query likelihood of query Q given document d, $P_{TNG}(Q|d)$ can be calculated[3] as $P_{TNG}(Q|d) = \prod_{i=1}^{L_Q} P_{TNG}(q_i|q_{i-1}, d)$, where $P_{TNG}(q_i|q_{i-1}, d) = \sum_{z_i=1}^{T}(P(x_i = 0|\hat{\psi}_{q_{i-1}})P(q_i|\hat{\phi}_{z_i}) + P(x_i = 1|\hat{\psi}_{q_{i-1}})P(q_i|\hat{\sigma}_{z_i q_{i-1}}))P(z_i|\hat{\theta}^{(d)})$, and we define $P(x_i|\hat{\psi}_{q_{i-1}}) = \sum_{z_{i-1}=1}^{T} P(x_i|\hat{\psi}_{z_{i-1}q_{i-1}})P(z_{i-1}|\hat{\theta}^{(d)})$.

[3]A dummy q_0 is assumed at the beginning of every query, for the convenience of mathematical presentation.

No.	Query	LDA	TNG	Change
053	Leveraged Buyouts	0.2141	0.3665	71.20%
097	Fiber Optics Applications	0.1376	0.2321	68.64%
108	Japanese Protectionist Measures	0.1163	0.1686	44.94%
111	Nuclear Proliferation	0.2353	0.4952	110.48%
064	Hostage-Taking	0.4265	0.4458	4.52%
125	Anti-smoking Actions by Government	0.3118	0.4535	45.47%
145	Influence of the "Pro-Israel Lobby"	0.2900	0.2753	-5.07%
148	Conflict in the Horn of Africa	0.1990	0.2788	40.12%

Table 5.2. Comparison of LDA and TNG on TREC retrieval performance (average precision) of eight queries on the SJMN dataset. The top four queries obviously contain phrase(s), and thus TNG achieves much better performance. On the other hand, the bottom four queries do not contain common phrase(s) after preprocessing (stopword and punctuation removal). Surprisingly, TNG still outperforms LDA on some of these queries.

Due to stopword and punctuation removal, we may simply set $P(x_i = 0|\hat{\psi}_{q_{i-1}}) = 1$ and $P(x_i = 1|\hat{\psi}_{q_{i-1}}) = 0$ at corresponding positions in a query. Note here in the above calculation, the bag-of-words assumption is not made any more.

Similar to the method in [67], we can combine the query likelihood from the basic language model and the likelihood from the TNG model in various ways. One can combine them at query level, i.e., $P(Q|d) = \lambda P_{LM}(Q|d) + (1 - \lambda)P_{TNG}(Q|d)$, where λ is a weighting factor between the two likelihoods.

Alternatively (used in this chapter), under first order Markov assumption, $P(Q|d) = P(q_1|d)\prod_{i=2}^{L_Q} P(q_i|q_{i-1}, d)$, and one can combine the query likelihood at query term level, that is, $P(q_i|q_{i-1}, d) = \lambda P_{LM}(q_i|d) + (1 - \lambda)P_{TNG}(q_i|q_{i-1}, d)$.

To illustrate the difference of TNG and LDA in IR applications, we select a few of the 100 queries that clearly contain phrase(s), and another few of them that do not contain phrase due to stopword and punctuation removal, on which we compare the IR performance (average precision)[4] as shown in Table 5.2.

[4]The results reported in [67] is a little better since they did stemming.

No.	Query	TNG	BTM	Change	LDACOL	Change
061	Israeli Role in Iran-Contra Affair	0.1635	0.1104	-32.47%	0.1316	-19.49%
069	Attempts to Revive the SALT II Treaty	0.0026	0.0071	172.34%	0.0058	124.56%
110	Black Resistance Against the South African Government	0.4940	0.3948	-20.08%	0.4883	-1.16%
117	Capacity of the U.S. Cellular Telephone Network	0.2801	0.3059	9.21%	0.1999	-28.65%
130	Jewish Emigration and U.S.-USSR Relations	0.2087	0.1746	-16.33%	0.1765	-15.45%
138	Iranian Support for Lebanese Hostage-takers	0.4398	0.4429	0.69%	0.3528	-19.80%
146	Negotiating an End to the Nicaraguan Civil War	0.0346	0.0682	97.41%	0.0866	150.43%
150	U.S. Political Campaign Financing	0.2672	0.2323	-13.08%	0.2688	0.59%
	All Queries	0.2122	0.1996	-5.94%*	0.2107	-0.73%*

Table 5.3. Comparison of the bigram topic model ($\lambda = 0.7$), LDA collocation model ($\lambda = 0.9$) and the topical n-gram Model ($\lambda = 0.8$) on TREC retrieval performance (average precision) on the SJMN dataset. * indicates statistically significant differences in performance with 95% confidence according to the Wilcoxon test. TNG performs significantly better than other two models overall.

5.2.3 Comparison of BTM, LDACOL and TNG on TREC Ad-hoc Retrieval

In this section, we compare the IR performance of the three n-gram based topic models on the SJMN dataset[5], as shown in Table 5.3. For a fair comparison, the weighting factor λ (reported in Table 5.3) are independently chosen to get the best performance from each model. Under the Wilcoxon test with 95% confidence, TNG significantly outperforms BTM and LDACOL on this standard retrieval task.

It is interesting to see that different models are good at quite different queries. For some queries (such as No. 117 and No. 138), TNG and BTM perform similarly, and better than LDACOL, and for some other queries (such as No. 110 and No. 150), TNG and LDACOL perform similarly, and better than BTM. There are also queries (such as No. 061 and No. 130) for which TNG performs better than both BTM and LDACOL. We believe that they are clear empirical evidence that our TNG model are more generic and powerful than BTM and LDACOL.

It is true that for certain queries (such as No. 069 and No. 146), TNG performs worse than BTM and LDACOL, but we notice that all models perform badly on these queries and the behaviors are more possibly due to randomness.

[5]The running times of our C implementation on a dual-processor Opteron for the three models are 11.5, 17, 22.5 hours, respectively.

Dataset	TNG	BTM	Change	LDACOL	Change
SJMN	0.2122	0.1996	-5.94%*	0.2107	-0.73%*
WSJ	0.3051	0.2863	-6.16%*	0.2999	-1.70%*

Table 5.4. Comparison of the bigram topic model ($\lambda = 0.7$), LDA collocation model ($\lambda = 0.9$) and the topical n-gram Model ($\lambda = 0.8$) on TREC retrieval performance (average precision). The values of λ were tuned on the SJMN dataset. * indicates statistically significant differences in performance with 95% confidence according to the Wilcoxon test. TNG performs significantly better than other two models on both datasets.

As one might notice, we tuned the λ parameter on the SJMN dataset. As a validation, we conduct the same experiments on the WSJ dataset using the exactly same setting, shown in Table 5.4 (we show the results on SJMN again for comparison). The WSJ dataset is roughly twice as large as the SJMN dataset, and we can see that TNG outperform BTM and LDACOL as well on the WSJ dataset.

5.3 Summary

In this chapter, we have presented the topical n-gram model. The TNG model automatically determines whether to form an n-gram (and further assign a topic) or not, based on its surrounding context. Examples of topics found by TNG are more interpretable than its LDA counterpart. We also demonstrate how TNG can help improve retrieval performance in standard ad-hoc retrieval tasks on TREC collections over its two special-case n-gram based topic models.

Unlike some traditional phrase discovery methods, the TNG model provides a systematic way to model (topical) phrases and can be seamlessly integrated with many probabilistic frameworks for various tasks such as phrase discovery, ad-hoc retrieval, machine translation, speech recognition and statistical parsing.

Evaluating n-gram based topic models is a big challenge. As reported in [59], the bigram topic models have only been shown to be effective on hundreds of documents,

and also we have not seen a formal evaluation of the LDA collocation models. To the best of our knowledge, our work presents the very first application of all three n-gram based topic models on Gigabyte collections, and a novel way to integrate n-gram based topic models into the language modeling framework for information retrieval tasks.

CHAPTER 6

CONCLUSIONS

This thesis has developed sophisticated topic models for analyzing text collections with multiple modality attributes. Directed graphical models were employed as a flexible framework to describe modeling assumptions about the data. Furthermore, we adopted collapsed Gibbs sampling inference techniques which free us from having to specify tractable models. These methods support a systemic approach to handling large datasets with multiple modalities.

With the general principle in mind, we developed several multi-modality topic models that could be mapped to one of the three meaningful configurations in Figure 1.2 (i.e., ART to configuration (A), GT and TOT to configuration (B), and TNG to configuration (C)). As we have shown, all the models presented are capable to discover interesting topics that could be not be found via text alone, or from multiple modalities without joint inference. In various real world applications such as information retrieval, timestamp prediction, these models outperform significantly their counterparts, respectively.

In terms of inference, we have used collapsed Gibbs sampling in the thesis. One concern is how long it will take for a Gibbs chain to reach equilibrium in complicated topic models on massive text collections. The general practice is to let the chain run long enough, say, 10,000 iterations. We did some formal analysis about the autocorrelation of perplexity [10]. Using ART model on the McCallum dataset as an example, the autocorrelation plot of perplexity on iterations 10001-10500 is shown in Figure 6.1. As we can see, the autocorrelations are near zero for almost all time-lag

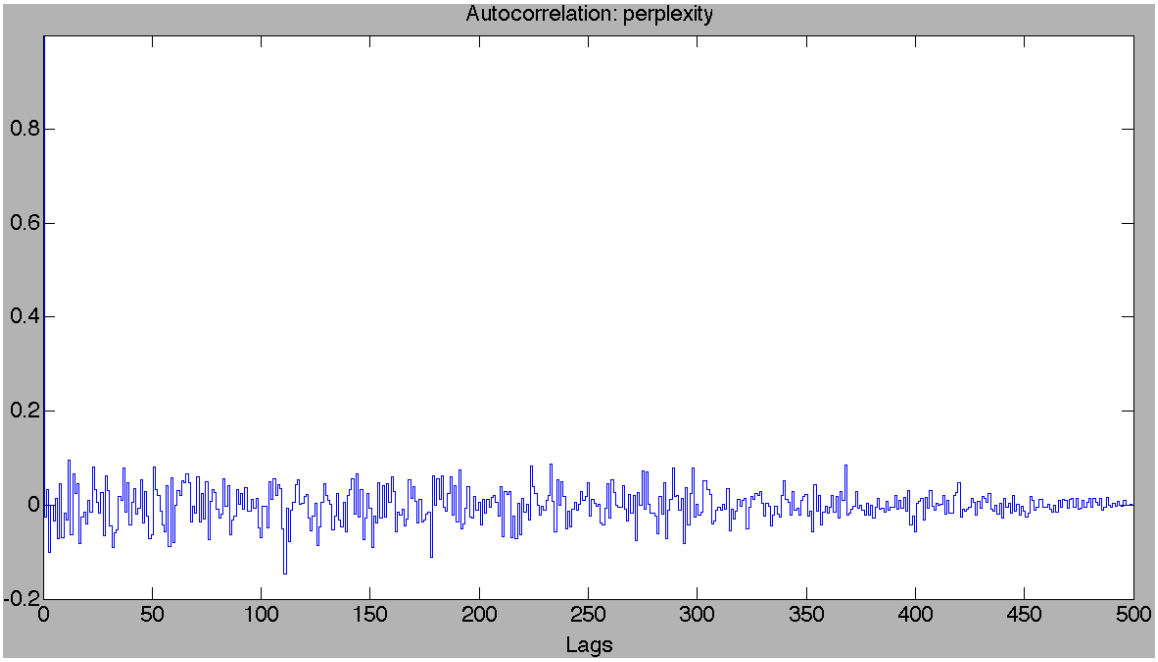


Figure 6.1. The autocorrelation plot of perplexity in the Gibbs Chain for iterations 10001-10500 of the ART Model on the McCallum dataset. Y-axis is the autocorrelations for perplexity at varying time lags. The randomness of Gibbs samples is ascertained by near-zero autocorrelations for any and all time-lag separations.

separations, i.e., the chain is mixed very well. Furthermore, we check the perplexity around 20,000 iterations and do not find significant difference in perplexity compared to 10,000 iterations. A useful future direction would be to utilize the "Gelman-Rubin" convergence test [19] to analyze the Gibbs chain from a different perspective.

We identify three areas of future work:

- *More conditional dependencies.* For simplicity, in this thesis, we limited the number of conditional dependencies to two, and we have demonstrated the usefulness of such a design in certain applications. However, three or more conditional dependencies do emerge in real world. For example, When consumers shop online, they are not just looking for attractive websites and easy navigation, they are looking for content. They want to search for the products that interest them, compare them to similar products, and ensure that they have all

the details they need to make informed buying decisions. Product specifications are one of the most important information in online catalogs for retailers.

In their specifications, products have different attributes and/or values for these attributes. Many attributes are pretty common across different categories, such as *size* and *weight*. On the other hand, some attributes are somewhat category specific, such as *max shutter speed*. In another word, the co-occurrence patterns of attributes are dramatically different across different categories.

At the same time, words in attribute values also have co-occurrence patterns like in other kind of text documents. One could run a topic model without distinguishing attributes and their values, treating such pairs as an atomic textual description. However, obviously, treating products as atomic entities hinders the effectiveness of many applications such as demand forecasting, product recommendations, and product supplier selection. If we can utilize the rich information contained in these pairs, all these application can be significantly improved.

Furthermore, the co-occurrence patterns over words in attribute values are not flat across the whole product specifications, and instead, many patterns are attribute dependent. For example, *Mac*, *MS*, and *Windows* probably only appear with OS-related attributes.

We would like models that captures both of the co-occurrences of words and attributes. In one possible scenario, attributes and words in their values are all generated from a hidden topic which in turn comes from a category specific distribution. To capture the association of attributes and their values, the words in attribute values are generated from a topic and attribute dependent distribution. As we can see, words are not independent to the attributes give

the hidden topics, another conditional dependency arrow in the graphical model representation.

Similar ideas can be applied to research articles as well if we consider a section header as an attribute, and text within a section as its value. In addition to the language usage difference, we might be able to discover different writing styles for different subjects, such as biology and computer science.

- *More modalities.* Directed graphical models can be described as generative processes and thus enjoy modeling and computational benefits conferred from conditional independencies, such as simple sampling procedures. However, in many applications, the dependency between two random variables in directed models can be difficult to describe and specify as a generative process and the direction of directed edges in the underlying graph can arguably be set either way. For example, when considering the authors and topics of documents, one can give reasonable arguments about either authors $\rightarrow$ topics or topics $\rightarrow$ authors. Particularly, when dealing with more modalities, the huge number of possible configurations of these directions between a large number of random variables have complicated the application of directed models to more complex multimodal, heterogeneous textual data.

  Furthermore, in state-of-the-art hierarchical Bayesian models such as LDA, exact posterior inference over hidden topic variables and parameters is typically intractable and approximate inference techniques such as variational methods [27], Gibbs sampling [4] and expectation propagation [42] are employed to address these issues. As a result, the inference for obtaining a topic decomposition for a previously unseen document can be slow and troublesome.

  Recently, a class of structured *undirected* latent variable models have gained attention for topic modeling – largely due to the fact that once model parameters

have been optimized, inference of hidden topics for a new document has the complexity of a matrix multiplication, which is fast compared to hierarchical Bayesian models.

Several pieces of work in this direction rely on a two-layer structure [18, 64, 68, 72, 73] that has an important property: the random variables at the two layers are conditionally independent given each other, which provides the property that the mapping from one layer to the other layer can be done by a simple matrix multiplication (and possibly some trivial follow-up transformations). However, there is no free lunch. The faster inference leads to more difficult learning due to the intractable normalizing constant in these types of undirected models. Fortunately, the contrastive divergence [23] approach has been shown to be efficient for inference and effective for learning in these models. Further and more importantly, in many situations involving document processing, training can be done off-line, which gives us more freedom in learning.

Undirected models of this structure have another important property that directed models lack: a more accurate characterization of rare words. As discussed in [72], in directed models such as latent Dirichlet allocation, a word is always generated from a single topic. When its count is low, this behavior becomes a very strong assumption or limitation. In the harmonium-structured models, a word arises from a distribution influenced by all the topics. This different mechanism might play a crucial role in certain applications.

- *Faster inference.* As we face more and more data in applications such as Web search, speeding up the inference procedure in topic models is in great need.

  One kind of the accelerated procedures slices up the sampling probability in different ways [48]. For any particular word and document, the distributions we want to draw sample from are often skewed such that most of the probability

mass is concentrated on a few topics, which leads to a possibility that on average only a small fraction of the topic probabilities need to be actually computed. With this principle in hand, a bound and refine procedure could be depicted.

Similarly, sparsity could be utilized to limit word inclusion on per-topic basis. For example, for the topic "Panama Canal", we do not care about the probabilities for the words "Enron" or "LDA". However, the plain model insists on probabilities for every word in every topic, no mater how they are irrelevant. Also, tricks such as storing pre-computed statistics could certainly help to some extent.

Another direction is parallelism. Due to the intrinsic sequential nature of inference procedures such as Gibbs sample, standard map-reduce could not be directly applied here. Nevertheless, limited parallelism has been demonstrated to be effective [44, 45].

APPENDIX A

COLLAPSED GIBBS SAMPLING DERIVATION FOR ART

We need to derive $P(x_{di}, z_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r})$, the conditional distribution of a topic and recipient for the word w_{di} given all other words' topic and recipient assignments, $\mathbf{x}_{-di}$ and $\mathbf{z}_{-di}$, to carry out the collapsed Gibbs sampling procedure for ART. We begin with the joint probability of the whole dataset. Note here that we can take advantage of conjugate priors to simplify the integrals.

$$
\begin{aligned}
& P(\mathbf{x}, \mathbf{z}, \mathbf{w}|\alpha, \beta, \mathbf{a}, \mathbf{r}) \\
= & \iint \prod_{i=1}^{A}\prod_{j=1}^{A} p(\theta_{ij}|\alpha) \prod_{t=1}^{T} p(\phi_t|\beta) \prod_{d=1}^{D}\prod_{i=1}^{N_d} P(x_{di}|\mathbf{r}_d) \cdot P(z_{di}|\theta_{a_d x_{di}}) P(w_{di}|\phi_{z_{di}}) \mathrm{d}\Phi \mathrm{d}\Theta \\
= & \prod_{d=1}^{D} (\frac{1}{|\mathbf{r}_d|})^{N_d} \int \prod_{i=1}^{A}\prod_{j=1}^{A} \left(\frac{\Gamma(\sum_{t=1}^{T}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_t)} \prod_{t=1}^{T} \theta_{ijt}^{\alpha_t - 1} \right) \prod_{i=1}^{A}\prod_{j=1}^{A}\prod_{t=1}^{T} \theta_{ijt}^{n_{ijt}} \mathrm{d}\Theta \\
& \times \int \prod_{t=1}^{T} \left(\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)} \prod_{v=1}^{V} \phi_{tv}^{\beta_v - 1} \right) \prod_{t=1}^{T}\prod_{v=1}^{V} \phi_{tv}^{m_{tv}} \mathrm{d}\Phi \\
\propto & \prod_{i=1}^{A}\prod_{j=1}^{A} \int \prod_{t=1}^{T} \theta_{ijt}^{\alpha_t + n_{ijt} - 1} \mathrm{d}\theta_{ij} \prod_{t=1}^{T} \int \prod_{v=1}^{V} \phi_{tv}^{\beta_v + m_{tv} - 1} \mathrm{d}\phi_t \\
\propto & \prod_{i=1}^{A}\prod_{j=1}^{A} \frac{\prod_{t=1}^{T}\Gamma(\alpha_t + n_{ijt})}{\Gamma(\sum_{t=1}^{T}(\alpha_t + n_{ijt}))} \prod_{t=1}^{T} \frac{\prod_{v=1}^{V}\Gamma(\beta_v + m_{tv})}{\Gamma(\sum_{v=1}^{V}(\beta_v + m_{tv}))}
\end{aligned}
$$

where $|\mathbf{r}_d|$ is the number of recipients in message d, n_{ijt} is the number of tokens assigned to topic t and the author-recipient pair (i, j), and m_{tv} represent the number of tokens of word v assigned to topic t.

Using the chain rule, we can obtain the conditional probability conveniently. We define $\mathbf{w}_{-di}$ as all word tokens except the token w_{di}.

$$
\begin{aligned}
P(x_{di}, z_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r}) &= \frac{P(x_{di}, z_{di}, w_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}_{-di}, \alpha, \beta, \mathbf{a}, \mathbf{r})}{P(w_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}_{-di}, \alpha, \beta, \mathbf{a}, \mathbf{r})} \\
\propto \quad \frac{P(\mathbf{x}, \mathbf{z}, \mathbf{w}|\alpha, \beta, \mathbf{a}, \mathbf{r})}{P(\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}_{-di}|\alpha, \beta, \mathbf{a}, \mathbf{r})} &\propto \frac{\frac{\Gamma(\alpha_{z_{di}}+n_{a_d x_{di} z_{di}})}{\Gamma(\alpha_{z_{di}}+n_{a_d x_{di} z_{di}}-1)}}{\frac{\Gamma(\sum_{t=1}^{T}(\alpha_t+n_{a_d x_{di} t}))}{\Gamma(\sum_{t=1}^{T}(\alpha_t+n_{a_d x_{di} t})-1)}} \frac{\frac{\Gamma(\beta_{w_{di}}+m_{z_{di} w_{di}})}{\Gamma(\beta_{w_{di}}+m_{z_{di} w_{di}}-1)}}{\frac{\Gamma(\sum_{v=1}^{V}(\beta_v+m_{z_{di} v}))}{\Gamma(\sum_{v=1}^{V}(\beta_v+m_{z_{di} v})-1)}} \\
\propto \quad \frac{\alpha_{z_{di}} + n_{a_d x_{di} z_{di}} - 1}{\sum_{t=1}^{T}(\alpha_t + n_{a_d x_{di} t}) - 1} &\frac{\beta_{w_{di}} + m_{z_{di} w_{di}} - 1}{\sum_{v=1}^{V}(\beta_v + m_{z_{di} v}) - 1} \\
\propto \quad \frac{\alpha_{z_{di}} + n'_{a_d x_{di} z_{di}}}{\sum_{t=1}^{T}(\alpha_t + n'_{a_d x_{di} t})} &\frac{\beta_{w_{di}} + m'_{z_{di} w_{di}}}{\sum_{v=1}^{V}(\beta_v + m'_{z_{di} v})}
\end{aligned}
$$

In the above, for simplicity, we redefine n and m as n' and m', respectively, to exclude the assignments of token w_{di}. If one wants, further manipulation can turn the above formula into separated update equations for the topic and recipient of each token, suitable for random or systematic scan updates:

$$
\begin{aligned}
P(x_{di}|\mathbf{x}_{-di}, \mathbf{z}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r}) &\propto \frac{\alpha_{z_{di}} + n'_{a_d x_{di} z_{di}}}{\sum_{t=1}^{T}(\alpha_t + n'_{a_d x_{di} t})} \\
P(z_{di}|\mathbf{x}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r}) &\propto \frac{\alpha_{z_{di}} + n'_{a_d x_{di} z_{di}}}{\sum_{t=1}^{T}(\alpha_t + n'_{a_d x_{di} t})} \frac{\beta_{w_{di}} + m'_{z_{di} w_{di}}}{\sum_{v=1}^{V}(\beta_v + m'_{z_{di} v})}
\end{aligned}
$$

APPENDIX B

COLLAPSED GIBBS SAMPLING DERIVATION FOR GT

Begin with the joint distribution $P(\mathbf{g}, \mathbf{V}, \mathbf{w}, \mathbf{t}|\alpha, \beta, \eta)$, we can take the advantages of conjugate priors to simplify the formulae. All symbols are defined in Sec. 3.1.

$$
\begin{aligned}
& P(\mathbf{g}, \mathbf{V}, \mathbf{w}, \mathbf{t}|\alpha, \beta, \eta) = \iiint p(\mathbf{g}, \mathbf{V}, \mathbf{w}, \mathbf{t}, \theta, \gamma, \phi|\alpha, \beta, \eta)\mathrm{d}\theta\mathrm{d}\gamma\mathrm{d}\phi \\
= & \iiint \prod_{b=1}^{B} P(t_b) \prod_{t=1}^{T} \left(p(\theta_t|\alpha) \prod_{s=1}^{S} P(g_{st}|\theta_t) p(\phi_t|\eta) \right) \\
& \times \prod_{b=1}^{B} \prod_{g=1}^{G} \prod_{h=g}^{G} p(\gamma_{gh}^{(b)}|\beta) \prod_{b=1}^{B} \prod_{i=1}^{N_b} P(w_i^{(b)}|\phi_{t_b}) \times \prod_{b=1}^{B} \prod_{i=1}^{S} \prod_{j=i+1}^{S} P(V_{ij}^{(b)}|\gamma_{g_i g_j}^{(b)}))\mathrm{d}\theta\mathrm{d}\gamma\mathrm{d}\phi \\
= & \iiint \left(\frac{1}{T}\right)^{B} \prod_{t=1}^{T} \left(\frac{\Gamma(\sum_{g=1}^{G} \alpha_g)}{\prod_{g=1}^{G} \Gamma(\alpha_g)} \prod_{g=1}^{G} \theta_{tg}^{\alpha_g - 1} \prod_{g=1}^{G} \theta_{tg}^{n_{tg}} \right) \times \prod_{t=1}^{T} \left(\frac{\Gamma(\sum_{v=1}^{V} \eta_v)}{\prod_{v=1}^{V} \Gamma(\eta_v)} \prod_{v=1}^{V} \phi_{tv}^{\eta_v - 1} \right) \\
& \times \prod_{b=1}^{B} \prod_{g=1}^{G} \prod_{h=g}^{G} \left(\frac{\Gamma(\sum_{k=1}^{2} \beta_k)}{\prod_{k=1}^{2} \Gamma(\beta_k)} \prod_{k=1}^{2} (\gamma_{ghk}^{(b)})^{\beta_k - 1} \right) \times \prod_{t=1}^{T} \prod_{v=1}^{V} \phi_{tv}^{c_{tv}} \prod_{b=1}^{B} \prod_{g=1}^{G} \prod_{h=g}^{G} \prod_{k=1}^{2} (\gamma_{ghk}^{(b)})^{m_{ghk}^{(b)}} \mathrm{d}\theta\mathrm{d}\gamma\mathrm{d}\phi \\
\propto & \iiint \prod_{t=1}^{T} \prod_{g=1}^{G} \theta_{tg}^{\alpha_g + n_{tg} - 1} \prod_{t=1}^{T} \prod_{v=1}^{V} \phi_{tv}^{\eta_v + c_{tv} - 1} \times \prod_{b=1}^{B} \prod_{g=1}^{G} \prod_{h=g}^{G} \prod_{k=1}^{2} (\gamma_{ghk}^{(b)})^{\beta_k + m_{ghk}^{(b)} - 1} \mathrm{d}\theta\mathrm{d}\gamma\mathrm{d}\phi \\
\propto & \prod_{t=1}^{T} \left(\frac{\prod_{g=1}^{G} \Gamma(\alpha_g + n_{tg})}{\Gamma(\sum_{g=1}^{G} (\alpha_g + n_{tg}))} \frac{\prod_{v=1}^{V} \Gamma(\eta_v + c_{tv})}{\Gamma(\sum_{v=1}^{V} (\eta_v + c_{tv}))} \right) \times \prod_{b=1}^{B} \prod_{g=1}^{G} \prod_{h=g}^{G} \frac{\prod_{k=1}^{2} \Gamma(\beta_k + m_{ghk}^{(b)})}{\Gamma(\sum_{k=1}^{2} (\beta_k + m_{ghk}^{(b)}))}
\end{aligned}
$$

Using the chain rule, we can get the conditional probability conveniently,

$$
\begin{aligned}
& P(g_{st}|\mathbf{V}, \mathbf{g}_{-st}, \mathbf{w}, \mathbf{t}, \alpha, \beta, \eta) = \frac{P(\mathbf{g}, \mathbf{V}, \mathbf{w}, \mathbf{t}|\alpha, \beta, \eta)}{P(\mathbf{g}_{-st}, \mathbf{V}, \mathbf{w}, \mathbf{t}|\alpha, \beta, \eta)} \propto \frac{P(\mathbf{g}, \mathbf{V}, \mathbf{w}, \mathbf{t}|\alpha, \beta, \eta)}{P(\mathbf{g}_{-st}, \mathbf{V}_{-st}, \mathbf{w}, \mathbf{t}|\alpha, \beta, \eta)} \\
\propto & \frac{\alpha_{g_{st}} + n_{tg_{st}} - 1}{\sum_{g=1}^{G} (\alpha_g + n_{tg}) - 1} \prod_{b=1}^{B} \left(I(t_b = t) \times \prod_{h=1}^{G} \frac{\prod_{k=1}^{2} \prod_{x=1}^{d_{g_{st}hk}^{(b)}} \left(\beta_k + m_{g_{st}hk}^{(b)} - x \right)}{\prod_{x=1}^{\sum_{k=1}^{2} d_{g_{st}hk}^{(b)}} \left((\sum_{k=1}^{2} (\beta_k + m_{g_{st}hk}^{(b)}) - x \right)} \right)
\end{aligned}
$$

and,

$$P(t_b|\mathbf{V},\mathbf{g},\mathbf{w},\mathbf{t}_{-b},\alpha,\beta,\eta) = \frac{P(\mathbf{g},\mathbf{V},\mathbf{w},\mathbf{t}|\alpha,\beta,\eta)}{P(\mathbf{g},\mathbf{V},\mathbf{w},\mathbf{t}_{-b}|\alpha,\beta,\eta)} \propto \frac{P(\mathbf{g},\mathbf{V},\mathbf{w},\mathbf{t}|\alpha,\beta,\eta)}{P(\mathbf{g},\mathbf{V}_{-b},\mathbf{w}_{-b},\mathbf{t}_{-b}|\alpha,\beta,\eta)}$$

$$\propto \frac{\prod_{v=1}^{V}\prod_{x=1}^{e_v^{(b)}}(\eta_v + c_{t_b v} - x)}{\prod_{x=1}^{\sum_{v=1}^{V} e_v^{(b)}}\left(\sum_{v=1}^{V}(\eta_v + c_{t_b v}) - x\right)} \times \prod_{g=1}^{G}\prod_{h=g}^{G}\frac{\prod_{k=1}^{2}\Gamma(\beta_k + m_{ghk}^{(b)})}{\Gamma(\sum_{k=1}^{2}(\beta_k + m_{ghk}^{(b)}))}$$

APPENDIX C

COLLAPSED GIBBS SAMPLING DERIVATION FOR TOT

We begin with the joint distribution $P(\mathbf{w}, \mathbf{t}, \mathbf{z}|\alpha, \beta, \Psi)$. We can take advantage of conjugate priors to simplify the integrals. All symbols are defined in Section 4.1.

$$
\begin{aligned}
& P(\mathbf{w}, \mathbf{t}, \mathbf{z}|\alpha, \beta, \Psi) = P(\mathbf{w}|\mathbf{z}, \beta)p(\mathbf{t}|\Psi, \mathbf{z})P(\mathbf{z}|\alpha) \\
= & \int P(\mathbf{w}|\Phi, \mathbf{z})p(\Phi|\beta)\mathrm{d}\Phi p(\mathbf{t}|\Psi, \mathbf{z}) \int P(\mathbf{z}|\Theta)p(\Theta|\alpha)\mathrm{d}\Theta \\
= & \int \prod_{d=1}^{D}\prod_{i=1}^{N_d} P(w_{di}|\phi_{z_{di}}) \prod_{z=1}^{T} p(\phi_z|\beta)\mathrm{d}\Phi \prod_{d=1}^{D}\prod_{i=1}^{N_d} p(t_{di}|\psi_{z_{di}}) \\
& \times \int \prod_{d=1}^{D} \left(\prod_{i=1}^{N_d} P(z_{di}|\theta_d)p(\theta_d|\alpha) \right) \mathrm{d}\Theta \\
= & \int \prod_{z=1}^{T}\prod_{v=1}^{V} \phi_{zv}^{n_{zv}} \prod_{z=1}^{T} \left(\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)} \prod_{v=1}^{V} \phi_{zv}^{\beta_v - 1} \right) \mathrm{d}\Phi \\
& \times \int \prod_{d=1}^{D}\prod_{z=1}^{T} \theta_{dz}^{m_{dz}} \prod_{d=1}^{D} \left(\frac{\Gamma(\sum_{z=1}^{T}\alpha_z)}{\prod_{z=1}^{T}\Gamma(\alpha_z)} \prod_{z=1}^{T} \theta_{dz}^{\alpha_z - 1} \right) \mathrm{d}\Theta \times \prod_{d=1}^{D}\prod_{i=1}^{N_d} p(t_{di}|\psi_{z_{di}}) \\
= & \left(\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)} \right)^T \left(\frac{\Gamma(\sum_{z=1}^{T}\alpha_z)}{\prod_{z=1}^{T}\Gamma(\alpha_z)} \right)^D \prod_{d=1}^{D}\prod_{i=1}^{N_d} p(t_{di}|\psi_{z_{di}}) \\
& \times \prod_{z=1}^{T} \frac{\prod_{v=1}^{V}\Gamma(n_{zv} + \beta_v)}{\Gamma(\sum_{v=1}^{V}(n_{zv} + \beta_v))} \prod_{d=1}^{D} \frac{\prod_{z=1}^{T}\Gamma(m_{dz} + \alpha_z)}{\Gamma(\sum_{z=1}^{T}(m_{dz} + \alpha_z))}
\end{aligned}
$$

Using the chain rule, we can obtain the conditional probability conveniently,

$$
\begin{aligned}
& P(z_{di}|\mathbf{w},\mathbf{t},\mathbf{z}_{-di},\alpha,\beta,\Psi) = \frac{P(z_{di},w_{di},t_{di}|\mathbf{w}_{-di},\mathbf{t}_{-di},\mathbf{z}_{-di},\alpha,\beta,\Psi)}{P(w_{di},t_{di}|\mathbf{w}_{-di},\mathbf{t}_{-di},\mathbf{z}_{-di},\alpha,\beta,\Psi)} \\
\propto \quad & \frac{P(\mathbf{w},\mathbf{t},\mathbf{z}|\alpha,\beta,\Psi)}{P(\mathbf{w}_{-di},\mathbf{t}_{-di},\mathbf{z}_{-di}|\alpha,\beta,\Psi)} \propto \frac{n_{z_{di}w_{di}}+\beta_{w_{di}}-1}{\sum_{v=1}^{V}(n_{z_{di}v}+\beta_v)-1}(m_{dz_{di}}+\alpha_{z_{di}}-1)p(t_{di}|\psi_{z_{di}}) \\
\propto \quad & (m_{dz_{di}}+\alpha_{z_{di}}-1)\frac{n_{z_{di}w_{di}}+\beta_{w_{di}}-1}{\sum_{v=1}^{V}(n_{z_{di}v}+\beta_v)-1} \times \frac{(1-t_{di})^{\psi_{z_{di}1}-1}t_{di}^{\psi_{z_{di}2}-1}}{B(\psi_{z_{di}1},\psi_{z_{di}2})}
\end{aligned}
$$

In practice, the balancing hyperparameter often appears as an exponential power of the last term above. Since timestamps are drawn from continuous Beta distributions, sparsity is not a big problem for parameter estimation of Ψ. For simplicity, we update Ψ after each Gibbs sample by the method of moments, detailed as follows:

$$
\begin{aligned}
\hat{\psi}_{z1} &= \bar{t}_z\left(\frac{\bar{t}_z(1-\bar{t}_z)}{s_z^2}-1\right) \\
\hat{\psi}_{z2} &= (1-\bar{t}_z)\left(\frac{\bar{t}_z(1-\bar{t}_z)}{s_z^2}-1\right)
\end{aligned}
$$

where $\bar{t}_z$ and s_z^2 indicate the sample mean and the biased sample variance of the timestamps belonging to topic z, respectively.

APPENDIX D

COLLAPSED GIBBS SAMPLING DERIVATION FOR TNG

We begin with the joint distribution $P(\mathbf{w}, \mathbf{x}, \mathbf{z}|\alpha, \beta, \gamma, \delta)$. We can take advantage of conjugate priors to simplify the integrals. All symbols are defined in Section 5.1.

$$
\begin{aligned}
& P(\mathbf{w}, \mathbf{z}, \mathbf{x}|\alpha, \beta, \gamma, \delta) \\
= & \iiiint \prod_{d=1}^{D} \prod_{i=1}^{N_d} (P(w_i^{(d)}|x_i^{(d)}, \phi_{z_i^{(d)}}, \sigma_{z_i^{(d)} w_{i-1}^{(d)}}) P(x_i^{(d)}|\psi_{z_{i-1}^{(d)} w_{i-1}^{(d)}})) \\
& \prod_{z=1}^{T} \prod_{v=1}^{W} p(\sigma_{zv}|\delta) p(\psi_{zv}|\gamma) \mathrm{d}\Sigma \mathrm{d}\Psi \prod_{z=1}^{T} p(\phi_z|\beta) \mathrm{d}\Phi \int \prod_{d=1}^{D} \left(\prod_{i=1}^{N_d} P(z_i^{(d)}|\theta_d) p(\theta_d|\alpha) \right) \mathrm{d}\Theta \\
= & \int \prod_{z=1}^{T} \left(\prod_{v=1}^{W} \phi_{zv}^{n_{zv}} \frac{\Gamma(\sum_{v=1}^{W} \beta_v)}{\prod_{v=1}^{W} \Gamma(\beta_v)} \prod_{v=1}^{W} \phi_{zv}^{\beta_v - 1} \right) \mathrm{d}\Phi \\
& \times \int \prod_{z=1}^{T} \prod_{w=1}^{W} \left(\prod_{v=1}^{W} \sigma_{zwv}^{m_{zwv}} \frac{\Gamma(\sum_{v=1}^{W} \delta_v)}{\prod_{v=1}^{W} \Gamma(\delta_v)} \prod_{v=1}^{W} \sigma_{zwv}^{\delta_v - 1} \right) \mathrm{d}\Sigma \\
& \times \int \prod_{z=1}^{T} \prod_{w=1}^{W} \left(\prod_{k=0}^{1} \psi_{zwk}^{p_{zwk}} \frac{\Gamma(\sum_{k=0}^{1} \gamma_v)}{\prod_{k=0}^{1} \Gamma(\gamma_k)} \prod_{k=0}^{1} \psi_{zwk}^{\gamma_k - 1} \right) \mathrm{d}\Psi \\
& \times \int \prod_{d=1}^{D} \left(\prod_{z=1}^{T} \theta_{dz}^{q_{dz}} \frac{\Gamma(\sum_{z=1}^{T} \alpha_z)}{\prod_{z=1}^{T} \Gamma(\alpha_z)} \prod_{z=1}^{T} \theta_{dz}^{\alpha_z - 1} \right) \mathrm{d}\Theta \\
\propto & \prod_{z=1}^{T} \frac{\prod_{v=1}^{W} \Gamma(n_{zv} + \beta_v)}{\Gamma(\sum_{v=1}^{W} (n_{zv} + \beta_v))} \prod_{z=1}^{T} \prod_{w=1}^{W} \frac{\prod_{v=1}^{W} \Gamma(m_{zwv} + \delta_v)}{\Gamma(\sum_{v=1}^{W} (m_{zwv} + \delta_v))} \\
& \prod_{z=1}^{T} \prod_{w=1}^{W} \frac{\prod_{k=0}^{1} \Gamma(p_{zwk} + \gamma_k)}{\Gamma(\sum_{k=0}^{1} (p_{zwk} + \gamma_k))} \prod_{d=1}^{D} \frac{\prod_{z=1}^{T} \Gamma(q_{dz} + \alpha_z)}{\Gamma(\sum_{z=1}^{T} (q_{dz} + \alpha_z))}
\end{aligned}
$$

Using the chain rule and $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, we can obtain the conditional probability conveniently,

$$P(z_i^{(d)}, x_i^{(d)} | \mathbf{w}, \mathbf{z}_{-i}^{(d)}, \mathbf{x}_{-i}^{(d)}, \alpha, \beta, \gamma, \delta) = \frac{P(w_i^{(d)}, z_i^{(d)}, x_i^{(d)} | \mathbf{w}_{-i}^{(d)}, \mathbf{z}_{-i}^{(d)}, \mathbf{x}_{-i}^{(d)}, \alpha, \beta, \gamma, \delta)}{P(w_i^{(d)} | \mathbf{w}_{-i}^{(d)}, \mathbf{z}_{-i}^{(d)}, \mathbf{x}_{-i}^{(d)}, \alpha, \beta, \gamma, \delta)}$$

$$\propto \quad (\gamma_{x_i^{(d)}} + p_{z_{i-1}^{(d)} w_{i-1}^{(d)} x_i} - 1)(\alpha_{z_i^{(d)}} + q_{d z_i^{(d)}} - 1) \times \begin{cases} \frac{\beta_{w_i^{(d)}} + n_{z_i^{(d)} w_i^{(d)}} - 1}{\sum_{v=1}^{W} (\beta_v + n_{z_i^{(d)} v}) - 1} & \text{if } x_i^{(d)} = 0 \\ \frac{\delta_{w_i^{(d)}} + m_{z_i^{(d)} w_{i-1}^{(d)} w_i^{(d)}} - 1}{\sum_{v=1}^{W} (\delta_v + m_{z_i^{(d)} w_{i-1}^{(d)} v}) - 1} & \text{if } x_i^{(d)} = 1 \end{cases}$$

Or equivalently,

$$P(z_i^{(d)} | \mathbf{w}, \mathbf{z}_{-i}^{(d)}, \mathbf{x}, \alpha, \beta, \gamma, \delta)$$

$$\propto \quad (\alpha_{z_i^{(d)}} + q_{d z_i^{(d)}} - 1) \times \begin{cases} \frac{\beta_{w_i^{(d)}} + n_{z_i^{(d)} w_i^{(d)}} - 1}{\sum_{v=1}^{W} (\beta_v + n_{z_i^{(d)} v}) - 1} & \text{if } x_i^{(d)} = 0 \\ \frac{\delta_{w_i^{(d)}} + m_{z_i^{(d)} w_{i-1}^{(d)} w_i^{(d)}} - 1}{\sum_{v=1}^{W} (\delta_v + m_{z_i^{(d)} w_{i-1}^{(d)} v}) - 1} & \text{if } x_i^{(d)} = 1 \end{cases}$$

And,

$$P(x_i^{(d)} | \mathbf{w}, \mathbf{z}, \mathbf{x}_{-i}^{(d)}, \alpha, \beta, \gamma, \delta)$$

$$\propto \quad (\gamma_{x_i^{(d)}} + p_{z_{i-1}^{(d)} w_{i-1}^{(d)} x_i} - 1) \times \begin{cases} \frac{\beta_{w_i^{(d)}} + n_{z_i^{(d)} w_i^{(d)}} - 1}{\sum_{v=1}^{W} (\beta_v + n_{z_i^{(d)} v}) - 1} & \text{if } x_i^{(d)} = 0 \\ \frac{\delta_{w_i^{(d)}} + m_{z_i^{(d)} w_{i-1}^{(d)} w_i^{(d)}} - 1}{\sum_{v=1}^{W} (\delta_v + m_{z_i^{(d)} w_{i-1}^{(d)} v}) - 1} & \text{if } x_i^{(d)} = 1 \end{cases}$$

APPENDIX E

ALL 50 ART TOPICS FOR ENRON DATASET

Topic 1		**Topic 2**	
folder	0.2363	report	0.1029
synchronizing	0.2102	website	0.0859
offline	0.0890	published	0.0569
item	0.0770	named	0.0558
added	0.0446	viewing	0.0487
updated	0.0293	class	0.0361
deleted	0.0255	propt	0.0274
notes	0.0171	pm	0.0197
calendar	0.0151	call	0.0175
mailbox	0.0145	material	0.0175
debra.perlingiere@enron.com	0.2461	errol.mclaughlin@enron.com	0.0679
richard.sanders@enron.com		sara.shackleton@enron.com	
gerald.nemec@enron.com	0.2127	sara.shackleton@enron.com	0.0471
mark.taylor@enron.com		errol.mclaughlin@enron.com	
steven.kean@enron.com	0.1310	john.arnold@enron.com	0.0454
richard.sanders@enron.com		susan.scott@enron.com	
Topic 3		**Topic 4**	
trading	0.0596	system	0.0281
master	0.0318	existing	0.0266
financial	0.0306	agreement	0.0259
trade	0.0258	questions	0.0253
counterparty	0.0247	opportunity	0.0222
online	0.0226	facilities	0.0213
place	0.0178	capacity	0.0197
company	0.0167	compression	0.0191
database	0.0152	compressor	0.0184
credit	0.0150	basis	0.0172
tana.jones@enron.com	0.3471	chris.germany@enron.com	0.1460
mark.taylor@enron.com		scott.hendrickson@enron.com	
tana.jones@enron.com	0.0784	jeff.dasovich@enron.com	0.0641
sara.shackleton@enron.com		richard.shapiro@enron.com	
mark.taylor@enron.com	0.0497	drew.fossum@enron.com	0.0388
tana.jones@enron.com		steven.harris@enron.com	

Topic 5		Topic 6	
section	0.0299	texas	0.0738
party	0.0265	america	0.0681
language	0.0226	houston	0.0672
contract	0.0203	north	0.0661
date	0.0155	smith	0.0661
enron	0.0151	fax	0.0655
parties	0.0149	debra	0.0637
notice	0.0126	street	0.0634
days	0.0112	corp	0.0585
include	0.0111	phone	0.0542
mary.hain@enron.com	0.0549	debra.perlingiere@enron.com	0.1651
james.steffes@enron.com		dan.hyvl@enron.com	
jeff.dasovich@enron.com	0.0377	debra.perlingiere@enron.com	0.1359
richard.shapiro@enron.com		stacy.dickson@enron.com	
dan.hyvl@enron.com	0.0362	debra.perlingiere@enron.com	0.1037
kim.ward@enron.com		gerald.nemec@enron.com	

Topic 7		Topic 8	
bill	0.0439	time	0.0457
edison	0.0310	july	0.0408
assembly	0.0284	month	0.0375
senate	0.0238	june	0.0339
dwr	0.0206	total	0.0330
direct	0.0131	december	0.0302
access	0.0129	october	0.0299
legislature	0.0126	august	0.0282
committee	0.0116	amount	0.0279
vote	0.0114	payment	0.0277
jeff.dasovich@enron.com	0.2446	jeff.dasovich@enron.com	0.0330
james.steffes@enron.com		richard.shapiro@enron.com	
jeff.dasovich@enron.com	0.1851	phillip.allen@enron.com	0.0308
richard.shapiro@enron.com		john.lavorato@enron.com	
jeff.dasovich@enron.com	0.1048	jeff.dasovich@enron.com	0.0279
john.lavorato@enron.com		james.steffes@enron.com	

Topic 9		Topic 10	
deal	0.0493	numbers	0.0677
zone	0.0254	data	0.0661
fuel	0.0200	day	0.0251
ces	0.0178	monthly	0.0251
capacity	0.0176	show	0.0233
demand	0.0176	file	0.0222
storage	0.0137	question	0.0214
transport	0.0135	spreadsheet	0.0214
contract	0.0133	flow	0.0206
transco	0.0124	include	0.0198
chris.germany@enron.com	0.3337	jeff.dasovich@enron.com	0.0867
judy.townsend@enron.com		james.steffes@enron.com	
chris.germany@enron.com	0.1087	jeff.dasovich@enron.com	0.0330
scott.neal@enron.com		richard.shapiro@enron.com	
chris.germany@enron.com	0.0872	vince.kaminski@enron.com	0.0309
scott.hendrickson@enron.com		matt.smith@enron.com	

Topic 11		Topic 12	
information	0.0822	price	0.0623
access	0.0299	trading	0.0534
including	0.0232	market	0.0431
financial	0.0166	pricing	0.0289
software	0.0151	prices	0.0261
events	0.0143	physical	0.0236
systems	0.0143	energy	0.0203
investment	0.0143	index	0.0189
management	0.0143	commodity	0.0167
system	0.0140	offer	0.0158
vince.kaminski@enron.com	0.0919	louise.kitchen@enron.com	0.1821
fletcher.sturm@enron.com		lynn.blair@enron.com	
jeff.dasovich@enron.com	0.0713	shelley.corman@enron.com	0.0859
james.steffes@enron.com		vince.kaminski@enron.com	
fletcher.sturm@enron.com	0.0637	jeff.dasovich@enron.com	0.0531
richard.shapiro@enron.com		james.steffes@enron.com	

Topic 13		Topic 14	
ll	0.1243	business	0.0413
ve	0.0854	opportunity	0.0255
don	0.0548	commercial	0.0254
won	0.0208	unit	0.0207
week	0.0179	interest	0.0198
couple	0.0173	development	0.0181
thing	0.0158	support	0.0174
problem	0.0151	directly	0.0164
show	0.0141	participate	0.0158
didn	0.0140	position	0.0157
jeff.dasovich@enron.com	0.1548	vince.kaminski@enron.com	0.0916
james.steffes@enron.com		joe.parks@enron.com	
jeff.dasovich@enron.com	0.0920	jeff.dasovich@enron.com	0.0711
richard.shapiro@enron.com		richard.shapiro@enron.com	
jeff.dasovich@enron.com	0.0313	jeff.dasovich@enron.com	0.0690
steven.harris@enron.com		james.steffes@enron.com	

Topic 15		Topic 16	
agreement	0.0552	contract	0.0532
master	0.0515	dated	0.0319
corp	0.0437	executed	0.0304
north	0.0355	referenced	0.0270
america	0.0323	copy	0.0263
entity	0.0299	list	0.0250
executed	0.0296	received	0.0233
stephanie	0.0282	confidentiality	0.0220
received	0.0265	llc	0.0213
transactions	0.0242	copies	0.0209
tana.jones@enron.com	0.0726	tana.jones@enron.com	0.1691
susan.bailey@enron.com		mark.taylor@enron.com	
stephanie.panus@enron.com	0.0699	tana.jones@enron.com	0.1198
mark.taylor@enron.com		louise.kitchen@enron.com	
stephanie.panus@enron.com	0.0690	tana.jones@enron.com	0.1091
susan.bailey@enron.com		carol.clair@enron.com	

Topic 17		**Topic 18**	
attached	0.0742	gas	0.3183
agreement	0.0493	pipeline	0.0387
review	0.0340	storage	0.0324
questions	0.0257	natural	0.0280
draft	0.0245	week	0.0232
letter	0.0239	end	0.0192
comments	0.0207	supply	0.0190
copy	0.0165	summary	0.0181
revised	0.0161	set	0.0151
document	0.0156	demand	0.0132
gerald.nemec@enron.com	0.0737	jeff.dasovich@enron.com	0.0834
barry.tycholiz@enron.com		james.steffes@enron.com	
gerald.nemec@enron.com	0.0551	jeff.dasovich@enron.com	0.0454
mark.whitt@enron.com		richard.shapiro@enron.com	
barry.tycholiz@enron.com	0.0325	drew.fossum@enron.com	0.0297
gerald.nemec@enron.com		steven.harris@enron.com	

Topic 19		**Topic 20**	
employees	0.0453	cut	0.0523
working	0.0391	schedule	0.0487
team	0.0369	power	0.0398
year	0.0194	number	0.0260
join	0.0194	put	0.0201
hr	0.0147	transmission	0.0201
level	0.0144	tag	0.0174
compensation	0.0137	bert	0.0148
job	0.0137	sold	0.0141
bonus	0.0134	cuts	0.0141
sally.beck@enron.com	0.1343	albert.meyers@enron.com	0.1398
stacey.white@enron.com		bill.williams@enron.com	
vince.kaminski@enron.com	0.1078	albert.meyers@enron.com	0.0434
darrell.schoolcraft@enron.com		kate.symes@enron.com	
shelley.corman@enron.com	0.0337	bill.williams@enron.com	0.0391
lynn.blair@enron.com		albert.meyers@enron.com	

Topic 21		**Topic 22**	
tw	0.0502	bankruptcy	0.0287
rate	0.0272	money	0.0258
capacity	0.0266	million	0.0256
fuel	0.0260	years	0.0184
lindy	0.0161	court	0.0156
michelle	0.0160	committee	0.0138
contract	0.0144	order	0.0133
year	0.0132	companies	0.0131
cost	0.0122	spent	0.0116
transport	0.0114	told	0.0107
kevin.hyatt@enron.com	0.0969	jeff.dasovich@enron.com	0.3265
michelle.lokay@enron.com		richard.sanders@enron.com	
lindy.donoho@enron.com	0.0924	jeff.dasovich@enron.com	0.1477
steven.harris@enron.com		james.steffes@enron.com	
michelle.lokay@enron.com	0.0796	jeff.dasovich@enron.com	0.1299
kevin.hyatt@enron.com		richard.shapiro@enron.com	

Topic 23		**Topic 24**	
plant	0.0317	area	0.0265
oneok	0.0301	environmental	0.0223
meters	0.0236	project	0.0211
test	0.0220	construction	0.0181
dave	0.0193	south	0.0168
northern	0.0188	impact	0.0152
effect	0.0177	amount	0.0139
points	0.0177	john	0.0131
force	0.0172	existing	0.0122
hpl	0.0145	miles	0.0105
drew.fossum@enron.com	0.2175	larry.campbell@enron.com	0.2759
lynn.blair@enron.com		kevin.hyatt@enron.com	
jim.schwieger@enron.com	0.0414	larry.campbell@enron.com	0.1352
thomas.martin@enron.com		steven.harris@enron.com	
drew.fossum@enron.com	0.0376	jeff.dasovich@enron.com	0.0518
stanley.horton@enron.com		james.steffes@enron.com	

Topic 25		Topic 26	
don	0.0407	customers	0.0541
love	0.0327	rate	0.0436
night	0.0320	increase	0.0201
good	0.0268	rates	0.0193
work	0.0239	end	0.0191
play	0.0239	utility	0.0179
guys	0.0201	continue	0.0173
email	0.0192	decision	0.0172
great	0.0180	credit	0.0159
tonight	0.0144	tomorrow	0.0149
matthew.lenhart@enron.com	0.0869	jeff.dasovich@enron.com	0.3751
eric.bass@enron.com		james.steffes@enron.com	
eric.bass@enron.com	0.0772	james.steffes@enron.com	0.0831
matthew.lenhart@enron.com		jeff.dasovich@enron.com	
susan.scott@enron.com	0.0419	jeff.dasovich@enron.com	0.0776
monique.sanchez@enron.com		richard.shapiro@enron.com	

Topic 27		Topic 28	
day	0.0419	ferc	0.0851
friday	0.0418	iso	0.0350
morning	0.0369	information	0.0229
monday	0.0282	px	0.0195
office	0.0282	market	0.0179
wednesday	0.0267	attached	0.0170
tuesday	0.0261	filing	0.0157
time	0.0218	order	0.0141
good	0.0214	epmi	0.0138
thursday	0.0191	section	0.0138
jeff.dasovich@enron.com	0.0340	mary.hain@enron.com	0.2703
richard.shapiro@enron.com		james.steffes@enron.com	
jeff.dasovich@enron.com	0.0289	james.steffes@enron.com	0.0678
james.steffes@enron.com		richard.shapiro@enron.com	
carol.clair@enron.com	0.0175	mary.hain@enron.com	0.0644
mark.taylor@enron.com		jeff.dasovich@enron.com	

Topic 29		Topic 30	
request	0.0947	credit	0.0540
tana	0.0643	master	0.0529
pm	0.0380	agreement	0.0446
jones	0.0377	isda	0.0382
subject	0.0364	guaranty	0.0291
link	0.0361	form	0.0208
questions	0.0292	send	0.0175
cc	0.0292	swap	0.0171
requests	0.0177	copy	0.0168
click	0.0168	legal	0.0164
mark.taylor@enron.com	0.1157	tana.jones@enron.com	0.0767
tana.jones@enron.com		sara.shackleton@enron.com	
sara.shackleton@enron.com	0.0637	stephanie.panus@enron.com	0.0712
tana.jones@enron.com		sara.shackleton@enron.com	
lynn.blair@enron.com	0.0409	sara.shackleton@enron.com	0.0689
tana.jones@enron.com		susan.bailey@enron.com	

Topic 31		Topic 32	
president	0.0440	company	0.0963
vice	0.0340	skilling	0.0284
general	0.0311	people	0.0223
power	0.0240	lay	0.0217
counsel	0.0226	business	0.0168
esq	0.0214	stock	0.0157
reference	0.0196	houston	0.0140
executive	0.0185	world	0.0138
page	0.0148	profile	0.0129
houston	0.0137	change	0.0124
mark.taylor@enron.com	0.5327	shelley.corman@enron.com	0.3256
louise.kitchen@enron.com		lynn.blair@enron.com	
joe.stepenovitch@enron.com	0.0433	jeff.dasovich@enron.com	0.0925
don.baughman@enron.com		richard.shapiro@enron.com	
jeff.dasovich@enron.com	0.0107	vince.kaminski@enron.com	0.0509
richard.shapiro@enron.com		fletcher.sturm@enron.com	

Topic 33		Topic 34	
contracts	0.0464	operations	0.0321
puc	0.0452	team	0.0234
edison	0.0324	office	0.0173
costs	0.0179	list	0.0144
utility	0.0150	bob	0.0129
contract	0.0144	open	0.0126
lynch	0.0124	meeting	0.0107
pay	0.0123	gas	0.0107
purchases	0.0119	business	0.0106
deal	0.0118	houston	0.0099
jeff.dasovich@enron.com	0.2412	sally.beck@enron.com	0.2158
steven.kean@enron.com		louise.kitchen@enron.com	
jeff.dasovich@enron.com	0.2259	sally.beck@enron.com	0.0826
james.steffes@enron.com		john.lavorato@enron.com	
jeff.dasovich@enron.com	0.2101	sally.beck@enron.com	0.0530
richard.shapiro@enron.com		stacey.white@enron.com	

Topic 35		Topic 36	
risk	0.0857	system	0.1026
meeting	0.0331	make	0.0274
rac	0.0212	include	0.0270
management	0.0209	rights	0.0239
view	0.0154	based	0.0222
board	0.0148	production	0.0209
portfolio	0.0144	process	0.0205
systems	0.0138	offering	0.0191
review	0.0138	release	0.0181
capital	0.0132	opportunity	0.0178
vince.kaminski@enron.com	0.1383	jeff.dasovich@enron.com	0.2232
james.steffes@enron.com		richard.shapiro@enron.com	
rick.buy@enron.com	0.0841	chris.germany@enron.com	0.0759
david.delainey@enron.com		scott.hendrickson@enron.com	
rick.buy@enron.com	0.0619	jeff.dasovich@enron.com	0.0735
sally.beck@enron.com		james.steffes@enron.com	

Topic 37		Topic 38	
market	0.0567	utilities	0.0528
power	0.0563	governor	0.0462
price	0.0280	commission	0.0450
system	0.0206	state	0.0432
prices	0.0182	utility	0.0235
high	0.0124	rate	0.0226
based	0.0120	percent	0.0226
buy	0.0117	california	0.0197
customers	0.0110	crisis	0.0153
costs	0.0106	gov	0.0127
jeff.dasovich@enron.com	0.1231	jeff.dasovich@enron.com	0.3040
james.steffes@enron.com		richard.shapiro@enron.com	
jeff.dasovich@enron.com	0.1133	jeff.dasovich@enron.com	0.2979
richard.shapiro@enron.com		james.steffes@enron.com	
mark.taylor@enron.com	0.0218	jeff.dasovich@enron.com	0.1030
elizabeth.sager@enron.com		richard.sanders@enron.com	

Topic 39		Topic 40	
var	0.0514	issues	0.0641
greg	0.0465	process	0.0443
position	0.0294	group	0.0369
jan	0.0257	discuss	0.0319
million	0.0224	project	0.0310
interest	0.0216	plan	0.0305
current	0.0163	work	0.0283
resume	0.0159	additional	0.0224
positions	0.0147	projects	0.0148
interview	0.0106	update	0.0134
john.lavorato@enron.com	0.0861	jeff.dasovich@enron.com	0.0700
john.arnold@enron.com		james.steffes@enron.com	
john.lavorato@enron.com	0.0355	vince.kaminski@enron.com	0.0528
rick.buy@enron.com		jeffrey.shankman@enron.com	
john.lavorato@enron.com	0.0273	richard.shapiro@enron.com	0.0298
greg.whalley@enron.com		steven.kean@enron.com	

Topic 41		Topic 42	
state	0.0404	blackberry	0.0726
california	0.0367	net	0.0557
power	0.0337	www	0.0409
energy	0.0239	website	0.0375
electricity	0.0203	report	0.0373
davis	0.0183	wireless	0.0364
utilities	0.0158	handheld	0.0362
commission	0.0136	stan	0.0282
governor	0.0132	fyi	0.0271
prices	0.0089	named	0.0260
jeff.dasovich@enron.com	0.3338	rod.hayslett@enron.com	0.1432
richard.shapiro@enron.com		tracy.geaccone@enron.com	
jeff.dasovich@enron.com	0.2440	tracy.geaccone@enron.com	0.0737
james.steffes@enron.com		rod.hayslett@enron.com	
jeff.dasovich@enron.com	0.1394	rod.hayslett@enron.com	0.0420
richard.sanders@enron.com		drew.fossum@enron.com	

Topic 43		Topic 44	
list	0.0767	business	0.0476
change	0.0494	presentation	0.0308
make	0.0486	businesses	0.0285
people	0.0454	document	0.0256
give	0.0390	mike	0.0250
put	0.0343	goals	0.0221
person	0.0338	specific	0.0180
send	0.0312	activities	0.0174
add	0.0262	discuss	0.0157
wanted	0.0258	information	0.0151
tana.jones@enron.com	0.0460	mike.mcconnell@enron.com	0.2555
mark.taylor@enron.com		jeffrey.shankman@enron.com	
jeff.dasovich@enron.com	0.0361	mike.mcconnell@enron.com	0.0929
richard.shapiro@enron.com		greg.whalley@enron.com	
james.steffes@enron.com	0.0279	mike.mcconnell@enron.com	0.0662
jeff.dasovich@enron.com		steven.kean@enron.com	

Topic 45		Topic 46	
game	0.0170	expense	0.0732
draft	0.0156	report	0.0708
week	0.0135	phone	0.0614
team	0.0135	info	0.0555
eric	0.0130	gerald	0.0508
make	0.0125	personal	0.0496
free	0.0107	list	0.0449
year	0.0106	vacation	0.0401
pick	0.0097	nemec	0.0331
phillip	0.0095	travel	0.0295
eric.bass@enron.com	0.3050	gerald.nemec@enron.com	0.4156
matthew.lenhart@enron.com		lynn.blair@enron.com	
eric.bass@enron.com	0.0780	john.hodge@enron.com	0.0248
phillip.love@enron.com		gerald.nemec@enron.com	
matt.motley@enron.com	0.0522	dan.hyvl@enron.com	0.0236
mike.grigsby@enron.com		stacy.dickson@enron.com	

Topic 47		Topic 48	
group	0.0379	oneok	0.0715
number	0.0326	meters	0.0443
process	0.0245	test	0.0318
people	0.0203	plant	0.0284
back	0.0175	majeure	0.0284
moving	0.0172	force	0.0227
start	0.0165	quality	0.0227
move	0.0161	northern	0.0227
plan	0.0158	special	0.0216
note	0.0158	measurement	0.0216
louise.kitchen@enron.com	0.1739	drew.fossum@enron.com	0.4767
john.lavorato@enron.com		lynn.blair@enron.com	
louise.kitchen@enron.com	0.0372	drew.fossum@enron.com	0.0647
geoff.storey@enron.com		stanley.horton@enron.com	
vince.kaminski@enron.com	0.0365	carol.clair@enron.com	0.0250
sally.beck@enron.com		susan.bailey@enron.com	

Topic 49		Topic 50	
transwestern	0.0546	issue	0.0674
parties	0.0487	issues	0.0583
project	0.0308	long	0.0250
settlement	0.0298	based	0.0249
affiliates	0.0258	specific	0.0241
supply	0.0238	general	0.0227
igs	0.0209	problems	0.0179
fawcett	0.0179	due	0.0177
scott	0.0169	position	0.0174
adequate	0.0169	understand	0.0172
susan.scott@enron.com steven.harris@enron.com	0.3555	jeff.dasovich@enron.com james.steffes@enron.com	0.0683
jeff.dasovich@enron.com susan.scott@enron.com	0.1231	tana.jones@enron.com mark.taylor@enron.com	0.0651
susan.scott@enron.com jeff.dasovich@enron.com	0.0665	vince.kaminski@enron.com mark.taylor@enron.com	0.0398

BIBLIOGRAPHY

[1] Airoldi, Edoardo, Blei, David, Fienberg, Stephen, and Xing, Eric. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research 9* (2008), 1981–2014.

[2] Airoldi, Edoardo, Blei, David, Xing, Eric, and Fienberg, Stephen. A latent mixed-membership model for relational data. In *Proceedings of the 3rd International Workshop on Link Discovery* (2005), pp. 82–89.

[3] Albert, Réka, and Barabási, Albert-László. Statistical mechanics of complex networks. *Reviews of Modern Physics 74*, 1 (2002), 47–97.

[4] Andrieu, Christophe, de Freitas, Nando, Doucet, Arnaud, and Jordan, Michael. An introduction to MCMC for machine learning. *Machine Learning 50* (2003), 5–43.

[5] Beeferman, Doug, and Berger, Adam. Agglomerative clustering of a search engine query log. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2000), pp. 407–416.

[6] Bhattacharya, Indrajit, and Getoor, Lise. Deduplication and group detection using links. In *Proceedings of the ACM SIGKDD Workshop on Link Analysis and Group Detection* (2004).

[7] Blei, David, and Mcauliffe, Jon. Supervised topic models. In *Advances in Neural Information Processing Systems 20* (2008), pp. 121–128.

[8] Blei, David, Ng, Andrew, and Jordan, Michael. Latent Dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.

[9] Blei, David M., and Lafferty, John D. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* (2006), pp. 113–120.

[10] Box, George, and Jenkins, Gwilym. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

[11] Carley, Kathleen. A comparison of artificial and human organizations. *Journal of Economic Behavior and Organization 56* (1996), 175–191.

[12] Carlin, Brad, and Chib, Siddhartha. Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society B 57* (1995), 473–484.

[13] Chemudugunta, Chaitanya, Smyth, Padhraic, and Steyvers, Mark. Modeling general and specific aspects of documents with a probabilistic topic model. In *Advances in Neural Information Processing Systems 19* (2007), pp. 241–248.

[14] Cox, Gary, and Poole, Keith. On measuring the partisanship in roll-call voting: The U.S. House of Represenatatives, 1887-1999. *American Journal of Political Science 46*, 1 (2002), 477–489.

[15] Erosheva, Elena, Fienberg, Steve, and Lafferty, John. Mixed membership models of scientific publications. *Proceedings of the National Academy of Sciences 101(Suppl. 1)* (2004), 5220–5227.

[16] Evans, David, Ginther-Webster, Kimberly, Hart, Mary, Lefferts, Robert, and Monarch, Ira. Automatic indexing using selective NLP and first-order thesauri. In *Proceedings of Intelligent Multimedia Information Retrieval Systems and Management* (1991), pp. 624–643.

[17] Fagan, Joel. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science 40*, 2 (1989), 115–139.

[18] Gehler, Peter, Holub, Alex, and Welling, Max. The rate adapting Poisson model for information retrieval and object recognition. In *Proceedings of the 23rd International Conference on Machine Learning* (2006), pp. 337–344.

[19] Gelman, Andrew, and Rubin, Donald. Inference from iterative simulation using multiple sequences. *Statistical Science 7* (1992), 457–511.

[20] Griffiths, Thomas, and Steyvers, Mark. Finding scientific topics. *Proceedings of the National Academy of Sciences 101(suppl. 1)* (2004), 5228–5235.

[21] Griffiths, Thomas, Steyvers, Mark, Blei, David, and Tenenbaum, Joshua. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17* (2005), pp. 537–544.

[22] Griffiths, Thomas, Steyvers, Mark, and Tenenbaum, Joshua. Topics in semantic representation. *Psychological Review 114* (2007), 211–244.

[23] Hinton, Geoffrey. Training products of experts by minimizing contrastive divergence. *Neural Computation 14* (2002), 1771–1800.

[24] Hix, Simon, Noury, Abdul, and Roland, Gerard. Power to the parties: Cohesion and competition in the European Parliament, 1979-2001. *British Journal of Political Science 35*, 2 (2005), 209–234.

[25] Hofmann, Thomas. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning 42*, 1 (2001), 177–196.

[26] Jakulin, Aleks, and Buntine, Wray. Analyzing the US Senate in 2003: Similarities, networks, clusters and blocs. http://eprints.fri.uni-lj.si/146/, 2004.

[27] Jordan, Michael, Ghahramani, Zoubin, Jaakkola, Tommi, and Saul, Lawrence. An introduction to variational methods for graphical models. In *Proceedings of the NATO Advanced Study Institute on Learning in graphical models* (1998), pp. 105–161.

[28] Kemp, Charles, Griffiths, Thomas, and Tenenbaum, Joshua. Discovering latent classes in relational data. *MIT AI Memo 2004-019* (2004).

[29] Kemp, Charles, Tenenbaum, Joshua, Griffiths, Thomas, Yamada, Takeshi, and Ueda, Naonori. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence* (2006), pp. 381–388.

[30] Kubica, Jeremy, Moore, Andrew, Schneider, Jeff, and Yang, Yiming. Stochastic link and group detection. In *Proceedings of the 18th National Conference on Artificial Intelligence* (2002), pp. 798–804.

[31] Kumaraswamy, P. A generalized probability density function for double-bounded random processes. *Journal of Hydrology 46* (1980), 79–88.

[32] Kurihara, Kenichi, Kameya, Yoshitaka, and Sato, Taisuke. A frequency-based stochastic blockmodel. In *Proceedings of the Workshop on Information Based Induction Sciences* (2006).

[33] Lorrain, Francois, and White, Harrison. The structural equivalence of individuals in social networks. *Journal of Mathematical Sociology 1* (1971), 49–80.

[34] MacKay, David, and Peto, Linda. A hierarchical Dirichlet language model. *Natural Language Engineering 1*, 3 (1994), 1–19.

[35] Madsen, Rasmus, Kauchak, David, and Elkan, Charles. Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the 22nd International Conference on Machine Learning* (2005), pp. 489–498.

[36] Manning, Chris, and Schutze, Hinrich. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[37] McCallum, Andrew. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[38] McCallum, Andrew. Multi-label text classification with a mixture model trained by EM. In *Proceedings of the 16th National Conference on Artificial Intelligence Workshop on Text Learning* (1999).

[39] McCallum, Andrew, Corrada-Emanuel, Andres, and Wang, Xuerui. Topic and role discovery in social networks. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (2005), pp. 786–791.

[40] McCallum, Andrew, Wang, Xuerui, and Corrada-Emmanuel, Andres. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research 30* (2007), 249–272.

[41] Mimno, David, and McCallum, Andrew. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence* (2008), pp. 411–418.

[42] Minka, Thomas, and Lafferty, John. Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence* (2002), pp. 352–359.

[43] Mitra, Mandar, Buckley, Christopher, Singhal, Amit, and Cardie, Claire. An analysis of statistical and syntactic phrases. In *Proceedings of the 5th International RIAO Conference* (1997), pp. 200–214.

[44] Nallapati, Ramesh, Cohen, William, and Lafferty, John. Parallelized variational EM for latent Dirichlet allocation: An experimental evaluation of speed and scalability. In *Proceedings of the 7th IEEE International Conference on Data Mining Workshop on High Performance Data Mining* (2007), pp. 349–354.

[45] Newman, David, Asuncion, Arthur, Smyth, Padhraic, and Welling, Max. Distributed inference for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 20* (2008), pp. 1081–1088.

[46] Nowicki, Krzysztof, and Snijders, Tom. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association 96*, 455 (2001), 1077–1087.

[47] Pajala, Antti, Jakulin, Aleks, and Buntine, Wray. Parliamentary group and individual voting behavior in Finnish Parliamentin Year 2003: A group cohesion and voting similarity analysis. http://vanha.soc.utu.fi/valtio-oppi/mopi/misc/pajala_jakulin_buntine_vers_1.pdf, 2004.

[48] Porteous, Ian, Newman, David, Ihler, Alexander, Asuncion, Arthur, Smyth, Padhraic, and Welling, Max. Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2008), pp. 569–577.

[49] Rosen-Zvi, M., Griffiths, T., Smyth, P., and Steyvers, M. Learning author-topic models from text corpora. *ACM Transactions on Information System 27*, 3 (2009).

[50] Rosen-Zvi, Michal, Griffiths, Thomas, Steyvers, Mark, and Smyth, Padhraic. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (2004), pp. 487–494.

[51] Shetty, Jitesh, and Adibi, Jafar. The Enron email dataset database schema and brief statistical report. Tech. rep., Information Sciences Institute, 2004.

[52] Sparrow, Malcolm. The application of network analysis to criminal intelligence: an assessment of prospects. *Social Networks 13* (1991), 251–274.

[53] Steyvers, Mark, Smyth, Padraic, Rosen-Zvi, Michal, and Griffiths, Thomas. Probabilistic author-topic models for information discovery. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2004), pp. 306–315.

[54] Strzalkowski, Tomek. Natural language information retrieval. *Information Processing and Management 31*, 3 (1995), 397–417.

[55] Swan, Russell, and Allan, James. Automatic generation of overview timelines. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2000), pp. 49–56.

[56] Swan, Russell, and Jensen, David. Timemines: Constructing timelines with statistical models of word usage. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Text Mining* (2000), pp. 73–80.

[57] Teh, Yee Whye, Jordan, Michael, Beal, Matthew, and Blei, David. Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*, 476 (2006), 1566–1581.

[58] Voeten, Eric. Documenting votes in the UN General Assembly. http://home.gwu.edu/∼voeten/UNVoting.htm, 2003.

[59] Wallach, Hanna. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning* (2006), pp. 977–984.

[60] Wang, Xuerui, and McCallum, Andrew. Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2006), pp. 424–433.

[61] Wang, Xuerui, McCallum, Andrew, and Wei, Xing. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE International Conference on Data Mining* (2007), pp. 697–702.

[62] Wang, Xuerui, Mohanty, Natasha, and McCallum, Andrew. Group and topic discovery from relations and text. In *The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Link Discovery: Issues, Approaches and Applications* (2005), pp. 28–35.

[63] Wang, Xuerui, Mohanty, Natasha, and McCallum, Andrew. Group and topic discovery from relations and their attributes. In *Advances in Neural Information Processing Systems 18* (2006), pp. 1449–1456.

[64] Wang, Xuerui, Pal, Chris, and McCallum, Andrew. Generalized component analysis for text with heterogeneous attributes. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2007), pp. 794–803.

[65] Wasserman, Stanley, and Faust, Katherine. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[66] Watts, Duncan. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company, 2003.

[67] Wei, Xing, and Croft, W. Bruce. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval* (2006), pp. 178–185.

[68] Welling, Max, Rosen-Zvi, Michal, and Hinton, Geoffrey. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems 17* (2004), pp. 1481–1488.

[69] Wolfe, Alicia, and Jensen, David. Playing multiple roles: Discovering overlapping roles in social networks. In *Proceedings of the 21st International Conference on Machine Learning Workshop on Statistical Relational Learning and its Connections to Other Fields* (2004).

[70] Woodrow Denham, Chad McDaniel, and Atkins, John. Aranda and alyawarra kinship : A quantative argument for a double helix model. *American Ethnologist 6*, 1 (1979), 1–24.

[71] Wu, Fang, Huberman, Bernardo, Adamic, Lada, and Tyler, Joshua. Information flow in social groups. http://arXiv.org/abs/cond-mat/0305305, 2003.

[72] Xing, Eric, Yan, Rong, and Hauptmann, Alexander. Mining associated text and images with dual-wing harmoniums. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence* (2005), pp. 633–641.

[73] Yang, Jun, Liu, Yan, Xing, Eric, and Hauptmann, Alexander. Harmonium-based models for semantic video representation and classification. In *Proceedings of the Seventh SIAM International Conference on Data Mining* (2007), pp. 378–389.

[74] Zhai, Chengxiang, and Lafferty, John. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information System 22*, 2 (2004), 179–214.