TOPIC MODELS FOR TAGGED TEXT

by

Zhiqiang Ma

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2014

Approved by:

_____
Dr. Srinivas Akella

_____
Dr. Wenwen Dou

_____
Dr. Jianping Fan

_____
Dr. Jing Xiao

_____
Dr. Jiancheng Jiang

ABSTRACT

ZHIQIANG MA. Topic models for tagged text. (Under the direction of DR. SRINIVAS AKELLA)

Our world has been experiencing a dramatic and continually increasing growth of digital textual information. This phenomenon raises challenges in analyzing, understanding, organizing, and summarizing these large bodies of textual information. A large portion of the textual information contains meta-data, such as user-annotated tags, which provides useful information and could help improve the current text mining results. Thus, this thesis focuses on handling tagged text using topic modeling techniques.

We start from the Latent Dirichlet Allocation (LDA) model and introduce a Trivial Tag-Latent Dirichlet Allocation (TriTag-LDA) model, which directly connects the tags to the topics via an improved two-layer LDA model. Specifically, the bottom layer is the standard LDA, while the upper layer is a constrained LDA with the topics coming from the bottom layer. After that, we propose a new topic model, Tag-Latent Dirichlet Allocation (Tag-LDA), which more naturally integrates tags into the generative process. In Tag-LDA, a document is viewed as a mixture of tags rather than topics, and topics are generated from multinomial distributions under tags. TriTag-LDA and Tag-LDA bridge the user-generated tags and the latent topics. In both these models, a tag is described in the form of a mixture of shared topics. This representation enables the analysis of the relationships between tags. We provide quantitative and qualitative comparisons between our models and related work, and

show that Tag-LDA is superior under the perplexity criterion. We also apply Tag-LDA to explain hashtags on Twitter and discover their relationships.

We then develop two extensions of Tag-LDA: Tag-Dirichlet Processes (Tag-LDP) and Tag-Dirichlet Allocation with concepts (ConceptTag-LDA). Tag-LDP utilizes the Dirichlet process in modeling so that the number of topics can be decided automatically based on the data. Our experiments demonstrate that Tag-LDP can infer the number of topics from the data and that the quality of topics is as good as Tag-LDA. ConceptTag-LDA provides a mechanism where users' prior knowledge can be incorporated in learning the topics. Users' knowledge represented as pre-defined concepts is modeled through the Dirichlet Tree prior which replaces the original Dirichlet prior in Tag-LDA. Our experiments study the influence of the concepts on the topics, and demonstrate that the input concepts can influence the topics toward users' prior knowledge.

Finally we present the dynamic Twitter topic model (DTTM), a specialized temporal topic model tailored for the short messages in social media. On social media such as Twitter, people's discussions are constantly evolving with many discussions centering around events. A major event usually involves twists and turns reflected by multiple sub-events as it develops over time. This temporal event development is in turn reflected by people's discussions on Twitter. In DTTM, we assume an event can be modeled by a mainstream topic plus several facets and that each tweet is a mixture of two topics: the mainstream topic and one facet topic. To capture the temporal dynamics of the discussions, DTTM models the temporal evolution of the mainstream topic and the facet topics. To demonstrate the effectiveness of DTTM in

modeling the temporal dynamics of topics, we did two case studies with our model using Twitter data and show that our model performs better in summarizing the discussions than existing topic models.

ACKNOWLEDGMENTS

I thank my advisor Srinivas Akella. Without his diligent advising and support, this dissertation would not have been possible. For the past five years, Srinivas has generously spent a great amount of effort in helping me improve my research skills and solve the problems I met. I thank my dissertation committee, Wenwen Dou, Jianping Fan, Jing Xiao, and Jiancheng Jiang, for their helpful suggestions and comments. I particularly thank Wenwen Dou who collaborated with me and helped me improve my work. I also would like to thank all the faculty members who have taught or helped me during my study.

I am very lucky to have several great labmates at UNC Charlotte. At this time, I would like to express my sincere appreciation to them. They are, in alphabetical order, Jinglin Li, Junjie Shan, Yi Shen, Zhou Teng, Chunlei Yang, and Ning Zhou. They not only contributed constructive comments and suggestions to my work, but also enriched my daily life in Charlotte.

Thanks to the College of Computing and Informatics and Graduate School at UNC Charlotte for providing the generous financial support. Without the support, I could not finish my study.

I am incredible grateful for the life-long support from my parents and family. I have seldom been with them since I studied overseas, however they always support me. Finally, I deeply thank my wife, Yunfei. Her encouragement, patience, and love made all this possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Overview and Research Motivation

The world has been experiencing a dramatic growth of digital textual information, including news articles, digital books, and reports. In 2006, 1.35 million research publications were published with a yearly growth rate of 2.5% [76]. In addition, the prevalence of social media contributes significantly to the generation of unstructured text. For instance, 200 million tweets were sent out on Twitter every day by the middle of 2011, and this number doubled after two years in 2013[1]. Organizing, exploring, and summarizing these vast and fast growing text collections has become an important and challenging task.

Topic modeling provides us a way to model text corpora by discovering the statistical relationships of terms and "latent topics" that pervade the document collections [18]. The terms in the latent topics often reflect semantically meaningful subjects [1] that we call topics. The representation of text data by applying topic modeling is useful for a large variety of tasks, such as classification [18, 48, 90], summarization [59, 78, 73], and similarity measurement [65]. Popular topic modeling algorithms include early algorithms such as probabilistic latent semantic indexing (pLSI) [34] and more recent algorithms such as Latent Dirichlet Allocation (LDA) [18]. Compared

[1] http://articles.washingtonpost.com/2013-03-21/business/37889387_1_tweets-jack-dorsey-twitter

Table 1: Four topics extracted from 74 New York Times news articles related to the US presidential election.

Topic1		Topic2		Topic3		Topic4	
Terms	Prob.	Terms	Prob.	Terms	Prob.	Terms	Prob.
news	0.036	clinton	0.026	bush	0.048	percent	0.029
cbs	0.024	york	0.019	president	0.024	election	0.026
report	0.018	pataki	0.019	democrats	0.019	voters	0.015
documents	0.016	abortion	0.015	party	0.018	vote	0.013
panel	0.015	conservative	0.010	campaign	0.015	kerry	0.013
segment	0.015	nation	0.009	democratic	0.015	states	0.011
president	0.012	speech	0.008	senator	0.014	results	0.010
mapes	0.011	administration	0.007	presidential	0.013	ohio	0.010
broadcast	0.011	called	0.007	house	0.013	voting	0.009
national	0.009	faith	0.007	political	0.011	day	0.009
sept	0.009	times	0.007	republican	0.011	american	0.008
minutes	0.009	putin	0.007	dean	0.010	iraq	0.008
wednesday	0.008	america	0.006	republicans	0.010	elections	0.007
network	0.008	family	0.006	kerry	0.009	time	0.006
service	0.008	conservatives	0.006	security	0.009	won	0.006

to pLSI, LDA is a better defined and more complete generative model, so we build on LDA in this thesis. For example, given a set of 74 New York Times news articles about the US presidential election, four latent topics obtained from LDA are shown in Table 1. In each topic, terms are ranked based on their probabilities on the right.

A large fraction of textual information contains user-annotated meta-data. For example, blog writers apply tags to label what their blogs are about; online news websites adopt keywords to annotate the corresponding news articles. A snapshot of a news article from Yahoo is shown in Figure 1 indicating how the tags are used in practice[2]. The user-annotated data usually captures users' knowledge and provides a high-level summarization of the documents. In the above examples, blogs or news can be easily organized and retrieved based on the tags or keywords. In terms of

[2]http://news.yahoo.com/incredible-technology-supercomputers-solve-giant-problems-153719212.html

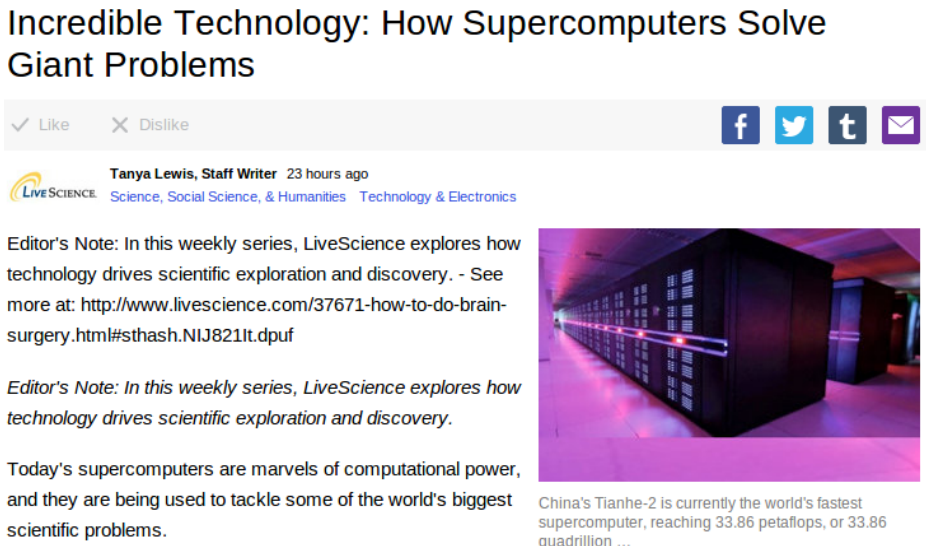


Figure 1: Snapshot of a news article from Yahoo news channel. Two tags "Science, Social Science, & Humanities" and "Technology & Electronics" are applied by Yahoo to categorize this article.

social media data, hashtags are widely used to tag what topic(s) or event(s) a tweet is related to. These tags provide extra information that could help us organize and summarize text corpora [89, 86, 39, 70, 87] and help us improve the text mining results. This thesis focuses on the problem of incorporating user-generated tags in topic models. Here the user-generated tags do not just strictly refer to the tags used in blogs, but more generally to the meta-data associated with the documents like labels, keywords, hashtags, etc.

1.1.1 Research Motivation

The motivation for our work originates from four questions:

1. How can we understand and interpret the user-generated tags?

2. Can we discover which words in a document should be attributed to which tags?

3. Can users' prior knowledge be incorporated in the data modeling procedure?

4. Can we summarize people's discussions in social media and also capture the temporal dynamics of the discussions?

So why are these four questions important to us? Let us take a look at social media as an example.

Social media such as Twitter captures moment-by-moment updates of discussions among people. To indicate the theme of the messages, people use a hashtag to label them. Hashtags, commonly used on Twitter and Google+, have become a unique tagging convention to organize social media content and associate events, trends, or topic information. According to Twitter, a hashtag is comprised of the symbol # followed by a sequence of keywords or phrases (without spaces) and is used to mark keywords or topics in tweets[3]. Over the years, the number of hashtags created and used has been on the rise. On the Twitter platform, for example, our observations suggest that one out of nine tweets now contains a hashtag. Such an elaborate tagging system establishes a bi-directional interaction between users and the online information. It enables the retrieval of all posts that include a specified hashtag, and it empowers users to follow conversations of interest.

The fact is that there are no restrictions on how a hashtag can be constructed, resulting in various lengths, forms, or structures of hashtags. Some of the existing hashtags are constructed in an intuitive manner, serving as meta-data to categorize what that post is about. For instance, #grammys and #ImmigrationStory, two trending hashtags in February 2013, denote the Grammy music awards and President Obama's pledge to share stories of immigration families to support his immigration

[3]https://support.twitter.com/articles/49309-what-are-hashtagssymbols

reforms, respectively. Other hashtags, however, are not as easy to make sense of. For instance, hashtag #NatGat (national gathering) and #tcot (Top Conservatives on Twitter) are difficult to decode by reading just the hashtag themselves. Similarly, hashtags, such as #Jan25, are also challenging to comprehend since they are too general to infer their significance without knowing the relevant context[4].

Ideally, one would like to maintain a clear relation structure for hashtags, with only a one-to-one relationship to the corresponding topics or events. However, in practice, this is not an option due to the creativity of the users. Given the ease and flexibility in creating hashtags, social media users can and often construct multiple hashtags for the same event or topic. For example, hashtags related to #MichaelJackson can also be seen in the form of #KingOfPop or #MJ; whereas #occupywallstreet and #OWS are both used to characterize the same event, but with different expressions. Sometimes, multiple hashtags are created to denote different aspects of a certain event. This is exemplified in the discussions of the Occupy movement on Twitter. Various hashtags are used to denote information about the who, what, when, and where in this movement. Specifically, such hashtags include #usdor denoting the organizing party of OWS, #sep17 denoting the date information about the Occupy movement, #occupywallstreet or #occupyChicago denoting where the protests occurred, and #pepperspray denoting a significant event in the movement.

Although the use of hashtags has become a convention, how well the users understand and use the hashtag information is still unclear. Sifting through trending hashtags on social media has become a popular way to learn what events have occurred,

[4]#Jan25 is used to indicate January 25, 2011, the date that Egyptian revolution began.

as shown in the up-to-the-minute trending topics listed on Twitter. The intrinsically polylingual, fragmented, unvetted, and dynamic nature of hashtags, however, also presents a disadvantage in depicting valuable information. Users can be overwhelmed with the noise of unrelated messages and conflicting information. Therefore, it is necessary for us to develop a solution that can help users effectively make sense of tags. One of our research goals is therefore to enable users to understand the meaning of the tags as well as the relationships between tags.

User-generated content on Twitter captures minute-by-minute updates of public and private snippets of information. Many of the discussions on Twitter center on events of interest to people, evolving rapidly over time. Using hashtags, the discussions centered around events can be easily located. To analyze and make sense of the wealth of information on Twitter, summarizing the user-generated content is a necessary step. More interesting, given the velocity of tweets, around 400 million tweets per day, it is beneficial to summarize the content in a way that highlights the ebb and flow of the moment-by-moment discussions. Therefore, it is essential to consider the temporal dynamics when summarizing and analyzing tweets.

Moreover, when modeling text data in LDA, the results only reflect the statistical relationship of terms and topics embedded in the data. It is also not easy to allow user customization of the models, like injecting users' prior knowledge into the term relationships. For example, the term "market" and the term "finance" might have similar probabilities under a given topic, since these two terms are semantically closely related and often appear together in the articles according to users' knowledge. Part of the reason is due to the unsupervised learning property of LDA. Besides that, asking

non-expert users to understand the mathematical principles underlying a model and learn to tune a model is quite overwhelming. In this thesis, we therefore also explore how to incorporate users' prior knowledge expressed as user-defined concepts in our topic model, such that the data can be modeled in a customizable way based on users' knowledge.

1.1.2 Thesis Contributions

The contribution of this thesis lies in developing statistical models for mining tagged text data, and in particular, focusing on applications of the models to social media data. The contributions of this thesis are listed below:

1. We present an approach to interpret user-generated tags using topics. We develop a new topic model, which views a tag as a distribution of topics. With our model, the meaning of the tags can be explained and the relationships between the tags can be discovered.

2. Inspired by recent work on Hierarchical Dirichlet Processes, we utilize Dirichlet processes as the priors to infer the number of topics from the data automatically.

3. To make the model customizable to non-expert users, we additionally extend our model to allow users' prior knowledge to be incorporated.

4. Finally, we propose a temporal topic model specially designed for short messages in social media to summarize the discussions in social media.

1.2 Background

This section is a brief overview of relevant statistics knowledge that is required in the subsequent discussion. We begin with the Dirichlet distribution and its conjugate distribution, the multinomial distribution (categorical distribution). Then we introduce collapsed Gibbs sampling, which is employed to learn the model. In the last section, the standard LDA is briefly introduced.

1.2.1 Dirichlet Distribution

The Dirichlet distribution is often used in Bayesian inference as a prior distribution to model the proportions of events [10]. Basically, the Dirichlet distribution is a multivariate generalization of the Beta distribution. Let $\boldsymbol{\theta}$ denote an $n > 2$ dimensional random variable defined in the $n-1$ simplex, so it implies $\sum_{i=1}^{n} \theta_i = 1$ and $\theta_i > 0$. In the Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, the probability density of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}) \sim \mathrm{Dir}(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}, \tag{1}$$

where every element in $\boldsymbol{\alpha}$ is a positive real number and $\Gamma(x)$ is the Gamma function. α_i is often considered as a "pseudocount" for each θ_i. The expectation of θ_i is given by [6]

$$E(\theta_i) = \frac{\alpha_i}{\sum_{j=1}^{n} \alpha_j}. \tag{2}$$

Figure 2 illustrates the support (log of probability density) of three 3-dimensional Dirichlet distributions with different parameters $\boldsymbol{\alpha}$ on the simplex. It can be observed that the density for $\alpha_i < 1$ concentrates near the vertices, while the density locates

somewhere inside the simplex when $\alpha_i \geq 1$.

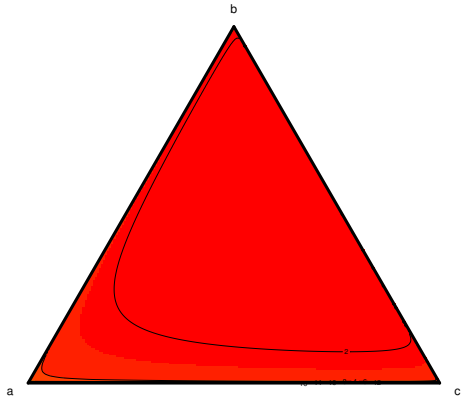
An important property of the Dirichlet distribution is its conjugacy to the multinomial distribution. The definition of conjugate prior (family) is described as "Let $\mathcal{F}$ denote the class of probability mass functions or probability density functions $f(x|\theta)$. A class Π of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \Pi$, and all $x \in \mathcal{X}$" [20]. We assume $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$ and $X|\boldsymbol{\theta} \sim \text{Multinomial}(\boldsymbol{\theta})$. Bayes' theorem gives us the relationship below

$$\begin{aligned} p(\boldsymbol{\theta}|x=j) &= \frac{p(x=j|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(x=j)} \\ &= \left(\frac{1}{p(x=j)}\right)\left(\frac{\Gamma(\sum_{i=1}^{n}\alpha_i)}{\prod_{i=1}^{n}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_j^{\alpha_j-1}\cdots\theta_n^{\alpha_n-1}\right)\theta_j \\ &= C\cdot(\theta_1^{\alpha_1-1}\cdots\theta_j^{\alpha_j}\cdots\theta_n^{\alpha_n-1}) \\ &\sim \text{Dir}(\alpha_1,\ldots,\alpha_j+1,\cdots,\alpha_n) \end{aligned} \tag{3}$$

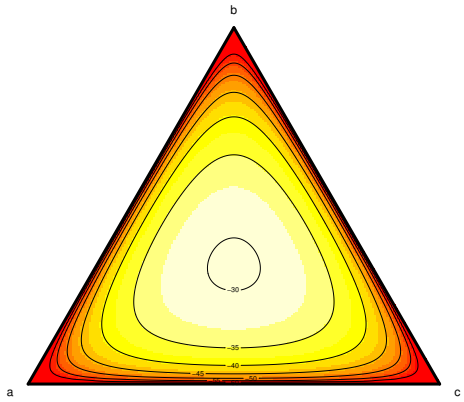
It is not difficult to derive $C = \frac{\Gamma(\sum_{i=1}^{n}\alpha_i+1)}{\prod_{i=1,j}^{n}\Gamma(\alpha_i)\Gamma(\alpha_j+1)}$ based on the property of the integral of probability density. Thus, Dirichlet and multinomial distributions are a conjugate prior pair. Equation 3 reflects that, with a Dirichlet prior, given a new observation from the multinomial distribution, the posterior distribution is still a Dirichlet with updated parameters.

1.2.2 Gibbs Sampling

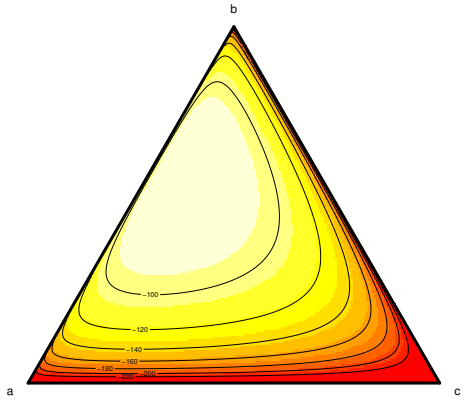
In Bayesian inference, it is often intractable to obtain the posterior due to the high dimensional integral. Markov Chain Monte Carlo (MCMC) is a widely used alternative approach to avoid the integration; MCMC approximates the distribution by constructing a Markov chain and using the samples after the Markov chain becomes



(a) $\boldsymbol{\alpha} = (0.6, 0.3, 0.9)$



(b) $\boldsymbol{\alpha} = (10, 10, 10)$



(c) $\boldsymbol{\alpha} = (10, 50, 30)$

Figure 2: The plots of the supports for three Dirichlet distributions with different parameters.

stationary [14]. In MCMC the next sample is generated based on the current sample, because the transition probability only depends on the variable's current state in a Markov process.

Gibbs sampling is a special case of MCMC, which was originally introduced for image processing [27]. The basic idea of Gibbs sampling, loosely speaking, is sampling each dimension of the random variable alternately while conditioned on the current observations of all other dimensions. This sampling scheme would construct a Markov chain that could lead to a stationary distribution after the "burn-in" period, as it fits the Metropolis-Hastings algorithm [30, 55]

When we use Gibbs sampling to approximate the posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$ given observation $\boldsymbol{x}$, the algorithm iteratively draws sample values for each dimension of $\boldsymbol{\theta}$ as below:

Algorithm 1 Gibbs sampler

initialization
repeat
  for $i \leftarrow 1, n$ **do**
    $\tilde{\theta}_i \sim p(\theta_i|\boldsymbol{\theta}_{-i}, \boldsymbol{x}) = \dfrac{p(\boldsymbol{\theta}, \boldsymbol{x})}{p(\boldsymbol{\theta}_{-i}, \boldsymbol{x})} = \dfrac{p(\boldsymbol{\theta}, \boldsymbol{x})}{\int p(\boldsymbol{\theta}, \boldsymbol{x}) d\theta_i}$
  end for
until end

In Algorithm 1, the subscript $-i$ indicates that the ith dimension is excluded. After the sampling sequence enters the stationary distribution and then obtains M samples $\tilde{\boldsymbol{\theta}}$, the posterior is estimated by [31]

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{1}{M} \sum_{i=1}^{M} \delta(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_{\boldsymbol{i}}), \tag{4}$$

where $\delta(\cdot)$ is the Kronecker delta.

1.2.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation [18] is a popular model in probabilistic topic modeling. LDA is a hierarchical Bayesian model that is used to model collections of text or other discrete data. Built on the "bag of words" assumption which ignores the order of words in the document, LDA assumes each word of a document is generated from one selected "latent" topic from a set of unobserved topics. In other words, a document can be viewed as a mixture of the latent topics, and each topic is a distribution over the vocabulary [72]. Figure 3 depicts the basic idea, with the shaded color of the words indicating the latent topics from which the words come.

LDA is a generative probabilistic model simulating the generative process of the documents. Like other topic models, LDA captures the statistical relationships among words, topics, and documents, such that the documents can be described and represented with the topics [18, 29]. Not only has LDA been successfully applied to perform classification, summarization, and similarity measurement on text data, but also to recommend tags for images [15], predict protein-protein relationships [5], and segment and classify objects in pictures [19].

LDA is a generative model, which means the model can describe how the document is being produced. To generate a document d containing N_d words $\boldsymbol{w} = \{w_1, w_2, \ldots, w_{N_d}\}$ for a corpus of M documents, the generative model views this document as a mixture of T topics following a multinomial distribution Multinomial($\boldsymbol{\theta}$), and their proportions $\boldsymbol{\theta}$ follow a Dirichlet prior Dir($\boldsymbol{\alpha}$). Each topic is also a multinomial distribution Multinomial($\boldsymbol{\beta}$) over the vocabulary. When generating each w_i,

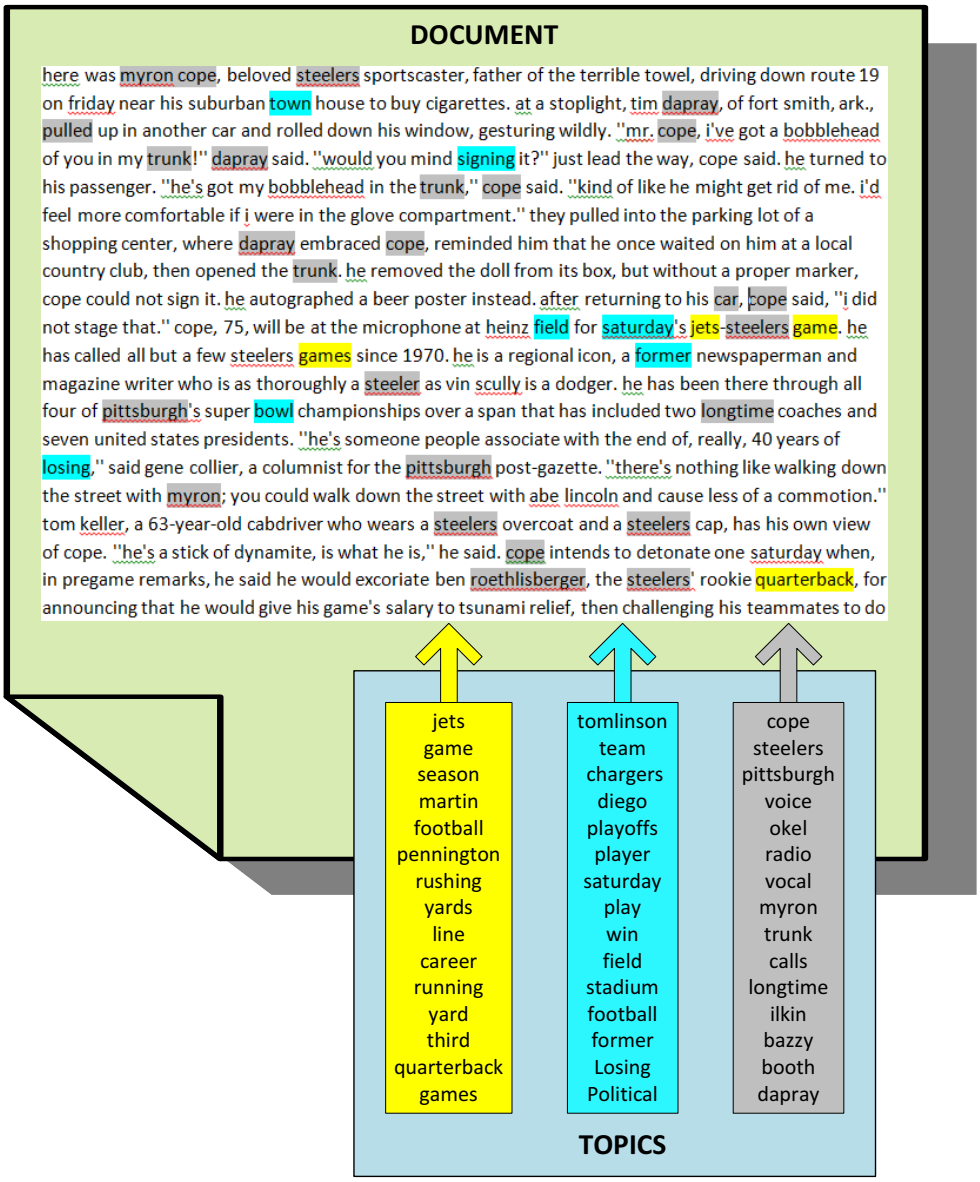


Figure 3: LDA views a document as a mixture of latent topics from which terms are chosen. The shaded color highlighting a word indicates which topic the word is selected from. The proportion of one topic is decided by the total counts of the terms selected from it in the document.

a sample topic z_i is first drawn from the Multinomial($\boldsymbol{\theta}$); then a term is chosen to be w_i according to the multinomial distribution Multinomial($\boldsymbol{\beta}$) of the z_i. Similarly, each $\boldsymbol{\beta}$ is drawn from a Dirichlet prior Dir($\boldsymbol{\phi}$). The details of the generative steps are listed below:

1. For each topic $t = \{1, 2, \ldots, T\}$, sample a topic-word multinomial distribution $\boldsymbol{\beta}_t$ over vocabulary from Dir($\boldsymbol{\phi}$).

2. For each document d_j:

   (a) Sample a document-topic multinomial distribution $\boldsymbol{\theta}_j$ from Dir($\boldsymbol{\alpha}$).

   (b) For every w_i in d_j:

      i. Sample a topic $z_i \sim$ Multinomial($\boldsymbol{\theta}_j$).

      ii. Sample a term $w_i \sim$ Multinomial($\boldsymbol{\beta}_{z_i}$).

The LDA model can be illustrated as a probabilistic graphical model (see Figure 4). In the graphical model, a circle denotes a random variable; if a variable is observed, the circle is shaded, otherwise the variable is latent. So only the variable w is observed in the LDA model, and others need to be inferred. The plate notation groups the repeated variables together to make the graph illustration concise, and the dimension of the group is indicated in the corner of the plate. For instance, a total of T multinomial parameters $\boldsymbol{\beta}$ are combined in the $\boldsymbol{\beta}$ plate. The directed edge between variables indicates the dependency relationship. For example, in Figure 4, w depends on both the topic assignment z and $\boldsymbol{\beta}$, while $\boldsymbol{\beta}$ only depends on $\boldsymbol{\phi}$.

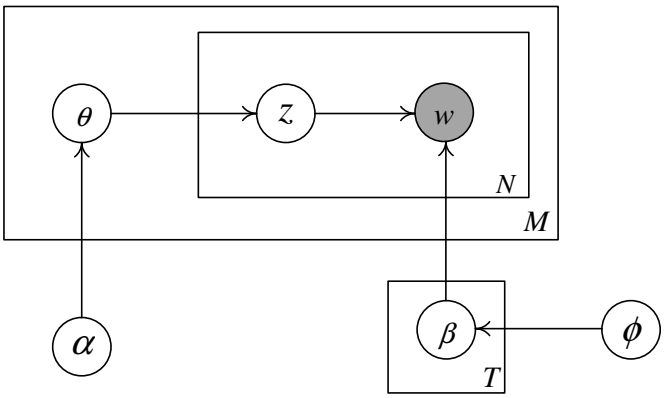


Figure 4: Graphical model for LDA. θ is the document-topic distribution, z is the topic assignment for word w, β is the topic-term distribution, and α and ϕ are the Dirichlet parameters.

The likelihood for the document $d = \{w_1, w_2, \ldots, w_{N_d}\}$ given hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ can be written down as

$$\begin{aligned} p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta} | \boldsymbol{\alpha}, \boldsymbol{\phi}) &= p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta}) p(\boldsymbol{z}|\boldsymbol{\theta}_d) p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) p(\boldsymbol{\beta}|\boldsymbol{\phi}) \\ &= \overbrace{\underbrace{\prod_{i=1}^{N_d} p(w_i|z_i, \boldsymbol{\beta}_{z_i}) p(z_i|\boldsymbol{\theta}_d)}_{\text{word plate}} p(\boldsymbol{\theta}_d|\boldsymbol{\alpha})}^{\text{document plate}} \underbrace{p(\boldsymbol{\beta}|\boldsymbol{\phi})}_{\text{topic plate}} \end{aligned} \tag{5}$$

Integrating out $\boldsymbol{z}$, $\boldsymbol{\theta}$, and $\boldsymbol{\phi}$, and expanding the probabilities, Equation 5 leads to

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\phi}) &= \int_{\boldsymbol{\beta}} \int_{\boldsymbol{\theta}_d} \sum_{\boldsymbol{z}} p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta}) p(\boldsymbol{z}|\boldsymbol{\theta}_d) p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) p(\boldsymbol{\beta}|\boldsymbol{\phi}) d\boldsymbol{\beta} d\boldsymbol{\theta}_d \\ &= \int_{\boldsymbol{\beta}} \int_{\boldsymbol{\theta}_d} \prod_{i=1}^{N_d} (\sum_{z_i} p(w_i|z_i, \boldsymbol{\beta}_{z_i}) p(z_i|\boldsymbol{\theta}_d)) p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) p(\boldsymbol{\beta}|\boldsymbol{\phi}) d\boldsymbol{\beta} d\boldsymbol{\theta}_d \\ &= \frac{1}{B(\boldsymbol{\phi})} \frac{1}{B(\boldsymbol{\alpha})} \int_{\boldsymbol{\beta}} \prod_{i=1}^{\mathcal{V}} \beta_i^{\phi_i - 1} \int_{\boldsymbol{\theta}_d} \prod_{i=1}^{T} \theta_i^{\alpha_i - 1} (\prod_{i=1}^{N_d} \sum_{z_i} (\beta_{z_i w_i} \theta_{d z_i})) d\boldsymbol{\beta} d\boldsymbol{\theta}_d. \end{aligned} \tag{6}$$

In Equation 6, $B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{T} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{T} \alpha_i)}$ and $B(\boldsymbol{\phi}) = \frac{\prod_{i=1}^{V} \Gamma(\phi_i)}{\Gamma(\sum_{i=1}^{\mathcal{V}} \phi_i)}$ are the Dirichlet normalization constants.

To estimate $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, we need to obtain the posterior $p(\boldsymbol{z}|\boldsymbol{w})$ [1]. However, it is

intractable to compute the posterior due to the coupling between $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ [14] in the summation over all possible z [1]. Researchers mainly use two types of approaches to indirectly tackle this problem: 1) variational inference [18], and 2) Gibbs sampling [29]. Parts of our subsequent work are based on Gibbs sampling, so we introduce how Gibbs sampling is adopted to solve the posterior.

Based on our introduction above in Section 1.2.2, $\boldsymbol{z}$ is a hidden variable and our goal is to approximate the posterior $p(\boldsymbol{z}|\boldsymbol{w})$. $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can be integrated out during the sampling, which is called "collapsed" Gibbs sampling [31, 61]. Gibbs sampling tells us the pursued posterior can be approximated by running an iterative sampler for each z_i. Based on Algorithm 1, the joint distribution is required to compute

$$p(z_i|\boldsymbol{z}_{-i}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\phi}) = \frac{p(\boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\phi})}{p(\boldsymbol{z}_{-i}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\phi})}. \tag{7}$$

From Equation 5, the joint distribution is computed by integrating out $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$

$$p(\boldsymbol{w}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}) = \int_{\boldsymbol{\beta}} p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{\phi}) d\boldsymbol{\beta} \cdot \int_{\boldsymbol{\theta}} p(\boldsymbol{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta}. \tag{8}$$

Since $p(\boldsymbol{\beta}|\boldsymbol{\phi})$ is a Dirichlet prior and $p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta})$ is a multinomial distribution, applying the property of conjugate prior, the first factor of Equation 8 can be derived as follows:

$$\int_{\boldsymbol{\beta}} p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{\phi}) d\boldsymbol{\beta} = \prod_{t=1}^{T} \frac{B(\boldsymbol{n}_t + \boldsymbol{\phi})}{B(\boldsymbol{\phi})}, \tag{9}$$

where $\boldsymbol{n}_t = (n_t^1, n_t^2, \ldots, n_t^{|\mathcal{V}|})$ is a vector consisting of the number of occurrence of each term in the vocabulary assigned to topic t. Likewise, the second factor is expanded as:

$$\int_{\boldsymbol{\theta}} p(\boldsymbol{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta} = \prod_{m=1}^{M} \frac{B(\boldsymbol{n}_m + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}, \tag{10}$$

where $\boldsymbol{n}_m = (n_d^1, n_d^2, \ldots, n_d^T)$ is the occurrence of every topic in document d. Substitute the joint distribution in Equation 7 with Equation 9 and Equation 10, and the derivation leads us to reach

$$p(z_i = t|\boldsymbol{z}_{-i}, \boldsymbol{w}) \propto \frac{n_{t-}^{w_i} + \phi_{w_i}}{\sum_w^{|\mathcal{V}|}(n_t^w + \phi_w)} \cdot (n_{d-}^{t-} + \alpha_t), \tag{11}$$

where the subscript "-" means the current term is excluded, as the sampler is updating the topic assignment for it. The explanation for the sampling update equation is quite straightforward: the first term on the right of Equation 11 denotes the probability of term w_i under topic t with prior ϕ_{w_i}, and the second term is directly proportional to the number of topics t in this document d.

The estimates of the multinomial parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are their expectations computed according to Equation 2:

$$\hat{\theta}_{dt} = \frac{n_d^t + \alpha_t}{\sum_k^T (n_d^k + \alpha_k)} \tag{12}$$

$$\hat{\beta}_{kt} = \frac{n_k^t + \phi_t}{\sum_w^{|\mathcal{V}|}(n_k^w + \phi_w)} \tag{13}$$

Usually $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ are given by users, although they can be inferred [18]. For the Dirichlet distribution, if the hyperparameters are identical, it is called a symmetric Dirichlet. The priors could influence the performance of LDA; relevant discussion can be found in [80]. Another parameter that needs to be specified is the number of topics T. There are several approaches that can be employed here to decide it: (1) Compare the perplexity (inverse of the geometric mean per-word likelihood) of the held-out data. Lower perplexity means better generalization of the model to the data. (2) Use secondary task based metrics. For example, if the output of LDA is

applied to perform a classification task, the best classification result can be used to select T. (3) Decide using algorithms. Teh et al. [75] proposed Hierarchical Dirichlet Processes where the number of topics is set by the algorithm based on data.

1.3 Thesis Organization

This thesis is organized as follows:

- In Chapter 1, we have introduced our research motivation. This chapter also introduces the Dirichlet distribution and Gibbs sampling, which provide the background statistical knowledge for LDA.

- In Chapter 2, we first review some extensions and modifications made to LDA in recent years. Additionally, we also survey some related research work on social media and applying topic models to process social media data.

- In Chapter 3, we introduce the *trivial Tag-LDA*, our first attempt to connect tags with topics. We then develop *Tag-LDA*, which naturally connects tags with topics in the generative model. We qualitatively and quantitatively evaluate our models and present results on explaining hashtags in tweets and learning the relationships between tags.

- In Chapter 4, we extend Tag-LDA, demonstrating that the structure of Tag-LDA makes it amenable to similar extensions as LDA. We propose *Tag-Latent Dirichlet Process* to address the problem of setting the number of topics. Additionally, we propose *ConceptTag-LDA* as a mechanism to incorporate users' pre-defined concepts when learning topics.

- In Chapter 5, we develop a specialized temporal topic model, called *dynamic Twitter topic model*, tailored for the short messages in Twitter. It is able to model the temporal dynamics of topics extracted from the tweets discussing the same event and thus facilitates event analysis. Experiments are conducted with tweet data crawled from Twitter to demonstrate the performance.

- Finally, in Chapter 6, we summarize the contributions of the thesis and outline future research directions.

CHAPTER 2: RELATED WORK

LDA has been reviewed in the previous chapter. This chapter mainly reviews the variations of LDA. Additionally work related to the topic models used in mining social media is briefly discussed.

2.1 Extensions of Latent Dirichlet Allocation

This section briefly reviews the related research on LDA. As a topic model, LDA discovers statistical relationships of words and "latent topics" that pervade the document collections. LDA, extended from the probabilistic latent semantic indexing (pLSI) [34], is an unsupervised algorithm, which does not account for tag information. To integrate meta-data, such as tags or labels into the unsupervised algorithm, researchers have proposed a few new approaches.

Blei and McAuliffe [13] introduced a supervised topic model. They took into account a response variable, which could be a rating or a category associated with a document. Their goal is to predict the response variables for new documents. Ramage et al. [65] developed a labeled LDA. Rather than predicting a response variable, labeled LDA links the latent topics to the labels in a one-to-one mapping, given a user labeled document collection. Therefore, the relationship between label and words can be established since one topic corresponds to one label. This one-to-one constraint is relaxed in their later work on partially labeled topic models [66], where one label

contains multiple topics. However, in partially labeled LDA, one topic can only be assigned to one label exclusively. In contrast, our work does not have this restriction; the tags are expressed as a distribution over all the topics and the topics are sharable among all tags.

One work similar to ours is author-topic model (ATM) introduced by Rosen-Zvi et al. [69]. ATM assumes the distribution of authors' contributions in a given document is uniform, but our models assume the distribution is multinomial with a Dirichlet prior, which is the major difference. We compared our models with ATM in the experiments and found the advantage of our model for certain data. Another similar work is Dependency-LDA proposed by Rubin et al. [70]. The authors discovered the dependency relationships between labels, assuming "topic" (note that their definition of topic is different from the standard LDA) is a distribution over observed labels. The main differences between our work and theirs are: 1) Dependency-LDA defines both the document-label distribution and document-topic distribution, while our model defines the document-tag distribution; 2) Dependency-LDA learns the dependency of labels via a higher level of latent topics, while we model the relationship of labels using latent topics.

Besides the research motioned above, there are a few other studies on topic relationships. Hierarchical topic models [14] could discover the usage of topics among the document collection via a topic hierarchy. The higher level topics are more semantically general compared to the lower level topics. The correlated topic models [17] model the per-document topic proportions with a logistic normal to reflect the correlation of topics from the covariance matrix of the logistic normal. Dynamic

topic models [16] explore the topic evolution over time. These three works explore the relationships of topics from different aspects.

The topics extracted from LDA are based on the co-occurrence of words in documents, therefore the topics are not guaranteed to satisfy people's semantic perceptions [56]. This is a result of the unsupervised learning procedure which lacks the input of human knowledge [23]. Involving the human knowledge as a prior, several research works [23, 4, 3, 37] proposed learning more meaningful and semantically coherent latent topics. In [3], a Dirichlet Tree prior [57] replaces the original Dirichlet prior to model Must-Link and Cannot-Link relationships between words. In [37], word constraints are built into the models using Dirichlet Tree priors too. Our work in Chapter 4 is inspired by [3] and [37].

Temporal topic models integrate temporal information in the models, which therefore are able to capture the temporal dynamics of topics, i.e., evolution and change of topics along the time line. The dynamic topic model (DTM) [16] chains the topics across time, and assumes current topics depend on their previous states. The continues dynamic topic model (cDTM) [82] applies the Brownian motion model, instead of the discrete state space model in DTM, so as to allow handling of data in continuous time. Instead of making Markov assumption as in DTM and cDTM, Wang and McCallum [83] in the Topics Over Time (TOT) model introduced a timestamp variable added to LDA to influence the document-topic proportion. An extension of TOT is introduced in the Trend Detection Model [44], where trends are represented by topic distributions rather than word distributions. Masada et al. [53] proposed to use a function of document timestamps to replace the topic Dirichlet priors in LDA.

Iwata et al. [38] assume topics evolve over multiple timescales, and develop an online inference procedure updating the model with newly obtained data.

2.2 Related Research on Social Media

Social media has attracted a great amount of research attention. There has been quite a lot of research on social network structure and identifying influential users [40][45][28][21]. Researchers also have tried employing published topic models to mine Twitter data. Twitter data has special characteristics compared to normal document data, because it is more noisy, of short length, and with high volume and velocity. Hong et al. [35] did an empirical study of topic modeling in Twitter applying LDA and the author-topic model [69] to predict popular Twitter messages and classify users and messages. They claim aggregation of the data is necessary, and the length of the "document" indeed influences the effectiveness of the models. Zhao et al. [88] proposed Twitter-LDA, in which each user possesses a topic distribution and each tweet can only be assigned with one topic. So words of a tweet message are generated from one topic selected from the user topic distribution or the common background model. LDA is also used in user recommendation systems to analyze users' interests [62].

Next, we review the recent related research on hashtags in Twitter. There is some work focusing on hashtags [24, 68, 43, 77], but very little of it focuses on the content analysis of hashtags. Most work considers hashtags as ideas, opinions, or information that flows and propagates over the social media network via the interaction of users. To the best of our knowledge, very little work has discussed the semantic meaning

of the hashtags. In this thesis, we propose new topic models and leverage the topic modeling techniques to analyze the hashtags.

Cunha et al. [24] studied how the hashtags are created and used from the perspective of linguistic theory. Romero et al. [68] studied the widely used hashtags and found different hashtags exhibit different spreading patterns, and they further explained that the difference of the patterns not only depends on the exposures of the hashtags but also the speed of decay. To research the spatial spread of social media, geo-tagged hashtags can also be adopted. Kamath et al. [43] combined two hypotheses of information spread and developed a probabilistic model to understand the global spread of social media. They found hashtags have local characteristics and therefore distance is the most significant factor influencing the spread. Tsur and Rappoport [77] developed a linear regression based approach to predict the spread of hashtags, and they found that content features combined with temporal and topological features would deliver the best prediction performance.

Lin et al. [50] assume hashtags are indicators of topics of interest and they track the topics in continuous streams of Twitter by integrating a "foreground" model and a "background" model. One popular function of hashtags is to track "trending topics" [47]. To identify the trendsetters, Saez-Trumper et al. [71] proposed a ranking algorithm in an information network with temporal factors integrated. Sentiment analysis on tweets can also be performed with hashtags [46][25].

There has been some research on mining events and tracking event trends in the discussions of social media. Some work applies ideas from previous work to social media, such as detecting burstiness of words or phrases [54, 49, 84], extracting entities

[67], and clustering similar tweets [8, 7]. Some work has been specially developed for microblog data. Vosecky et al. [79] emphasize the function of specific entities, and propose to jointly model entities and general terms together in latent topics with time characteristics. In [26], personal related posts are separated out to better detect bursty global topics.

CHAPTER 3: CONNECTING TAGS WITH TOPICS

In this chapter, we describe our work on connecting tags with topics; we do this by integrating tags with LDA. We introduce two models, Trivial Tag-LDA (TriTag-LDA) and Tag-LDA; both connect tags with topics assuming tags are a mixture of topics. We compare these two models with experiments and also present applications of the models. Portions of the work were introduced in this chapter is published in [52].

3.1 Trivial Tag-LDA: Starting from LDA

In this section, we first introduce a straightforward extension, Trivial Tag-LDA, to the standard LDA so as to incorporate tags. We detail our model and its generative procedures in Section 3.1.1, and then describe how to learn the model in Section 3.1.2.

3.1.1 Trivial Tag-LDA Model

As an extension of LDA, our topic model is also a probabilistic generative model. It simulates the procedure of generating documents and further discovers the relationship between tags and topics. We assume a tag is represented as a multinomial distribution over topics. Here, tags are observed and can be described by topics. Specifically, we are modeling over a document corpus $\mathcal{C} = \{d_1, d_2, \ldots, d_M\}$; each document contains a collection of words $\boldsymbol{w} = \{w_1, w_2, \ldots, w_{N_d}\}$ following the bag-of-words assumption, and is further associated with a set of tags $\boldsymbol{\delta}_d = \{p_1, p_2, \ldots, p_{L_d}\}$. We also define a set $\boldsymbol{\Delta} = \{p_1, p_2, \ldots, p_L\}$ which contains all tags without duplica-

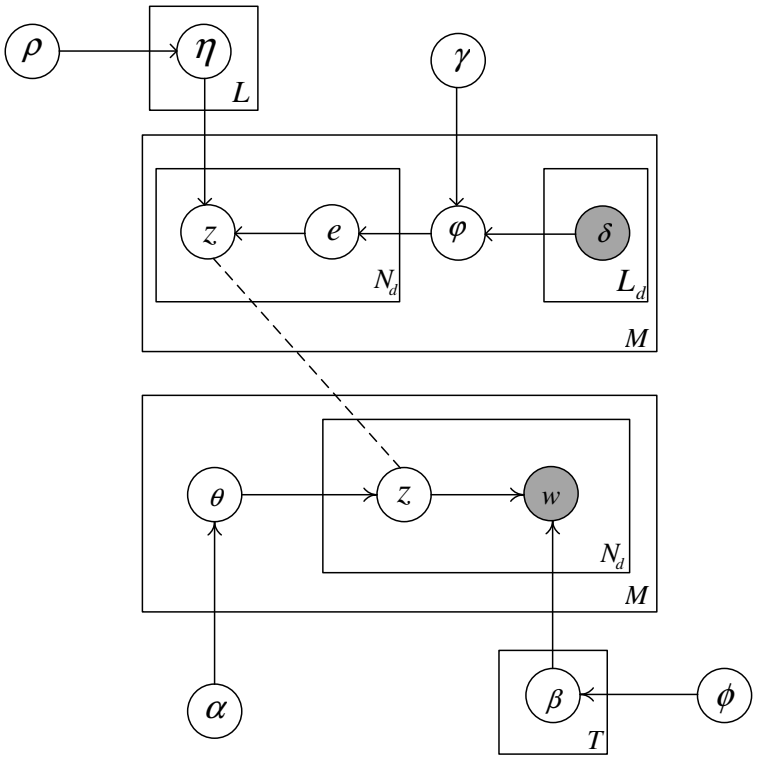


Figure 5: Graphical model of TriTag-LDA. Grey circles δ and w are observed variables for each document, others are latent variables. Note, z is shared between the two LDA components, representing the topic assignment for a word. While z is a latent variable in the bottom LDA model, it is regarded as an observed variable in the top model. η denotes the tag-topic distribution. ϕ denotes the document-tag distribution. e is the tag assignment. ρ and γ are the Dirichlet parameters.

tion. Therefore $\boldsymbol{\delta}_d$ is a subset of $\boldsymbol{\Delta}$. We assume all the elements of $\boldsymbol{w}$ come from a corpus-wide vocabulary $\mathcal{V}$.

For a document d with a text body $\boldsymbol{w}$ of length N_d, a set of tags $\boldsymbol{\delta}_d$ is observed. When generating the ith word in $\boldsymbol{w}$, a topic is chosen based on its multinomial probability $\boldsymbol{\theta}_d$ on this document. Topics are described as multinomial distributions $\boldsymbol{\beta}$ over vocabulary terms, and the distributions are independently drawn from a Dirichlet($\boldsymbol{\phi}$). Similarly, the topic distribution $\boldsymbol{\theta}_d$ of d is also sampled from Dirichlet($\boldsymbol{\alpha}$). To discover the tag-topic relationship, $\boldsymbol{w_d}$ can be treated as a mixture of observed tags, given by a multinomial distribution $\boldsymbol{\varphi}_d$, where $\boldsymbol{\varphi}_d \sim$ Dirichlet($\boldsymbol{\gamma}$). Let T be the predefined

total number of topics. Each tag p in $\boldsymbol{\delta}_d$ is a multinomial distribution $\boldsymbol{\eta}_p$ over all topics, and $\boldsymbol{\eta}_p$ with length equal to T is a sample drawn from a symmetric Dirichlet distribution with parameter $\boldsymbol{\rho}$, i.e., $\boldsymbol{\eta}_p \sim \text{Dirichlet}(\boldsymbol{\rho})$. So with the notation given in Table 2, the generative steps of our model are listed together below:

1. For each topic $t \in \{t_1, t_2, \ldots, t_T\}$, sample $\boldsymbol{\beta}_t$ over $|\mathcal{V}|$ terms from $\boldsymbol{\beta}_t \sim \text{Dirichlet}(\boldsymbol{\phi})$.

2. For each document d with $\boldsymbol{w}_d$ from the corpus:

   (a) Sample a distribution over topics $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha})$.

   (b) For the ith word in document d:

      i. Sample a topic $z \sim \text{Multinomial}(\boldsymbol{\theta}_d)$.

      ii. Sample a term $w \sim \text{Multinomial}(\boldsymbol{\beta}_z)$.

3. For each tag $p \in \boldsymbol{\Delta} = \{p_1, p_2, \ldots, p_L\}$, sample $\boldsymbol{\eta}_p$ over all topics $\boldsymbol{\eta}_p \sim \text{Dirichlet}(\boldsymbol{\rho})$.

4. For each document d with $\boldsymbol{w}_d$ from the corpus:

   (a) Sample a distribution over observed tags from $\boldsymbol{\varphi}_d \sim \text{Dirichlet}(\boldsymbol{\gamma})$.

   (b) For the ith topic in document d:

      i. Sample a tag $e \sim \text{Multinomial}(\boldsymbol{\varphi}_d)$.

      ii. Sample a topic $z \sim \text{Multinomial}(\boldsymbol{\eta}_e)$.

We need to point out that this generative procedure implies a two layer structure. First, a standard LDA is applied to generate the topic assignment $\boldsymbol{z}$ (bottom part in

Table 2: Notation table for TriTag-LDA

Symbol	Size	Description
M	scalar	number of documents in the corpus
L	scalar	number of distinct tags
$\mathcal{C}$	$1 \times M$	corpus
L_d	scalar	number of tags in document $d \in \mathcal{C}$
$\boldsymbol{w}_d$	$1 \times N_d$	words of document $d \in \mathcal{C}$
$\boldsymbol{\delta}_d$	$1 \times L_d$	tags of document $d \in \mathcal{C}$
$\boldsymbol{\eta}_p$	$1 \times T$	tag-topic multinomial distribution for tag p
$\boldsymbol{\beta}_t$	$1 \times \lvert\mathcal{V}\rvert$	topic-term multinomial distribution for topic t
$\boldsymbol{\varphi}_d$	$1 \times L_d$	document-tag multinomial distribution in d
$\boldsymbol{\theta}_d$	$1 \times T$	document-topic multinomial distribution in d
$\boldsymbol{\rho}$	$1 \times T$	Dirichlet distribution parameters
$\boldsymbol{\phi}$	$1 \times \lvert\mathcal{V}\rvert$	Dirichlet distribution parameters
$\boldsymbol{\alpha}$	$1 \times T$	Dirichlet distribution parameters
$\boldsymbol{\gamma}$	$1 \times L_d$	Dirichlet distribution parameters

Figure 5). Given the topic assignment and considering it already known in the second layer (top part in Figure 5), the generative process proceeds.

3.1.2 Learning Parameters

For the learning process, we mainly focus on estimating $\boldsymbol{\eta}$, $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{\varphi}$. We assume other parameters, $\boldsymbol{\phi}$, $\boldsymbol{\alpha}$, $\boldsymbol{\rho}$, and $\boldsymbol{\gamma}$, are symmetric and choose them in a heuristic manner. We modified the collapsed Gibbs sampling (detailed by Griffiths and Steyvers in [29]) to learn our model estimating the variables.

Note, tuning the Dirichlet parameters is not the main focus of this work; details on this can be referred in [18]. Here we also adopt and modify the collapsed Gibbs sampling to learn these parameters.

3.1.2.1 Estimating the Topic-Term Distribution and Document-Topic Distribution

Although the entirety of our model differs from the original version of LDA, it retains the same relationship between topics and vocabulary terms as in LDA. In this process, the estimate of $\beta_{k,w}$, the probability of term w given topic k, can be computed by Equation 14, according to the expectation of the Dirichlet distribution

$$\hat{\beta}_{k,w} = \frac{n_k^w + \phi}{\sum_{w=1}^{|\mathcal{V}|}(n_k^w + \phi)}, \tag{14}$$

where n_k^w denotes the number of occurrences of term w assigned to topic k.

Following the method in [29], the estimate of the multinomial $\boldsymbol{\theta}_d$ of the topic distributions in document d is written similarly as

$$\hat{\theta}_{k,d} = \frac{n_k^d + \alpha}{\sum_{k=1}^{T}(n_k^d + \alpha)}, \tag{15}$$

where n_k^d is the number of words assigned to topic k respectively in document d.

When training the model, Gibbs sampling updates the topic assignment of each word, one at a time. The update probability of assigning topic k to word $w_i^d = w$ in document d is

$$P(z_i = k | w_i = w, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}, \boldsymbol{\phi}, \boldsymbol{\alpha}) \propto \frac{n_{wk,-i}^{\mathcal{C}} + \phi}{\sum_{w=1}^{|\mathcal{V}|}(n_{wk,-i}^{\mathcal{C}} + \phi)} \cdot (n_{k,-i}^d + \alpha). \tag{16}$$

In Equation 16, $n_{wk,-i}^{\mathcal{C}}$ is the count of term w assigned to topic k in the whole corpus and $n_{k,-i}^{\mathcal{C}}$ represents the count of topic k assigned in document d, after removing the current topic assignment of the word w_i. The subscript $-i$ refers to excluding the ith word.

3.1.2.2 Estimating the Tag-Topic Distribution and Document-Tag Distribution

It is easy to see that the tag-topic and topic-term relationships are similar. So we derive the estimate of the tag-topic distribution $\boldsymbol{\eta}$ similarly to the estimation of $\boldsymbol{\beta}$, as

$$\hat{\eta}_{p,k} = \frac{n_k^p + \rho}{\sum_{k=1}^{T}(n_k^p + \rho)}, \tag{17}$$

where n_k^p is the frequency of topic k assigned to tag p in the entire corpus. Very similarly, the estimate of the tag distribution in a document is

$$\hat{\varphi}_{p,d} = \frac{n_p^d + \gamma}{\sum_{p=1}^{L_d}(n_p^d + \gamma)}, \tag{18}$$

where n_p^d is the occurrences of tag p in d.

Now we need to update the tag assignment for topics. Note we can only use the observed tags of the current document during the update, even though we have the full set of tags. The Gibbs sampling update equation is given by

$$P(e_i^d = p | z_i^d = k, \boldsymbol{z}_{-i}, \boldsymbol{e}_{-i}, \boldsymbol{\rho}, \boldsymbol{\gamma}) \propto \frac{n_{pk,-i}^{\mathcal{C}} + \rho}{\sum_{k=1}^{T}(n_{pk,-i}^{\mathcal{C}} + \rho)} \cdot (n_{p,-i}^d + \gamma), \tag{19}$$

where $n_{pk,-i}^{\mathcal{C}}$ is the frequency of topic k being assigned to tag p over the corpus and $n_{p,-i}^d$ is the frequency of tag p in the current document d excluding the current topic.

3.2 Tag-LDA

3.2.1 Tag-LDA Model

The TriTag-LDA model we have introduced is built on LDA, and adds one more layer to incorporate the tag information. We can notice that in TriTag-LDA the tag assignment has no influence on the generation of the topic and afterwards the word,

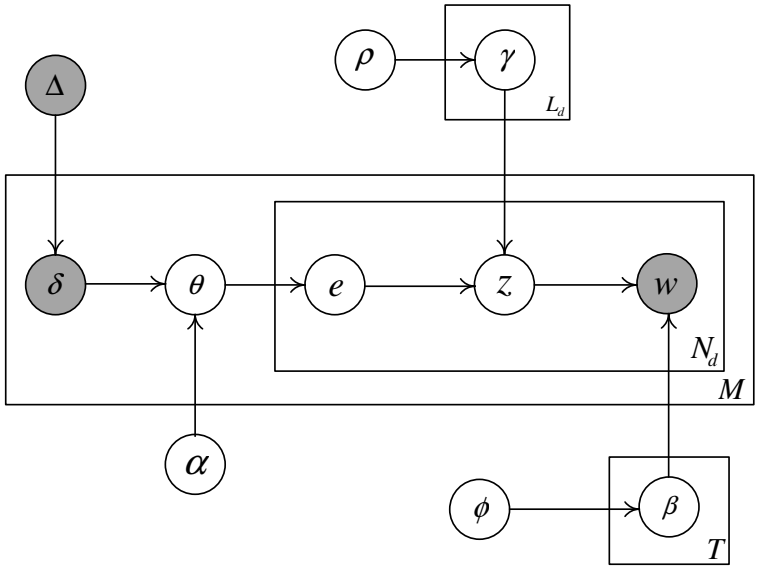


Figure 6: Graphical model of Tag-LDA. Word w and tags $\boldsymbol{\delta}$ are observed. Latent variables e and z are the tag and topic assignment to the word. Variables θ, γ, and β are latent variables. Tag set Δ is included so as to keep the completeness of the generative process.

as the topic assignment is already decided in the first layer. So in this section, we introduce another model, Tag-LDA, that improves this point.

Tag-LDA is also a probabilistic generative model. It simulates the procedure of generating documents and further discovers the relationship between tags and topics. Tags are observed and associated with documents. We define a tag as a multinomial distribution over topics. To describe the model, we introduce some notation. We have a corpus of documents $\mathcal{C} = \{d_1, d_2, \ldots, d_M\}$. Each document d consists of a set of words $\boldsymbol{w}_d = \{w_1, w_2, \ldots, w_{N_d}\}$, which we assume meets the "bag of words" assumption, and a set of tags $\boldsymbol{\delta}_d = \{p_1, p_2, \ldots, p_{L_d}\}$. We also define a set $\Delta = \{p_1, p_2, \ldots, p_L\}$ containing all tags without duplication in the corpus. $\boldsymbol{\delta}_d$ therefore is a subset of Δ. In addition, we assume all the elements of $\boldsymbol{w}$ come from a corpus-wide vocabulary $\mathcal{V}$.

More formally, when generating document d, a subset of tags $\boldsymbol{\delta}_d$ are first selected

from Δ. For the ith word in $\boldsymbol{w}_d$ of document d, a tag e is chosen from $\boldsymbol{\delta}_d$ based on the multinomial distribution $\boldsymbol{\theta}_d$ of tags on this document. The multinomial distribution $\boldsymbol{\theta}_d$ is sampled from a symmetric Dirichlet distribution with hyperparameter $\boldsymbol{\alpha}$. Then, under the chosen tag, a topic is sampled from a multinomial distribution $\boldsymbol{\gamma}_e$, where $\boldsymbol{\gamma}_e$ is also assumed to be generated from a symmetric Dirichlet distribution. As the LDA model, topics are described as multinomial distributions $\boldsymbol{\beta}$ over vocabulary terms, and the distributions are independently drawn from a Dirichlet($\boldsymbol{\phi}$). To discover the tag-topic relationship, $\boldsymbol{w}_d$ can be explicitly thought as a mixture of observed tags and implicitly as a mixture of topics, because words are in fact generated from the topics under the tags. Let T be the total number of topics predefined. So with the notation given in Table 3, the generative steps of our model are listed below:

1. For each tag $p \in \Delta$, sample $\boldsymbol{\gamma}_p$ over all topics from $\boldsymbol{\gamma}_p \sim$ Dirichlet($\boldsymbol{\rho}$).
2. For each topic $t \in \{t_1, t_2, \ldots, t_T\}$, sample $\boldsymbol{\beta}_t$ over $|\mathcal{V}|$ terms from $\boldsymbol{\beta}_t \sim$ Dirichlet($\boldsymbol{\phi}$).
3. For each document d with $\boldsymbol{w}_d$ from the corpus:
   (a) Sample a distribution over observed tags from $\boldsymbol{\theta}_d \sim$Dirichlet($\boldsymbol{\alpha}$).
   (b) For ith word in document d:
      i. Sample a tag $e \sim$ Multinomial($\boldsymbol{\theta}_d$).
      ii. Sample a topic $z \sim$ Multinomial($\boldsymbol{\gamma}_e$) a Multinomial probability conditioned on current tag assignment e.
      iii. Sample a term $w \sim$ Multinomial($\boldsymbol{\beta}_z$) a Multinomial probability conditioned on current topic assignment. z

Table 3: Notation table for Tag-LDA.

symbol	size	description
M	scalar	number of documents
L	scalar	number of distinct tags
Δ	$1 \times L$	complete set of tags
$\mathcal{C}$	$1 \times M$	corpus
$\boldsymbol{w}_d$	$1 \times N_d$	words of document d
$\boldsymbol{\delta}_d$	$1 \times L_d$	tags of document d
$\boldsymbol{\gamma}_p$	$1 \times T$	tag-topic multinomial dist.
$\boldsymbol{\beta}_t$	$1 \times \lvert\mathcal{V}\rvert$	topic-term multinomial dist.
$\boldsymbol{\theta}_d$	$1 \times L_d$	document-tag multinomial dist.
$\boldsymbol{\rho}$	$1 \times T$	Dirichlet hyperparameters
$\boldsymbol{\phi}$	$1 \times \lvert\mathcal{V}\rvert$	Dirichlet hyperparameters
$\boldsymbol{\alpha}$	$1 \times L_d$	Dirichlet hyperparameters

The graphical model in Figure 6 demonstrates our model. Since $\boldsymbol{\delta}_d$ is already observed, the selection of it from Δ is not mathematically modeled, but for completeness, it is kept in the graphical model.

3.2.2 Learning Parameters

Given a document with words $\boldsymbol{w}_d$, the associated tags $\boldsymbol{\delta}_d$, and all the hyperprameters, we would like to compute the posterior distribution of the latent variables

$$P(\boldsymbol{e}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho}) = \frac{P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})}{\sum_{\boldsymbol{z}} P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}, |\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})}.$$

However, this posterior distribution is not computable, because the denominator is intractable to compute. In other LDA related work, the researchers mainly use two types of approaches to indirectly tackle this problem: 1) variational inference [18], and 2) Monte Carlo Markov chain (MCMC) sampling [29]. We adopt Gibbs sampling, a MCMC sampling method, in our work.

To build the Gibbs sampler, we require the joint distribution of observed words

and their tag and topic assignments $P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})$. This joint distribution can be factorized as below

$$P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho}) = P(\boldsymbol{w}|\boldsymbol{\phi}, \boldsymbol{z}) \cdot P(\boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\rho}), \tag{20}$$

based on the independence of variables. We analyze the two terms of the right side of Equation 20 one by one. First, we compute the first term. In fact, the first term is the same as the LDA model (Equation 9). We simply write down the derivations as follows and do not introduce detailed explanation. Interested readers could refer to references [31] and [29].

$$\begin{aligned} P(\boldsymbol{w}|\boldsymbol{\phi}, \boldsymbol{z}) &= \int_{\boldsymbol{\beta}} P(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta}) P(\boldsymbol{\beta}|\boldsymbol{\phi}) d\boldsymbol{\beta} \\ &= \prod_{t=1}^{T} \frac{B(\boldsymbol{n}_t + \boldsymbol{\phi})}{B(\boldsymbol{\phi})}, \end{aligned}$$

where $\boldsymbol{n}_t = (n_t^1, n_t^2, \cdots, n_t^{|V|})$ is a vector of length $|\mathcal{V}|$ consisting of the number of occurrences of each term assigned to topic t, and $B(\cdot)$ is a multinomial beta function.

Now we turn to computing the second term in Equation 20. The second term can be further factorized as $P(\boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\rho}) = P(\boldsymbol{z}|\boldsymbol{e}, \boldsymbol{\rho}) P(\boldsymbol{e}|\boldsymbol{\alpha})$ by applying Bayes rule and the independence assumption. Let us first look at $P(\boldsymbol{z}|\boldsymbol{e}, \boldsymbol{\rho})$. We notice that $P(\boldsymbol{z}|\boldsymbol{e}, \boldsymbol{\rho})$ can be obtained after integrating out $\boldsymbol{\gamma}$:

$$P(\boldsymbol{z}|\boldsymbol{e}, \boldsymbol{\rho}) = \int_{\boldsymbol{\gamma}} P(\boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{e}) P(\boldsymbol{\gamma}|\boldsymbol{\rho}) d\boldsymbol{\gamma}. \tag{21}$$

For word i in document d, given its tag assignment e_{di}, $P(z_{di} = t|e_{di})$ is a multinomial

distribution with parameter $\gamma_{e_{di}t}$. So we can obtain

$$P(\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{e}) = \prod_{d=1}^{M}\prod_{i=1}^{|\boldsymbol{w}_d|} P(z_{di}|e_{di}) = \prod_{d=1}^{M}\prod_{i=1}^{|\boldsymbol{w}_d|} \gamma_{e_{di}z_{di}} = \prod_{p=1}^{L_d}\prod_{t=1}^{T} \gamma_{pt}^{n_p^t}.$$

We already assume $P(\boldsymbol{\gamma}|\boldsymbol{\rho})$ follows a Dirichlet distribution. Substitute $P(\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{e})$ and $P(\boldsymbol{\gamma}|\boldsymbol{\rho})$ in Equation 21 and apply Dirichlet integrals:

$$\begin{aligned} P(\boldsymbol{z}|\boldsymbol{e},\boldsymbol{\rho}) &= \frac{1}{B(\boldsymbol{\rho})}\int_{\boldsymbol{\gamma}} \prod_{p=1}^{L_d}\prod_{t=1}^{T} \gamma_{pt}^{n_p^t+\rho-1} d\boldsymbol{\gamma} \\ &= \prod_{p=1}^{L_d} \frac{B(\boldsymbol{n}_p+\boldsymbol{\rho})}{B(\boldsymbol{\rho})}, \end{aligned}$$

where $\boldsymbol{n}_p = (n_p^1, n_p^2, \cdots, n_p^T)$ contains the occurrences of each topic assigned to tag p. The derivation of the tag distribution $P(\boldsymbol{e}|\boldsymbol{\alpha})$ is quite similar to that of $P(\boldsymbol{z}|\boldsymbol{e},\boldsymbol{\rho})$ via integrating out $\boldsymbol{\theta}$. Therefore, the derivation similarly yields

$$\begin{aligned} P(\boldsymbol{e}|\boldsymbol{\alpha}) &= \int_{\boldsymbol{\theta}} P(\boldsymbol{e}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{\alpha})d\boldsymbol{\theta} \\ &= \prod_{d=1}^{M} \frac{B(\boldsymbol{n}_d+\boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}, \end{aligned}$$

and $\boldsymbol{n}_d = (n_d^1, n_d^2, \cdots, n_d^{L_d})$ are the occurrences of each tag present in document d. Note that n_d^p will always be zero if tag p is not associated with document d, i.e., the tag distribution per document must be over the associated observed tags only. Finally, the joint distribution of Equation 20 can be written down:

$$P(\boldsymbol{w},\boldsymbol{e},\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\phi},\boldsymbol{\rho}) = \prod_{t=1}^{T} \frac{B(\boldsymbol{n}_t+\boldsymbol{\phi})}{B(\boldsymbol{\phi})} \cdot \prod_{p=1}^{L_d} \frac{B(\boldsymbol{n}_p+\boldsymbol{\rho})}{B(\boldsymbol{\rho})} \cdot \prod_{d=1}^{M} \frac{B(\boldsymbol{n}_d+\boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}. \tag{22}$$

In Gibbs sampling, the value of each variable is sampled sequentially conditioned on the current values of all other variables. Therefore the update equation that the

sampler uses to update the topic and tag assignment for the ith word in document d is a conditional distribution:

$$P(e_{di} = p, z_{di} = t|\boldsymbol{w}, \boldsymbol{e}_{-}, \boldsymbol{z}_{-}, \boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\phi}) = \frac{P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})}{P(\boldsymbol{w}, \boldsymbol{e}_{-}, \boldsymbol{z}_{-}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})}, \tag{23}$$

where the subscript "$-$" indicates the current updating assignments of topic and tag of word w_{di} are excluded. Equation 23 can be further factored as:

$$\begin{aligned} P(e_{di} = p, z_{di} = t|\boldsymbol{w}, \boldsymbol{e}_{-}, \boldsymbol{z}_{-}, \boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\phi}) &= \frac{P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})}{P(\boldsymbol{w}_{-}, \boldsymbol{e}_{-}, \boldsymbol{z}_{-}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})P(w_{di})} \\ &\propto \frac{P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})}{P(\boldsymbol{w}_{-}, \boldsymbol{e}_{-}, \boldsymbol{z}_{-}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})} \end{aligned} \tag{24}$$

Substituting the joint probabilities in the numerator and denominator above with Equation 20 and canceling out the common terms, Equation 24 yields

$$\begin{aligned} \frac{P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})}{P(\boldsymbol{w}_{-}, \boldsymbol{e}_{-}, \boldsymbol{z}_{-}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})} &= \frac{B(\boldsymbol{n}_t + \boldsymbol{\phi})}{B(\boldsymbol{n}_{t-} + \boldsymbol{\phi})} \cdot \frac{B(\boldsymbol{n}_p + \boldsymbol{\rho})}{B(\boldsymbol{n}_{p-} + \boldsymbol{\rho})} \cdot \frac{B(\boldsymbol{n}_d + \boldsymbol{\alpha})}{B(\boldsymbol{n}_{d-} + \boldsymbol{\alpha})} \\ &= \frac{\Gamma(n_t^{w_{di}} + \phi)\Gamma(\sum_w^{|\mathcal{V}|}(n_{t-}^w + \phi))}{\Gamma(n_{t-}^{w_{di}} + \phi)\Gamma(\sum_w^{|\mathcal{V}|}(n_t^w + \phi))} \cdot \frac{\Gamma(n_p^t + \rho)\Gamma(\sum_{k=1}^T(n_{p-}^k + \rho))}{\Gamma(n_{p-}^t + \rho)\Gamma(\sum_{k=1}^T(n_p^k + \rho))} \\ &\cdot \frac{\Gamma(n_d^p + \alpha)\Gamma(\sum_l^{|\boldsymbol{\delta}_d|}(n_{d-}^l + \alpha))}{\Gamma(n_{d-}^p + \alpha)\Gamma(\sum_l^{|\boldsymbol{\delta}_d|}(n_d^l + \alpha))} \end{aligned}$$

Finally, the update equation is given by

$$\begin{aligned} P(e_{di} = p, z_{di} = t|\boldsymbol{w}, \boldsymbol{e}_{-}, \boldsymbol{z}_{-}, \boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\phi}) \propto \\ \frac{n_{t-}^{w_{di}} + \phi}{\sum_w^{|\mathcal{V}|}(n_{t-}^w + \phi)} \cdot \frac{n_{p-}^t + \rho}{\sum_{k=1}^T(n_{p-}^k + \rho)} \cdot \frac{n_{d-}^p + \alpha}{\sum_l^{|\boldsymbol{\delta}_d|}(n_{d-}^l + \alpha)}. \end{aligned} \tag{25}$$

In Equation 25, n_{k-}^w is the count of term w under topic t excluding the current topic assignment of this term. Very similarly, n_{p-}^k denotes the count of tag p assigned to topic k excluding the current tag assignment to the topic, and n_{d-}^l is the count of words in document d assigned to tag l excluding the current word. The explanation of

the update equation is straightforward. The first term in the right part of Equation 25 represents the probability of term w_{di} under topic t with prior ϕ, and the second term is the probability of topic t under tag p with prior ρ, while the last term is the probability of tag p in document d with prior α. Thus, the current topic and tag assignment to a word is proportional to the tag proportion in the document, the topic proportion under the tag, and the term proportion under the topic.

To estimate these three multinomial parameters after sampling, compute their expectations in the Dirichlet distributions. The estimate of topic-term distribution $\boldsymbol{\beta}$ is

$$\hat{\beta}_{tw} = \frac{n_t^w + \phi}{\sum_w^{|\mathcal{V}|}(n_t^w + \phi)},$$

where the estimation is identical to the LDA model. The estimates of document-tag distribution $\boldsymbol{\theta}$ and tag-topic distribution $\boldsymbol{\gamma}$ can be similarly derived and expressed as follows:

$$\hat{\theta}_{dp} = \frac{n_d^p + \alpha}{\sum_l^{|\boldsymbol{\delta}_d|}(n_d^l + \alpha)},$$

$$\hat{\gamma}_{pt} = \frac{n_p^t + \rho}{\sum_{k=1}^{T}(n_p^k + \rho)}.$$

3.3 Experiments

TriTag-LDA and Tag-LDA can find the correlation of tags and topics. In this section, we evaluate the performance difference of these two models with different datasets. Besides that, a comparison to author-topic model is provided. Representing tags as a distribution of topics provides us an approach to explore the relationships of tags, so we provides two examples, using Tag-LDA in Twitter tweet dataset and

Nation Science Foundation proposal dataset, to demonstrate the application of Tag-LDA.

For both TriTag-LDA and Tag-LDA in the experiments below, Gibbs sampling runs iterations between burn-in and sampling after a random initialization. For each Markov Chain Monte Carlo (MCMC) chain, each sample is drawn after 500 burn-in steps. For the convergence of the MCMC chain, we hard define an iteration length as the convergence criterion. Other more complicated convergence criteria can be found in reference [51].

3.3.1 Quantitative Comparison of Topic Models

We first introduce the datasets used to perform the quantitative studies on TriTag-LDA and Tag-LDA. Here we employ two different datasets: 1) New York Times Corpus and 2) proposal abstract data from National Science Foundation (NSF) awards database. The New York Times Corpus is a news collection of New York Times. It comes with General Online Descriptors, assigned automatically and verified by nytimes.com production staff, for each news article. A news article might have multiple descriptors associated with it. Examples of individual descriptor include Crime and Criminals, Animals, Computers and the Internet. We consider these descriptors as the tags in our models. For the purpose of experiments, we selected the news published in January and February 2005, and filtered out descriptors that occur less than 150 times, which finally results in 67 descriptors (tags). Overall 9964 news articles are left in the dataset. On average, there are 2.39 descriptors attached for each news, and the average length of each news article is 296.9 words after the stop words are

eliminated.

The NSF proposal abstract dataset is crawled from the NSF awards database[5] by ourself. NSF has seven directorates, one of which is the Directorate for Computer & Information Science & Engineering (CISE) [6]. There are three divisions under the CISE directorate: the Division of Computing & Communication Foundations (CCF); the Division of Computer and Network Systems (CNS); and the Division of Information and Intelligent Systems (IIS). The abstract data is between the years 2000 to 2009. Under each division, there are more specific programs, for instance, computer vision, human-centered computing, and robotics, to which the proposals are assigned. These programs are considered as tags in this dataset. We keep the top 44 most used programs and obtain a total of 4811 corresponding proposal abstract. Note that one proposal abstract may be assigned to several programs together. Specially, one proposal abstract is assigned to 1.26 programs on average, and the average length of each proposal abstract is 138.5 words excluding the stop words.

We adopt the perplexity to quantitatively evaluate the models and to profile the performance variation of the models as the number of topics varies. The perplexity measures the generalization performance of the model, with a lower score indicating better generalization of the model and better distribution prediction on the unseen data. The perplexity of the testing data $\mathcal{D}_{test}$ is mathematically expressed as:

$$Perplexity(\mathcal{D}_{test}) = \exp\left[-\frac{\log(P(\boldsymbol{w}|\mathcal{D}_{test}))}{N_{\mathcal{D}_{test}}}\right] \tag{26}$$

[5]http://www.nsf.gov/awardsearch/
[6]http://www.nsf.gov/dir/index.jsp?org=CISE

where $P(\boldsymbol{w}|\mathcal{D}_{test})$ is the likelihood of the testing data and given by

$$P(\boldsymbol{w}|\mathcal{D}_{test}) = \prod_{w} \sum_{z \in T} \sum_{e \in \Delta} P(w|z)P(z|e)P(e|d).$$

After obtaining the trained model, it is employed on the testing data to infer the likelihood. $N_{\mathcal{D}_{test}}$ is the total number of words in the testing data.

We split each of the datasets mentioned above into a training set and a testing set. The setup is that 10 percent of each dataset is drawn randomly as the corresponding testing data and the remainder is used to train.

Besides TriTag-LDA and Tag-LDA, we added author-topic model (ATM) [69] in the quantitative evaluation as a comparison, since ATM assumes each author can be viewed a distribution of topics, which is a similar assumption to ours. However, ATM does not have a Dirichlet prior in the model, which is the major difference between our models and ATM.

We compare the perplexities of TriTag-LDA, Tag-LDA and ATM on these two text datasets. The number of topics was varied along during the evaluation of the perplexities, so that we could profile the performance of the models. Figure 7 demonstrates the testing perplexity of Tri-LDA, Tag-LDA, and ATM on the NSF proposal abstract data as the number of topics was varied from 5 to 150. The results shows Tri-LDA gains the best generalization at 5 topics and 20 topics, but its performance drops with an increase in the number of topics. Through comparison, we can clearly note that the overall performance of Tag-LDA is the best. The testing perplexities on the New York Times data are shown in Figure 8. Comparing to the model performance in the prior experiment, we can see a similar conclusion that TriTag-LDA becomes worse as

the number of topics grows, but it is still superior to ATM. Tag-LDA produces the lowest perplexity when the number of topics exceeds 60. Basically we conclude that introducing a Dirichlet prior for the distribution of tag in the documents in Tag-LDA, improves the perplexity on the unseen data. Please note, if each document only contains only one tag, there would be no difference no matter which prior is assumed, therefore our Tag-LDA would become the same as ATM.

We select several tags listed in Table 4 and Table 5 to show the corresponding topics extracted by TriTag-LDA and Tag-LDA from New York Times corpus. We can observe that there are differences in the topics between these two methods. For instance, the topic for "Crime and Criminals" looks to be inclined to justice and jury in TriTag-LDA, while the topic in Tag-LDA is closer to law enforcement and crime. Another example is the tag "Advertising and Marketing", it can be noticed that these two topics are quite related to this tag, but they focus on different perspectives.

3.3.2 Understanding Hashtags in Twitter

To demonstrate the capability of our proposed methods, we conducted experiments on a dataset of tweets. In our experiments, we extracted hashtags from individual tweets as the tags. We try to understand the hashtags by modeling the tweet content, and further discover the relationships between the hashtags.

The tweet dataset used here is drawn from the TREC 2011[7] microblog data, which contains 16 million tweets sampled between January 23rd and February 8th, 2011. In this dataset, there are around 1.78 million tweets containing at least one hashtag.

[7] http://trec.nist.gov/data/tweets/

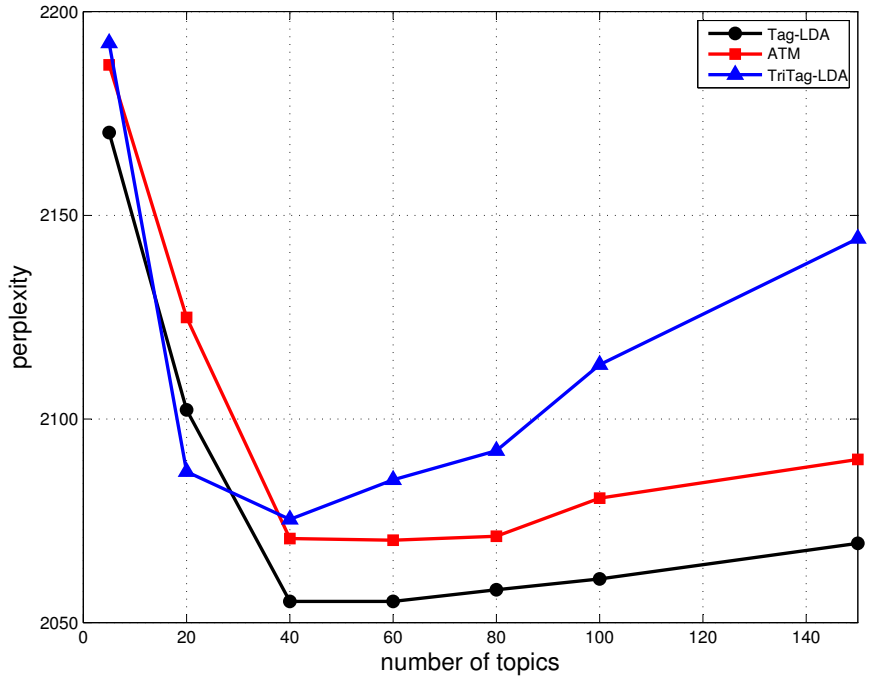


Figure 7: Perplexity comparison among ATM, Tag-LDA, and TriTag-LDA on NSF proposal abstract data. Lower perplexity value denotes a better generalization on the testing data of the model.

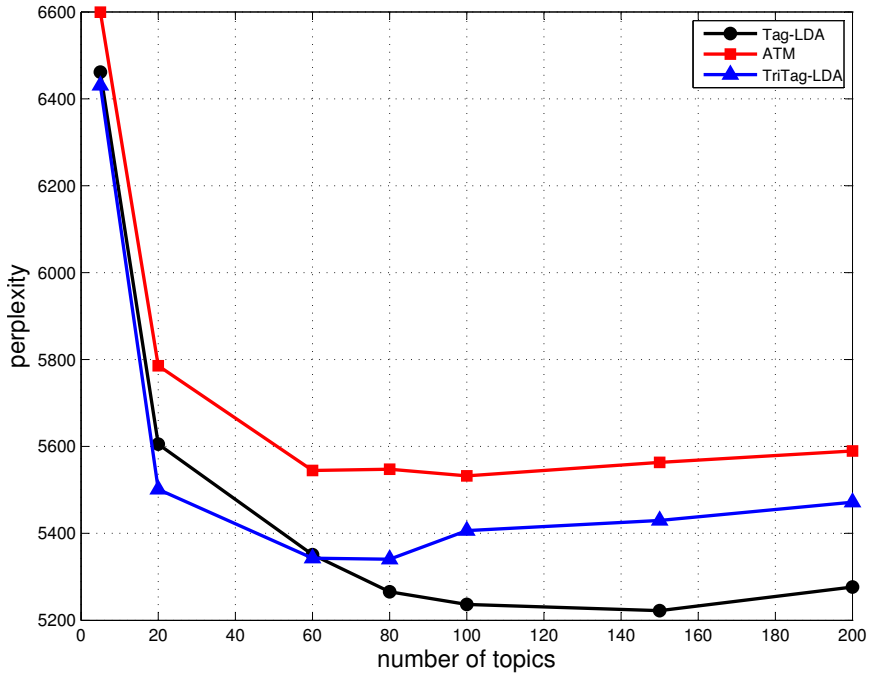


Figure 8: Perplexity comparison among ATM, Tag-LDA, and TriTag-LDA on the New York Times data. Lower perplexity value denotes a better generalization on the testing data of the model.

Among tweets containing hashtags, 19.9% use two or more hashtags. We ranked and selected the 200 most frequent hashtags. We then filtered out the hashtags that mainly appeared in non-English tweets, which left 161 hashtags and 150K tweets; nearly 22.0% of these tweets have more than one hashtag. For the purpose of cleaning the data, the words in the tweets are stemmed by utilizing the NLTK package [9], and the resulting vocabulary size is 21,139.

To understand the hashtags in the TREC2011 dataset, 140 topics are extracted from the corpus. The number of topics was determined by the lowest perplexity. Each hashtag is then epresented as a distribution over all topics.

To illustrate our results, we selected a few hashtags belonging to different categories (sports, politics, world, etc.) and list the topic with the highest probability for each hashtag in Table 6. At a first glance, one can see that some hashtags are difficult to interpret. The hashtag #tcot (Top Conservatives on Twitter) for example, it is a coalition of conservatives on the Internet. However, without reading the topic, the meaning of the hashtag is hard to infer. The keywords of the highest probability topic for hashtag #tcot capture terms related to conservative American political parties. Another interesting example is the hashtag #tahrir, which is an Arabic word meaning liberation. The topic results for the hashtag clearly indicate that the hashtag refers to a specific location – Tahrir Square in Cairo, Egypt. In addition, the topic also contains information regarding possible protests in Tahrir Square that might warrant further investigation. The examples demonstrate that topic terms could greatly help with the interpretation of the hashtags, which could otherwise be difficult to decode through reading the hashtags alone or merely examining a few tweets containing the

hashtag.

In addition to understanding the meaning and contexts of the hashtags, knowing the relationships among hashtags also contributes to proper categorization of tweets using hashtags. However, even with the topic results for each hashtag, the discovery of similar hashtags through a manual process is still a challenging and laborious task. Such a task can be adequately addressed through combining the Tag-LDA results with proper distance measures and visual representations.

As mentioned in Section 2.1, the difference between Tag-LDA and a previously proposed similar model, partially labeled LDA [66], is that in our model all tags are modeled as a distribution over a shared topic space. Therefore, the computation of the similarity between every pair of tags becomes straightforward. Since a tag is in the form of a probabilistic distribution over topics, we utilize the Hellinger distance [33] to measure the distance between a pair of tags. Given any two discrete probability distributions $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$, the Hellinger distance is defined as $H(P, Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i (\sqrt{p_i} - \sqrt{q_i})^2}$. Therefore a distance matrix can be constructed by measuring the distance between every pair of hashtag distributions over topics. We invert the distances between the hashtags to get the similarity measurement.

To enable users to discover similar hashtags, we visualize the similarities using a matrix metaphor. Figure 9 illustrates the similarities between pairs of hashtags in the TREC2011 data. At a glance, one can detect two major clusters in the visualization (lower left and upper right), with a larger dot size denoting higher similarity. The upper right cluster in Figure 9 highlights a group of hashtags related to football teams

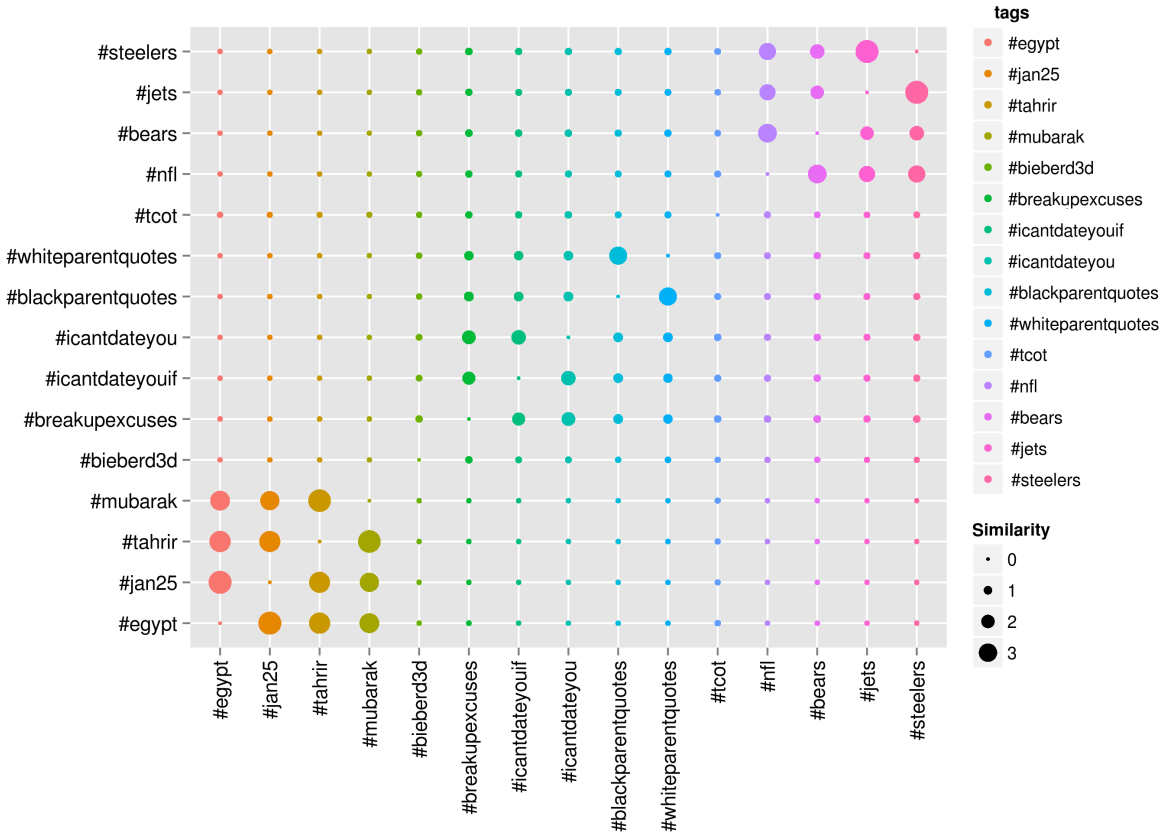


Figure 9: Visualization of the hashtag similarity matrix. Dots with larger size represents higher similarity; the larger the radius, the greater the similarity. Similarity values are scaled and self similarities on the diagonal are removed for clarity of the display.

and events. The two biggest dots in pink denote the high similarity between #steelers and #jets. Indeed, there are a lot of discussions in the tweets about the game between the New York Jets and Pittsburgh Steelers on that Sunday weekend.

A more interesting and intriguing example is the lower left cluster of hashtags including #egypt, #tahrir, #mubarak, #jan25. Coupling the results with the main topic for hashtag #tahrir and #mubarak (Table 6), which includes keywords such as “protest”, “protestors”, “anti”, “government”, “police”, one can infer that there may be a big protest event occurring at Tahrir Square in Egypt on January 25. Indeed, according to Wikipedia, the event was the 2011 Egyptian Revolution against for-

mer president Hosni Mubarak, and over 50,000 protestors occupied Tahrir Square in Cairo[8]. The hashtags in the lower left cluster are related to one event and different hashtags are created to describe temporal, geospatial, and people information. This example illustrates that by coupling the hashtag similarity results with the topics for hashtags, one can conduct deep analysis of events discussed on Twitter.

In summary, the above examples illustrate that visually presenting the similarities among hashtags could help users identify groups of hashtags used to characterize similar topics or events. In addition, combining the hashtag similarity results with the hashtag-topic results supports the development of comprehensive understanding of events discussed on Twitter. Retrospective examination of the hashtags and tweet content provides an overview of what has been discussed on Twitter. However, if the modeling and visualization can be done in real time, the implication is that one can monitor and even respond to certain events such as protests and gatherings.

3.4 Summary

In this chapter, we focused on integrating tag information in topic models. Our contribution are summarized below. We proposed two models, TriTag-LDA and Tag-LDA, to incorporate tags. The key idea is that we assume a tag is a multinomial distribution of topics. TriTag-LDA adds an extra layer to standard LDA, to discover the topic representations of tags. Tag-LDA inserts the tags in the generative process of the document. A Gibbs sampling based algorithm is adopted to learn the models. In the experiments, we compared these two models with the author-topic model using

[8]http://en.wikipedia.org/wiki/Tahrir_Square

perplexity. The comparison shows Tag-LDA is superior to the other two models. As an application of our Tag-LDA model, we used Tag-LDA to analyze the hashtags in Twitter. We demonstrated that Tag-LDA is capable of solving two challenges: first, understanding how hashtags can be interpreted, and second, elucidating the relationships between hashtags.

Table 4: The highest probability topic for each of several tags extracted from the New York Times dataset by TriTag-LDA.

Tag	Prominent Topic
Medicine and Health	health drug medical patient doctor disease study hospital heart care aids treatment percent risk patient medicine pain merck doctor university blood center cancer surgery
Crime and Criminals	court judge law lawyer lawyers federal trial justice prison legal jury charges death attorney prosecutors district supreme criminal filed states decision appeals lawsuit ruling
Art	art museum artists artist gallery exhibition collection painting paintings design gates modern century park contemporary arts sculpture photographs images street glass york
Terrorism	officials public department government time president federal agency report states office united board director law national former million decision called administration commission program
Football	game season football jets team bowl super coach patriots play eagles field yards quarterback games players england steelers sunday edwards pro philadelphia touchdown victory
Airlines and Airplanes	fashion airlines airline designer airport flights airways plane flight delta designers fares dress air passengers clothes travel collection southwest wear fare business aviation class jet runway york seats elite clothing
Religion and Churches	church gay religious jewish catholic marriage christian jews religion sex faith rabbi churches holocaust orthodox conservative rev community nazi lesbian evangelical roman
Advertising and Marketing	advertising media marketing campaign ads york commercials magazine super brand business bowl commercial agency advertisers spot president division creative worldwide
Stocks and Bonds	percent oil market prices china growth funds investors price dollar stocks rates average stock fund economy economic trade markets quarter billion rose energy american rise
Books and Literature	book novel author world life story history published wrote american century writing writes writer read war stories written self literary fiction society readers writers

Table 5: The highest probability topic for each of several tags extracted from the New York Times dataset by Tag-LDA.

Tag	Prominent Topic
Medicine and Health	health medical drug patient doctor care hospital heart study research company disease pain medicine risk cancer companies treatment surgery percent medicare
Crime and Criminals	police death officer prison murder ross court crime drug family charges prosecutors county arrest shot killing officer killed jury charged woman authorities home
Art	art museum artists artist paintings painting gates exhibition gallery park collection sculpture project modern drawings city christo century building heizer central design arts curator
Terrorism	united bush states american officials president administration iraq security palestinian iran intelligence north nuclear government military china nations israel world korea war foreign
Football	season game team jets football yankees bowl players mets super coach play baseball league patriots field games eagles yards ball quarterback teams player giambi beltran
Airlines and Airplanes	airlines airline airport air flights airways united passengers fares delta plane flight boeing travel pilots southwest fare bankruptcy aviation industry aircraft planes class miles airbus jet business airports carriers elite travelers
Religion and Churches	church religious jewish jews christian religion pope israel holocaust rabbi catholic faith muslim nazi orthodox churches muslims marriage anti war father evangelical community islam survivors rev death auschwitz prayer
Advertising and Marketing	company internet online computer web technology site software advertising video business media digital service sites google mail customers sales marketing services companies industry apple consumers microsoft computers
Stocks and Bonds	percent stocks market china funds growth oil economic investors fund rates prices economy inflation dollar stock average index countries markets term poverty poor earnings energy rose treasury bonds world rise rising mutual
Books and Literature	life american world time love black house york street woman women story white home company father family mother music art play children makes wife name london century self live book review culture night history set war

Table 6: The list of high probability terms of the highest probability topic for each hashtag. We manually added the categories to aid understanding.

Categories	Hashtag	Topic with the highest probability
Politics	#tcot	teaparty gop obama ocra tlot sgp obamacare bill palin twisters tpp vote repeal reagan repeal vote tpp conservative gore republicans constitution
Sport	#steeler	game yellow afc black championship beat fan lose nfl nyjets steelernation jersey blackandyellow sunday twitpic pittsburgh picks
World	#mubarak	egypt jan mubarak people pro square thugs protesters internet protests egyptian cnn tahrir watch government police news anti egyptians feb violence
	#tahrir	square thugs pro jan liberation clashes blessed cairo live protesters aje egyptian breaking mubarak armed twitpic

CHAPTER 4: TAG-LATENT DIRICHLET PROCESSES AND TAG-LDA WITH CONCEPTS

In the previous chapter, we discussed incorporating tags in the topic models, and specifically proposed Tag-LDA to describe the tags using a distribution of topics. We develop two extensions of Tag-LDA in this chapter. Our goal is to make the model more complete, and also to illustrate Tag-LDA can be extended like LDA.

One natural question for Tag-LDA is whether the number of topics can be decided automatically and how we can achieve that. In this chapter, we develop the Tag-Latent Dirichlet Processes (Tag-LDP), by modifying Tag-LDA with Dirichlet processes, so that the number of topics can be inferred from the data.

Furthermore, the topics extracted from Tag-LDA are based on the co-occurrence of words in the documents, so there is no assurance that they are consistent with the semantic perception of users. We thus introduce the second extension of Tag-LDA, Tag-LDA with concepts, which incorporates users' prior knowledge by employing the Dirichlet Tree priors. Users can influence the learned topics to some extent by providing their own knowledge expressed as concepts, which is an easy scheme for non-expert users to express their prior knowledge in topic models.

4.1 Tag-Latent Dirichlet Processes

LDA provides a way to learn the unobserved topic structure in the corpus, and views the documents as mixtures of latent topics and topics as mixture of terms. In LDA,

the number of topics is an important parameter that needs to be specified in advance. The setting of the number of topics normally depends on the user's experience and is determined in a heuristic way. So one question that might arise unavoidably when applying LDA is: How many topics should be set and is there an automatic method to decide the number? Fortunately, non-parametric statistical methods can help infer the number of topics from data. Teh et al. [75] developed Hierarchical Dirichlet Processes (HDP) which assumes the number of mixture component is an unknown prior and is to be decided from the data. Unlike the parametric Dirichlet prior in LDA, HDP have non-parametric Dirichlet process priors. Inspired by HDP, our work on Tag-Latent Dirichlet Processes (Tag-LDP) utilizes Dirichlet processes so as to decide the topic size automatically from data. Thus we first briefly introduce HDP below.

The Dirichlet process (DP) is a stochastic process, which generalizes the Dirichlet distribution, generating discrete multinomial parameters [32]. Formally, G_0 is a probability measure on a measurable space $(\Theta, \mathcal{B})$ and $(A_1, A_2, \ldots, A_r)$ is a finite measurable partition of Θ. A Dirichlet process $\mathrm{DP}(\alpha_0, G_0)$ is defined as the distribution of random probability measure G over the space $(\Theta, \mathcal{B})$, such that $(G(A_1), G(A_2), \ldots, G(A_r))$ is distributed as a Dirichlet distribution with parameters $(\alpha_0 G_0(A_1), \alpha_0 G_0(A_2), \ldots, \alpha_0 G_0(A_r))$ [75]. G_0 is called the base probability measurement [75] or base distribution [32].

Samples from a DP are discrete and show a cluster property [75]. Let $\theta_1, \theta_2, \ldots$ be a sequence of samples drawn from G. The distribution of θ_k given $\theta_1, \theta_2, \ldots \theta_{k-1}$

follows [11]

$$\theta_k|\theta_1,\ldots,\theta_{k-1},\alpha_0,G_0 \sim \sum_{l=1}^{k-1}\frac{1}{k-1+\alpha_0}\delta_{\theta_l}+\frac{\alpha_0}{k-1+\alpha_0}G_0 \tag{27}$$

where δ_{θ_i} is a probability measure at θ_i . To better exhibit the clustering property, Equation 27 can be rewritten as

$$\theta_k|\theta_1,\ldots,\theta_{k-1},\alpha_0,G_0 \sim \sum_{m=1}^{M}\frac{n_m}{k-1+\alpha_0}\delta_{\phi_m}+\frac{\alpha_0}{k-1+\alpha_0}G_0, \tag{28}$$

where $\phi_1,\ldots,\phi_M$ are the distinct values that $\theta_1,\ldots,\theta_{k-1}$ take, and n_m is the amount of θ_l taking value ϕ_m. We can obtain from Equation 28 that a new sample can take the same value as the previous draws with probability proportional to the number of times the value that has previously been taken. In addition, with probability proportional to α_0, the new sample will take a new value. Comparing to the Dirichlet prior with finite number of components, DP can allow an infinite number of components. There exists a metaphor for this phenomenon known as the *Pólya urn scheme*. In a urn, each atom is associated with a ball painted with distinct color. Every ball can be drawn with equal probability. If a ball is drawn from the urn, it will be put back with an extra one of the same color, i.e. sampling with over-replacement. Additionally, there is certain probability that a new atom can be generated and thus a ball painted with a new color is added to the urn [75]. Alternatively, the sampling behavior can be described by another metaphor called Chinese restaurant process (CRP) [75]. In CRP, we assume there is a Chinese restaurant which can hold infinite number of tables. Guest θ_i enters this restaurant, and could share a table ϕ_m with probability proportional to n_m, or could sit at a new table sampled from the base distribution

G_0.

When using DP to model the mixture structure of LDA, it becomes more complicated due to the topics shared among documents. To overcome this problem, Teh et al. [75] proposed that the document DPs use a shared base distribution which is sampled from another DP, i.e. chain a DP over the document DPs in a hierarchical structure. Teh et al. also described the generative process of the HDP using the metaphor of a Chinese restaurant franchise. In their metaphor, a restaurant franchise (corpus) comprises of multiple restaurants (documents) with a shared menu (topics). Each restaurant can hold an infinite number of tables. Each table can serve only one dish from the menu ordered by the first guest siting at it, and all guests sitting at the table share the same dish. Each guest (word) that enters a restaurant, could either sit at an already occupied table and enjoy the dish with the other guests, or sit at an unoccupied table and order a dish that has never been ordered before (new topic) or has already been ordered by other guests. Figure 10 demonstrates the metaphor.

Our work on Tag-LDP is inspired by the idea of HDP, so that the number of topics can be decided from the data. In Tag-LDP, the tag-topic mixture is expressed by a DP mixture model. Formally the generative process of Tag-LDP is described as follows:

1. Define a base distribution H.

2. Draw $G_0 \sim \mathrm{DP}(H, \epsilon)$.

3. For each tag $p \in \Delta$:

   (a) Draw $G_p \sim \mathrm{DP}(G_0, \gamma)$.

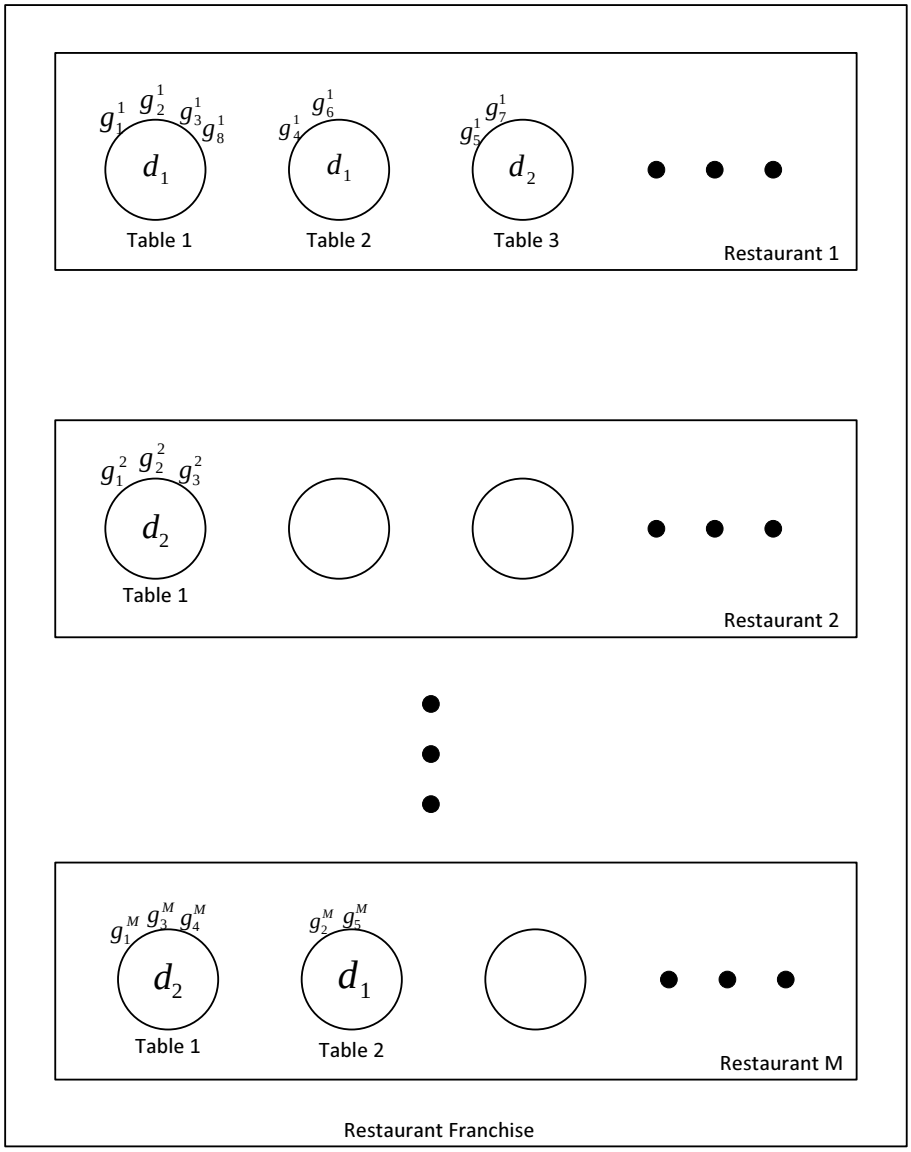


Figure 10: Chinese restaurant franchise. There are M restaurants in this franchise. Each restaurant is able to hold an infinite number of tables and each table can only serve one dish d. Dish d is ordered from the shared menu in this franchise by the first guest g sitting at this table.

4. For each document d with $\boldsymbol{w}_d$ from the corpus:

   (a) Sample a distribution over observed tags from $\boldsymbol{\theta}_d \sim$Dirichlet($\boldsymbol{\alpha}$).

   (b) For ith word in document d:

      i. Sample a tag $e \sim$ Multinomial($\boldsymbol{\theta}_d$).

      ii. Sample a topic $z \sim G_e$, where z could be an existing topic or a newly generated one.

      iii. Sample a term $w \sim \boldsymbol{\phi}_z$, where $\boldsymbol{\phi}_z | G_0$ is drawn from G_0.

In the generative process, ϵ and γ are two parameters required to be set; ϵ influences the new topic generation and γ influences the new table generation as in Chinese restaurant process. The graphical model is shown in Figure 11. Please note the probability of generating a new topic is proportional to ϵ and the probability selecting an existing one is proportional to the count of the topic previously used. With the replacement of the Dirichlet prior by the DP, the Gibbs sampling update equation can be written:

$$P(e_{di} = p, z_{di} = t | \boldsymbol{w}, \boldsymbol{e}_-, \boldsymbol{z}_-, \boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\phi}) \propto \frac{n^p_{d-} + \alpha}{\sum_l^{\boldsymbol{\delta}_d} (n^l_{d-} + \alpha)} \cdot \begin{cases} \frac{1}{|\mathcal{V}|} \cdot \frac{\epsilon H}{\sum_k^\Gamma n^k_{p-} + \epsilon}, & \text{new } t \\ \frac{n^{w_{di}}_{t-} + \phi}{\sum_w^\mathcal{V} (n^w_{t-} + \phi)} \cdot \frac{n^t_{p-} + G^t_0}{\sum_k^\Gamma n^k_{p-} + \epsilon}, & \text{existing } t \end{cases}.$$

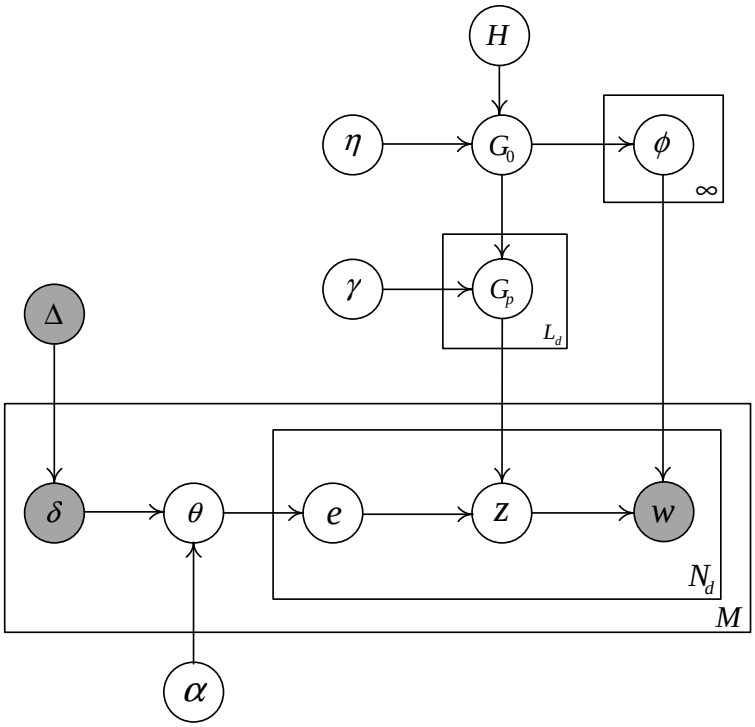


Figure 11: Graphical model of Tag-LDP.

4.2 Experimental Study on Tag-LDP

In our experiments with Tag-LDP, we used the NIPS conference paper dataset[9]. The NIPS paper dataset contains Volume 0 to Volume 12 of the conference proceedings, which consists of 1740 articles contributed by 2037 authors. The vocabulary size of the dataset is 13649. We further processed the dataset by keeping the authors publishing more than ten papers. The process finally retains 28 authors and 379 papers published by these authors. The average word count for each paper is 1378.9, and there are 1.15 authors in average associated with each paper. Our experiments are conducted on a server with Intel XEON E7540 2.0GHz CPU and 128G memory. Tag-LDA is coded in C++, while Tag-LDP is coded in Java. The number of Gibbs

[9]http://www.cs.nyu.edu/~roweis/data.html

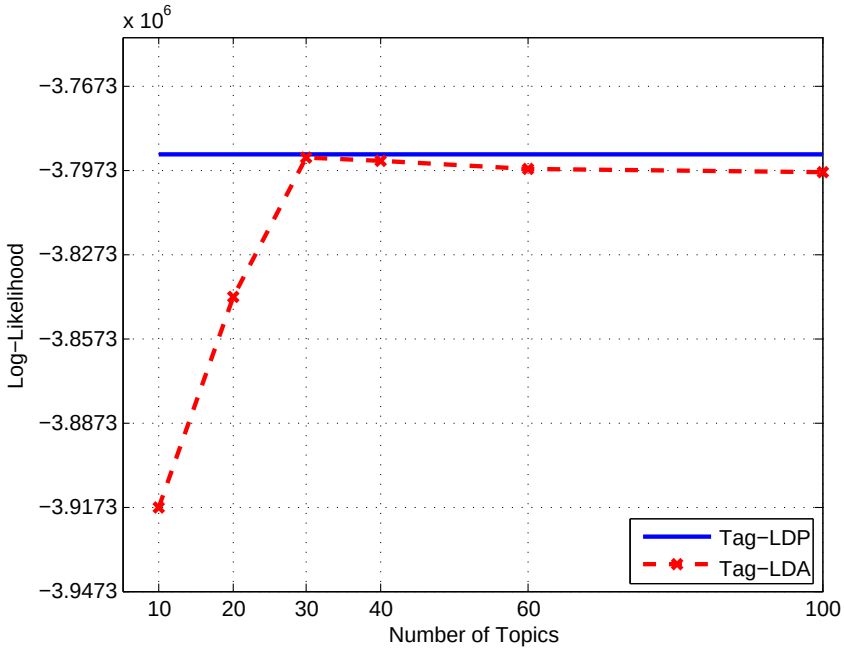


Figure 12: Log-likelihood for Tag-LDP and Tag-LDA with respect to different numbers of topics. Tag-LDP is drawn in the solid blue line. Please note it does not require pre-defined number of topics.

sampling iterations is fixed at 1000 times for both models.

We have mentioned that tags could be any type of meta-data coming with the documents. In this study, the tags refer to the authors. We compared Tag-LDA and Tag-LDP by plotting their log-likelihood on the NIPS dataset in Figure 12. The log-likelihood is given by $\log\left(\prod_w \sum_z \sum_e P(w|z)P(z|e)P(e|d)\right)$. When inferring Tag-LDA, the number of topics is varied from $T = 10$ to $T = 100$. For Tag-LDP, we do not need to set the number of topics, because the number can be inferred from the data. Figure 12 illustrates that the likelihood of Tag-LDP is comparable with the best case of Tag-LDA, so Tag-LDP is able to decide approximately the best number of topics from the data under the likelihood criterion.

To evaluate the quality of the topics, we examine the prominent topic, the topic with

Table 7: The prominent topics for ten authors discovered by Tag-LDP.

Author	Prominent Topic (Tag-LDA)
Jordan_M	jordan model networks field bound forward hidden variables mixture algorithm models likelihood boltzmann probability parameters tree distribution experts architecture em
Hinton_G	hidden hinton units image visible models model mixture data unit cost images vector control code energy weights digit space distribution
Bower_J	cortex neural bower simulation cell networks brain system overlap fiber neuron fig classification olfactory vor fish electric cerebellar cortical model
Lippmann_R	training classifiers classifier rbf speech error word decision figure lippmann gaussian layer classification centers performance nodes regions features rate mixture
Smola_A	kernel support sv space functions pca feature vector regularization data kernels smola training vapnik linear regression sch machines margin kopf
Williams_C	gaussian data distribution covariance posterior process model prior tree matrix noise williams models neural regression space hidden processes networks bayesian
Moody_J	moody data error prediction models decay committee variables smoothing networks trading weight nonlinear information risk inputs selection price stochastic validation
Koch_C	motion voltage koch velocity chip current circuit analog direction spike vlsi flow units pyramidal noise attention conductance image contrast conductances
Waibel_A	word speech recognition tdnn waibel phoneme speaker units connectionist training words multi time vocabulary hme delay networks system sentences connections
Tresp_V	data model gaussian density em variables missing tresp mixture likelihood bayesian models step conditional neural posterior distribution markov nonlinear gaussians

Table 8: The prominent topics for ten authors discovered by Tag-LDA.

Author	Prominent Topic (Tag-LDA)
Jordan_M	jordan model network bound state probability hidden forward models variables networks field tree algorithm likelihood distribution output mixture boltzmann approximation
Hinton_G	units network hidden hinton image object model weights visible unit images models input training mixture layer recognition data gaussian weight
Bower_J	cortex neural bower neuron cell simulation system fig fiber neurons electric brain stimulus overlap olfactory cerebellar classification fish cells spikes
Lippmann_R	training classifiers classifier rbf speech error layer word input decision performance gaussian figure lippmann state classification nodes centers output regions
Smola_A	kernel space support functions vector sv feature pca function regularization linear data kernels smola vapnik case regression sch problem machines
Williams_C	gaussian data distribution process prior covariance model function posterior matrix space tree williams regression processes noise models neural log hidden
Moody_J	moody data input prediction error units models decay committee training network variables weight smoothing functions inputs networks information trading nonlinear
Koch_C	motion voltage koch noise current velocity chip circuit spike analog direction vlsi time flow channel pyramidal synaptic neuron contrast fig
Waibel_A	word recognition speech training network tdnn waibel networks units phoneme input level speaker time hidden neural performance system connectionist layer
Tresp_V	data network model variables tresp gaussian missing density training input neural bayesian likelihood based networks models distribution conditional markov variable

the greatest probability, for each author. Table 7 and 8 show the prominent topics for ten authors as examples. We list the top 20 terms of each topic here. We review the first four authors in here. Jordan_M refers to Michael I. Jordan, and his prominent topic expresses his research related to graphical models with the representative terms: forward, hiden, variables, likelihood, boltzmann, etc., while James Bower (Bower_J), as a neuroscientist, seems to mainly works on neuron computing simulation and neural network related directions. Significant difference in the prominent topics from Tag-LDP and Tag-LDA is not identified from our observation. Therefore, Tag-LDP is able to discover topics of similar quality as Tag-LDA, and also can decide the number of topics automatically.

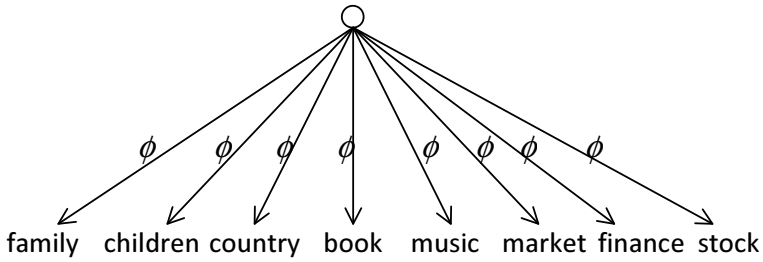
4.3 Tag-LDA with Concepts

Topic models discover the statistical pattern of topics from documents. However the topics might not satisfy people's semantic perception [56]. The semantic perception can be reflected by users' strong beliefs about the probabilities of terms in the topics. For example, the term "market" and term "finance" might have similar probability under a given topic, since these two terms often co-occur in users' minds.

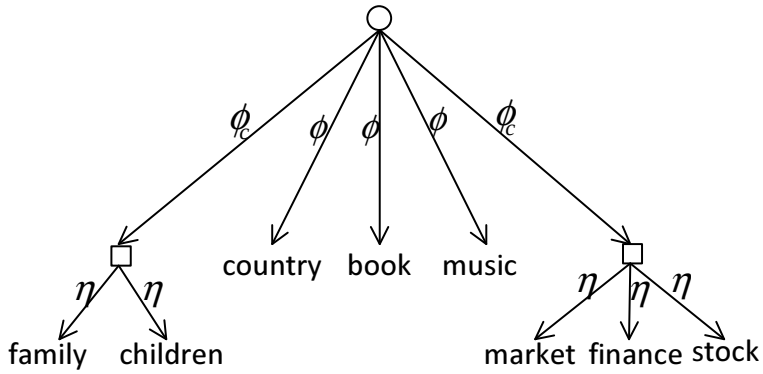
One possible reason for this weakness in topic models is the unsupervised learning procedure lacking input from human knowledge [23]. So several researchers [23, 4, 3, 37] have proposed learning more meaningful and semantically coherent latent topics by involving human knowledge as a prior. Our work is inspired by the previous work of [37, 3], and similarly allows concepts provided by users to be incorporated in Tag-LDA. We name this extension *Tag-LDA with concept* (ConceptTag-LDA).

Specifically, users might have pre-defined constraints on the terms in practice, which can be viewed as *concepts*. A concept is a set of terms that are considered semantically close to each other, pre-defined by the users with respect to the specific application domain. The concepts can be viewed as users' intuition or prior knowledge. Like the example we used above, "market" and "finance" are usually seen together in documents. Thus a concept is a set of terms that are considered semantically related to some extent, so they might have high probabilities in the same topic. Topic models with concepts inserted are therefore able to bias the topics based on terms that users view relevant, such that data is modeled with users' intuition encoded.

To model concepts in our topic model, we adopt the Dirichlet Tree prior [57, 3] to replace the original Dirichlet prior. The Dirichlet Tree distribution is able to describe the correlation of terms of the concepts in the topics. All terms in a concept are likely to share high probabilities together or low probabilities in the topics. Figure 13 shows the structure of a Dirichlet Tree prior with two concepts, and the corresponding Dirichlet prior for comparison. The Dirichlet Tree on the first level is a multinomial distribution over both terms and concepts, and the concepts on the second level can be viewed as multinomial distributions over terms. Although we only show an example of a two-level prior structure, theoretically the prior structure can be generalized to more than two levels [57]. That is, a concept can recursively have sub-concepts. In this paper, we only allow concepts of one level. For the sake of simplifying model inference and not introducing ambiguity in the generative process, we add a restriction to the concepts, which is that concepts are not allowed to share common terms, which means terms under each concept exclusively belong to their parent concept.



(a) Dirichlet prior structure.



(b) Dirichlet Tree prior structure with two concepts.

Figure 13: Dirichlet prior structure of LDA and Dirichlet Tree prior structure with concepts. Assume there are eight terms in the vocabulary, and two concepts are provided. (a) The Dirichlet prior for LDA, where each term has an equal prior. (b) The Dirichelt Tree prior with two concepts: {family, children} and {market, finance, stock}. On the first level, the priors of the concepts are not identical to those of the regular terms. Once a concept is selected, the probabilities of terms being emitted in the concepts are likely to be simultaneously high or low. Note that one concept is not correlated with the other.

Tag-LDA can be viewed as a three-layer structure. With the illustrations of the Dirichlet prior above, we draw Figure 14 as an example illustrating the three-layer structure for one tag. From the top to the bottom, the block denotes a tag; the middle circles denote topics; and the triangles at the leaves denote the terms. In the Dirichlet distribution, the terms under one topic are mutually independent with the constraint that their probabilities sum to one [58]. We want to capture the constraints on the terms in a concept that the terms have similar probabilities despite their mutual independence. There are internal nodes between topics and terms in Dirichlet Tree distribution, which are able to capture these constraints on the concepts. The shaded circle in Figure 15 is a concept consisting of two terms. The advantage of using a Dirichlet Tree prior is avoiding large probability of terms in concepts being forced in all topics [1]. Reversely, in Dirichlet distribution, increasing the priors of terms may generate similar probabilities for terms in topics, however, the increased priors would influence the probabilities of terms in every topic, which definitely is not what we like.

In standard LDA with a Dirichlet prior, each topic is a distribution of terms. After substitution of the Dirichlet Tree prior as in the example above, each topic is a distribution of terms and concepts on the first level, and a concept is a distribution over terms on the second level. As we discussed above, please note one term can only appear once in the structure, either under a concept on the second level or on the first level. When generating a document, if the word belongs to a certain concept, the corresponding concept is first selected and then the word is emitted. Otherwise, the word is emitted directly as in the standard LDA. More specifically, the generative

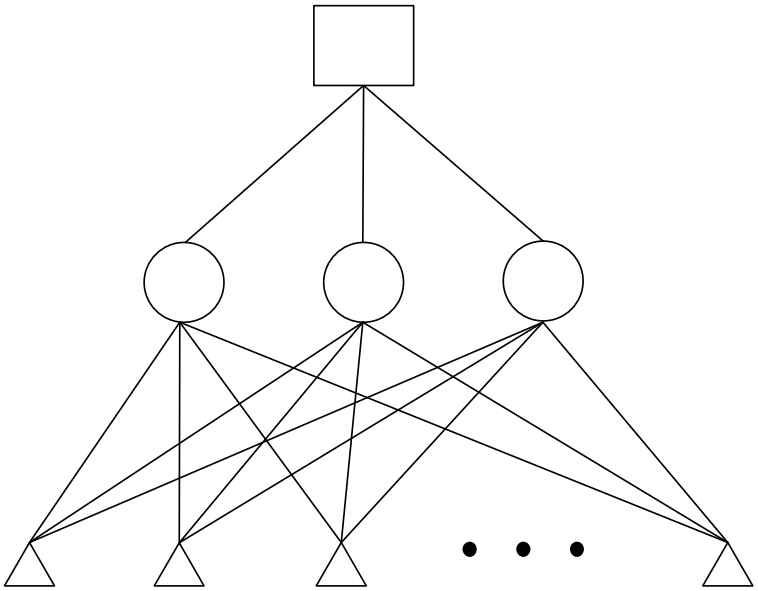


Figure 14: Tag-LDA structure. The block, the circle, and the triangles denotes tag, topic, and terms respectively.

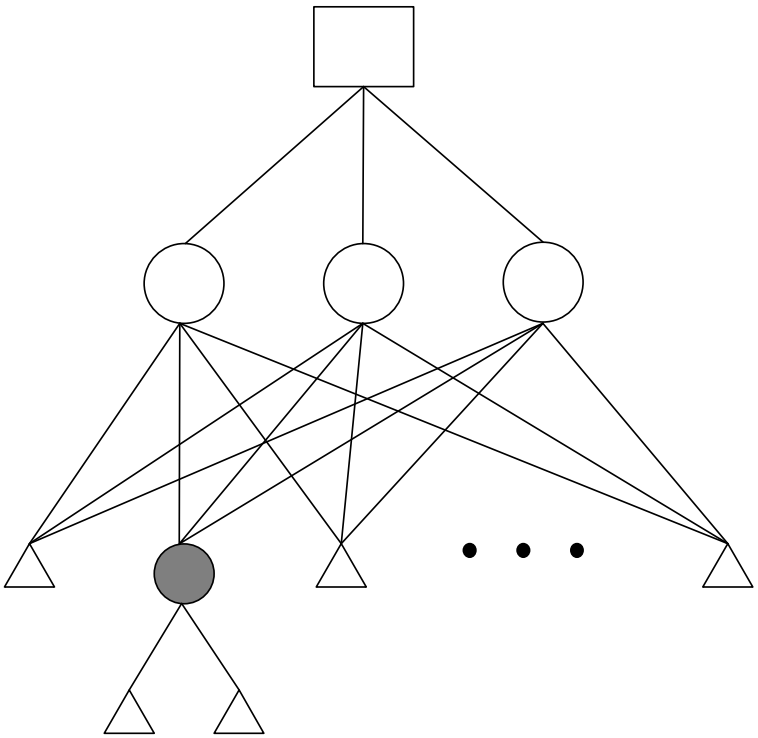


Figure 15: ConceptTag-LDA structure. The shadowed circle is a concept of two terms denoted by triangles. The blank circle denotes a topic and the block denotes a tag.

process is updated with concepts introduced as below:

1. For each tag $p \in \Delta$, sample $\boldsymbol{\gamma}_p$ over all topics from $\boldsymbol{\gamma}_p \sim \text{Dirichlet}(\boldsymbol{\rho})$.

2. For each topic $t \in \Gamma = \{t_1, t_2, \ldots, t_T\}$:

   (a) sample $\boldsymbol{\beta}_t$ over B edges (terms and concepts) from $\boldsymbol{\beta}_t \sim \text{Dirichlet}(\boldsymbol{\phi}')$.

   (b) for each concept $s \in \mathscr{C}$:

      i. sample $\boldsymbol{\beta}_s$ over the terms in the concept $\boldsymbol{\beta}_s \sim \text{Dirichlet}(\boldsymbol{\eta})$.

3. For each document d with $\boldsymbol{w}_d$ from the corpus:

   (a) Sample a distribution over observed tags from $\boldsymbol{\theta}_d \sim$Dirichlet($\boldsymbol{\alpha}$).

   (b) For ith word in document d:

      i. Sample a tag $e \sim \text{Multinomial}(\boldsymbol{\theta}_d)$.

      ii. Sample a topic $z \sim \text{Multinomial}(\boldsymbol{\gamma}_e)$ a multinomial probability conditioned on current tag assignment e.

         A. If emit a word, sample the word $w \sim \text{Multinomial}(\boldsymbol{\beta}_z)$.

         B. Otherwise, first sample a concept c, and then sample a term $w \sim$ Multinomial($\boldsymbol{\beta}_c$).

The Dirichlet Tree distribution is also conjugate to the multinomial distribution like the Dirichlet distribution. This property is helpful in deriving the Gibbs sample update equation. Please note that regardless of whether or not concepts are involved in the model learning procedure, the joint distribution $P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})$ in

Equation 20 still holds. It is easy to observe that introducing concepts would not influence the second term in the joint distribution but only the first term. Formally, we define the symmetric Dirichlet prior η for the terms in the concepts, and assume there are π given concepts $\mathscr{C} = \{s_1, s_2, \dots s_\pi\}$. The vocabulary $\mathcal{V}$ can be split into two subsets (1) $\mathcal{V}_r$ regular terms not in concepts and (2) $\mathcal{V}_{\mathscr{C}}$ terms belonging to concepts. $P(\boldsymbol{w}|\boldsymbol{\phi}', \boldsymbol{\eta}, \boldsymbol{z})$ can be be factored as below (interested readers please refer to the detailed mathematical derivation in Section 3 in [58])

$$P(\boldsymbol{w}|\boldsymbol{\phi}', \boldsymbol{\eta}, \boldsymbol{z}) = \prod_{t=1}^{T} \left(\frac{B(\boldsymbol{n}'_t + \boldsymbol{\phi}')}{B(\boldsymbol{\phi}')} \prod_{s=1}^{\pi} \frac{B(\boldsymbol{n}_{ts} + \boldsymbol{\eta})}{B(\boldsymbol{\eta})} \right), \tag{29}$$

where $\boldsymbol{n}'_t = (n_t^1, n_t^2, \dots n_t^{|\mathcal{V}_r|}, n_{ts_1}, n_{ts_2}, \dots n_{ts_\pi})$ is the vector of the counts of the individual terms in $\mathcal{V}_r$ and the counts of the concepts in $\mathscr{C}$ assigned to topic t; the size of $\boldsymbol{n}'_t$ is $|\mathcal{V}_r| + \pi$; $\boldsymbol{n}_{ts} = (n_t^{w_1^s}, n_t^{w_2^s}, \dots, n_t^{w_m^s})$ contains the counts of terms in concept s assigned to topic t; $\boldsymbol{\phi}'$ and $\boldsymbol{\eta}$ are their corresponding Dirichlet priors. When counting n_{ts_i}, it is the sum of $n_t^{w^{s_i}}$. Substituted $P(\boldsymbol{w}|\boldsymbol{\phi},\boldsymbol{z})$ in Equation 22, the joint distribution for ConceptTag-LDA is

$$P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}', \boldsymbol{\eta}, \boldsymbol{\rho}) = \prod_{t=1}^{T} \left(\frac{B(\boldsymbol{n}'_t + \boldsymbol{\phi}')}{B(\boldsymbol{\phi}')} \prod_{s=1}^{\pi} \frac{B(\boldsymbol{n}_{ts} + \boldsymbol{\eta})}{B(\boldsymbol{\eta})} \right) \cdot \prod_{p=1}^{L_d} \frac{B(\boldsymbol{n}_p + \boldsymbol{\rho})}{B(\boldsymbol{\rho})} \cdot \prod_{d=1}^{M} \frac{B(\boldsymbol{n}_d + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}. \tag{30}$$

Now we turn to the Gibbs sampling update equation. As before, we utilize Equation 23 to derive the update equation. We can easily see that the change to $P(\boldsymbol{w}|\boldsymbol{\phi}, \boldsymbol{z})$ in the joint distribution would not influence the second and third terms in the original update equation, so only the first term must be modified. Specifically, during

sampling iteration, if the word w_{di} being generated is a regular term in $\mathcal{V}_r$ the simplification to $P(e_{di} = p, z_{di} = t|\boldsymbol{w}, \boldsymbol{e}_-, \boldsymbol{z}_-, \boldsymbol{\alpha}, \boldsymbol{\phi}', \boldsymbol{\eta}, \boldsymbol{\rho})$ will result in a similar update equation as Equation 25. However, if the word w_{di} is from concept c in $\mathcal{V}_{\mathscr{C}}$, updating the topic assignment might alter both the count of the concept it belongs to and its own count. After simplification, the Gibbs sampling update equation is rewritten as

$$P(e_{di} = p, z_{di} = t|\boldsymbol{w}, \boldsymbol{e}_-, \boldsymbol{z}_-, \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\phi}) \propto$$

$$\begin{cases} \frac{n_{t-}^{w_{di}}+\phi}{\sum_w^{\mathcal{V}_r}(n_{t-}^w+\phi)+\sum_s^{\mathscr{C}}(n_t^s+\phi_s)} \cdot \frac{n_{p-}^t+\rho}{\sum_k^{\Gamma}(n_{p-}^k+\rho)} \cdot \frac{n_{d-}^p+\alpha}{\sum_l^{\boldsymbol{\delta}_d}(n_{d-}^l+\alpha)}, & \text{if } w_{di} \in \mathcal{V}_r \\ \frac{n_{tc-}+\phi_c}{\sum_w^{\mathcal{V}_r}(n_t^w+\phi)+\sum_s^{\mathscr{C}}(n_{ts-}+\phi_s)} \cdot \frac{n_{tc-}^{w_{di}}+\eta}{\sum_w^c(n_{tc-}^w+\eta)} \cdot \frac{n_{p-}^t+\rho}{\sum_k^{\Gamma}(n_{p-}^k+\rho)} \cdot \frac{n_{d-}^p+\alpha}{\sum_l^{\boldsymbol{\delta}_d}(n_{d-}^l+\alpha)}, & \text{if } w_{di} \in \mathcal{V}_{\mathscr{C}} \end{cases}, \quad (31)$$

Equation 31 is the simplified update equation for the Dirichlet Tree prior with concepts of one level. A Generalized updating equation for multiple level concepts is not difficult to derive by directly rewriting Equation 29.

The estimates of document-tag distribution $\boldsymbol{\theta}$ and tag-topic distribution $\boldsymbol{\gamma}$ are the same as those of Tag-LDA, because the Dirichlet Tree prior has no influence on them in the model. The estimate of topic-term distribution $\boldsymbol{\beta}$ is updated as follows:

$$\hat{\beta} = \begin{cases} \frac{n_t^w+\phi}{\sum_w^{\mathcal{V}_r}(n_t^w+\phi)+\sum_s^{\mathscr{C}}(n_t^s+\phi_s)} & \text{if } w \in \mathcal{V}_r \\ \frac{n_t^c+\phi_c}{\sum_w^{\mathcal{V}_r}(n_t^w+\phi)+\sum_s^{\mathscr{C}}(n_t^s+\phi_s)} \cdot \frac{n_{tc}^w+\sigma}{\sum_w^c(n_{tc}^w+\sigma)} & \text{if } w \in \mathcal{V}_{\mathscr{C}} \end{cases} \quad (32)$$

4.4 ConceptTag-LDA Evaluation Experiments

We conducted several experiments to evaluate ConceptTag-LDA. Our experiments were conducted on a server with Intel XEON E7540 2.0GHz CPU and 128G memory. Both Tag-LDA and ConceptTag-LDA are coded in C++. The hyper-parameter

settings are configured as follows: $\alpha = 0.1$, $\rho = 0.1$, $\eta = 100$, unless stated otherwise. The number of Gibbs sampling iterations is fixed at 1000 for both models.

The proceeding discussion of ConceptTag-LDA assumes the concepts are already known. How the concepts are initially defined has not been touched so far. In practice, when defining the concepts, users normally apply their subjective knowledge [74]. There is no doubt that the concept reflects users knowledge very well in this way, but the quality and rationality of the concepts cannot be guaranteed due to this subjective approach, which presumably depends on how well users understand the data.

In our work, we extract concepts automatically from the corpus with respect to the tags. Specifically, the extraction steps are as follows:

1. Extract top five important words out of each document based on word's TF-IDF.

2. Compute the co-occurrence frequency of each pair of important words.

3. Select the most frequently co-occurring pair as the concept seeds.

4. Expand the concept by adding words from the remaining three most important words which co-occur mostly with either one of the seeds. In our experiments, we added one word to make the concepts be of three words.

5. Check whether there are common words between concepts.

4.4.1 Qualitative Comparison on Topics

To explore the quality of topics generated by ConceptTag-LDA and to compare it to Tag-LDA, we present results on the New York Times corpus[10]. The New York Times Corpus is a news collection of the New York Times. It comes with General Online Descriptors, assigned automatically and verified by nytimes.com production staff, for each news article. A news article might have multiple descriptors associated with it. Examples of individual descriptors include "Finances", "International relations", "Computers and the Internet", etc.. We consider these descriptors as the tags in our models. For the purpose of our experiments, we randomly selected the news articles published in the period October 1–15, 2006, and filtered out descriptors that occur infrequently, which finally results in 24 descriptors (tags). Overall, 1933 news articles remain left in the dataset. On average, there are 2.14 descriptors attached for each news article, and the average length of the article is 323.9 words after the stop words are eliminated.

We generated six concepts from the dataset following the approach above for six randomly selected tags. The concepts are listed in Table 9. Among these six concepts, every concept consists of three terms except C4 where the term "percent" is removed since it is shared with C3.

Inferring topic models by Gibbs sampling methods usually takes some time, because Gibbs sampling is a sequential procedure where the topic and tag assignment are sampled for each word in the documents until the convergence is reached. For

[10]https://catalog.ldc.upenn.edu/LDC2008T19

Table 9: Six concepts automatically extracted for six tags from the New York Times dataset. The second column is the terms in the concept and the third column is the corresponding tag.

Concept ID	Concept Terms	Tag
C1	series league manager	baseball
C2	life novel set	books and literature
C3	executive percent web	computers and the internet
C4	money billion	finances
C5	children food family	medicine and health
C6	president bush united	politics and government

ConceptTag-LDA inferring on this dataset with 30 topics, the running time is 14 hours in average, while Tag-LDA cost less time at four hours 18 minutes.

We trained 30 topics using Tag-LDA and ConceptTag-LDA. We list the topics by showing the top terms for several topics below in Figure 10. We can see that among the example topics, terms belonging to the same concept are ranked high together. Please note although the concepts are extracted with respect to the tags, it does not mean the concepts must be ranked high in the prominent topics. This reflects the advantage of the Dirichlet Tree prior, which is able to restrict the co-occurrence of the terms inside of concepts without increasing the probability of the concepts in every topic. One future research direction is to tie the concepts with the tags, which might increase the efficiency of the topics in explaining the tags.

To provide a comparison showing the topic differences between Tag-LDA and ConceptTag-LDA, we also present the topics extracted by Tag-LDA in Figure 11. Without the concept restriction in Tag-LDA, parts of the concepts are ranked high in the prominent topics. Taking the topic of “baseball” as an example, “manager” and “league” are shown but not “series”.

Table 10: Prominent topics of several tags extracted by ConceptTag-LDA. Terms in each topic are ordered based on their probability. Concepts terms show up in the prominent topics of the first three tags but not the last one.

Tag	Prominent Topic
books and literature	book books story author life novel bosch stone set hepburn writes history father woman pamuk street reader mother read hughes death literary writer american louis black
baseball	game mets yankees series baseball season team league manager tigers torre games rodriguez play postseason run runs left ball pitch inning division players cardinals
finances	company companies funds money fund billion percent market million stock investors executive prices web pay chief stocks investment oil financial business shares firm markets price
politics and government	court law government religious political tax justice church federal party hamas states supreme palestinian judges judge israel courts county city country legal officials justices
medicine and health	health patients drug food care medical disease drugs cancer doctors children treatment family patient study hospital pigeons school eat milk heart fats researchers fat trans
computers and internet	company google youtube video site internet computer microsoft software online yahoo companies web technology media music friendster windows parks sites executive ads percent vista sandy

Table 12 illustrates the statistics of concept term co-occurrence in the topics. For these two models, we checked the top 20 terms of each topic and count how many topics contain only one concept term, two concept terms, and three concept terms. For instance, all of the three terms of concept C1 are observed in one topic in ConceptTag-LDA, while no topic contains those three terms of C1 in Tag-LDA. From this table, we see that most of concept terms are ranked high simultaneously in ConceptTag-LDA, which means ConceptTag-LDA can integrate the human's concepts in the topic representation to some extent.

A natural question that arises is: What topics would ConceptTag-LDA generate if the concepts have low quality? This can occur if the concepts fit the users intuition but deviate too much from the real data. This is a valid question and it could happen

Table 11: Prominent topics of several tags extracted by Tag-LDA. Terms in each topic are ordered based on their probability.

Tag	Prominent Topic
books and literature	book books novel life story author stone bosch world father love hepburn writes history writer mother war reader woman pamuk death writing read film literary published street
baseball	torre yankees baseball team rodriguez players fans steinbrenner season manager league sports yankee fan played leyland base mets teams piniella neil cashman joe jeter world
finances	percent company companies funds money fund billion market million stock investors prices executive investment chief pay financial stocks oil business private firm shares price markets
politics and government	time york day city american public home world million life percent five president family country left days month national set states times united office university director former
medicine and health	health patients drug medical food care disease drugs cancer treatment doctors study patient hospital pigeons states percent milk researchers medicine research pigeon hospitals companies heart
computer and internet	company google youtube video site web internet computer microsoft software online companies yahoo media technology sites friendster windows parks music billion executive vista sandy ads

Table 12: Concept term co-occurrence statistics of Tag-LDA and ConceptTag-LDA. The left column is the concept ID. For each concept, we count the number of topics where only one concept term, two concept terms, and three concept terms appear together in the first 20 terms of the topic.

Concept ID	Tag-LDA			ConceptTag-LDA		
	1	2	3	1	2	3
C1		2				1
C2	1	2				1
C3	6	1			1	2
C4		2			1	
C5	4	1		1	1	1
C6	2	2			1	1

in practice. Users may not understand the data very well so the concepts defined by them could be very biased. Taking an extreme case as an example, consider what the topics would be if the concepts are made up arbitrarily. To test this scenario, we made up two fake concepts by mixing the concepts above. The two concepts are $FC1 = \{\text{league, web, money, children}\}$ and $FC2 = \{\text{food, series, united, novel}\}$. We conducted the experiment using these two concepts with the same experiment setup as above. We checked the top 50 terms in each topics, and Table 13 shows the distribution of concept terms in the topics. We observe that concept terms still appear up together in most cases as expected. However, the introduction of the fake concepts seems to deteriorate the quality of the topics. Table 14 lists Topic 0 which contains all terms of $FC1$ and Topic 8 which contains all terms of $FC2$. Most terms of Topic 0 appear to discuss a family related theme, but "league" and "web" do not seem to belong in this theme. Topic 8 illustrates the same phenomenon; "series", "food", and "novel" do not appear relevant in a topic related to international relationships. Therefore, although ConceptTag-LDA provides an intuitive method for non-expert users to model data in a customized way, improper selection of on the concepts will lead to the generated topics deviating from the data too much.

4.4.2 Quantitative Comparison Between ConceptTag-LDA and Tag-LDA

In our next experiment, we studied the influence of the concepts in modeling data. We compare ConceptTag-LDA and Tag-LDA by comparing their log-likelihood. In these experiments, we use the same New York Time dataset and those concepts as in Section 4.4.1.

Table 13: Topics that contain the terms of the fake concepts. The symbol * indicates that term appears in the top 50 terms of that topic.

	Concept \ Topics	Topic 0	Topic 26	Topic 8	Topic 16	Topic 22	Topic 26	Topic 27
$FC1$	league	*						
	web	*						
	money	*	*					
	children	*	*					
$FC2$	food			*	*	*	*	
	series			*		*	*	
	united			*		*	*	*
	novel			*		*	*	

Table 14: Two topics extracted from ConceptTag-LDA with two fake concepts.

	Terms
Topic 0	school parents parenting child mother roberts kids jabari children police amish teenagers night money web league youth alexander mcdonald girl love family families care parent daughter chess baby miller life evangelical college son nickelodeon day street minutes hour football brewster ben play game fleming boys husband home friends arrived crew
Topic 8	north korea nuclear american states military security officials test iraq weapons iran china bush iraqi war south korean government nations country administration baghdad united president international japan forces council russia killed sanctions bomb countries troops police official world threat series food novel shiite minister foreign policy intelligence monday afghanistan sunni

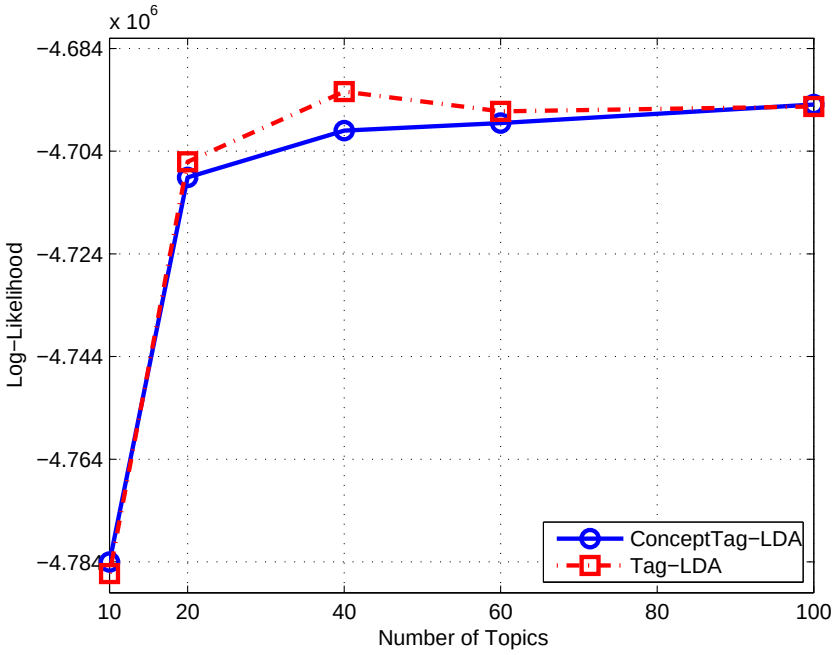


Figure 16: Log-likelihood for ConceptTag-LDA and Tag-LDA with number of topics $T = \{10, 20, 40, 60, 100\}$. The blue solid line is ConceptTag-LDA, and the red dotted line is Tag-LDA.

First we learned both models with varying number of topics $T = \{10, 20, 40, 60, 100\}$ and output the corresponding log-likelihood. Figure 16 illustrates the log-likelihood for both models corresponding to different T. We can observe that the difference in log-likelihood is insignificant between these two models, especially for small and high values of T. So, with only six concepts containing 17 terms, it is not surprising that they have a minor influence in modeling the data overall. We can also observe that the log-likelihood of ConceptTag-LDA is slightly lower than Tag-LDA for $T = 40$. Our explanation of the discrepancy is that, although the introduction of concepts makes the learned topics semantically closer to users prior knowledge, the given concepts may violate the true statistical relationship of the terms in the data. So users should be cautious when defining the concepts.

In the next experiment, we examined the parameter η, which is the prior for concept terms. Its value influences the co-occurrence of the concept terms in the topics. We compare the log-likelihood of the model with multiple η settings, changing from 50 to 1200; other hyper-parameters were kept the same as in the previous experiments and the number of topics T is also fixed at 30. The log-likelihood is shown in Figure 17. The blue solid curve is ConceptTag-LDA, and the red dashed line is Tag-LDA with fixed $T = 30$ as a reference. We see that the likelihood gets lower as η increases. As a prior, η influences the co-occurrence of concept terms in the topics. Large η encourages equivalent probabilities of the terms, but may not represent the true statistical relationship of the terms and eventually leads to inaccurate modeling of the data, which might be the reason for inducing the lower log-likelihood. Again, from the likelihood criterion perspective, users pay a small price to integrate their prior knowledge in the model.

4.5 Summary

This chapter has presented two extensions building on Tag-LDA: Tag-LDP and ConceptTag-LDA. To avoid the users having to set the number of topics, we propose Tag-LDP model, which is inspired by Hierarchial Dirichlet Processes. Tag-LDP can infer the number of topics from the data automatically and can generate topics with quality comparable to Tag-LDA. ConceptTag-LDA replaces the Dirichlet prior by the Dirichlet Tree prior, which models the concepts in the form of a group of terms given by the users. The concepts reflect users' prior knowledge regarding the terms, so ConceptTag-LDA enables customized modeling of the text by users. We provide

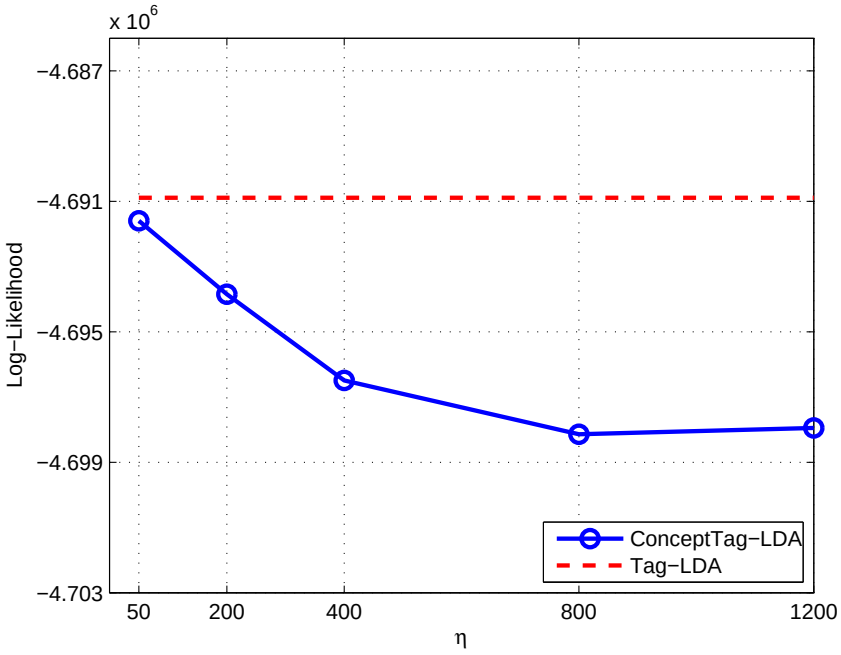


Figure 17: Log-likelihood values for ConceptTag-LDA and Tag-LDA. The number of topics is fixed for both models. ConceptTag-LDA has a varying $\eta = \{50, 200, 400, 800, 1200\}$.

Gibbs sampling based solutions to learn the models and our experiments show the characteristics and the usefulness of these two models.

CHAPTER 5: A DYNAMIC TWITTER TOPIC MODEL

Social media such as Twitter captures moment-by-moment updates of discussions among people. The discussions are constantly evolving with many discussions centering around events. Therefore, it is essential to consider the temporal dynamics when summarizing and analyzing the discussions. A major event usually involves twists and turns reflected by multiple sub-events. This temporal event development is in turn reflected by people's discussions on Twitter. To highlight this temporal dynamics, we propose the dynamic Twitter topic model (DTTM), a specialized topic model tailored for the short messages in Twitter. We assume an event can be modeled by a mainstream theme plus several facets. We are inspired by temporal topic models and model the mainstream theme and these facets to evolve with time.

5.1 Introduction

User-generated content on Twitter captures minute- by-minute updates of public and private snippets of information. Many of the discussions on Twitter center on events of interest to people, evolving rapidly over time. To analyze and make sense of the wealth of information on Twitter, summarizing the user-generated content is a necessary step. More interesting, given the velocity of tweets, it is beneficial to summarize the content in a way that also highlights the ebb and flow of the moment-by-moment discussions.

One way to summarize social media content and capture the major themes is to leverage topic models [12]. Most of the topic models are designed for regular, well-written text such as news, blogs, and scientific papers [18, 29, 70, 69, 66, 64]. Others have developed specialized topic models for the analysis of social media such as Twitter [63, 88]. Since the short and noisy nature of social media messages differs from other regular text resources, researchers have developed ways of assembling social media content in order to better utilize topic models designed for regular texts. Such methods include aggregation strategies [35, 85] for shorter, more fragmented data. Hong and Davison in [35] empirically examined three different aggregation schemes where the standard Latent Dirichlet Allocation (LDA) model [18] was applied.

Although the above topic model with aggregation strategies can extract topics as meaningful summaries of social media data, they did not consider the temporal evolution of the discussions, which is an essential characteristic of communications on social media. To incorporate the temporal evolution of topics, some previous work, e.g., [16, 36], model topics along time in a Markovian manner, in that the current state of a topic is dependent on its previous state. Wang et al. in the continuous dynamic topic model (cDTM) [82] replaced the state space model in the dynamic topic model (DTM) [16] with the Brownian motion model such that continuous time-series data can be modeled with arbitrary granularity. However, the aforementioned topic models are designed for regular texts, and thus may be less effective at modeling the dynamics of short messages such as tweets.

To explore the temporal evolution of topics that pervade a collection of tweets, we propose the dynamic Twitter topic model (DTTM), a new model that not only models

the dynamic nature of the topics, but is also tailored for shorter tweets. Specifically, for a collection of related tweets discussing one major event, we assume it can be described by one mainstream topic plus multiple facet topics, and each tweet can be viewed as a mixture of two topics: the mainstream topic and one facet topic. In addition, our work is inspired by previous dynamic topic models and assumes these topics evolve as time proceeds.

To evaluate the quality of topics from DTTM, we compared it to the dynamic topic model (DTM) [16] and LDA. Our experiment results show that the DTTM model provides more distinctive topics and better coverage of the details of the events. To enable the understanding of the changes in topics over time, we used a visualization to convey the statistical results. The visualization enables the analysis of topic trends over time, as well as highlights the topic term difference in time. By combining the change in topic trends (such as bursts) and the change in topic terms, the visualization illustrates the evolution of events that may otherwise be buried in the topics.

5.2 Dynamic Twitter Topic Model

In this section, we introduce a dynamic Twitter topic model particularly designed to capture the dynamic property of topics. The special characteristic of tweets, which is of much shorter length than regular articles, is considered in designing DTTM. Usually, topic models assume each document is a mixture of topics across all topics, while DTTM limits the number of topics in tweets due to their short length. To discover the topic trends over time, in DTTM we adopt the idea of modeling temporally varying topics from DTM. This will potentially enable us to analyze the development of

related events and evolution of people's discussion.

To describe the multiple aspects of an event, we intuitively assume there is a mixture of multiple facet topics for that event at every time slice. Each tweet at a certain time slice is generated by one of these facet topics and the shared mainstream topic between tweets. Formally, an event p is modeled as a multinomial distribution θ_t of K facet topics plus a shared mainstream topic c between tweets at time t, with a total of $K+1$ topics in the model. Topics are latent and normally are defined as a multinomial distribution over all terms in vocabulary $\mathcal{V}$. A tweet at time t is composed of words generated by facet topic s, where $s \sim \text{Multinomial}(\theta_t)$, and the shared mainstream topic c. So as to model the temporal dynamics of topics, we adopt the idea of DTM [16] mentioned before and assume $\boldsymbol{\beta}_t$, the topic-term distribution for time t, depends on $\boldsymbol{\beta}_{t-1}$ with Gaussian noise in a state space model. In other words, we assume $\beta^k_{w,t}$, the probability of a term w in topic k at time slice t follows a Gaussian distribution with mean $\beta^k_{w,t-1}$ and a constant variance. So topics evolve as time proceeds, and words in tweets are generated by the corresponding topics at the same time slice. In cDTM [82], this assumption is generalized by using a Brownian motion model where the variance linearly depends on the time lag.

DTTM is also a generative model like other topic models. The generative process for it is described as follows:

1. For each topic k in $K+1$:

   (a) Draw topic $\beta^k_t|\beta^k_{t-1} \sim \mathcal{N}(\beta^k_{t-1}, \sigma^2 I)$.

2. For the event at time slice t:

Table 15: Notation table for DTTM.

symbol	size	description
t	scalar	time slice
$\boldsymbol{w}_d$	$1 \times N_d$	words of tweet d
s	scalar	topic assignment for each tweet
$\boldsymbol{\epsilon}_d$	1×2	tweet-topic multinomial distribution
z	scalar	topic assignment for each word
$\boldsymbol{\beta}_t$	$(K+1) \times \lvert\mathcal{V}\rvert$	topic-term multinomial distribution at time t
θ_t	$1 \times K$	event-topic multinomial distribution at time t
$\boldsymbol{\eta}$	1×2	Dirichlet hyperparameters
$\boldsymbol{\alpha}$	$1 \times K$	Dirichlet hyperparameters

(a) Draw $\theta_t \sim \text{Dirichlet}(\boldsymbol{\alpha})$.

3. For each tweet d at time slice t:

(a) Draw $s \sim \text{Multinomial}(\theta_t)$.

(b) Draw $\boldsymbol{\epsilon} \sim \text{Dirichlet}(\boldsymbol{\eta})$.

(c) For each word :

i. Draw $z \sim \text{Multinomial}(\boldsymbol{\epsilon})$.

ii. Draw $w \sim \text{Multinomial}(f(\beta_t^z))$.

Function f maps the multinomial natural parameters to mean parameters and we adopt $f = \frac{\exp(\beta_t^{z,w})}{\sum_{w'}^{\mathcal{V}} \exp(\beta_t^{z,w'})}$. The graphical model is demonstrated in Figure 18 and the notation in Table 15. Note that the model, for simplicity, does not allow the hyperparameter $\boldsymbol{\alpha}$ to evolve unlike in DTM [16].

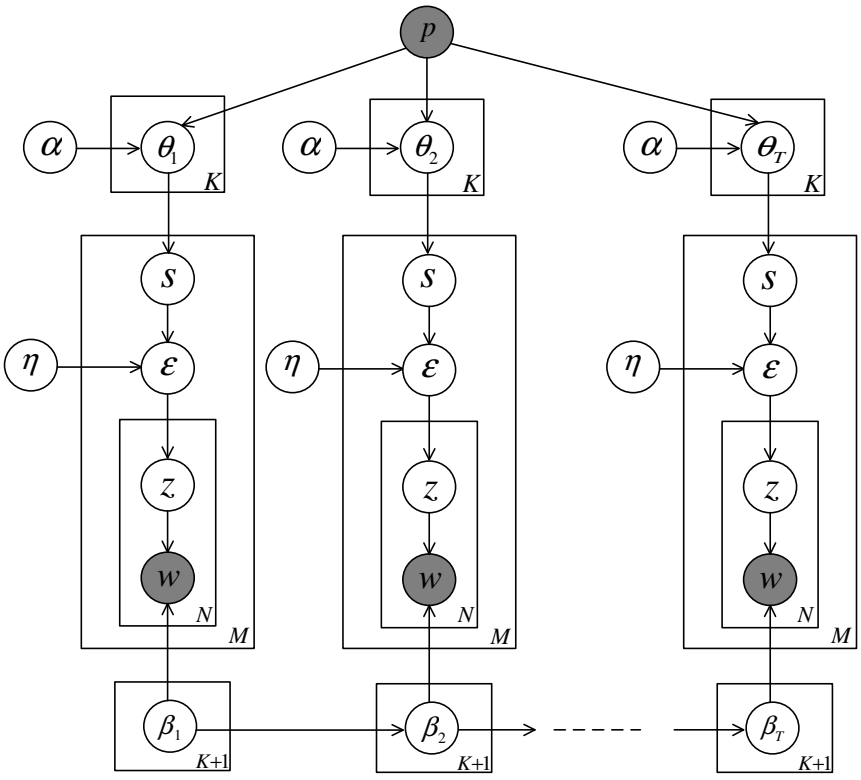


Figure 18: Dynamic Twitter Topic Model.

5.2.1 Approximate Inference with Kalman Filtering

In this section, we are going to estimate the distribution of the latent variables in the model via posterior inference. The coupling of latent variables in the multinomial models makes the posterior inference intractable to compute in topic models. Researchers mainly use two types of approaches to indirectly tackle this problem by approximation: 1) variational inference [18, 82, 16], and 2) Monte Carlo Markov chain (MCMC) sampling [29, 83]. In our case, due to the nonconjugacy of Gaussian and multinomial models, we employ variational inference rather than the sampling approach. Variational inference method has been applied in modeling temporal data [82, 16, 60]. In variational methods, the true posterior is approximated by finding a tractable family of distributions over the latent variables, which is closest to the true

posterior in Kullback-Liebler (KL) divergence [41, 16]. These distributions are called variational distributions and indexed by a set of free variational parameters.

The latent variables in our models include the event-topic proportions $\boldsymbol{\theta}$, per-tweet topic assignment s_d, per-tweet topic proportions $\boldsymbol{\epsilon}_d$, per-word topic assignment z_{dw}, and the $K+1$ sequences of topics $\boldsymbol{\beta}_t$. By breaking the coupling between latent variables and introducing free variational parameters below, we adopt the following variational distribution:

$$
\begin{aligned}
&q(\boldsymbol{\beta}_{1:T}, \boldsymbol{\theta}, \boldsymbol{s}, \boldsymbol{\epsilon}, \boldsymbol{z} | \hat{\boldsymbol{\beta}}_{1:T}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\phi}) = \\
&\prod_{k}^{K+1} q(\beta_1^k, \cdots, \beta_T^k | \hat{\beta}_1^k, \cdots, \hat{\beta}_T^k) \times \\
&\prod_{t}^{T} \left[q(\theta_t | \boldsymbol{\gamma}_t) \prod_{d}^{D_t} \left(q(s_d | \tau_d) q(\boldsymbol{\epsilon}_d | \boldsymbol{\rho}_d) \prod_{w}^{N_d} p(z_{dw} | \phi_{dw}) \right) \right],
\end{aligned}
\tag{33}
$$

where $\hat{\boldsymbol{\beta}}_{1:T}$ are the "variational observations" of topic $\boldsymbol{\beta}_{1:T}$; to incorporate the temporal dynamics of the topics, variational Kalman filter is applied to approximately model the sequential structure [16]. In variational Kalman filter, the variational $\hat{\boldsymbol{\beta}}_{1:T}$ are viewed as "observations" while the true parameters are viewed as latent states, and the chained structure of the true parameters are kept. The probability of variational observations given the true parameters is also a Gaussian distribution:

$$
\hat{\beta}_t^k | \beta_t^k \sim \mathcal{N}(\beta_t^k, \hat{\sigma}_t^2 I), \tag{34}
$$

where $\hat{\sigma}_t$ is another variational parameter. Besides that, Dirichlet hyperparameters $\boldsymbol{\gamma}$ are the variational parameters for the event-topic proportions, multinomials $\boldsymbol{\tau}$ for

the per-tweet facet topic assignments, Dirichlet hyperparameters $\boldsymbol{\rho}$ for the per-tweet topic proportions, and multinomials $\boldsymbol{\phi}$ for word topic assignments.

Now we turn to the forward-backward algorithm, which is used to compute the variational parameters so as to obtain the lower bound in the variational inference eventually. According to the standard Kalman filter calculations [42], the variational forward distribution is a Gaussian $\beta_t^k|\hat{\boldsymbol{\beta}}_{1:t}^k \sim \mathcal{N}(m_t^k, V_t^k)$ and the forward mean and variance of the variational posterior are characterized by:

$$m_t^k = \mathbb{E}(\beta_t^k|\hat{\boldsymbol{\beta}}_{1:t}^k) = h_t^k m_{t-1}^k + \left(1 - h_t^k\right)\hat{\beta}_t^k, \tag{35}$$

$$V_t^k = \mathbb{E}((\beta_t^k - m_t^k)^2|\hat{\boldsymbol{\beta}}_{1:t}^k) = h_t^k(V_{t-1}^k + \sigma^2), \tag{36}$$

$$\text{and } h_t^k = \left(\frac{\hat{\sigma}_t^2}{V_{t-1}^k + \sigma^2 + \hat{\sigma}_t^2}\right).$$

Constants m_0 and V_0 are set as the initial status in the forward. Similarly, the variational backward distribution β_t given variational $\hat{\boldsymbol{\beta}}_{1:T}$ is $\beta_t^k|\hat{\boldsymbol{\beta}}_{1:T}^k \sim \mathcal{N}(\tilde{m}_t^k, \tilde{V}_t^k)$ with the mean and variance characterized by

$$\tilde{m}_{t-1}^k = \mathbb{E}(\beta_{t-1}^k|\hat{\boldsymbol{\beta}}_{1:T}^k) = \tilde{h}_t^k m_{t-1}^k + \left(1 - \tilde{h}_t^k\right)\tilde{m}_t^k, \tag{37}$$

$$\tilde{V}_{t-1}^k = \mathbb{E}((\beta_{t-1}^k - \tilde{m}_{t-1}^k)^2|\hat{\boldsymbol{\beta}}_{1:T}^k) = V_{t-1}^k + \left(\frac{V_{t-1}^k}{V_{t-1}^k + \sigma^2}\right)^2\left(\tilde{V}_t^k - V_{t-1}^k + \sigma^2\right), \tag{38}$$

where $\tilde{h}_t^k = \frac{\sigma^2}{V_{t-1}^k + \sigma^2}$ and the initial setting is $\tilde{m}_T^k = m_T^k$ and $\tilde{V}_T^k = V_T^k$.

Next, we compute the values of the variational parameters by bounding the log likelihood of documents. With the variational distribution in Equation 33 and using

Jensen's inequality [41], the log likelihood is bounded as:

$$\begin{aligned}
&\log p(\boldsymbol{w}|p,\boldsymbol{\alpha},\boldsymbol{\eta}) \geq \mathbb{E}_q(\log p(\boldsymbol{\beta}_{1:T},\boldsymbol{\theta},\boldsymbol{s},\boldsymbol{\epsilon},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\eta})) \\
&+ \mathbb{E}_q(\log q) \\
&= \mathbb{E}_q(\log p(\boldsymbol{\beta}_{1:T})) + \mathbb{E}_q(\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})) + \mathbb{E}_q(\log(\boldsymbol{s}|\boldsymbol{\theta})) \\
&+ \mathbb{E}_q(\log(\boldsymbol{\epsilon}|\boldsymbol{s},\boldsymbol{\eta})) + \mathbb{E}_q(\log p(\boldsymbol{z}|\boldsymbol{\epsilon})) + \mathbb{E}_q(p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta}_{1:T})) \\
&+ \mathbb{H}(q),
\end{aligned} \tag{39}$$

where $\mathbb{H}(q)$ is the entropy. Maximizing the lower bound is equivalent to minimizing the KL divergence between the variational posterior and the true posterior probability [18]. So we first introduce the derivation of these seven expectations and then discuss how to maximize the lower bound with respect to the variational parameters.

The first term $\mathbb{E}_q(\log p(\boldsymbol{\beta}_{1:T}))$ is the same as that in DTM, therefore we just write down the final result here after following similar derivation steps:

$$\begin{aligned}
&\mathbb{E}_q(\log p(\boldsymbol{\beta}_{1:T})) = -\frac{|\mathcal{V}|(K+1)T}{2}(\log\sigma^2 + \log 2\pi) \\
&- \frac{1}{2\sigma^2}\sum_{t}^{T}\sum_{k}^{K+1}||\tilde{m}_t^k - \tilde{m}_{t-1}^k||^2 - \frac{1}{\sigma^2}\sum_{t}^{T}\sum_{k}^{K+1}\mathrm{Tr}(\tilde{V}_t^k) \\
&+ \frac{1}{2\sigma^2}(\mathrm{Tr}(\tilde{V}_0^k) - \mathrm{Tr}(\tilde{V}_T^k)).
\end{aligned}$$

The second term $\mathbb{E}_q(\log p(\boldsymbol{\theta}|\boldsymbol{\alpha}))$ can be expanded as below:

$$\begin{aligned}\mathbb{E}_q(\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})) &= \sum_{t}^{T} \mathbb{E}_q \log p(\theta_t|\boldsymbol{\alpha}) \\ &= \sum_{t}^{T} \left[\log\Gamma(\sum_{j}^{K} \alpha_j) - \sum_{k}^{K} \log\Gamma(\alpha_k) + \right. \\ &\left. \sum_{k}^{K} (\alpha_k - 1) \left(\Psi(\gamma_k) - \Psi(\sum_{j}^{K} \gamma_j) \right) \right].\end{aligned}$$

Interested readers may refer to Section A.1 in [18] to find similar derivation steps. For $\mathbb{E}_q(\log(\boldsymbol{s}|\boldsymbol{\theta}))$, we have:

$$\mathbb{E}_q(\log(\boldsymbol{s}|\boldsymbol{\theta})) = \sum_{t}^{T} \sum_{d}^{D_t} \sum_{k}^{K} \tau_k (\Psi(\gamma_k) - \Psi(\sum_{j}^{K} \gamma_j)) \rho_{d,k}.$$

and similarly we have $\mathbb{E}_q(\log(\boldsymbol{\epsilon}|\boldsymbol{s}, \boldsymbol{\eta}))$ expanded as:

$$\begin{aligned}&\mathbb{E}_q(\log(\boldsymbol{\epsilon}|\boldsymbol{s}, \boldsymbol{\eta})) = \\ &= \sum_{t}^{T} \sum_{d}^{D_t} \left[\log\Gamma(\sum_{j\in\Lambda} \eta_{d,j}) - \sum_{i\in\Lambda} \log\Gamma(\eta_{d,i}) + \sum_{i\in\Lambda} (\eta_{d,i} - 1) \left(\Psi(\rho_{d,i}) - \Psi(\sum_{j\in\Lambda} \rho_{d,j}) \right) \right],\end{aligned}$$

where $\Lambda = \{s, c\}$ comprises the selected facet topic s and the shared topic c. For the fifth term, we have:

$$\mathbb{E}_q(\log p(\boldsymbol{z}|\boldsymbol{\epsilon})) = \sum_{t}^{T} \sum_{d}^{D_t} \sum_{w}^{N_d} \sum_{i\in\Lambda} \phi_{d,w} \left(\Psi(\rho_{d,i}) - \Psi(\sum_{j\in\Lambda} \rho_{d,j}) \right).$$

The term $\mathbb{E}_q(p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta}_{1:T}))$ can be expanded as

$$\mathbb{E}_q(p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta}_{1:T})) = \sum_{t}^{T} \sum_{d}^{D_t} \sum_{w}^{N_d} \left[\sum_{i\in\Lambda} \phi_{d,w} \tilde{m}_t^{i,w} - \sum_{i\in\Lambda} \mathbb{E}_q(\log \sum_{v}^{\mathcal{V}} \exp(\beta_t^{i,v})) \right] \tag{40}$$

Due to the non-conjugate mapping function, we have to further compute the third term in Equation 40. We apply Taylor expansion on it with another variational

parameter ς_t, so it is upper bounded as:

$$\mathbb{E}_q(\log \sum_{v}^{\mathcal{V}} \exp(\beta_t^{i,v})) \leq \phi_{d,w} \left(\varsigma_t^{-1} \sum_{v}^{\mathcal{V}} \mathbb{E}_q \exp(\beta_t^{i,v}) - 1 + \log(\varsigma_t) \right), \tag{41}$$

where $\exp(\beta_t^{i,v})$ is a log normal distribution and $\mathbb{E}_q(\exp(\beta_t^{i,v}))$ is its mean. With Equation 41, Equation 40 is rewritten as

$$\begin{aligned}
&\mathbb{E}_q p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta}_{1:T}) \geq \sum_{t}^{T} \sum_{d}^{D_t} \sum_{w}^{N_d} \left[\sum_{i \in \Lambda} \phi_{d,w} \tilde{m}_t^{i,w} - \right. \\
&\left. \sum_{i \in \Lambda} \phi_{d,w} \left(\varsigma_t^{-1} \sum_{v}^{\mathcal{V}} \exp(\tilde{m}_t^{i,v} + \tilde{V}_t^{i,v}/2) - 1 + \log(\varsigma_t) \right) \right].
\end{aligned}$$

Finally, we expand the entropy as:

$$\begin{aligned}
&\mathbb{H}(q) = -\mathbb{E}_q q(\boldsymbol{\beta}_{1:T} \boldsymbol{\theta}, \boldsymbol{s}, \boldsymbol{\epsilon}, \boldsymbol{z} | \hat{\boldsymbol{\beta}}_{1:T}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \boldsymbol{\eta}, \boldsymbol{\phi}) \\
&= -\sum_{t}^{T} \sum_{k}^{K+1} \left(\frac{|\mathcal{V}|}{2} \log 2\pi \right) - \frac{1}{2} \sum_{t}^{T} \sum_{k}^{K+1} \sum_{v}^{\mathcal{V}} \log(\hat{V}_v^k) - \\
&\sum_{t}^{T} \left[\log\Gamma(\sum_{j}^{K} \gamma_j) - \sum_{k}^{K} \log\Gamma(\gamma_k) + \right. \\
&\left. \sum_{k}^{K} (\gamma_k - 1) \left(\Psi(\gamma_k) - \Psi(\sum_{j}^{K} \gamma_j) \right) \right] - \\
&\sum_{t}^{T} \sum_{d}^{D_t} \sum_{k}^{K} \tau_d \log \tau_d - \\
&\sum_{t}^{T} \sum_{d}^{D_t} \left[\log\Gamma(\sum_{j \in \Lambda} \rho_j) - \sum_{i \in \Lambda} \log\Gamma(\rho_i) + \right. \\
&\left. \sum_{i \in \Lambda} (\rho_i - 1) \left(\Psi(\rho_i) - \Psi(\sum_{j \in \Lambda} \rho_j) \right) \right] - \\
&\sum_{t}^{T} \sum_{d}^{D_t} \sum_{w}^{N_d} \sum_{i \in \Lambda} \phi_{d,w} \log \phi_{d,w}.
\end{aligned}$$

We need to maximize the lower bound in Equation 39 so as to approximate the

true posterior. We take derivatives with respective to ς_t, set to zero, and solve for ς_t:

$$\varsigma_t = \frac{1}{\sum_d^{D_t} N_d} \sum_d^{D_t} \sum_w^{N_d} (\sum_{i \in \Lambda} \phi_{d,w} \sum_v^{\mathcal{V}} \exp(\tilde{m}_t^{i,v} + \tilde{V}_t^{i,v}/2)). \tag{42}$$

Then we follow similar optimization procedures as introduced in Section A.3 of [18] to solve other variational parameters. We have them listed below:

$$\tau_t^k \propto \exp(\Psi(\gamma_t^{p,k}) - \Psi(\sum_j^K \gamma_t^{p,j})) \sum_d^{D_t} \rho_{d,k} \tag{43}$$

$$\gamma_t^{p,k} = \alpha_k + \sum_d^{D_t} \tau_t^k \tag{44}$$

$$\phi_{d,w}^i \propto \exp(\Psi(\rho_{d,i}) - \Psi(\sum_{j \in \Lambda} \rho_{d,j})) \times \exp(\tilde{m}_t^{i,w} - \mathbb{E}_q \log \sum_v^{\mathcal{V}} \exp(\beta_t^{i,v})), \quad i \in \Lambda \tag{45}$$

$$\rho_{d,i} = \eta_i + \sum_w^{N_d} \phi_{d,w}, \quad i \in \Lambda \tag{46}$$

In DTTM, the facet topic s in Λ is selected for each tweet. So we have the restriction on i in Equation 45 and 46. This selection can be understood as "hard assignment". In implementation, rather than solving the hard assignment directly, we replace it with a soft assignment by introducing changes in Equation 45 below:

$$\phi_{d,w}^i \propto \exp(\Psi(\tau_t^i \rho_{d,i}) - \Psi(\sum_j^K \tau_t^j \rho_{d,j})) \times \exp(\tilde{m}_t^{i,w} - \mathbb{E}_q \log \sum_v^{\mathcal{V}} \exp(\beta_t^{i,v})), \; i \in [1, K]$$

With the change, the topic for the tweet is not exclusive to the selected one anymore, but can be the K facet topics weighted by their probabilities. Similarly, the constraint in Equation 46 is removed.

To maximize the lower bound with respect to $\hat{\beta}$, the partial derivative of the lower

bound with respect to $\hat{\beta}_t^{k,w}$ is obtained:

$$-\frac{1}{\sigma^2}\sum_t^T(\tilde{m}_t^{k,w}-\tilde{m}_{t-1}^{k,w})\left(\frac{\partial\tilde{m}_t^{k,w}}{\partial\hat{\beta}_t^{k,w}}-\frac{\partial\tilde{m}_{t-1}^{k,w}}{\partial\hat{\beta}_t^{k,w}}\right)$$
$$\sum_t^T\left(N_t^w\phi_w^k-\sum_v^{\mathcal{V}}N_t^v\phi_v^k\varsigma_t^{-1}\times\exp(\tilde{m}_t^{i,v}+\tilde{V}_t^{i,v}/2)\right)\frac{\partial\tilde{m}_t^{k,w}}{\partial\hat{\beta}_t^{k,w}}$$

The conjugate gradient algorithm is employed to find a local optimum of $\hat{\beta}$ [16, 36], where the gradients $\frac{\partial\tilde{m}_t^{k,w}}{\partial\hat{\beta}_t^{k,w}}$ are needed. The gradients can be computed with forward and backward mean in Equation 35 and 37:

$$\frac{\partial m_t^{k,w}}{\partial\hat{\beta}_s^{k,w}}=h_t\frac{\partial m_{t-1}^{k,w}}{\partial\hat{\beta}_s^{k,w}}+(1-h_t)I_t(s)$$

$$\frac{\partial\tilde{m}_{t-1}^{k,w}}{\partial\hat{\beta}_s^{k,w}}=\hat{h}_t\frac{\partial m_{t-1}^{k,w}}{\partial\hat{\beta}_s^{k,w}}+(1-\hat{h}_t)\frac{\partial m_t^{k,w}}{\partial\hat{\beta}_s^{k,w}}.$$

$I_t(s)$ is the indicator function, and the initial conditions for forward and backward recurrence are ${}^{\partial m_0^{k,w}}/_{\partial\hat{\beta}_s^{k,w}}=0$ and ${}^{\partial\tilde{m}_T^{k,w}}/_{\partial\hat{\beta}_s^{k,w}}={}^{\partial m_T^{k,w}}/_{\partial\hat{\beta}_s^{k,w}}$ respectively.

The overall variational inference procedure can be performed using the EM algorithm. With the derivation introduced above, we sketch the inference flow in Algorithm 2.

5.3 Case Studies

To demonstrate the capability of our proposed method, we conducted two case studies on Twitter data. In the first study, we examine the performance of our model on tweets related to one event and compare its performance with other topic models. We further evaluated our model on an unfiltered tweet dataset in the second study. The two studies show that our model can support both focused investigation and explorative analysis.

Algorithm 2 Variational inference algorithm.

Initialization parameters
repeat
  E step:
  for $t \leftarrow 1, T$ **do**
    Update $\boldsymbol{\tau}_t$ with Equation 43
    Update $\boldsymbol{\gamma}_t$ with Equation 44
    for $d \leftarrow 1, D_t$ **do**
      Update $\boldsymbol{\phi}_d$ with Equation 45
      Update $\boldsymbol{\rho}_d$ with Equation 46
    end for
    Update ς_t with Equation 42
  end for
  M step:
  Update $\hat{\boldsymbol{\beta}}$ using conjugate gradient descent
until converged

5.3.1 Occupy Movement

In this study we analyzed events related to the Occupy Movement, with ground truth available in Wikipedia. The Occupy Movement consisted of a series of demonstrations and was known to use social media to organize and attract protestors. It is interesting to summarize the major events throughout the long-running, widely participated movement. We illustrate how the topic results can explain the events and help make sense of the movement in a temporal manner. We visualize the results to help end users better understand the developing event trends. We also compare our model with DTM and LDA.

We first collected tweets through Twitter's public streaming API[11], which yields one percent samples. To identify a collection of tweets related to the Occupy Movement, we queried for tweets containing occupy-related hashtags such as "#occupy". The query captures a wide range of protests including "#OccupyWallSt", "#Occupy-

[11] https://dev.twitter.com/docs/streaming-apis/streams/pu-blic

London", "#OccupyTogether", etc. The resulting collection contains around 200,000 tweets posted between August 19, 2011 and July 2, 2012. We removed stop words and excluded tweets in non-English characters. Tweets of length less than five words are also removed. To remove noise words, terms appearing less than five times in total are considered noise and not included in the vocabulary. After the cleaning steps, 74,362 tweets are retained and the size of the vocabulary is 13,813 terms.

We use the methods described in Section 3 to learn topics with $K = 3$ and one shared mainstream topic. The parameter setting is: $\sigma^2 = 0.1$, $\hat{\sigma}^2 = 1$, $\alpha = 0.1$, and $\eta = 1$. Due to space constraints, we are able to list only a few topic samples in Figure 16 and Figure 17. For purpose of comparison, we also ran DTM program with the same setting on the dataset and list the topics together.

Based on our analysis of the topic results, the shared topic primarily captures the main story of the movement as well as the most significant events of the movement. More specifically, the shared topic always cover major events related to protests (sometimes international) and marches participated by the masses. For example, the shared topic captures multiple conflicts between protesters and police that occurred in November 2011 (Figure 16). Such conflicts include the NYPD (New York Police Department) raiding the protesters' camp in an effort to evict the protestors, and the LAPD (Los Angeles Police Department) using of pepper spray on protestors. As the movement progressed, in January 2012, the shared topic captured the breakout of a major international event "OccupyNigeria". As shown by multiple sources[12], the

[12]http://en.wikipedia.org/wiki/Timeline_of_Occupy_Wall_Street;
http://www.motherjones.com/mojo/2012/09/occupy-wall-street-anniversary-timeline

Table 16: Topics discovered by DTTM and DTM for November 2011.

DTTM (11/2011)			
Topic 1	**Topic 2**	**Topic 3**	**Shared**
envoyez	police	reading	city
soutenons	occupy	psychic	protesters
police	street	time	police
@boldprogressives	movement	police	park
time	live	rondo	eviction
protesters	wall	protesters	almudena
park	protesters	occupy	lapd
movement	video	voxer	cops
@robinsage	time	change	pepper
@providesecurity	protest	calling	occupy
@occupy	day	nyc	arrested
mayor	news	movement	raid
occupy	media	arnold	camp
eviction	tonight	country	mayor
stand	world	call	tents
pour	march	street	oakland
@kanyewest	park	day	hall
defend	nyc	protest	dbkl
protest	protests	video	nlc
urgent	check	wall	live
DTM (11/2011)			
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
city	lapd	movement	police
park	stand	time	protesters
almudena	eviction	occupy	encroaching
subsidy	unesco	cops	street
hall	democracy	protesters	occupy
catedral	nlc	protest	wall
caixa	real	tonight	live
twitter	demand	day	pepper
video	send	scanzi	video
zuccotti	nyc	@fattoquotidiano	raid
occupy	reading	world	camp
@kevskewl	street	mayor	arrested
@ruggedybaba	ojota	photo	riot
time	message	eviction	march
evict	friday	live	arrests
movement	psychic	media	fadzil
press	wall	night	femi
live	dbkl	love	@fckh
kami	cover	envoyez	@fahmi
obama	temp	check	@robinsage

Table 17: Topics discovered by DTTM and DTM for January 2012.

DTTM (01/2012)			
Topic 1	**Topic 2**	**Topic 3**	**Shared**
envoyez	protest	occupy	nigeria
soutenons	gej	protesters	protest
protesters	protesters	police	@omojuwa
@boldprogressives	police	live	gej
occupy	occupy	voxer	almudena
police	@omojuwa	city	@ogundamisi
protest	live	arnold	police
live	movement	movement	subsidy
@robinsage	day	protest	fuel
@providesecurity	news	oakland	nigerians
nigerian	join	street	strike
movement	peaceful	day	nlc
@rosanwo	street	arrested	@rosanwo
@kanyewest	time	video	protesters
day	change	watch	jonathan
@occupy	vacancy	wall	@elrufai
street	wall	nyc	govt
brt	@naijacyberhack	time	lagos
park	crowd	scanzi	nigerian
@anonlgrisback	video	adriana	protests

DTM (01/2012)			
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
almudena	unesco	gej	police
occupy	nigerian	occupy	nigeria
subsidy	@rosanwo	movement	encroaching
fuel	@ogundamisi	protest	protesters
president	protest	live	london
city	@omojuwa	time	arrested
house	nigerians	day	live
govt	protesters	scanzi	rally
news	nlc	government	occupy
catedral	lagos	@fattoquotidiano	video
@elrufai	strike	dey	fadzil
caixa	jonathan	envoyez	femi
removal	ojota	country	@fahmi
street	nyc	photo	@fckh
@kevskewl	dbkl	soutenons	march
hall	lapd	action	protest
park	temp	cops	dis
protest	lockouts	join	street
twitter	@eggheader	money	day
@omojuwa	forecast	bank	bakare

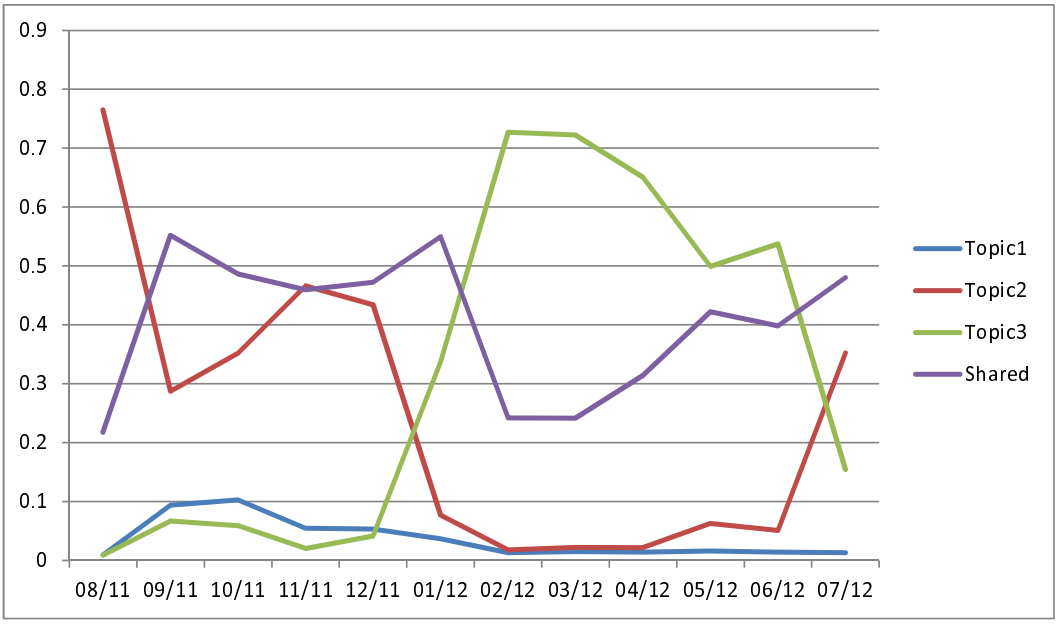


Figure 19: Topic proportions of DTTM at each time slice.

scale of discussions of the OccupyNigeria event on Twitter is much more signicant than any domestic Occupy events. As shown in Figure 17, terms including "nigeria", "fuel", "subsidy" in the shared topic highlight the "OccupyNigeria" event, and its cause, the ending of the government oil subsidy. In comparison, although the topics in DTM cover several words related to the event, coverage of the major "OccupyNigeria" event is divided into several topics, with term "nigeria" in Topic 4, terms "nigerian" and"lagos" appearing in Topic 2, and terms "fuel" and "subsidy" in Topic 1.

To provide evidence that DTTM captures more distinctive events that occur at different times, we present several quantitative measures. First, we visualize the event-topic proportions over time in Figure 19. The hypothesis is that if the topic proportions are more concentrated on a few of the topics, as opposed to relatively uniform trends, then those topics reflect themes that are the focus of more discussions on Twitter. For comparison purposes, we also compute the topic proportions for DTM

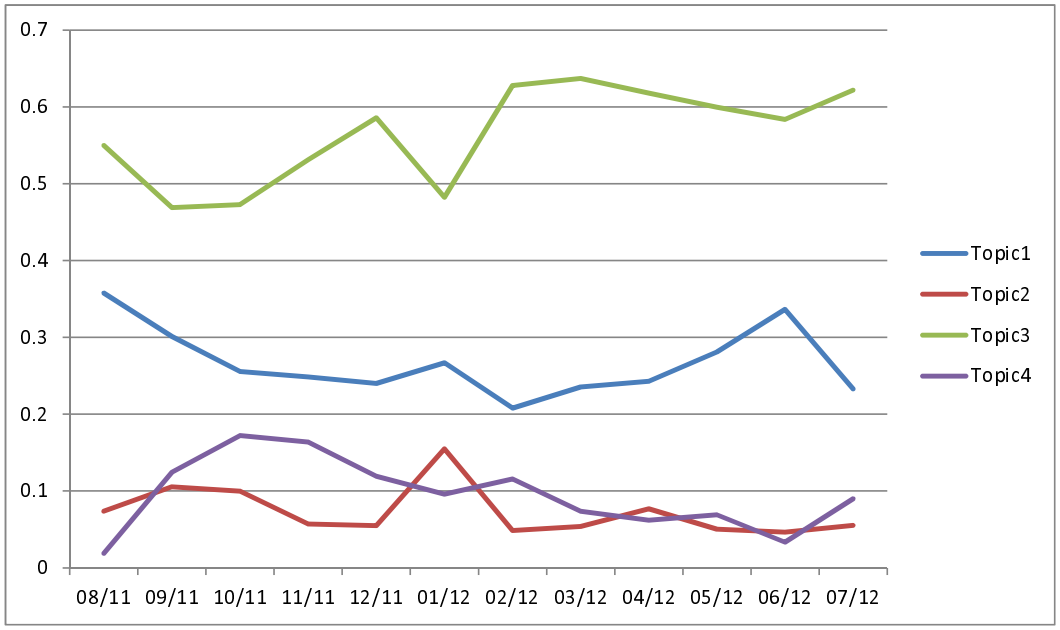


Figure 20: Topic proportions of DTM at each time slice.

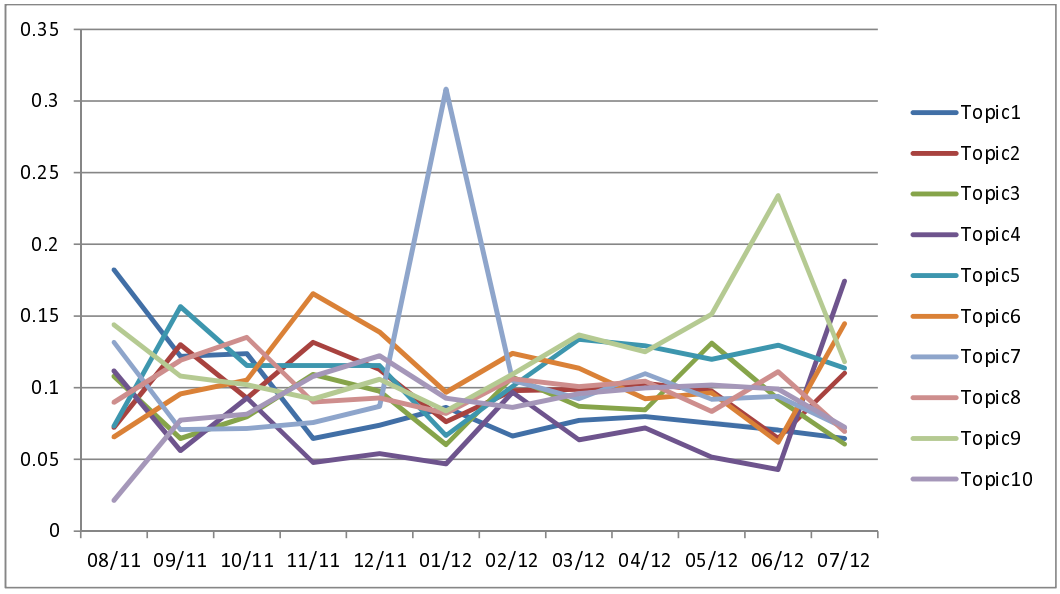


Figure 21: Topic proportions of LDA at each time slice.

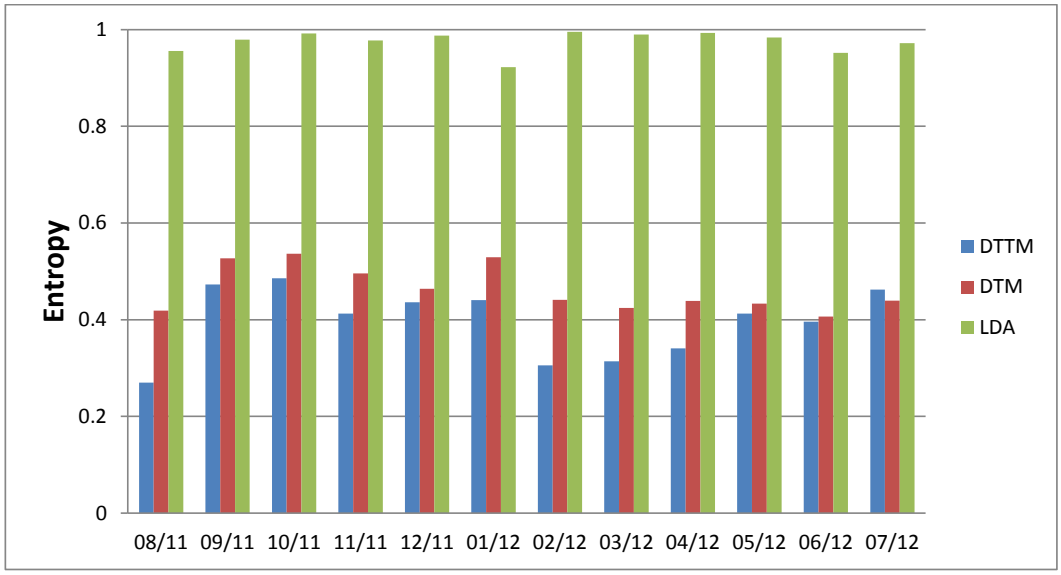


Figure 22: Entropy comparison for DTTM, DTM, and LDA.

and plot it in Figure 20. DTM does not have a variable for the topic proportion; we compute it by $\frac{\sum_{d_t} |d_t^k|}{\sum_{d_t} \sum_k |d_t^k|}$, the frequency of words assigned to topic k normalized by the total number of words. We ran LDA[13] on the data and extracted 10 topics. LDA natively cannot discover the temporal dynamics because time is not modeled. We inferred the per document-topic proportions, and aggregated and normalized the proportions for each month such that we obtain pseudo topic proportions for every month. Figure 21 shows the topic proportions for LDA.

To view the difference in topic proportions among these three models, we compute the entropy of topic proportions at each time slice. The results are shown in Figure 22. A lower entropy implies less uncertainty in the topic proportions, and thus indicates that a few topics are more informative. As shown in Figure 22, DTTM generally exhibits the lowest entropy compared to the other two models, thus it helps end users to better identify and analyze the event by observing the importance of the topics

[13]http://www.cs.princeton.edu/~blei/lda-c/index.html

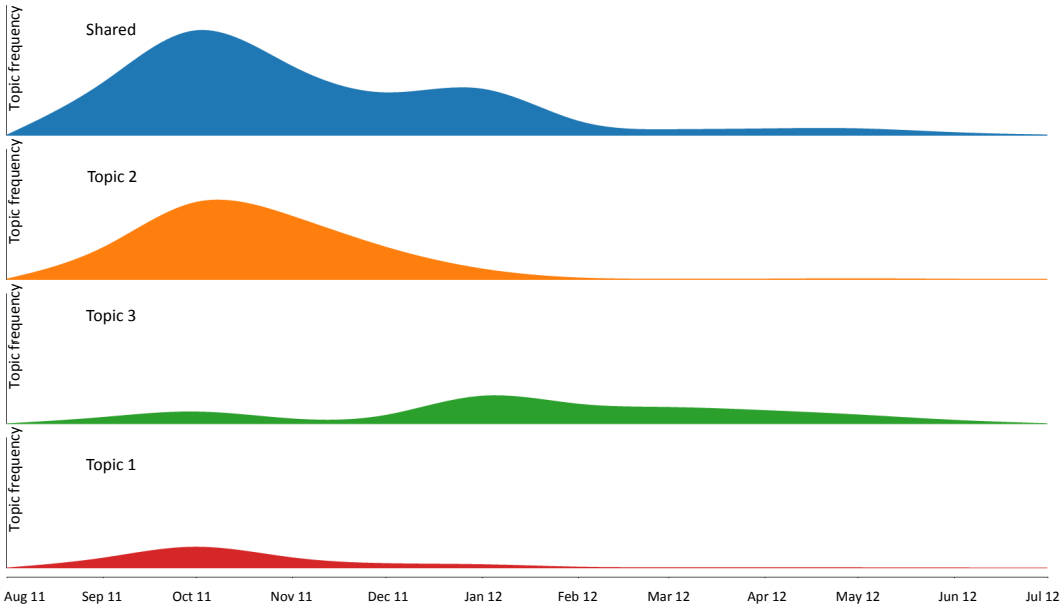


Figure 23: Variation in topic frequency with time for DTTM. The y-axis represents the topic frequency.

from their probabilities. Please note although the number of topics in LDA is more than in DTTM and DTM, we can see from Figure 21 that the topic proportions of different topics do not differ much, so it becomes harder for users to identify which topic conveys more information at a given time.

In addition to comparing the topic proportions produced by different models, we now evaluate the topic quality. We did not employ perplexity for this evaluation, since it may not be adequate [81, 22]. Instead we interpreted the topics and match to the timeline of the Occupy Movement in Wikipedia as the ground truth. We observed that the shared topic covers more terms related to the major events in the movement while facet topics can capture smaller yet still significant events different from the main storyline. For instance, Topic 2 and the shared topic dominate the topic proportions in November 2011. Topic 2 includes relevant terms “police”, “occupy”, “movement”, “wall”, “media”, etc.. The shared topic has been discussed above.

To facilitate the understanding of the topics with respect to their temporal evolution, we developed visualizations (Figure 23) to represent the topic trends, showing the ebb and flow for each topic. The area under the curve represents the frequency of the topics. We further extract the term difference between two consecutive time frames, with the new terms in a topic usually signaling an emergence of new sub-events. More specifically, we compare the topics at t with the corresponding topics at $t-1$. With the topic differences, we can notice emerging of sub-events and how the events develop. The visualizations for the topic differences are shown in Figure 24, 25, and 26, in which a group of unique terms are visualized together with the trends of the topics. In September 2011, two distinct events are captured by DTTM results. The shared topic (A in Figure 24) covers the big social media event "OccupySesameStreet", with the slogan "1% of the monsters consume 99% of the cookies", while Topic 2 captures the "virtual march movement on Wednesday" shown in A in Figure 25. In October–November 2011, the topic differences in the shared topic highlight the significant event of NYPD evicting the protesters from Zucotti Park, as well as other significant events in California, including police pepper spraying protesters and the Oakland port being occupied (B in Figure 24).

Visually representing the topic term differences in time together with the temporal trend of the topic helps users to make sense of the events. As we can see, although the Occupy Movement started in New York City, the movement quickly spread to the west coast of the US. After December 2011, although the number of tweets started to reduce in general, DTTM is still able to capture significant events. The shared topic covered the big international event "OccupyNigeria" (C in Figure 24), while

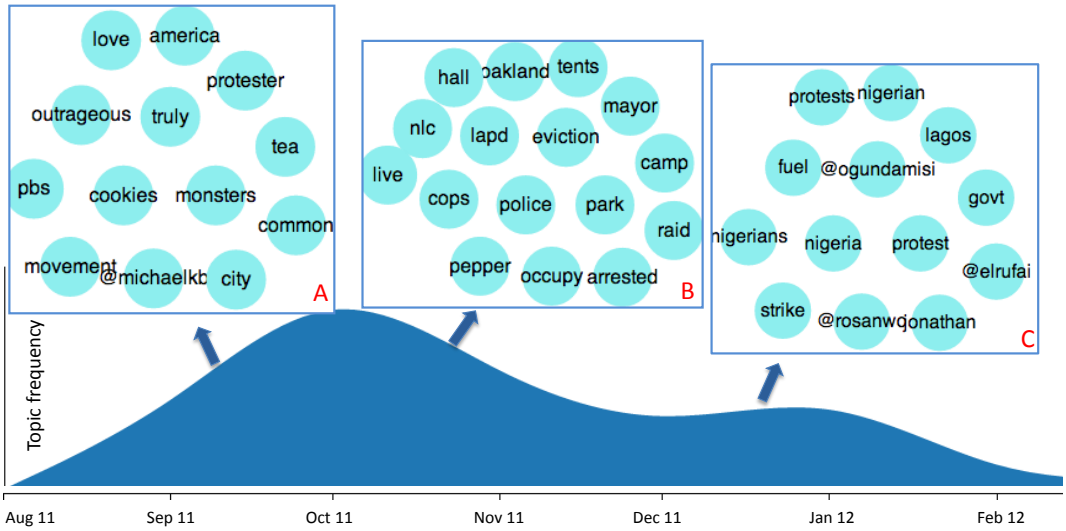


Figure 24: Topic differences for the shared topic in September–October 2011, October–November 2011, and December 2011–January 2012.

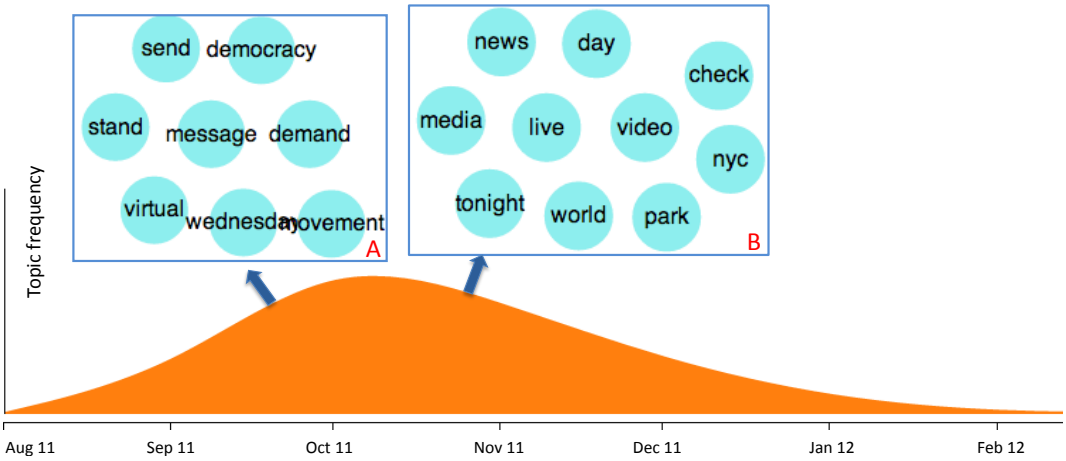


Figure 25: Topic differences for Topic 2 in September–October and October–November 2011.

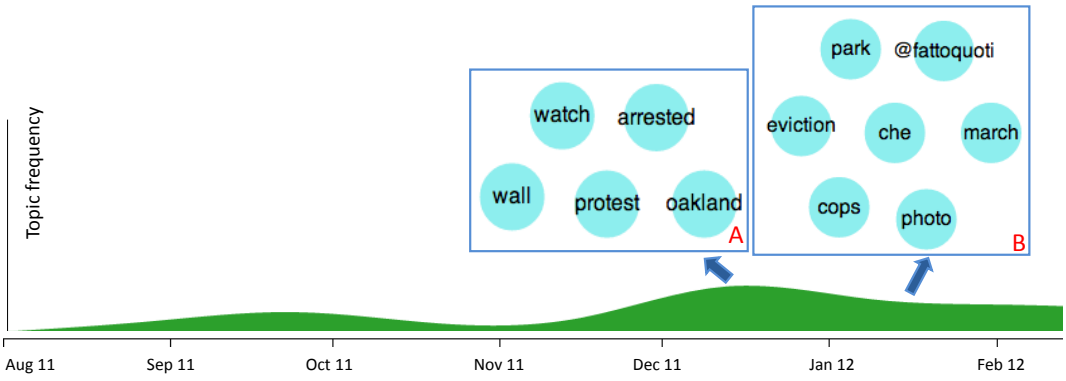


Figure 26: Topic differences for Topic 3 in December 2011–January 2012 and January–February 2012.

Topic 3 captures another violent event in Oakland that resulted in 400 arrested (A in Figure 26). As described, DTTM together with the visualization helps users observe and understand significant sub-events or facets throughout the Occupy Movement. We tried to perform a similar analysis using DTM and LDA, but found the results from these two models were not as intuitive. Therefore, we believe the proposed modeling approach of DTTM provides a better analysis and summarization of the movement.

5.3.2 Epidemic Spread

In the previous case study, we evaluated our model using tweets related to the Occupy Movement. In practice, it is usually difficult to know exactly what event occurred, and thus filtering for tweets discussing one event is infeasible. To demonstrate that our DTTM model can be applied to general tweet collections for analyzing major events, we applied it to unfiltered raw tweets that were collected during a three week period and report the results.

We used the benchmark dataset release by the IEEE VAST Challenge 2011 committee[14]. The dataset contains 1,023,077 tweets posted from April 30, 2011 to May 20, 2011. The dataset was generated by a group of experts by combining real tweets with manufactured tweets regarding a threat scenario. The dataset serves as a benchmark to evaluate DTTM. The size of the vocabulary is 45,185 after removing stopwords. Compared to the Occupy Movement dataset, this dataset is much noisier and closer to situations in practical tasks.

[14]IEEE VAST Challenge 2011. http://hcil.cs.umd.edu/localphp/hcil/vast11/index.php/

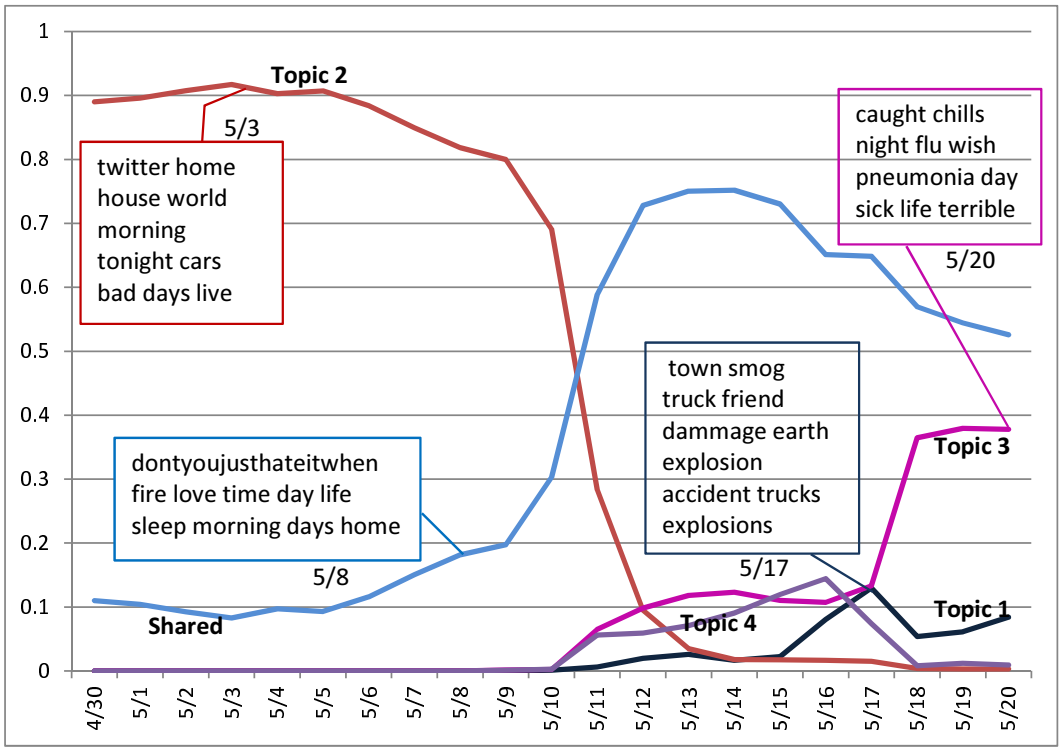


Figure 27: Topic examples and topic proportions of DTTM at each time slice on the epidemic dataset.

For analysis, we divided the tweets at daily intervals, and trained $K = 4$ facet topics and one shared mainstream topic. The topic proportions over time are shown in Figure 27. Additionally, several topics for different time slices are also displayed in the figure. Topic 2 (in red) and the mainstream topic (in blue) are dominant from the beginning until May 11. The two topics include terms "twitter", "cars", "world", "life", "sleep", "playing", etc. Topic 2 and the shared mainstream topic capture people's posts on Twitter on their daily life and personal status updates, as opposed to discussions driven by external events.

While the mainstream topic continues to be significant during the rest of the time, other topics (Topic 1, Topic 3, Topic 4) peak at different times as indicators of potential events. Topic 1 (in navy) peaks on May 17; the topic terms for that time include

"smog", "truck", "explosion", "accident". It clearly outlines a major traffic accident occurred on that day involving a truck that led to an explosion. This topic successfully captures this sub-event (facet) embedded in the tweets. The ground truth of this dataset indeed verified this facet. More interestingly, during the last three days of the 3-week period, the proportion of Topic 3 (in purple) increased significantly. By reading prominent terms of Topic 3, which include "chills", "flu", "sick", and "pneumonia", one can infer that people are reporting flu-like symptoms. In addition, the topic proportions and the temporal pattern can inform us of the first reports of symptoms and the magnitude of the disease spread. According to the ground truth, the facet described in Topic 3 is the inserted threat scenario that a sound model should be able to capture. Through deeper investigation of tweets related to Topics 1 and 3, it is not difficult for one to uncover the causal relationship between the truck accident event captured by Topic 1 and the disease spread event captured by Topic 3. It turns out the truck accident caused the chemicals carried by the truck to leak into a local river so people got sick from drinking the water.

In this study, we applied DTTM to analyze and summarize an unfiltered tweet collection. We observe that DTTM provides meaningful summarization of the data that correctly reflects the events according to the ground truth. It is worth noting that this case study demonstrates DTTM is also capable of analyzing tweets with diverse themes even though DTTM assumes the tweet data centers on one event. The topics plus their temporal dynamics learned from DTTM can facilitate event summarization on general tweet data.

5.4 Summary

We developed a new temporal topic model for analyzing and summarizing event development in Twitter. We designed the model specifically for the short length tweets. In our model, a event is described as a mixture of topics. To capture the mainstream theme of the event, we assume there is a topic shared by all tweets. The temporal dynamics of the topics are expressed in a state space model with Gaussian noise in a chained structure. A variational inference based approach is used to compute the posterior. With visualization of the results, it offers a new and intuitive view of the event development over time, and therefore eases user understanding of unstructured and noisy Twitter data.

Our focus for the future is a more comprehensive quantitative evaluation of the model and comparison with other topic models. We also would like to test if the model is extensible to perform event detection. Additionally, determining how to automatically decide the number of topics per time slice, i.e., how to model the disappearance and emergence of topics, is another direction for future work.

CHAPTER 6: CONCLUSION

This thesis has developed topic models for tagged text, with a special emphasis on processing social media data such as tweets with hashtags. Tags, as one kind of important meta-data, are primarily used to organize and cluster relevant documents. However, as large volumes of user-generated tags are generated especially in the social media community, understanding the meanings of tags and summarizing tagged text is becoming a challenging and increasingly important problem. This thesis explores using topic modeling techniques to address this problem. We summarize the contributions of the thesis in this chapter and present directions for future work.

6.1 Summary and Contributions

In Chapter 1, we first introduced the fundamental concepts of topic modeling and our research motivation for tagged text. We then briefly introduced Latent Dirichlet Allocation (LDA) along with essential mathematical background.

In Chapter 2, we surveyed recent research work on topic modeling, especially important extensions that build on LDA. Additionally, we also reviewed work applying topic models to mine social media data.

Starting from Chapter 3, we introduced our contributions for modeling tagged text. We first proposed TriTag-LDA, in which tags are represented as a multinomial distribution of topics. After the topic assignment for each term of the documents

is inferred by LDA, TriTag-LDA uses the same mechanism as LDA to infer the tag assignments for the topics the terms are assigned to. We improved TriTag-LDA by introducing Tag-LDA, which views a document as a mixture of observed tags. The tags are thus naturally involved in the generative process. To learn the models, we derived and used Gibbs sampling based solutions. We conducted experiments quantitatively comparing our models with the author-topic model using perplexity measurements. The experiments show that Tag-LDA is superior to the other two models. We applied Tag-LDA to a practical application, understanding the hashtags in tweets and the relationships between the hashtags, to demonstrate the capability of Tag-LDA.

Chapter 4 presents two extensions that build on Tag-LDA. The first extension is Tag-Latent Dirichlet Processes (Tag-LDP), which is inspired by prior work on Hierarchical Dirichlet Processes. Tag-LDP utilizes a nonparametric approach, the Dirichlet process, to answer the question of how many topics should be learned given a text corpus. The second extension is ConceptTag-LDA. We introduced the use of the Dirichlet Tree prior, which replaces the Dirichlet prior in LDA. The Dirichlet Tree prior allows users' prior knowledge to be conveyed in the form of a set of terms, which are called concepts, to be integrated in topic modeling. Simply speaking, ConceptTag-LDA constraints the co-occurrence of the concept terms in the topics; the probabilities of the concept terms are encouraged to be similar in the topics. Compared to Tag-LDA, ConceptTag-LDA provides an additional layer of flexibility for users under the modeling procedure. The extensibility of Tag-LDA is exemplified by these two extensions. We believe other extensions of LDA can be successfully

adapted, with reasonable effort, to Tag-LDA. For instance, Tag-LDA can adopt the essence of the work in [2] by specifying the topic constraint for each term.

Finally, in Chapter 5 we turned to applying topic modeling techniques in analyzing and summarizing discussions in Twitter. We emphasize the temporal dynamics of the events, since a major event usually involves twists and turns reflected by multiple sub-events throughout its development in different time periods. This temporal event development is in turn reflected by people's discussions on Twitter. We proposed the dynamic Twitter topic model (DTTM), which assumes an event can be modeled by a mainstream theme plus several facets. The mainstream theme of the event is a topic shared by all tweets. The temporal dynamics of the topics are expressed in a state space model with Gaussian noise in a chained structure. A variational inference based approach is used to compute the posterior. To demonstrate the effectiveness of DTTM in modeling the temporal dynamics of topics and its ability to facilitate event analysis, we conducted two case studies with our model using Twitter data and showed that our model performs better than the other general purpose topic models.

In this thesis we addressed the four motivating questions raised in Chapter 1 on utilizing topic modeling techniques for text with tags. The contributions of this thesis are summarized below:

1. This thesis models tagged text data. We developed a new topic model, Tag-LDA, to interpret user-generated tags using topics. Our model can make sense of the tags and also help to understand their relationships.

2. To avoid having the user set the number of topics, we extended our model

to Tag-Latent Dirichlet Processes to infer the number of topics from the data automatically.

3. We additionally extended our model to allow users' prior knowledge to be involved in ConceptTag-LDA, which helps non-expert users to model the data according to their prior knowledge.

4. We developed a temporal topic model focused on modeling the short messages in social media. Our model, DTTM, is capable of summarizing the discussions in social media and also capturing the temporal dynamics of the discussions.

6.2 Future Work

There are several directions for future work. We introduced ConceptTag-LDA, where users' prior knowledge is expressed by specifying constraints on the co-occurrence of concept terms in topics. It is a useful tool that enables users to have a certain degree of flexibility compared to LDA. However, users' prior knowledge usually is much more complex than constraining the co-occurrence of terms in topics. For example, when defining the concept terms, users can additionally provide the degree of the correlation between terms, e.g., "stock" and "money" are relatively highly correlated while "stock" and "sports" are less correlated. Another example is people's knowledge of many polysemic terms. "Apple" can appear in a fruit related concept, and also it could appear in a consumer electronics related concept. How to mathematically represent the additional knowledge and smartly incorporate the knowledge in topic modeling is an interesting direction. One potential solution is to enable interactive topic modeling by asking users to be involved in the learning procedure, such that

topics can be adjusted in realtime by users. However, the long training time as well as establishing an appropriate metric for evaluating topic quality are challenges. Furthermore, in ConceptTag-LDA, we did not explicitly constrain the linkage between concepts and the tags. So another potential direction is to link the concepts with the tags. In other words, the concept constraints only take effect in the dominant topics of the tag. For example, we define concept c ={stock, money} is linked with tag "finance", so terms "stock" and "money" are encouraged to have higher probabilities in the dominant topics of "finance"; otherwise there are no restrictions on c in the other topics.

For summarizing and analyzing discussions of events in social media, we could also incorporate other data sources, like traditional news media, e.g., newspapers. Traditional news media are much less noisy and normally provide more accurate information on event participants, locations, etc., which definitely helps better characterize and summarize the events. Additionally, a system that can predict the temporal development of the discussions based on the historical temporal dynamics obtained would be especially useful.

REFERENCES

[1] D. Andrzejewski. *Incorporating Domain Knowledge in Latent Topic Models*. PhD thesis, Department of Computer Science, University of Wisconsin-Madison, 2010.

[2] D. Andrzejewski and X. Zhu. Latent dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, SemiSupLearn '09, pages 43–48, Boulder, Colorado, 2009. Association for Computational Linguistics.

[3] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 25–32, Montreal, Quebec, Canada, 2009. ACM.

[4] D. Andrzejewski, X. Zhu, M. Craven, and B. Recht. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, IJCAI '11, pages 1171–1177, Barcelona, Catalonia, Spain, 2011. AAAI Press.

[5] T. Asou and K. Eguchi. Predicting protein-protein relationships from literature using collapsed variational latent dirichlet allocation. In *Proceedings of the 2nd International Workshop on Data and Text Mining in Bioinformatics*, DTMBIO '08, pages 77–80, Napa Valley, California, USA, 2008. ACM.

[6] N. Balakrishnan and V. B. Nevzorov. *A Primer on Statistical Distributions*. Wiley-Interscience, Hoboken, NJ, 2003.

[7] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 291–300, New York, New York, USA, 2010. ACM.

[8] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on Twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 438–441, Barcelona, Spain, 2011.

[9] S. Bird, E. Loper, and E. Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., Sebastopol, California, USA, 2009.

[10] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Secaucus, NJ, USA, 2007.

[11] D. Blackwell and J. B. MacQueen. Ferguson Distributions Via Pólya Urn Schemes. *Annals of Statistics*, 1:353–355, 1973.

[12] D. Blei and J. Lafferty. *Topic Models. Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC, Boca Raton, FL, 2009.

[13] D. Blei and J. McAuliffe. Supervised topic models. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. MIT Press, 2008.

[14] D. M. Blei. *Probabilistic Models of Text and Images*. PhD thesis, Department of Computer Science, University of California, Berkeley, 2004.

[15] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134, Toronto, Canada, 2003.

[16] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, Pittsburgh, Pennsylvania, 2006.

[17] D. M. Blei and J. D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1:17–35, 2007.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[19] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *IEEE 11th International Conference on Computer Vision*, ICCV '07, pages 1–8, Rio de Janeiro, Brazil, 2007.

[20] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, Pacific Grove, CA, 2nd edition, 2001.

[21] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *International AAAI Conference on Weblogs and Social Media*, ICWSM '10, pages 10–17, Washington, D.C., USA, May 2010. AAAI Press.

[22] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, pages 288–296, Vancouver, British Columbia, 2009.

[23] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Leveraging multi-domain prior knowledge in topic models. In *the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 2071–2077, Beijing, China, 2013.

[24] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Gonçalves, and F. Benevenuto. Analyzing the dynamic evolution of hashtags on Twitter: a language-based approach. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 58–65, Portland, Oregon, USA, 2011. Association for Computational Linguistics.

[25] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 241–249, Beijing, China, 2010. Association for Computational Linguistics.

[26] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL '12, pages 536–544, Jeju Island, Korea, 2012. Association for Computational Linguistics.

[27] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[28] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 241–250, New York, NY, USA, 2010. ACM.

[29] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

[30] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):pp. 97–109, 1970.

[31] G. Heinrich. Parameter estimation for text analysis. Technical report, Fraunhofer IGD, 2009.

[32] G. Heinrich. "Infinite LDA"—implementing the HDP with minimum code complexity. Technical report, arbylon.net, 2011.

[33] E. Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen (German). *Journal für die Reine und Angewandte Mathematik*, 136:210–271, 1909.

[34] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, Berkeley, California, United States, 1999. ACM.

[35] L. Hong and B. D. Davison. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, Washington, D.C., USA, 2010. ACM.

[36] L. Hong, D. Yin, J. Guo, and B. D. Davison. Tracking trends: Incorporating term volume into temporal topic models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 484–492, San Diego, California, USA, 2011.

[37] Y. Hu, J. Boyd-Graber, and B. Satinoff. Interactive topic modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 248–257, Portland, Oregon, 2011.

[38] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda. Online multiscale dynamic topic models. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 663–672, Washington, D.C., USA, 2010. ACM.

[39] T. Iwata, T. Yamada, and N. Ueda. Modeling social annotation data with content relevance using a topic model. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 835–843, Vancouver, B.C., Canada, 2009.

[40] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, San Jose, California, 2007. ACM.

[41] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[42] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME: Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[43] K. Y. Kamath, J. Caverlee, Z. Cheng, and D. Z. Sui. Spatial influence vs. community influence: modeling the global spread of social media. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 962–971, Maui, Hawaii, USA, 2012. ACM.

[44] N. Kawamae and R. Higashinaka. Trend detection model. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1129–1130, Raleigh, North Carolina, USA, 2010. ACM.

[45] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 137–146, Washington, D.C., 2003. ACM.

[46] E. Kouloumpis, T. Wilson, and J. Moore. Twitter sentiment analysis: The good the bad and the OMG! In *International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 538–541, Barcelona, Spain, 2011. AAAI Press.

[47] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, Raleigh, North Carolina, USA, 2010. ACM.

[48] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems*, pages 897–904, Vancouver, B.C., Canada, 2008.

[49] C. Li, A. Sun, and A. Datta. Twevent: Segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 155–164, Maui, Hawaii, USA, 2012. ACM.

[50] J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 422–429, San Diego, California, USA, 2011. ACM.

[51] J. Liu, R. Hu, M. Wang, Y. Wang, and E. Y. Chang. Web-scale image annotation. In *Proceedings of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, PCM '08, pages 663–674, Tainan, Taiwan, 2008. Springer-Verlag.

[52] Z. Ma, W. Dou, X. Wang, and S. Akella. Tag-latent Dirichlet Allocation: Understanding hashtags and their relationships. volume 1, pages 260–267, Atlanta, GA, Nov 2013. IEEE Computer Society.

[53] T. Masada, D. Fukagawa, A. Takasu, T. Hamada, Y. Shibata, and K. Oguri. Dynamic hyperparameter optimization for Bayesian topical trend analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1831–1834, Hong Kong, China, 2009. ACM.

[54] M. Mathioudakis and N. Koudas. Twittermonitor: Trend detection over the Twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, Indianapolis, Indiana, USA, 2010. ACM.

[55] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. Teller, and H. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.

[56] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[57] T. Minka. The Dirichlet-tree distribution. Technical report, Justsystem Pittsburgh Research Center, 1999.

[58] T. P. Minka. Estimating a Dirichlet distribution. 2004.

[59] K. Nagesh and M. Murty. Obtaining single document summaries using latent Dirichlet allocation. In T. Huang, Z. Zeng, C. Li, and C. Leung, editors, *Neural Information Processing*, volume 7666 of *Lecture Notes in Computer Science*, pages 66–74. Springer Berlin Heidelberg, Doha, Qatar, 2012.

[60] R. M. Nallapati, S. Ditmore, J. D. Lafferty, and K. Ung. Multiscale topic tomography. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 520–529, San Jose, California, USA, 2007. ACM.

[61] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):pp. 249–265, 2000.

[62] M. Pennacchiotti and S. Gurumurthy. Investigating topic models for social media user recommendation. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 101–102, Hyderabad, India, 2011. ACM.

[63] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *International AAAI Conference on Weblogs and Social Media*, ICWSM '10, pages 130–137, Washington, D.C., USA, 2010. AAAI Press.

[64] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, Singapore, August 2009. Association for Computational Linguistics.

[65] D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina. Clustering the tagged web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 54–63, Barcelona, Spain, 2009. ACM.

[66] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 457–465, San Diego, California, USA, 2011. ACM.

[67] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open domain event extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1104–1112, Beijing, China, 2012. ACM.

[68] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 695–704, Hyderabad, India, 2011. ACM.

[69] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Banff, Canada, 2004. AUAI Press.

[70] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88:157–208, December 2012.

[71] D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, and F. Benevenuto. Finding trendsetters in information networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1014–1022, Beijing, China, 2012. ACM.

[72] M. Steyvers and T. Griffiths. *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2007.

[73] J. Tang, L. Yao, and D. Chen. Multi-topic based query-oriented summarization. In *SIAM International Conference on Data Mining*, SDM '2009, pages 144–149, Sparks, NV, April 2009.

[74] K. Tatsukawa and I. Kobayashi. Topic extraction based on prior knowledge obtained from target documents. In *Proceedings of ACL 2012 Student Research Workshop*, pages 31–36, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[75] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

[76] The National Science Board. *Science and Engineering Indicators 2010*. National Science Foundation, 2010.

[77] O. Tsur and A. Rappoport. What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 643–652, Seattle, Washington, USA, 2012. ACM.

[78] R. K. Venkatesh and K. Raghuveer. Legal documents clustering and summarization using hierarchical latent Dirichlet allocation. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 2(1):27–35, 2013.

[79] J. Vosecky, D. Jiang, K. W.-T. Leung, and W. Ng. Dynamic multi-faceted topic discovery in Twitter. In *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, pages 879–884, San Francisco, California, USA, 2013. ACM.

[80] H. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta,

editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981, Vancouver, B.C., Canada, 2009.

[81] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, Montreal, Quebec, Canada, 2009. ACM.

[82] C. Wang, D. Blei, and D. Heckerman. Continuous time dynamic topic models. In *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 579–586, Helsinki, Finland, 2008.

[83] X. Wang and A. McCallum. 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*, KDD '06, pages 424–433, Philadelphia, PA, USA, 2006.

[84] J. Weng and B.-S. Lee. Event detection in Twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 401–408, Barcelona, Spain, July 2011.

[85] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, February 2010. ACM.

[86] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 417–426, Edinburgh, Scotland, 2006. ACM.

[87] S. Xu, S. Bao, Y. Cao, and Y. Yu. Using social annotations to improve language model for information retrieval. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, CIKM '07, pages 1003–1006, Lisbon, Portugal, 2007. ACM.

[88] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing Twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, pages 338–349, Dublin, Ireland, 2011. Springer-Verlag.

[89] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles. Exploring social annotations for information retrieval. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 715–724, Beijing, China, 2008. ACM.

[90] J. Zhu, A. Ahmed, and E. P. Xing. MedLDA: Maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1257–1264, Montreal, Quebec, Canada, 2009. ACM.