上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

学士学位论文

BACHELOR'S THESIS

论文题目: 学术大数据的主题分析

学生姓名: 何俊贤

学生学号: 5130309699

专　　业: 信息工程

指导教师: 王新兵

学院(系): 电子信息与电气工程学院

学术大数据的主题分析

摘　要

主题模型在过去的十年里被广泛应用，方便了人们对学术大数据的探索。但是已有的主题模型在应用来分析学术大数据时依然存在两个重大缺陷：（1）传统的主题用关键词表示，缺乏对学术数据在文章层面的分析；（2）具有代表性的相关性主题模型可以有效地提取出主题间关系，但其计算复杂度非常高，使之不能够在工业界投入实际使用。在这篇论文中，我们设计了两个不同的模型来分别解决这两个问题。

第一，我们引入了“论文主题”的概念，在两种不同主题的基础上构建了异构的主题网络去关联词层面分析和文章层面分析。为此我们提出了一个新的模型量化存在于两种不同主题之间的三种关系。除此之外，我们还开发了一个演示系统 TopicAtlas 来展示异构主题网络。量化实验部分，我们在真实的学术数据上进行了实验，证明了我们的模型相较于其他方法的优势。

第二，我们提出了一个新的模型去学习紧凑的主题向量，然后通过主题向量之间的距离来对主题相关度建模。我们的模型将此前针对主题数量立方或者平方的时间复杂度降到了**线性**。后续的实验证明我们的方法能够处理的数据和模型规模是之前的 **100 倍**，在这同时我们也提供了论文分类和检索的结果证明了我们在提高模型效率的同时也并没有牺牲模型的精确性。

关键词： 主题模型，异构网络，主题向量，主题相关性，线性复杂度

Topic Analysis of Big Scholarly Data

ABSTRACT

Topic models serve as a powerful tool to facilitate big scholarly data exploration. However, existing topic models have encountered two limitations when applied to analyze big scholarly data: (1) Traditional topics are composed of words and lack an insight for academic data on document level; (2) Existing expressive correlated topic models are computationally expensive and impractical for industry deployment. In this thesis, we propose two different models to address the problems respectively.

First, we introduce the concept of "*DocTopic*" and construct a heterogeneous web of topics to associate word level with document level. To achieve this, a new generative model is proposed, where three different relationships in the heterogeneous topic web are quantified. We also develop a prototype demo system named *TopicAtlas* to exhibit such heterogeneous topic web. Extensive qualitative analyses are included to verify the efficacy of this heterogeneous topic web. Besides, we validate our model on real-life academic citation networks, showing that it preserves good performance on objective evaluation metrics.

Second, we propose a new model which learns compact topic embeddings and captures topic correlations through the closeness between the topic vectors. Our method reduces previous cubic or quadratic time complexity to *linear* w.r.t the topic size. Extensive experiments show that our approach is capable of handling model and data scales which are several orders of magnitude larger than existing correlation results, without sacrificing modeling quality by providing competitive or superior performance in document classification and retrieval.

Key words: topic model, heterogeneous web, topic embedding, topic correlation, linear complexity

Content

Chapter 1 Introduction

1.1 Motivation

Large ever-growing academic document collections provide great opportunities, and pose compelling challenges, to infer rich semantic structures underlying the scholarly data for data management and utilization. When faced with a new or unfamiliar academic collections, people may first ask a basic question: "What is there?". To answer this question, we resort to the notion of exploratory search [1], which is proposed to help people develop a general sense of the properties of a new academic collections before embarking on more specific inquiries [2]. Topic models, particularly the Latent Dirichlet Allocation (LDA) model [3], have been one of the most popular statistical frameworks to identify latent semantics from text corpora and facilitate exploratory search. Nevertheless, existing topic models suffer from significant limitations and are far from adequate for advanced big scholarly data mining.

One limitation of LDA lies in its intrinsic assumption of "topic" as distribution over words, lacking an insight on document level for text corpora, which is sometimes significantly demanding (e.g., researchers might want to locate the influential papers related to some word-composed topics). It is thus expected to introduce a new variable which represents the document importance to enable advanced big scholarly data navigation.

Another drawback of LDA derives from the conjugate Dirichlet prior, as it models topic occurrence (almost) independently and fails to capture rich topical correlations (e.g., a document about virus may be likely to also be about disease while unlikely to also be about finance). Effective modeling of the pervasive correlation patterns is essential for structural topic navigation, improved document representation, and accurate prediction [4, 5, 6]. Correlated Topic Model (CTM) [5] extends LDA using a logistic-normal prior which explicitly models correlation patterns with a Gaussian covariance matrix. Despite the enhanced expressiveness and resulting

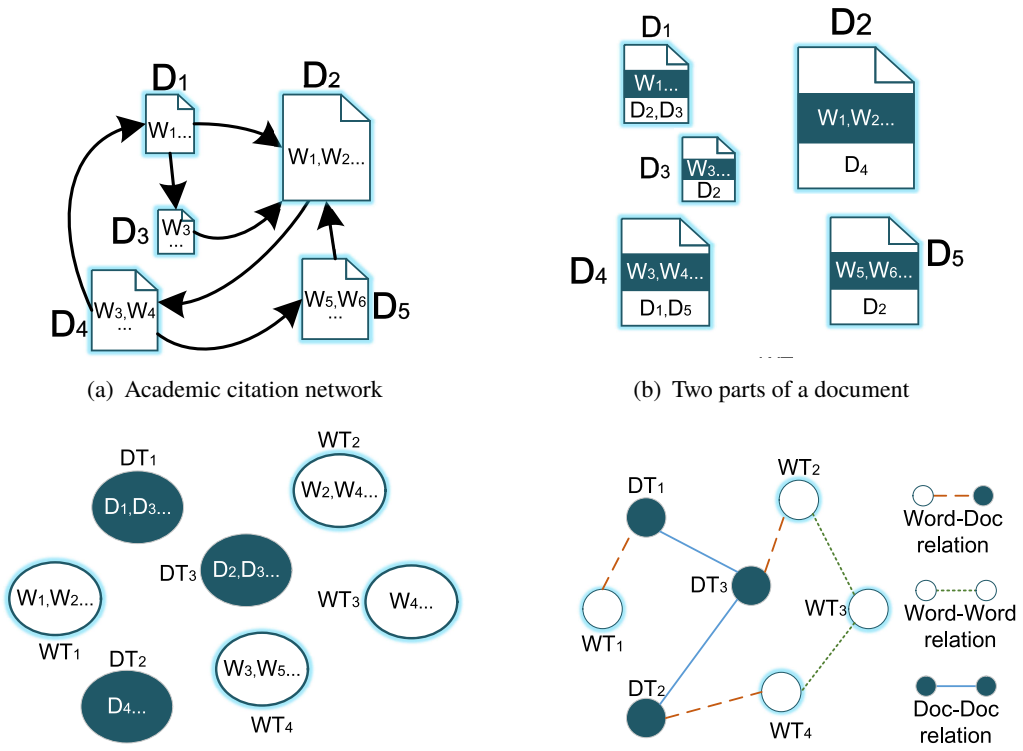
richer representations, practical applications of correlated topic modeling have unfortunately been limited due to high model complexity and poor scaling on large data. For instance, in CTM, direct modeling of pairwise correlations and the non-conjugacy of logistic-normal prior impose inference complexity of $\mathcal{O}(K^3)$, where K is the number of latent topics, significantly more demanding compared to LDA which scales only linearly. While there has been recent work on improved modeling and inference [6, 7, 8, 9], the model scale has still limited to less than 1000s of latent topics. This stands in stark contrast to recent industrial-scale LDA models which handle millions of topics on billions of documents [10, 11] for capturing long-tail semantics and supporting industrial applications [12], yet, such rich extraction task is expected to be better addressed with more expressive correlation models. It is therefore highly desirable to develop efficient correlated topic models with great representational power and highly scalable inference, for practical academic deployment.

In this thesis, we develop two different models to learn topic-related influential documents and extract correlation structures of industrial-scale latent topics, respectively.

1.2 Learning Topic-Related Influential Documents

In the topic-related influential documents learning task, we model the academic citation network as vertices associated with text and possessing high degrees of connectivity among themselves as shown Figure 1–1(a). We view each document as a "bag of references" [13, 14] inspired by popular "bag of words" assumption. For example, in the academic paper network, a paper with k references is viewed as a document with k "reference tokens" (or "document tokens"). Then, we can model these documents within a topic model framework where a new type of "topics" characterized by distributions over documents arises and important documents are assigned with high probabilities. By combining "word token" and "document token", each document is composed of two parts as shown in Figure 1–1(b), and two different types of topics are included as illustrated in Figure 1–1(c). To distinguish the two categories of topics, we call them *WordTopic* and *DocTopic* respectively.

However, it is still inconvenient to explore big scholarly data since users can only inspect



(a) Academic citation network

(b) Two parts of a document

(c) WordTopic and DocTopic

(d) Heterogeneous topic web

Figure 1–1: Illustration of some concepts. (a) Academic citation network. (b) Two parts of a document. W represents the "word token" part, and D below W represents "document token" part. (c) WordTopic (WT) and DocTopic (DT). (d) Heterogeneous topic web with two types of topics and three types of relationships.

the individual topic in isolation. Therefore, we expect to uncover the relations between topics to enable users to examine not only a topic itself but also the related fields and important documents. With that in mind, a complete heterogeneous topic web which displays three different types of relationships as described in Figure 1–1(d) is indispensable. Although the relationship between WordTopics (*Word-Word relation*) has been investigated previously [5, 15, 16, 17, 18, 14, 19, 20], the connections between DocTopic and DocTopic (*Doc-Doc relation*) and WordTopic and DocTopic (*Word-Doc relation*) have not been studied before.

To construct such heterogeneous topic web, we propose a probabilistic generative model called MHT (**M**odel for **H**eterogeneous **T**opic Web), where all three relationships are quantified. Our experiments on two academic citation networks demonstrate that MHT not only produces reliable heterogeneous topic web with high-quality topics but also preserves strong generaliz-

ability and predictive power.

Furthermore, we build **TopicAtlas**, a prototype demo system for convenient navigation in heterogeneous topic web. TopicAtlas displays Word-Word relation, Doc-Doc relation, and Word-Doc relation in a unified framework. With TopicAtlas, users are able to freely wander around the academic citation network via WordTopics and DocTopics.

1.3 Extracting Industrial-Scale Correlated Topics

To ease the high complexity of existing correlated topic models, we design a new model that extracts correlation structures of latent topics, sharing comparable expressiveness with the costly CTM model, while keeping as efficient as the simple LDA. We propose to learn a distributed representation for each latent topic, and characterize correlatedness of two topics through the closeness of respective topic vectors in the embedding space. Compared to previous pairwise correlation modeling, our topic embedding scheme is parsimonious with less parameters to estimate, yet flexible to enable richer analysis and visualization. Figure 1–2 illustrates the correlation patterns of 10K topics inferred by our model from two million NYTimes news articles, in which we can see clear dependency structures among the large collection of topics and grasp the semantics of the massive text corpus.

We further derive an efficient variational inference procedure combined with a fast sparsity-aware sampler for stochastic tackling of non-conjugacies. Our embedding based correlation modeling enables inference in the low-dimensional vector space, resulting in *linear* complexity w.r.t topic size as with the lightweight LDA. This allows us to discover 100s of 1000s of latent topics with their correlations on near 10 million articles, which is several orders of magnitude larger than prior work [6, 5].

Our topic embedding scheme differs from recent research which combines topic models with word embeddings [21, 22, 23, 24] for capturing word dependencies, as we instead focus on modeling dependencies in the latent topic space which exhibit uncertainty and are inferentially more challenging. To the best of our knowledge, this is the first work to incorporate distributed representation learning with topic correlation modeling, offering both intuitive geometric inter-

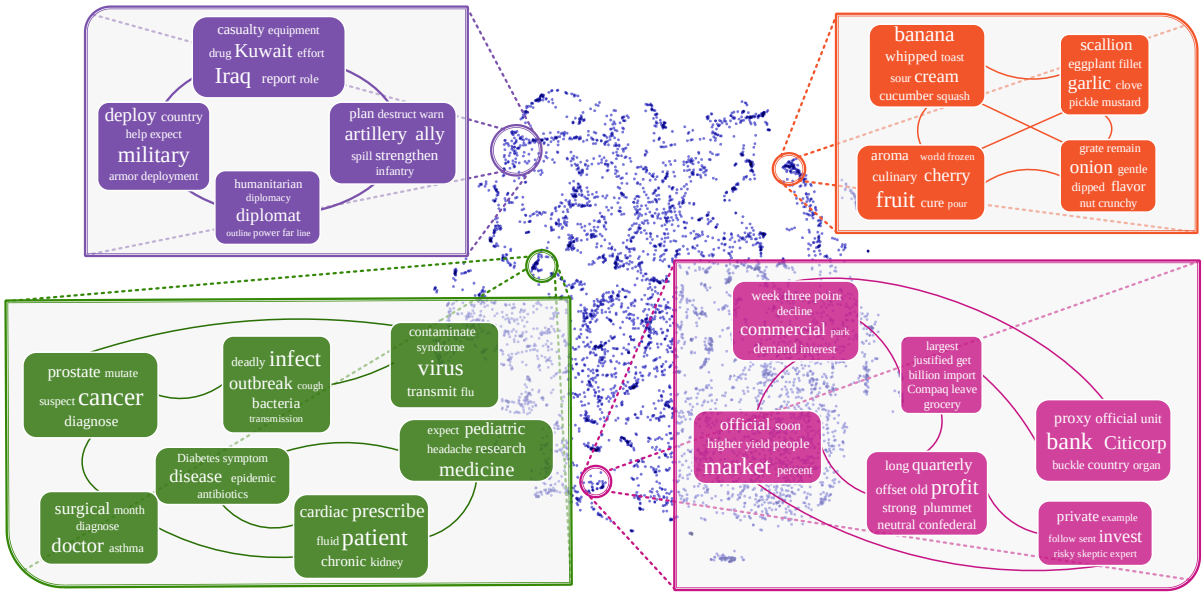


Figure 1–2: Visualization of 10K correlated topics on the NYTimes news corpus. The point cloud shows the 10K topic embeddings where each point represents a latent topic. Smaller distance indicates stronger correlation. We show four sets of topics which are nearby each other in the embedding space, respectively. Each topic is characterized by the top words according to the word distribution. Edge indicates correlation between topics with strength above some threshold.

pretation and theoretical Bayesian modeling advantages.

We demonstrate the efficacy of our method through extensive experiments on various large text corpora. Our approach shows greatly improved efficiency over previous correlated topic models, and scales well as with the much simpler LDA. This is achieved without sacrificing the modeling power—the proposed model extracts high-quality topics and correlations, obtaining competitive or better performance than CTM in document classification and retrieval tasks.

1.4 Contributions

To summarize, contributions of this thesis are three folds:

- We design a new WordTopic-DocTopic model to construct heterogeneous web of topics successfully.
- We develop TopicAtlas, a prototype system for academic citation network exploration, allowing users to investigate the heterogeneous topic web with details and explore big

scholarly data easily.

- We propose a new correlated topic model with topic embeddings, capable of handling model and data scales which are several orders of magnitude larger than existing correlation results.

The rest of the thesis is organized as follows: Chapter 2 briefly reviews related work; Chapter 3 presents the proposed WordTopic-DocTopic model and topic embedding model; Chapter 4 shows extensive experimental results; and finally we conclude the thesis.

Chapter 2 Related Work

2.1 Exploratory Search

When dealing with large collections of digitized historical documents, very often only little is known about the quantity, coverage, and relations of its content. In order to get an overview, exploring the data beyond simple "lookup" approaches is needed. The notion of exploratory search has been introduced to cover such cases [1].

Chaney and Blei [25] make an early effort in exploratory search via visualizing traditional topic models, where a navigator of documents is created and allows users to explore the hidden structure. Gretarsson et al. build a relatively mature system called TopicNets [26], which enables users to visualize individual document sections and their relations within the global topic document. Maiya et al. [27] build the topic similarity network for exploration and recognize how topics form large themes. Recently, Jahnichen et al. [28] develop a complete framework in this field, they depict probability distributions as tag clouds and permit the identification of related topic groups or outliers.

While the works mentioned above convey some information visually, these approaches consider the data as isolated-document corpus rather than linked text networks. With only text they cannot conduct a serious analysis for a text network on a document level. Specifically, although some of them are able to retrieve *topic-related* documents, there is no possibility for them to identify *topic-significant* documents, which are more crucial in exploratory search. We introduce DocTopic and propose the idea of heterogeneous topic web to enable users to keep track of related topic groups, relevant documents and significant documents.

2.2 Topic Modeling

Topic models represent a document as a mixture of latent topics. Among the most popular topic models is the LDA model [3] which assumes conjugate Dirichlet prior over topic mixing proportions for easier inference. Due to its simplicity and scalability, LDA has extracted broad interest for industrial applications [11, 12].

However, traditional topic models only consider text and ignore the significant link information. Recently, some variants of topic models are proposed for jointly analyzing text and links. A major part of them models the link information as evidence of content similarity between two documents [29, 30, 31, 16, 17, 32, 18, 33], but this kind of approach is not able to detect important documents with respect to a specific topic. Another categories of methods which generate the links from DocTopics can recognize significant documents [34, 35, 13, 15, 14]. These works, however, fail to construct a complete heterogeneous topic web composed of WordTopic, DocTopic and three different types of relations among them. Although the connection between WordTopics has been investigated before [5, 15, 16, 17, 18, 14, 19, 20], we are the first to model two types of topics and three types of relations jointly and build the heterogeneous topic web successfully.

On the other hand, the Dirichlet prior of LDA is incapable of capturing dependencies between topics. The classic CTM model provides an elegant extension of LDA by replacing the Dirichlet prior with a logistic-normal prior which models pairwise topic correlations with the Gaussian covariance matrix. However, the enriched extraction comes with computational cost. The number of parameters in the covariance matrix grows as square of the number of topics, and parameter estimation for the full-rank matrix can be inaccurate in high-dimensional space. More importantly, frequent matrix inversion operations during inference lead to $\mathcal{O}(K^3)$ time complexity, which has significantly restricted the model and data scales. To address this, [6] derives a scalable Gibbs sampling algorithm based on data augmentation. Though bringing down the inference cost to $\mathcal{O}(K^2)$ per document, the computation is still too expensive to be practical in real-world massive tasks. [8] reformulates the correlation prior with independent factor models for faster inference. However, similar to many other approaches, the problem

scale has still limited to thousands of documents and hundreds of topics. In contrast, we aim to scale correlated topic modeling to industrial level deployment by reducing the complexity to the LDA level which is linear to the topic size, while providing as rich extraction as the costly CTM model. We note that recent scalable extensions of LDA such as alias methods [36, 11] are orthogonal to our approach and can be applied in our inference for further speedup. We consider this as our future work.

Another line of topic models organizes latent topics in a hierarchy which also captures topic dependencies. However, the hierarchy structure is either pre-defined [37, 38, 39] or inferred from data using Bayesian nonparametric methods [40, 41] which are known to be computationally demanding [42, 43]. Our proposed topic embedding model is flexible without sacrificing scalability.

2.3 Distributed Representation Learning

There has been a growing interest in distributed representation that learns compact vectors (a.k.a embeddings) for words [44, 45], network nodes [46, 47], and others. The induced vectors are expected to capture semantic relatedness of the target items, and are successfully used in various applications. Compared to most work that induces embeddings for observed units, we learn distributed representations of latent topics which poses unique challenge for inference. Some previous work [48, 49] also induces compact topic manifold for visualizing large document collections. Our work is distinct in that we leverage the learned topic vectors for efficient correlation modeling and account for the uncertainty of correlations.

An emerging line of approaches [21, 22, 23, 24] incorporates word embeddings (either pre-trained or jointly inferred) with conventional topic models for capturing word dependencies and improving topic coherence. Our topic embedding model differs since we are interested in the topic level, aiming at capturing topic dependencies with learned topic embeddings.

Chapter 3 Models

This chapter proposes two different models for topic-specific influential documents mining and efficient topic correlation and embedding learning, respectively. We present the model structure in detail and derive the algorithm for inference.

3.1 Model for Heterogeneous Topic Web

In this part we describe the framework, generative process, and inference of MHT (Model for Heterogeneous Topic web).

3.1.1 Model Structure

In classical topic models each document is seen as "bag of words" and associated with a document specific topic distribution, which is used to draw a topic for each word in the generative process. Note that the "topic" here actually represents WordTopic in our notation framework and

Table 3–1: Notations used in this thesis.

Symbol	Description
D, K, V	number of documents, latent topics, and vocabulary words
N_d, L_d	number of words and references in document d
M	embedding dimension of topic and document
$\boldsymbol{u}_k$	embedding vector of topic k
$\boldsymbol{a}_d$	embedding vector of document d
$\boldsymbol{\eta}_d$	(unnormalized) topic weight vector of document d
w_{dn}	the nth word in document d
z_{dn}	the topic assignment of word w_{dn}
z'_{dl}	the DocTopic assignment of link y_{dl}
t_{dl}	the transitive WordTopic assignment of reference y_{dl}
$\boldsymbol{\pi}_k$	Transitive DocTopic distribution of WordTopic k
$\boldsymbol{\phi}_k$	word distribution of topic k
K_s	number of non-zero entries of document's topic proportion
V_s	number of non-zero entries of topic word distribution

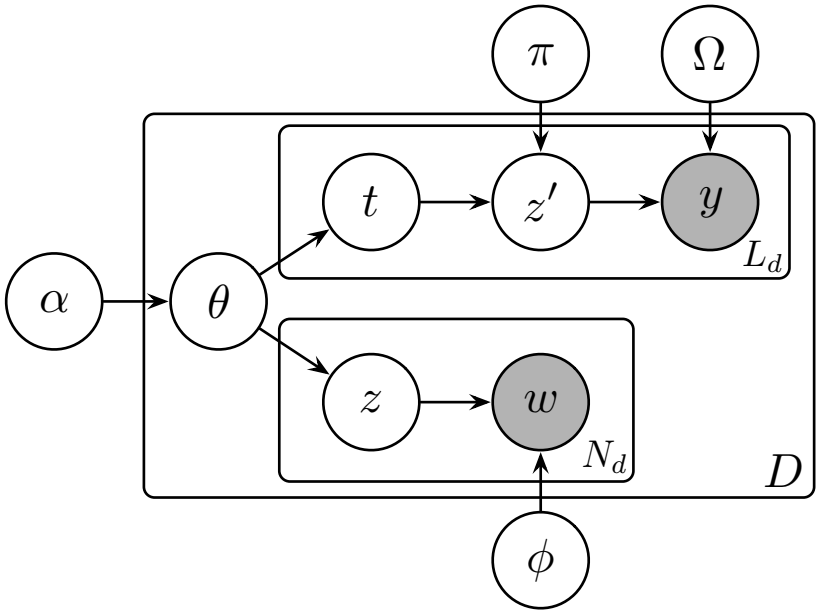


Figure 3–1: Graphical model representation of MHT.

is distribution over words. Inspired by previous topic models, we adopt the assumption of "bag of references" and produce a new "topic" which is distribution over references (documents), where each document assigns DocTopics for its references from DocTopic distribution. Since document specific WordTopic distribution and DocTopic distribution are totally different (e.g., a paper about disease is likely to cite quite a lot biology-related papers), some transition procedure between them is required to jointly model text and references.

Based on the discussion above, we employ a transition distribution $\boldsymbol{\pi}$ over DocTopics to depict the relation between the two types of topics. Details for complete generative process of our proposed model MHT are demonstrated in Algorithm 3–1 and Figure 3–1. Table 3–1 lists key notations.

In Algorithm 3–1, Step 1 and Step 2 are the same as classical topic model to generate words. A major distinction of MHT from other models is *Step 3*, where we employ a transitive latent WordTopic t as an "intermediary" from WordTopic domain to DocTopic domain. Intuitively, the WordTopics in a paper's content reminds authors of related DocTopics and enables them to locate references, it is therefore expected to design a "chain" from WordTopic to DocTopic and then references in the generative process. The transitive WordTopic t is the head of this

Algorithm 3–1 Generative Process of MHT

For each document $d = 1, 2 \cdots, D$,
1. Generate WordTopic distribution: $\theta_d \sim \text{Dir}(\alpha)$
2. For each word $n = 1, 2, \cdots, N_d$,
   (a) Draw the topic assignment $z_{dn} \sim \text{Multi}(\boldsymbol{\theta}_d)$
   (b) Draw the word $w_{dn} \sim \text{Multi}(\boldsymbol{\phi}_{z_{dn}})$
3. For each link $l = 1, \cdots, L_d$,
   (a) Draw a transition topic $t_{dl} \sim \text{Multi}(\boldsymbol{\theta}_d)$
   (b) Draw a DocTopic $z'_{dl} \sim \text{Multi}(\pi_{t_{dl}})$
   (c) Draw a linked document $y_{dl} \sim \text{Multi}(\boldsymbol{\Omega}_{z'_{dl}})$

"chain", and serves as the *reminder* WordTopic. In transition stage, we introduce a transition parameter $\boldsymbol{\pi}$ to express the connectivity strength between WordTopic and DocTopic so that the generation of DocTopic is equivalent to drawing it from $\boldsymbol{\theta\pi}$. Thus $\boldsymbol{\pi}$ serves as a transition matrix from $\boldsymbol{\theta}$ to a "spurious" underlying mixed DocTopic distribution $\boldsymbol{\theta'} = \boldsymbol{\theta\pi}$. Specifically, for a given WordTopic k, the value of $\pi_{kk'}$ indicates the probability for generating DocTopic k', i.e. $p(z' = k'|z = k) = \pi_{kk'}$. With that in mind, we can see how $\boldsymbol{\pi}$ works on transforming WordTopic domain into DocTopic domain.

3.1.2 Model Learning

To learn MHT, we resort to the mean-field variational EM inference method. For each document d, we use a fully factorized variational distribution to approximate the posterior distribution:

$$\begin{aligned} q(\boldsymbol{\theta}_d, \boldsymbol{z}_d, \boldsymbol{t}_d, \boldsymbol{z}'_d) = q(\boldsymbol{\theta}_d|\tau_d) \prod\nolimits_n q(z_{dn}|\boldsymbol{\kappa}_{dn}) \\ \times \prod\nolimits_l q(t_{dl}|\beta\nu_{dl}) \prod\nolimits_l q(z'_{dl}|\boldsymbol{\sigma}_{dl}), \end{aligned} \tag{3–1}$$

where $q(\boldsymbol{\theta}_d|\tau_d)$ is Dirichlet distribution and $q(z_{dn}|\boldsymbol{\kappa}_{dn})$, $q(t_{dl}|\boldsymbol{\nu}_{dl})$ and $q(z'_{dl}|\boldsymbol{\sigma}_{dl})$ are all multinomial distributions. Then we will try to maximize the evidence lower bound (ELBO) defined by:

$$\begin{aligned} \text{ELBO} = \sum\nolimits_d (\mathbb{E}_q[\log p(\boldsymbol{\theta}_d, \boldsymbol{z}_d, \boldsymbol{t}_d, \boldsymbol{z}'_d, \boldsymbol{w}_d, \boldsymbol{y}_d|\alpha, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\Omega})] \\ - \mathbb{E}_q[\log q(\boldsymbol{\theta}_d, \boldsymbol{z}_d, \boldsymbol{t}_d, \boldsymbol{z}'_d)]). \end{aligned} \tag{3–2}$$

Algorithm 3–2 Variational EM inference of MHT

```
Initialize variational parameters randomly. 𝒟 denotes the dataset
repeat
    for all d ∈ 𝒟 do
        repeat
            update κ_d, τ_d, ν_d, σ_d with Eqs.(3–3), (3–4), (3–5), (3–6)
        until convergence
        update ϕ, π, Ω, α with Eqs.(3–7), (3–8), (3–9)
    end for
until convergence
```

In the E-step, we update $\boldsymbol{\tau}, \boldsymbol{\kappa}, \boldsymbol{\nu}$ and $\boldsymbol{\sigma}$ iteratively to approximate the posterior distribution. Then, in the M-step, $\alpha, \boldsymbol{\phi}, \boldsymbol{\pi}$ and $\boldsymbol{\Omega}$ are renewed to maximize ELBO. Due to the limitation of space, we only provide crucial equations here.

$$\kappa_{dnk} \propto \phi_{kx} \exp(\Psi(\tau_{dk})). \tag{3–3}$$

$$\tau_{dk} = \alpha_k + \sum\nolimits_n \kappa_{dnk} + \sum\nolimits_n \nu_{dlk}. \tag{3–4}$$

$$\nu_{dlk} \propto \exp(\Psi(\tau_{dk}) + \sum\nolimits_{k'} \sigma_{dlk'} \log \pi_{kk'}). \tag{3–5}$$

$$\sigma_{dlk'} \propto \Omega_{k'd} \exp(\sum\nolimits_k \nu_{dlk} \log \pi_{kk'}). \tag{3–6}$$

$$\phi_{kv} \propto \sum\nolimits_{d,n} \kappa_{dnk} \cdot \mathbf{1}(w_{dn} = v). \tag{3–7}$$

$$\pi_{kk'} \propto \sum\nolimits_{d,l} \sigma_{dlk'} \nu_{dlk}. \tag{3–8}$$

$$\Omega_{k'i} \propto \sum\nolimits_{d,l} \sigma_{dlk'} \cdot \mathbf{1}(y_{dl} = i). \tag{3–9}$$

$\Psi(\cdot)$ denotes the digmma function, $\mathbf{1}()$ is the boolean operator. α is updated by Newton-Raphson

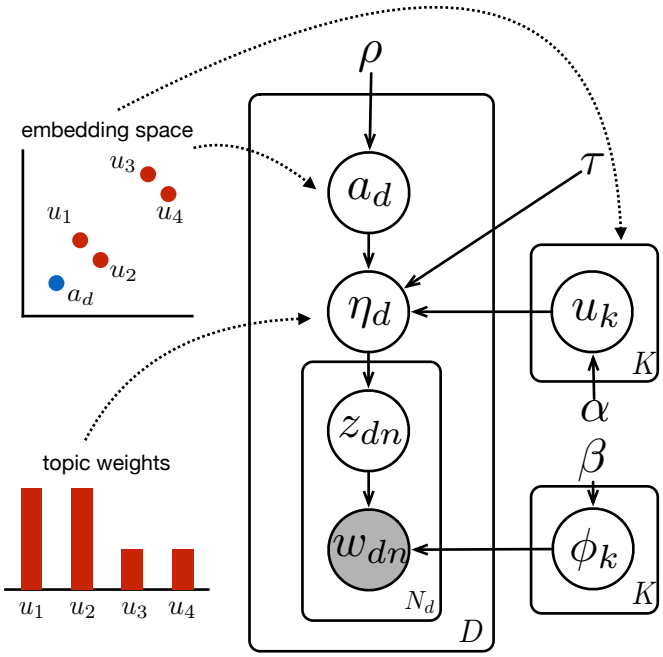


Figure 3–2: Graphical model representation of topic embedding model. The left part schematically shows our correlation modeling mechanism, where nearby topics tend to have similar (either large or small) weights in a document.

algorithm, the interested readers may refer to [3]. We summarize our variational inference algorithm in Algorithm 3–2.

3.2 Efficient Topic Embedding Model

This section proposes our topic embedding model for correlated topic modeling. We first give an overview of our approach, and present the model structure in detail. We then derive an efficient variational algorithm for inference.

3.2.1 Model Overview

We aim to develop an expressive topic model that discovers latent topics and underlying correlation structures. Despite this added representational power, we want to keep the model parsimonious and efficient in order to scale to large text data. As discussed above (Chapter 2), CTM captures correlations between topic pairs with a Gaussian covariance matrix, imposing $\mathcal{O}(K^2)$

parameter size and $\mathcal{O}(K^3)$ inference cost. In contrast, we adopt a new modeling scheme drawing inspiration from recent work on distributed representations, such as word embeddings [44] which learn low-dimensional word vectors and have shown to be effective in encoding word semantic relatedness.

We induce continuous distributed representations for latent topics, and, as in word embeddings, expect topics with relevant semantics to be close to each other in the embedding space. The contiguity of the embedding space enables us to capture topical co-occurrence patterns conveniently—we further embed documents into the same vector space, and characterize document's topic proportions with its distances to the topics. Smaller distance indicates larger topic weight. By the triangle inequality of distance metric, intuitively, a document vector will have similar (either large or small) distances to the vectors of two semantically correlated topics which are themselves nearby each other in the space, and thus tend to assign similar probability mass to the two topics. Figure 3–2, left part, schematically illustrates the embedding based correlation modeling.

We thus avoid expensive modeling of pairwise topic correlation matrix, and are enabled to perform inference in the low-dimensional embedding space, leading to significant reduction in model and inference complexity. We further exploit the intrinsic sparsity of topic occurrence, and develop stochastic variational inference with fast sparsity-aware sampling to enable high scalability. We derive the inference algorithm in section 3.2.3.

In contrast to word representation learning where word tokens are observed and embeddings can be induced directly from word collocation patterns, topics are hidden from the text, posing additional inferential challenge. We resort to generative framework as in conventional topic models by associating a word distribution with each topic. We also take into account uncertainty of topic correlations for flexibility. Thus, in addition to the intuitive geometric interpretation of our embedding based correlation scheme, the full Bayesian treatment also endows connection to the classic CTM model, offering theoretical insights into our approach. We present the model structure in the next section. (Table 3–1 lists key notations; Figure 3–2 shows the graphical model representation of our model.)

3.2.2 Model Structure

We first establish the notations. Let $\boldsymbol{W} = \{\boldsymbol{w}_d\}_{d=1}^{D}$ be a collection of documents. Each document d contains N_d words $\boldsymbol{w}_d = \{w_{dn}\}_{n=1}^{N_d}$ from a vocabulary of size V.

We assume K topics underlying the corpus. As discussed above, for each topic k, we want to learn a compact distributed representation $\boldsymbol{u}_k \in \mathbb{R}^M$ with low dimensionality ($M \ll K$). Let $\boldsymbol{U} \in \mathbb{R}^{K \times M}$ denote the topic vector collection with the kth row $\boldsymbol{U}_{k\cdot} = \boldsymbol{u}_k^T$. As a common choice in word embedding methods, we use the vector inner product for measuring the closeness between embedding vectors. In addition to topic embeddings, we also induce document vectors in the same vector space. Let $\boldsymbol{a}_d \in \mathbb{R}^M$ denote the embedding of document d. We now can conveniently compute the affinity of a document d to a topic k through $\boldsymbol{u}_k^T \boldsymbol{a}_d$. A topic k' nearby, and thus semantically correlated to topic k, will naturally have similar distance to the document, as $|\boldsymbol{u}_k^T \boldsymbol{a}_d - \boldsymbol{u}_{k'}^T \boldsymbol{a}_d| \leq \|\boldsymbol{u}_k - \boldsymbol{u}_{k'}\| \|\boldsymbol{a}_d\|$ and $\|\boldsymbol{u}_k - \boldsymbol{u}_{k'}\|$ is small.

We express uncertainty of the affinity by modeling the actual topic weights $\boldsymbol{\eta}_d \in \mathcal{R}^K$ as a Gaussian variable centered at the affinity vector, following $\boldsymbol{\eta}_d \sim \mathcal{N}(\boldsymbol{U}\boldsymbol{a}_d, \tau^{-1}\boldsymbol{I})$. Here τ characterizes the uncertainty degree and is pre-specified for simplicity. As in logistic-normal models, we project the topic weights into the probability simplex to obtain topic distribution $\boldsymbol{\theta}_d = \text{softmax}(\boldsymbol{\eta}_d)$, from which we sample a topic $z_{dn} \in \{1, \ldots, K\}$ for each word w_{dn} in the document. As in conventional topic models, each topic k is associated with a multinomial distribution $\boldsymbol{\phi}_k$ over the word vocabulary, and each observed word is drawn from respective word distribution indicated by its topic assignment.

Putting everything together, the generative process of the proposed model is summarized in Algorithm 3–3. A theoretically appealing property of our method is its intrinsic connection to conventional logistic-normal models such as the CTM model. If we marginalize out the document embedding variable $\boldsymbol{a}_d$, we obtain $\boldsymbol{\eta}_d \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{U}\boldsymbol{U}^T + \tau^{-1}\boldsymbol{I})$, recovering the pairwise topic correlation matrix with low rank constraint, where each element is just the closeness of respective topic embeddings, coherent to the above geometric intuitions. Such covariance decomposition has been used in other context, such as sparse Gaussian processes [50] for efficient approximation and Gaussian reparameterization [51] for differentiation and reduced variance.

Algorithm 3–3 Generative Process of Topic Embedding Model

1. For each topic $k = 1, 2, \cdots, K$,
   - Draw the topic word distribution $\boldsymbol{\phi_k} \sim \text{Dir}(\beta)$
   - Draw the topic embedding $\boldsymbol{u}_k \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\boldsymbol{I})$
2. For each document $d = 1, 2, \cdots, D$,
   - Draw the document embedding $\boldsymbol{a}_d \sim \mathcal{N}(\mathbf{0}, \rho^{-1}\boldsymbol{I})$
   - Draw the document topic weight $\boldsymbol{\eta}_d \sim \mathcal{N}(\boldsymbol{U}\boldsymbol{a}_d, \tau^{-1}\boldsymbol{I})$
   - Derive the distribution over topics $\boldsymbol{\theta}_d = \text{softmax}(\boldsymbol{\eta}_d)$
   - For each word $n = 1, 2, \cdots, N_d$,
     (a) Draw the topic assignment $z_{dn} \sim \text{Multi}(\boldsymbol{\theta}_d)$
     (b) Draw the word $w_{dn} \sim \text{Multi}(\boldsymbol{\phi}_{z_{dn}})$

Here we relate low-dimensional embedding learning with low-rank covariance decomposition and estimation.

The low-dimensional representations of latent topics enable parsimonious correlation modeling with parameter complexity of $\mathcal{O}(MK)$ (i.e., topic embedding parameters), which is efficient in terms of topic number K. Moreover, we are allowed to perform efficient inference in the embedding space, with inference cost *linear* in K, a huge advance compared to previous cubic complexity of vanilla CTM [5] and quadratic of recent improved version [6]. We derive our inference algorithm in the next section.

3.2.3 Model Inference

Posterior inference and parameter estimation is not analytically tractable due to the coupling between latent variables and the non-conjugate logistic-normal prior. This makes the learning difficult especially in our context of scaling to unprecedentedly large data and model sizes. We develop a stochastic variational method that (1) involves only compact topic vectors which are cheap to infer, and (2) includes a fast sampling strategy which tackles non-conjugacy and exploits intrinsic sparsity of both the document topic occurrence and the topical words.

We first assume a mean-field family of variational distributions:

$$q(\boldsymbol{u}, \boldsymbol{\phi}, \boldsymbol{a}, \boldsymbol{\eta}, \boldsymbol{z}) = \prod_k q(\boldsymbol{u}_k) q(\boldsymbol{\phi}_k) \prod_d q(\boldsymbol{a}_d) q(\boldsymbol{\eta}_d) \prod_n q(z_{dn}). \tag{3–10}$$

where the factors have the parametric forms:

$$\begin{aligned} q(\boldsymbol{u}_k) &= \mathcal{N}(\boldsymbol{u}_k|\boldsymbol{\mu}_k, \Sigma_k^{(u)}), \quad q(\boldsymbol{a}_d) = \mathcal{N}(\boldsymbol{a}_d|\boldsymbol{\gamma}_d, \Sigma_d^{(a)}), \\ q(\boldsymbol{\phi}_k) &= \text{Dir}(\boldsymbol{\phi}_k|\boldsymbol{\lambda}_k), \quad q(\boldsymbol{\eta}_d) = \mathcal{N}(\boldsymbol{\eta}_d|\boldsymbol{\xi}_d, \Sigma_d^{(\eta)}), \\ q(z_{dn}) &= \text{Multi}(z_{dn}|\boldsymbol{\kappa}_{dn}) \end{aligned} \tag{3–11}$$

Variational algorithms aim to minimize KL divergence from q to the true posterior, which is equivalent to tightening the evidence lower bound (ELBO):

$$\begin{aligned} \mathcal{L}(q) = &\sum\nolimits_k \mathbb{E}_q \left[\log \frac{p(\boldsymbol{u}_k)p(\boldsymbol{\phi}_k)}{q(\boldsymbol{u}_k)q(\boldsymbol{\phi}_k)}\right] + \\ &\sum\nolimits_{d,n} \mathbb{E}_q \left[\log \frac{p(\boldsymbol{a}_d)p(\boldsymbol{\eta}_d|\boldsymbol{a}_d, \boldsymbol{U})p(z_{dn}|\boldsymbol{\eta}_d)p(w_{dn}|z_{dn}, \boldsymbol{\phi})}{q(\boldsymbol{a}_d)q(\boldsymbol{\eta}_d)q(z_{dn})}\right] \end{aligned} \tag{3–12}$$

We optimize $\mathcal{L}(q)$ via coordinate ascent, interleaving the update of the variational parameters at each iteration. We employ stochastic variational inference which optimizes the parameters with stochastic gradients estimated on data minibatchs. Due to the space limitations, here we only describe key computation rules of the gradients (or closed-form solutions). These stochastically estimated quantities are then used to update the variational parameters after scaled by a learning rate.

Updating topic-word distribution $q(\boldsymbol{\phi}_k)$. For each topic k, we isolate only the terms that contain $q(\boldsymbol{\phi}_k)$,

$$\begin{aligned} q(\boldsymbol{\phi}_k) &\propto \exp\left\{\mathbb{E}_{-\boldsymbol{\phi}_k}(\log \prod\nolimits_v \phi_{kv}^{\beta-1}) + \mathbb{E}_{-\boldsymbol{\phi}_k}(\log \prod\nolimits_{d,n,v} \phi_{kv}^{\mathbf{1}(w_{dn}=v)\cdot\mathbf{1}(z_{dn}=k)})\right\} \\ &\propto \prod\nolimits_v \phi_{kv}^{\beta-1+\sum_{d,n}\mathbf{1}(w_{dn}=v)\cdot q(z_{dn}=k)}. \end{aligned} \tag{3–13}$$

Therefore,

$$q(\boldsymbol{\phi}_k) \sim \text{Dir}(\boldsymbol{\lambda}_k), \tag{3–14}$$

$$\lambda_{kv} = \beta + \sum\nolimits_{d,n} \mathbf{1}(w_{dn} = v) \cdot q(z_{dn} = k). \tag{3–15}$$

The cost for updating $q(\boldsymbol{\phi})$ is globally amortized across documents and words, and thus insignificant compared with other local parameter update.

Updating topic and document embeddings $q(\boldsymbol{u}_k), q(\boldsymbol{a}_d)$. For each topic k, we isolate only the terms that contain $q(\boldsymbol{u}_k)$,

$$\begin{aligned}\mathcal{L}(q(\boldsymbol{u}_k)) = \mathbb{E}_q\left[\log p(\boldsymbol{u}_k)\right] + \sum\nolimits_d \mathbb{E}_q\left[\log p(\boldsymbol{\eta}_d|\boldsymbol{a}_d, \boldsymbol{U})\right] \\ - \mathbb{E}_q\left[\log q(\boldsymbol{u}_k)\right].\end{aligned} \tag{3–16}$$

Then the variational distribution $q(\boldsymbol{u}_k)$ can be computed as:

$$q(\boldsymbol{u}_k) \propto \exp\left\{\mathbb{E}_{-\boldsymbol{u}_k}\left[\log p(\boldsymbol{u}_k|\alpha)\right] + \sum\nolimits_d \mathbb{E}_{-\boldsymbol{u}_k}\left[\log p(\boldsymbol{\eta}_d|\boldsymbol{a}_d, \boldsymbol{u}, \tau)\right]\right\}, \tag{3–17}$$

$$\begin{aligned}\mathbb{E}_{-\boldsymbol{u}_k}\left[\log p(\boldsymbol{u}_k|\alpha)\right] &= \mathbb{E}_{-\boldsymbol{u}_k}\left[\log\left\{\frac{1}{(2\pi)^{\frac{M}{2}}\alpha^{-\frac{M}{2}}}\exp(-\frac{\alpha}{2}\boldsymbol{u}_k^T\boldsymbol{u}_k)\right\}\right] \\ &\propto -\frac{\alpha}{2}\boldsymbol{u}_k^T\boldsymbol{u}_k,\end{aligned} \tag{3–18}$$

$$\begin{aligned}\mathbb{E}_{-\boldsymbol{u}_k}\left[\log p(\boldsymbol{\eta}_d|\boldsymbol{a}_d, \boldsymbol{u}, \tau)\right] &= \mathbb{E}_{-\boldsymbol{u}_k}\left[\log\left\{\frac{1}{(2\pi)^{\frac{M}{2}}\tau^{-\frac{M}{2}}}\exp(-\frac{\tau}{2}(\boldsymbol{\eta}_d - \boldsymbol{U}\boldsymbol{a}_d)^T(\boldsymbol{\eta}_d - \boldsymbol{U}\boldsymbol{a}_d))\right\}\right] \\ &= -\frac{\tau}{2}\boldsymbol{u}_k^T\left[\sum\nolimits_d(\Sigma_d^{(a)} + \boldsymbol{\gamma}_d\boldsymbol{\gamma}_d^T)\right]\boldsymbol{u}_k + \tau\sum\nolimits_d \xi_{dk}\boldsymbol{\gamma}_d^T\boldsymbol{u}_k + C.\end{aligned} \tag{3–19}$$

Therefore,

$$q(\boldsymbol{u}_k) \propto \exp\left\{-\frac{1}{2}\boldsymbol{u}_k^T\left[\alpha\boldsymbol{I} + \sum\nolimits_d(\tau\Sigma_d^{(a)} + \tau\boldsymbol{\gamma}_d\boldsymbol{\gamma}_d^T)\right]\boldsymbol{u}_k + \tau\sum\nolimits_d \xi_{dk}\boldsymbol{\gamma}_d^T\boldsymbol{u}_k\right\}, \tag{3–20}$$

where $\Sigma_d^{(a)}$ is the covariance matrix of $\boldsymbol{a}_d$. From Eq.(3–20), we know $q(\boldsymbol{u}_k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k^{(u)})$.

$$\begin{aligned}\boldsymbol{\mu}_k &= \tau\Sigma^{(u)} \cdot (\sum\nolimits_d \xi_{dk}\boldsymbol{\gamma}_d), \\ \Sigma_k^{(u)} &= \left[\alpha\boldsymbol{I} + \sum\nolimits_d(\tau\Sigma_d^{(a)} + \tau\boldsymbol{\gamma}_d\boldsymbol{\gamma}_d^T)\right]^{-1}.\end{aligned} \tag{3–21}$$

Notice that $\Sigma_k^{(u)}$ is unrelated to k, which means all topic embeddings share the same covariance matrix, we denote it as $\Sigma^{(u)}$.

Symmetrically,

$$\begin{aligned} \boldsymbol{\gamma}_d &= \tau\Sigma^{(a)} \cdot (\sum\nolimits_k \xi_{dk}\boldsymbol{\mu}_k), \\ \Sigma^{(a)} &= \left[\gamma\boldsymbol{I} + \tau K\Sigma^{(u)} + \sum\nolimits_k \tau\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T\right]^{-1}. \end{aligned} \tag{3–22}$$

Since $\Sigma^{(a)}$ is unrelated to d, we can rewrite equation(3–21) as

$$\begin{aligned} \boldsymbol{\mu}_k &= \tau\Sigma^{(u)} \cdot (\sum\nolimits_d \xi_{dk}\boldsymbol{\gamma}_d), \\ \Sigma^{(u)} &= \left[\alpha\boldsymbol{I} + \tau D\Sigma_d^{(a)} + \sum\nolimits_d \tau\boldsymbol{\gamma}_d\boldsymbol{\gamma}_d^T\right]^{-1}. \end{aligned} \tag{3–23}$$

where we have omitted the subscript k of the variational covariance matrix $\Sigma^{(u)}$ as it is independent with k. Intuitively, the optimal variational topic embeddings are the centers of variational document embeddings scaled by respective document topic weights and transformed by the variational covariance matrix.

Learning low-dimensional topic and document embeddings is computationally cheap. Specifically, by Eq.(3–23), updating the set of variational topic vector means $\{\boldsymbol{\mu}_k\}_{k=1}^K$ imposes complexity $\mathcal{O}(KM^2)$, and updating the covariance $\Sigma^{(u)}$ requires only $\mathcal{O}(M^3)$. Similarly, by Eq.(3–22), the cost of optimizing $\boldsymbol{\gamma}_d$ and $\Sigma^{(a)}$ is $\mathcal{O}(KM)$ and $\mathcal{O}(KM^2)$, respectively. Note that $\Sigma^{(a)}$ is shared across all documents and does not need updates per document. We see that all the updates cost only linearly w.r.t to the topic size K which is critical to scale to large-scale practical applications.

Sparsity-aware topic sampling. We next consider the optimization of the variational topic assignment $q(z_{dn})$ for each word w_{dn}. Letting $w_{dn} = v$, the optimal solution is:

$$q(z_{dn} = k) \propto \exp\{\boldsymbol{\xi}_{dk}\} \exp\left\{\Psi(\lambda_{kv}) - \Psi\left(\sum\nolimits_{v'} \lambda_{kv'}\right)\right\}, \tag{3–24}$$

where $\Psi(\cdot)$ is the digamma function; and $\boldsymbol{\xi}_d$ and $\boldsymbol{\lambda}_k$ are the variational means of the document's topic weights and the variational word weights (Eq.(3–11)), respectively. Direct computation of $q(z_{dn})$ with Eq.(3–24) has complexity of $\mathcal{O}(K)$, which becomes prohibitive in the presence of many latent topics. To address this, we exploit two aspects of intrinsic sparsity in the modeling:

(1) Though a whole corpus can cover a large diverse set of topics, a single document in the corpus is usually about only a small number of them. We thus only maintain the top K_s entries in each $\boldsymbol{\xi}_d$, where $K_s \ll K$, making the complexity due to the first term in the right-hand side of Eq.(3–24) only $\mathcal{O}(K_s)$ for all K topics in total; (2) A topic is typically characterized by only a few words in the large vocabulary, we thus cut off the variational word weight vector $\boldsymbol{\lambda}_k$ for each k by maintaining only its top V_s entries ($V_s \ll V$). Such sparse treatment helps enhance the interpretability of learned topics, and allows cheap computation with on average $\mathcal{O}(KV_s/V)$ cost for the second term[1]. With the above sparsity-aware updates, the resulting complexity for Eq.(3–24) with K topics is brought down to $\mathcal{O}(K_s + KV_s/V)$, a great speedup over the original $\mathcal{O}(K)$ cost. The top K_s entries of $\boldsymbol{\xi}_d$ are selected using a Min-heap data structure, whose computational cost is amortized across all words in the document, imposing $\mathcal{O}(K/N_d \log K_s)$ computation per word. The cost for finding the top V_s entries of $\boldsymbol{\lambda}_k$ is similarly amortized across documents and words, and becomes insignificant.

Updating the remaining variational parameters will frequently involve computation of variational expectations under $q(z_{dn})$. It is thus crucial to speedup this operation. To this end, we employ sparse approximation by sampling from $q(z_{dn})$ a single indicator $\tilde{z}_{dn}$, and use the "hard" sparse distribution $\tilde{q}(z_{dn} = k) := \mathbf{1}(\tilde{z}_{dn} = k)$ to estimate the expectations. Note that the sampling operation is cheap, having the same complexity with computing $q(z_{dn})$ as above. As shown shortly, such sparse computation will significantly reduce our running cost. Though stochastic expectation approximation is commonly used for tackling intractability [52, 53], here we instead apply the technique for fast estimation of tractable expectations.

We next optimize the variational topic weights $q(\boldsymbol{\eta}_d|\boldsymbol{\xi}_d, \Sigma_d^{(\eta)})$. Extracting only the terms in $\mathcal{L}(q)$ involving $q(\boldsymbol{\eta}_d)$, we get:

$$\begin{aligned}\mathcal{L}(q(\boldsymbol{\eta}_d)) = \mathbb{E}_q\left[\log p(\boldsymbol{\eta}_d|\boldsymbol{a}_d, \boldsymbol{U})\right] + \mathbb{E}_q\left[\log p(\boldsymbol{z}_d|\boldsymbol{\eta}_d)\right] \\ - \mathbb{E}_q\left[\log q(\boldsymbol{\eta}_d)\right],\end{aligned} \tag{3–25}$$

[1] In practice we also set a threshold s such that each word v needs to have at least s non-zero entries in $\{\lambda_k\}_{k=1}^K$. Thus the exact complexity of the second term is $\mathcal{O}(\max\{KV_s/V, s\})$.

where the second term

$$\mathbb{E}_q\left[\log p(\boldsymbol{z}_d|\boldsymbol{\eta}_d)\right] = \sum\nolimits_{k,n} q(z_{dn}=k)\mathbb{E}_q\left[\log(\text{softmax}_k(\boldsymbol{\eta}_d))\right]$$

involves variational expectations of the logistic transformation which does not have an analytic form. We construct a fast Monto Carlo estimator for approximation. Particularly, we employ reparameterization trick by first assuming a diagonal covariance matrix $\Sigma_d^{(\eta)} = \text{diag}(\boldsymbol{\sigma}_d^2)$ as is commonly used in previous work [5, 54], where $\boldsymbol{\sigma}_d$ denotes the vector of standard deviations, resulting in the following sampling procedure:

$$\boldsymbol{\eta}_d^{(t)} = \boldsymbol{\xi}_d + \boldsymbol{\sigma}_d \odot \boldsymbol{\epsilon}^{(t)}; \quad \boldsymbol{\epsilon}^{(t)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{3–26}$$

where $\odot$ is the element-wise multiplication. With T samples of $\boldsymbol{\eta}_d$, we can estimate the variational lower bound and the derivatives $\nabla\mathcal{L}$ w.r.t the variational parameters $\{\boldsymbol{\xi}_d, \boldsymbol{\sigma}_d\}$. For instance,

$$\begin{aligned}
&\nabla_{\boldsymbol{\xi}_d}\mathbb{E}_q\left[\log p(\boldsymbol{z}_d|\boldsymbol{\eta}_d)\right] \\
&\approx \sum\nolimits_{k,n} q(z_{dn}=k)\boldsymbol{e}_k - (N_d/T)\sum\nolimits_{t=1}^{T} \textbf{softmax}\left(\boldsymbol{\eta}_d^{(t)}\right) \\
&\approx \sum\nolimits_{k,n} \mathbf{1}(\tilde{z}_{dn}=k)\boldsymbol{e}_k - (N_d/T)\sum\nolimits_{t=1}^{T} \textbf{softmax}\left(\boldsymbol{\eta}_d^{(t)}\right)
\end{aligned} \tag{3–27}$$

where $\boldsymbol{e}_k$ is an indicator vector with the kth element being 1 and the rest 0. In practice $T = 1$ is usually sufficient for effective inference. The second equation applies the hard topic sample mentioned above, which reduces the time complexity $\mathcal{O}(KN_d)$ of the original standard computation (the first equation) to $\mathcal{O}(N_d + K)$ (i.e., $\mathcal{O}(N_d)$ for the first term and $\mathcal{O}(K)$ for the second).

The first term in Eq.(3–25) depends on the topic and document embeddings to encode topic correlations in document's topic weights. The derivative w.r.t to the variational parameter $\boldsymbol{\xi}_d$ is computed as:

$$\nabla_{\boldsymbol{\xi}_d}\mathbb{E}_q\left[\log p(\boldsymbol{\eta}_d|\boldsymbol{U}, \boldsymbol{a}_d)\right] = \tau(\tilde{\boldsymbol{U}}\boldsymbol{\gamma}_d - \boldsymbol{\xi}_d). \tag{3–28}$$

Algorithm 3–4 Stochastic variational inference of topic embedding model

Initialize variational parameters randomly
repeat
  compute learning rate $\iota_{iter} = 1/(1 + iter)^{0.9}$
  sample a minibatch of documents $\mathcal{B}$
  for all $d \in \mathcal{B}$ **do**
    repeat
      update $q(\boldsymbol{z}_d)$ with Eq.(3–24) and sample $\tilde{\boldsymbol{z}}_d$
      update $\boldsymbol{\gamma}_d$ with Eq.(3–22)
      update $q(\boldsymbol{\eta}_d)$ using respective gradients computed with Eqs.(3–27),(3–28).
    until convergence
    compute stochastic optimal values $\boldsymbol{\mu}^*, \Sigma^{(u)*}$ with Eq.(3–23)
    compute stochastic optimal values $\boldsymbol{\lambda}^*$ with Eq.(3–29)
    update $\boldsymbol{x} = (1 - \iota_{iter})\boldsymbol{x} + \iota_{iter}\boldsymbol{x}^*$ with $\boldsymbol{x} \in \{\boldsymbol{\mu}, \Sigma^{(u)}, \boldsymbol{\lambda}\}$
    update $\Sigma^{(a)}$ with Eq.(3–22)
  end for
until convergence

Here $\tilde{\boldsymbol{U}}$ is the collection of variational means of topic embeddings where the kth row $\tilde{\boldsymbol{U}}_{k\cdot} = \boldsymbol{\mu}_k^T$. We see that, with low-dimensional topic and document vector representations, inferring topic correlations is of low cost $\mathcal{O}(KM)$ which grows only linearly w.r.t to the topic size. The complexity of the remaining terms in Eq.(3–25), as well as respective derivatives w.r.t the variational parameters, has complexity of $\mathcal{O}(KM)$ (Please see the supplements [55] for more details). In summary, the cost of updating $q(\boldsymbol{\eta}_d)$ for each document d is $\mathcal{O}(KM + K + N_d)$.

Finally, the optimal solution of the variational topic word distribution $q(\boldsymbol{\phi}_k|\boldsymbol{\lambda}_k)$ is given by:

$$\lambda_{kv} = \beta + \sum\nolimits_{d,n} \mathbf{1}(w_{dn} = v)\mathbf{1}(\tilde{z}_{dn} = k). \tag{3–29}$$

Algorithm summarization. We summarize our variational inference in Algorithm 3–4. As analyzed above, the time complexity of our variational method is $\mathcal{O}(KM^2+M^3)$ for inferring topic embeddings $q(\boldsymbol{u}_d)$. The cost per document is $\mathcal{O}(KM)$ for computing $q(\boldsymbol{a}_d)$, $\mathcal{O}(KM)$ for updating $q(\boldsymbol{\eta}_d)$, and $\mathcal{O}((K_s + KV_s/V)N_d)$ for maintaining $q(\boldsymbol{z}_d)$. The overall complexity for each document is thus $\mathcal{O}(KM + (K_s + KV_s/V)N_d)$, which is linear to model size (K),

comparable to the LDA model while greatly improving over previous correlation methods with cubic or quadratic complexity.

The variational inference algorithm endows rich independence structures between the variational parameters, allowing straightforward parallel computing. In our implementation, updates of variational topic embeddings $\{\boldsymbol{\mu}_k\}$ (Eq.(3–23)), topic word distributions $\{\boldsymbol{\lambda}_k\}$ (Eq.(3–29)), and document embeddings $\{\boldsymbol{\gamma}_d\}$ (Eq.(3–22)) for a data minibatch, are all computed in parallel across multiple CPU cores.

Chapter 4 Experiments

4.1 Evaluation of MHT

In this section, we first describe the experiment setups such as dataset selection and parameter settings. Then, we show how to construct the heterogeneous topic web for TopicAtlas, and present some qualitative analysis of the constructed network. Besides, the demo system TopicAtlas is displayed as well. Finally, we validate the effectiveness of MHT, the backbone method for TopicAtlas, as a topic model for text network. Compared with some representative baseline methods, MHT achieves the best averaged performance in terms of topic interpretability and generalizability.

4.1.1 Setup

Dataset. We use the following two datasets in our experiments:

ACL Anthology Network (AAN). AAN [56] is a public scientific literature dataset in the Natural Language Processing (NLP) field with $20,989$ abstracts of papers and $125,934$ citations.

CiteseerX. CiteseerX[1] is a well-known scientific literature digital library that primarily focuses on the literature in computer and information science. We collect a subset of CiteseerX dataset, which includes the abstracts of $716,800$ documents and $1,760,574$ links.

Parameter setting. On the task of exploring heterogeneous topic web, we first need to select a reasonable topic number, which is a non-trivial task in topic models. To achieve this, we first preprocess the data using classical LDA model with varying topic numbers and evaluate the topic interpretability in terms of the topic coherence score [57]. Among the candidate topic numbers 50, 70, 90, 110, 130, and 150, topic number 70 leads to the highest topic coherence score for both AAN and CiteseerX. For simplicity, we set the topic number of WordTopic and

[1] http://citeseer.ist.psu.edu/oai.html

DocTopic equal. Therefore, we implement MHT with 70 WordTopics and 70 DocTopics to explore the text networks in the two datasets. In addition, we follow the convention of [58] and initialize $\alpha = 0.01$. The parameters $\boldsymbol{\pi}$, $\boldsymbol{\phi}$ and $\boldsymbol{\Omega}$ are randomly initialized since we do not have any prior knowledge.

Furthermore, as discussed above, we use variational EM inference to learn the parameters in MHT. In our experiments, for both datasets the inner variational inference loop terminates when the fractional increase of ELBO is less than 10^{-9} in two successive iterations, or the number of iterations exceeds 100. For the outer EM loop, we stop it when the relative increment ratio is less than 10^{-4}, or the number of iterations exceeds 50.

4.1.2 Heterogeneous Topic Web Construction

We use co-occurrence probability to quantify the strength of the three types of relations in heterogeneous topic web, and thus our goal is to figure out $p(z_1 = k_1, z_2 = k_2|\mathcal{D})$, $p(z'_1 = k'_1, z'_2 = k'_2|\mathcal{D})$ and $p(z = k, z' = k'|\mathcal{D})$.

Word-Word Relation Strength. Since we assume the generation of WordTopics is independent with each other given document d, the Word-Word relation strength can be calculated as follows:

$$\begin{aligned} p(z_1 = k_1, z_2 = k_2|\mathcal{D}) = \sum\nolimits_{d,z'} & p(z'|\mathcal{D})p(d|z') \\ & \times p(z_1 = k_1|d) \\ & \times p(z_2 = k_2|d), \end{aligned} \tag{4–1}$$

where $p(z|d)$ and $p(d|z')$ can be obtained from $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ respectively. Posterior expectation of $\boldsymbol{\theta}$ is given by:

$$\theta_{ik} = \frac{\#(d = i, z = k) + \alpha_k}{\sum_{k^*=1}^{K_w}(\#(d = i, z = k^*) + \alpha_{k^*})}, \tag{4–2}$$

where $\#(d = i, z = k)$ represents the number of words assigned with WordTopic k in document i and the assignment can be obtained from $\boldsymbol{\kappa}$. K_w is the number of WordTopics.

In addition, the empirical posterior distribution over DocTopics can be computed as:

$$p(z' = k'|\mathcal{D}) = \frac{\#(z' = k')}{\sum_{k^*} \#(z' = k^*)}, \tag{4–3}$$

where $\#(z' = k')$ represents the number of references assigned with DocTopic k' and can be obtained from $\boldsymbol{\sigma}$.

Doc-Doc Relation Strength. Based on the assumption that DocTopics are generated independently given a WordTopic, we can compute Doc-Doc relation strength as:

$$\begin{aligned} p(z'_1 = k'_1, z'_2 = k'_2|\mathcal{D}) = \sum\nolimits_z p(z|\mathcal{D}) p(z'_1 = k'_1|z;\mathcal{D}) \\ \times p(z'_2 = k'_2|z;\mathcal{D}). \end{aligned} \tag{4–4}$$

$\boldsymbol{\pi}$ represents $p(z'|z;\mathcal{D})$ and similarly the empirical posterior distribution over WordTopics is given by:

$$p(z = k|\mathcal{D}) = \frac{\#(z = k)}{\sum_{k^*} \#(z = k^*)}. \tag{4–5}$$

Word-Doc Relation Strength. Word-Doc relation strength can be easily computed by Bayes' theorem:

$$p(z = k, z' = k'|\mathcal{D}) = p(z' = k'|z = k;\mathcal{D}) p(z = k|\mathcal{D}). \tag{4–6}$$

Summarizing DocTopic. While top words are able to represent WordTopic explicitly, on the document side there are only distributions over documents to express DocTopics, yet generally it would be preferable to summarize topics with a few words [59]. Therefore, we leverage the words in abstracts to summarize DocTopics. Specifically, for a given DocTopic k', we compute the expectancy of word w as:

$$\mathbb{E}(w|z' = k') = \sum\nolimits_d \Omega_{k'd} \cdot \#(w, d). \tag{4–7}$$

Then the words with high expectancy are selected as *indicative words* of this DocTopic, which will be displayed in our demo system TopicAtlas.

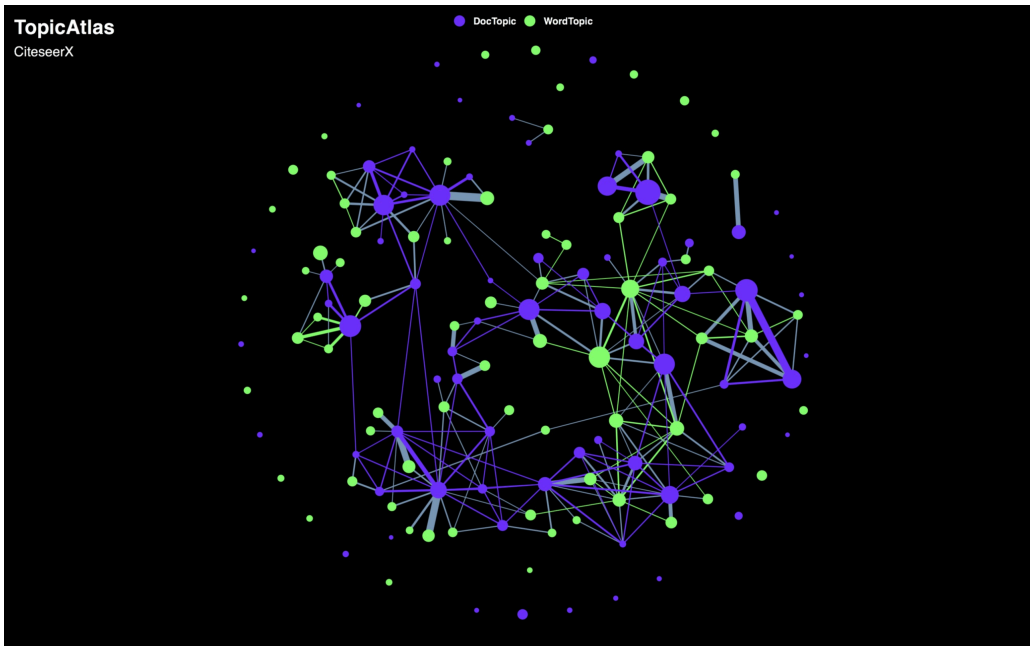


Figure 4–1: An overview of TopicAtlas. Different colors represent different types of topics, and the vertex size expresses the dominance of corresponding topic. Thickness of edges is proportionate to relation strength (best seen in color).

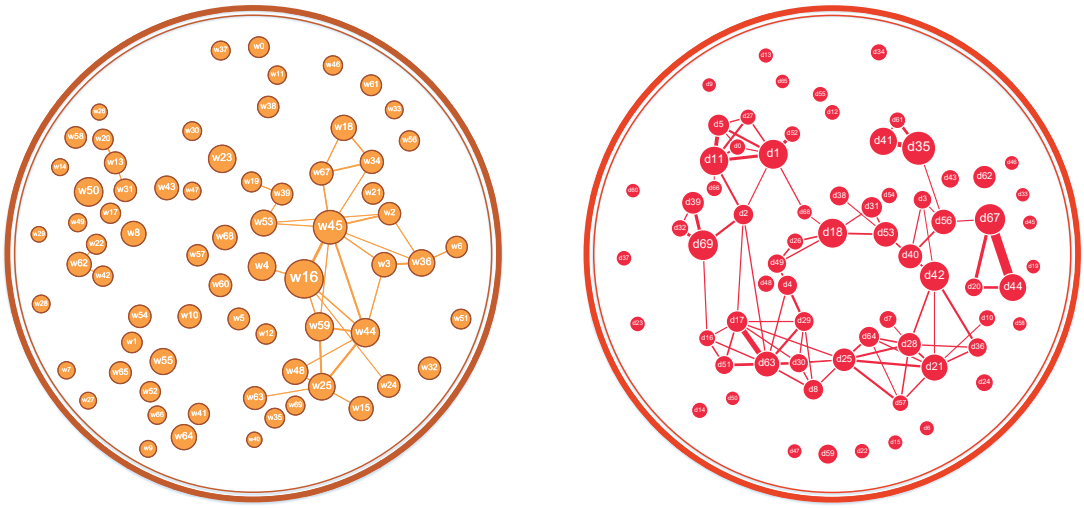
4.1.3 TopicAtlas

We design TopicAtlas based on the constructed heterogeneous topic web. An overview of TopicAtlas on CiteseerX dataset is displayed in Figure 4–1, and the TopicAtlas demo is available for public[1]. Aiming to help users navigate in an unfamiliar academic citation network, TopicAtlas has the following features:

Topic Landscape Exhibition. We display top 10 keywords for each WordTopic, and top 5 representative documents and top 10 indicative words for each DocTopic. The diameters of topic vertices express their corresponding *topic dominance* or *topic importance*, which is computed by $p(z|\mathcal{D})$ for each WordTopic and $p(z'|\mathcal{D})$ for each DocTopic. With TopicAtlas, we are able to answer the questions of "what is there" and "what is important" in an academic document collections.

Accurate Relationship. The three types of relations correspond to three types of edges in

[1] https://jxhe.github.io/demo/TopicAtlas/CiteseerX.html



(a) Word-Word subgraph (b) Doc-Doc subgraph

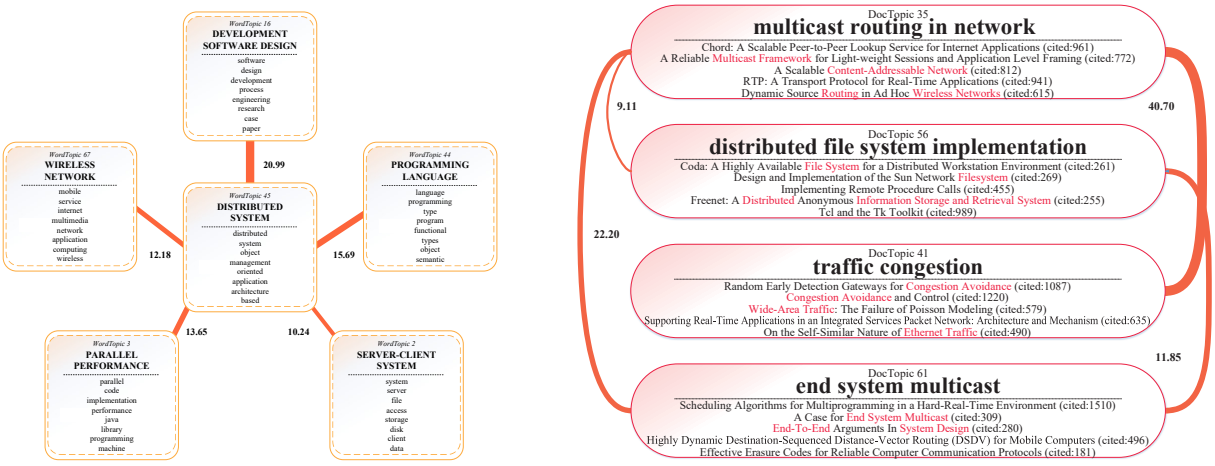
Figure 4–2: Subgraphs of heterogeneous topic web: (a) Word-Word Subgraph and (b) Doc-Doc Subgraph (best seen in color). The yellow vertices in the Word-Word subgraph represent WordTopics and the red vertices in the Doc-Doc subgraph represent DocTopics

the graph. The weights of these edges are the ratio of the co-occurrence probability calculated to the prior probability of a random edge (0.0002). The thickness of the edges is proportionate to these values and we remove those whose weights are negligible. The edges provide us with the connectivity between different topics, allowing us to track topic correlation and locate topic specific influential documents, and enabling alternatively navigating big scholarly data on word and document level.

Topic Community Identification. As shown in Figure 4–1, while some topics are relatively isolated, other topics hold strong connections between each other and form *topic communities*. Several related topics are able to represent more general *parent topic* (e.g., topics about disease, medicine, virus are heavily correlated and can represent parent topic "medical science"). Therefore, we can explore topic hierarchy structures from TopicAtlas.

4.1.4 Text Network Exploration via Heterogeneous Topic Web

In this part, we engage in an in-depth exploration of the heterogeneous topic web. To facilitate the analytic reasoning, three auxiliary subgraphs of TopicAtlas are presented here: *Word-*



(a) WordTopic example (b) DocTopic example

Figure 4–3: WordTopic "*Distributed system*" example and DocTopic "*Multicast routing in network* example. These topics are labeled manually.

Word subgraph, *Doc-Doc* subgraph and *Word-Doc* subgraph. As the name suggests, Word-Word subgraph only includes the edges between WordTopics, Doc-Doc subgraph contains merely the edges between DocTopics, and Word-Doc subgraph displays edges between WordTopics and DocTopics. Due to the limitation of the space, we only give analysis for CiteseerX here and interested readers can refer to the public demo for the AAN TopicAtlas.

Word-Word Relation. As shown in Figure 4–2(a), 62.87% of WordTopic nodes have no connection with other WordTopic nodes, which implies that one paper mainly focuses on one WordTopic. This phenomenon agrees with our intuition: most of high quality scientific papers show clear themes.

Though the connection between WordTopics is not strong, there are still a few nodes which link to multiple WordTopics worth investigating. On the basis of previous recognition that the content of documents is generally "pure", we believe that those WordTopics which enjoy high co-occurrence probability with various other WordTopics are foundation of certain scientific fields. In Figure 4–2(a), WordTopic w45 (degree: 9), w44(degree: 6), w16 (degree: 5), and w25 (degree: 5) have the highest degrees. The corresponding WordTopics are "*distributed system*", "*programming language*" , "*software design*", and "*semantic reasoning*". Obviously they are all general and basic. Take "*distributed system*" as an example, distributed system achieves

efficiency improvement of solving computational problems and therefore has broad applications in different fields such as telephone networks, routing algorithms, network file system, etc. As a case study, we show WordTopic w45 "*distributed system*" and its related WordTopics in Figure 4–3(a), from which we can see our word-word edges successfully capture the relation between WordTopics.

Doc-Doc Relation. The DocTopics are closely connected as shown in Figure 4–2(b), which indicates that authors tend to cover multiple DocTopics in the reference list. It is coherent with our intuition since a comprehensive reference section is desirable for most authors. Furthermore, since ubiquitous techniques are likely to be cited in a variety of distinct domains, we expect nodes with high degrees in the Doc-Doc subgraph represent DocTopics about universal principle and method. In Figure 4–2(b), the top four highest-degree nodes are DocTopic d63 (degree: 11), d28 (degree:7), d21 (degree:7), d17 (degree:7) and they represent "*linear system method*", "*logic programming*", "*model checking*" and "*conservation law*" respectively. Unsurprisingly, these DocTopics are basic techniques and laws.

In addition to examining DocTopics from a global perspective, inspecting details of specific DocTopic provides insight into an academic citation network on the document level. The DocTopic allows us to assess topic-aware impact of papers given that the top documents in one DocTopic are generally the most popular and representative ones. In Figure 4–3(b) we list top 5 documents in the most dominant DocTopic d35 and its neighbours d41, d56, d61. We also give the document citation numbers for reference, in which while top ranked papers in our DocTopics generally enjoy high citations, the influence ranking of papers does not exactly follows the order of citation number. Such phenomenon derives from the topic-awareness in our model scheme. Specifically, high citation number of a paper might come from citations from various fields or topics, thus the influence ranking within one *specific topic* is not completely reflected by citation numbers.

Word-Doc Relation. We summarize the contributions of Word-Doc relation from three perspectives. These examples are illustrated in Figure 4–4.

Connect WordTopic and DocTopic reasonably. As Figure 4–4 suggests, the DocTopic d17 is about "*conservation law*", and its neighbouring WordTopics are w54 "*particle phase energy*",

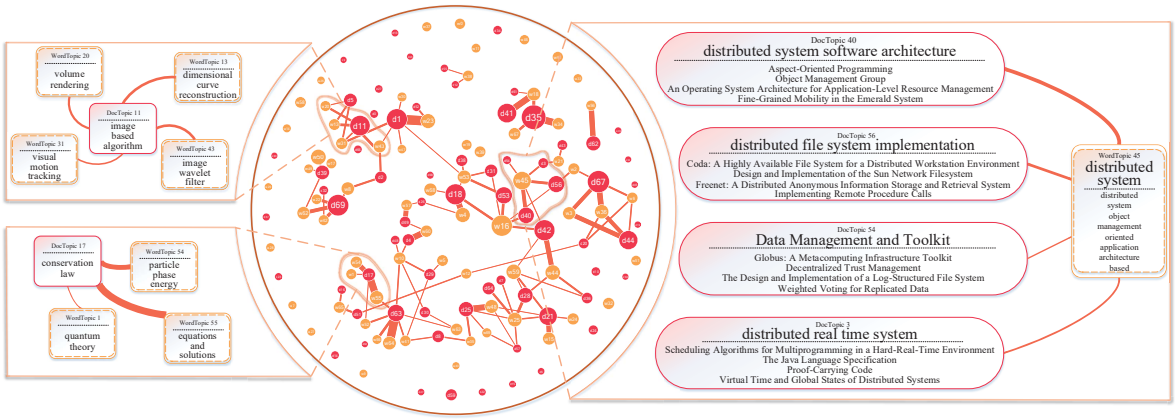


Figure 4–4: Word-Doc Subgraph and some instances. The red nodes represent DocTopics and the orange nodes indicate WordTopics. Only the edges between WordTopics and DocTopics are displayed. Doctopic 11 and Doctopic 17 are expressed by indicative words.

w1 "*quantum theory*" and w55 "*equations and solutions*". These topics cover some basic components of quantum mechanics. In addition, WordTopic w36 is about "*shared memory processor*", and it has a strong link with DocTopic d44 "*shared memory system*" and d67 "*cache performance*". Also, it connects with DocTopic d20 "*power analysis of design*" through a edge weighting about 15 since energy reduction plays an important role in shared memory processor. Besides, WordTopic w57 "*mobile robot navigation*" is connected with DocTopic d49 "*mobile robot localization*" and d26 "*motion planning*". These connections expose the main structure of "*mobile navigation*". There are a lot of other examples in our heterogeneous topic web, readers can check them in our demo TopicAtlas.

Link WordTopics indirectly. The missing co-occurrence phenomenon between WordTopics results in difficulty in spotting relevant WordTopics. However, DocTopics can serve as intermediaries between WordTopics and uncover the hidden relationship. More specifically, if two WordTopics co-occur frequently with the same DocTopic, then we can say the two WordTopics are related. For example, WordTopic w43 "*image wavelet filter*" is connected with WordTopics w13 "*dimensional curve reconstruction*", w20 "*volume rendering*" and w31 "*visual motion tracking*" through DocTopic d11 "*image based algorithm*", which agrees with the fact that many volume rendering and visual motion tracking models are wavelet-based. There are other exam-

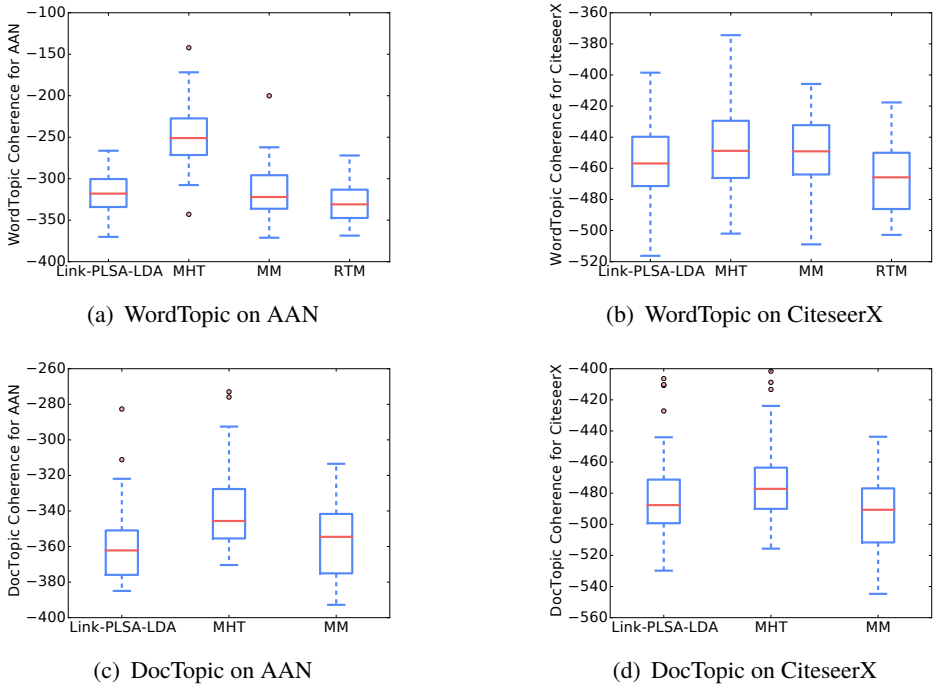
ples: WordTopic w1 "*quantum theory*", w54 "*particle phase energy*" and w55 "*equations and solutions*" are connected through DocTopic d17 "*conservation law*", WordTopic w41 "*random number set*", w64 "*numerical method*", w66 "*matrix factorization*", w52 "*dynamical model simulation*" and w55 "*equations and solutions*" are connected by the general and dominant DocTopic d63 "*linear system algorithm*".

Locate Relevant Documents. Through establishing connection between DocTopics and WordTopics, users can investigate relevant documents for WordTopics. Note that instead of simply recognizing all related documents for WordTopics, TopicAtlas organizes the relevant documents according to DocTopics and allows for inspecting them in different aspects . If a researcher aims to find relevant documents for WordTopic w45 "*distributed system*", he can locate papers about the implementation of distributed file or network system in d56, examine distributed system architecture stuff in d40, get to know some data management or toolkit documents in distributed system from d54, or explore papers about distribution application in real-time system from d3. With the relevant documents sorted, the researcher is less prone to be swamped by the flood of information.

4.1.5 Topic Modeling

Since we aim to obtain effective heterogeneous topic web, it is important to ensure that the introduction of the transition parameter has not come at the expense of the semantic quality of topics and the generalizability of the topic model.

Baselines. We compare our method MHT with mixed-membership model (MM) [13], Link-PLSA-LDA [31] and RTM [16], all of which are joint models for both text and links. Mixed membership model is proposed by Erosheva et al. to classify documents [13]. Nallapati et al. [31] propose two well-known joint topic models Pairwise-Link-LDA and Link-PLSA-LDA. Pairwise-Link-LDA models the presence and absence of links in a pairwise manner while Link-PLSA-LDA views links as "link tokens". Since Link-PLSA-LDA outperforms Pairwise-Link-LDA with respect to heldout likelihood and recall, we only include Link-PLSA-LDA in our baseline methods. The core idea of RTM is that topic relations directly account for the presence of links. To guarantee the justness, all these models are inferred through variational



(a) WordTopic on AAN

(b) WordTopic on CiteseerX

(c) DocTopic on AAN

(d) DocTopic on CiteseerX

Figure 4–5: Topic coherence for WordTopic and DocTopic in two datasets (higher is better).

EM algorithm and parameters are initialized with the same way as MHT.

Topic Interpretability. There are some metrics for evaluating topic interpretability such as *PMI* [60], *word intrusion* [59], and *topic coherence* [57]. We adopt *topic coherence* in our experiment. For one thing, while word intrusion needs expert annotations, topic coherence is an automated evaluation metric and does not rely on human annotators. For another, topic coherence does not reference collections outside the training data as PMI dose. Also, topic coherence is proven more closely associated with the expert annotations than PMI [57].

Letting $D(w)$ be the *document frequency* of word type w and $D(w, w')$ be *co-document frequency* of word types w and w', *topic coherence* is defined as

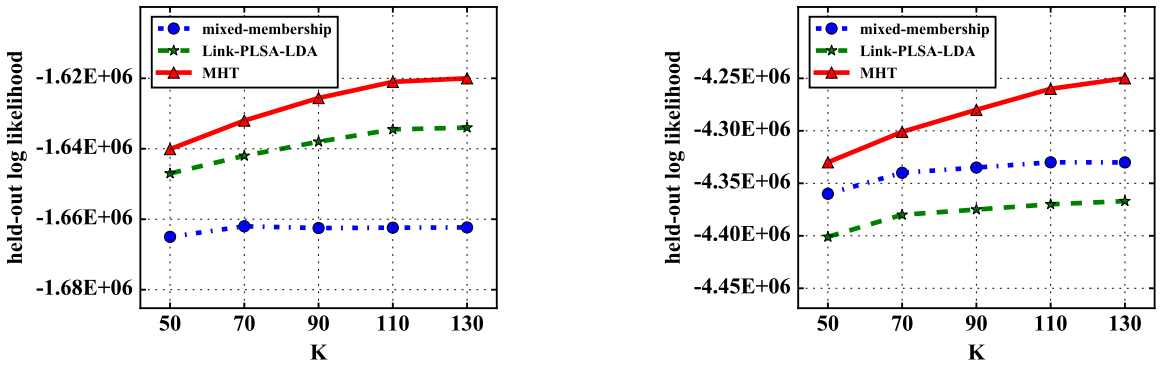
$$C(k; W^{(k)}) = \sum_{m=2}^{M} \sum_{n=1}^{m-1} \log \frac{D(w_m^{(k)}, w_n^{(k)}) + 1}{D(w_n^{(k)})} \tag{4–8}$$

where $W^{(k)} = (w_1^{(k)}, \cdots, w_M^{(k)})$ is a list of the M most probable words in topic k. In our experiment, we choose $M = 10$.

Although it is originally designed for WordTopics, by using the indicative words as keywords, we can also calculate the topic coherence for DocTopics. To distinguish the two different topic coherence score, we denote them as *WordTopic coherence* and *DocTopic coherence*.

We compare the topic coherence score of different methods for all topics, and the averaged results are illustrated in Figure 4–5. As RTM does not produce DocTopics, it is not included in the DocTopic coherence comparison. Obviously, our model achieves superior topic qualities to the baseline methods.

Held-Out Log Likelihood. Held-out Log Likelihood is a well-accepted metric to measure the generalizability and predictive power of topic models. To ease the favor for text and obtain a convincing result, we filter out the documents with less than 3 links and 8 links for AAN and CiteseerX respectively, and get a collection of AAN with $16,350$ documents and CiteseerX with $61,901$ documents.



(a) AAN (b) CiteseerX

Figure 4–6: Held-out log likelihood for both text and links on two datasets. (higher is better)

Our experimental set-up is as follows. We randomly split data into five folds and repeat the experiment for five times, for each time we use one fold for test, four folds for training, and we report the averaged values in Figure 4–6. The performance of MHT is better than the baseline methods. Note that we exclude RTM in this part since held-out log likelihood favors RTM significantly due to its pairwise manner. More Specifically, if links are generated without any training stages and prior knowledge (i.e. links are generated uniformly), the probability for

generating a link in RTM (0.5) is much larger than other models (0.00006 in AAN and 0.00002 in CiteseerX).

4.2 Evaluation of Topic Embedding Model

We demonstrate the efficacy of our approach with extensive experiments. (1) We evaluate the extraction quality in the tasks of document classification and retrieval, in which our model achieves similar or better performance than existing correlated topic models, significantly improving over simple LDA. (2) For scalability, our approach scales comparably with LDA, and handles massive problem sizes orders-of-magnitude larger than previously reported correlation results. (3) Qualitatively, our model reveals very meaningful topic correlation structures.

4.2.1 Setup

Datasets. We use three public corpora provided in the UCI repository[1] for the evaluation: **20Newsgroups** is a collection of news documents partitioned (nearly) evenly across 20 different newsgroups. Each article is associated with a category label, serving as ground truth in the tasks of document classification and retrieval; **NYTimes** is a widely-used large corpus of New York Times news articles; and **PubMed** is a large collection of academic medical paper abstracts. The detailed statistics of the datasets are listed in Table 4–1. We removed a standard list of 174 stop words and performed stemming. For NYTimes and Pubmed, we kept the top 10K frequent words in vocabulary, and selected 10% documents uniformly at random as test sets, respectively. For 20Newsgroups, we followed the standard training/test splitting, and performed the widely-used pre-processing[2] by removing indicative meta text such as headers and footers so that document classification is forced to be based on the semantics of plain text.

Baselines. We compare the proposed model with a set of carefully selected competitors:

- **Latent Dirichlet Allocation (LDA)** [3] uses conjugate Dirichlet priors and thus scales linearly w.r.t the topic size but fails to capture topic correlations. Inference is based on

[1] http://archive.ics.uci.edu/ml

[2] http://scikit-learn.org/stable/datasets/twenty_newsgroups.html

Table 4–1: Statistics of the three datasets, including the number of documents (D), vocabulary size (V), and average number of words in each document.

Dataset	#doc (D)	vocab size (V)	doc length
20Newsgroups	18K	30K	130
NYTimes	1.8M	10K	284
PubMed	8.2M	10K	77

the stochastic variational algorithm [61]. When evaluating scalability, we leverage the same sparsity assumptions as in our model for speeding up.

- **Correlated Topic Model (CTM)** [5] employs standard logistic-normal prior which captures pairwise topic correlations. The model uses stochastic variational inference with $\mathcal{O}(K^3)$ time complexity.
- **Scalable CTM (S-CTM)** [6] developed a scalable sparse Gibbs sampler for CTM inference with time complexity of $\mathcal{O}(K^2)$. Using distributed inference on 40 machines, the method discovers 1K topics from millions of documents, which to our knowledge is the largest automatically learned topic correlation structures so far.

Parameter Setting. Throughout the experiments, we set the embedding dimension to $M = 50$, and sparseness parameters to $K_s = 50$ and $V_s = 100$. We found our modeling quality is robust to these parameters. The hyper-parameters are fixed to $\beta = 1/K, \alpha = 0.1, \rho = 0.1$, and $\tau = 1$.

All experiments were performed on a Linux machine with 24 4.0GHz CPU cores and 128GB RAM. All models are implemented using C/C++, and parallelized whenever possible using the OpenMP library.

4.2.2 Document Classification

We first evaluate the performance of document classification based on the learned document representations. We evaluate on the 20Newsgroups dataset where ground truth class labels are available. We compare our proposed model with LDA and CTM. For LDA and CTM, a multi-class SVM classifier with linear kernel[1] is trained for each of them based on the topic

[1] https://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html

distributions of the training documents, while for the proposed model, the SVM classifier takes the document embedding vectors as input. Generally, more accurate modeling of topic correlations enables better document modeling and representations, resulting in improved document classification accuracy.

Figure 4–7 shows the classification accuracy as the number of topics varies. We see that the proposed model performs best in most of the cases, indicating that our method can discover high-quality latent topics and correlations. Both CTM and our model significantly outperforms LDA which treats latent topics independently, validating the importance of topic correlation for accurate text semantic modeling. Compared to CTM, our method achieves better or competitive accuracy as K varies, which indicates that our model, though orders-of-magnitude faster (as shown in the next), does not sacrifice modeling power compared to the complicated and computationally demanding CTM model.

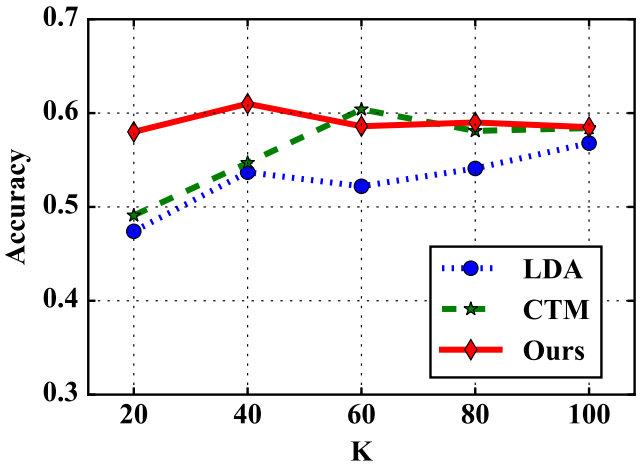


Figure 4–7: Classification accuracy on 20newsgroup.

4.2.3 Document Retrieval

We further evaluate the topic modeling quality by measuring the performance of document retrieval [62]. We use the 20Newsgroups dataset. A retrieved document is relevant to the query document when they have the same class label. For LDA and CTM, document similarity is

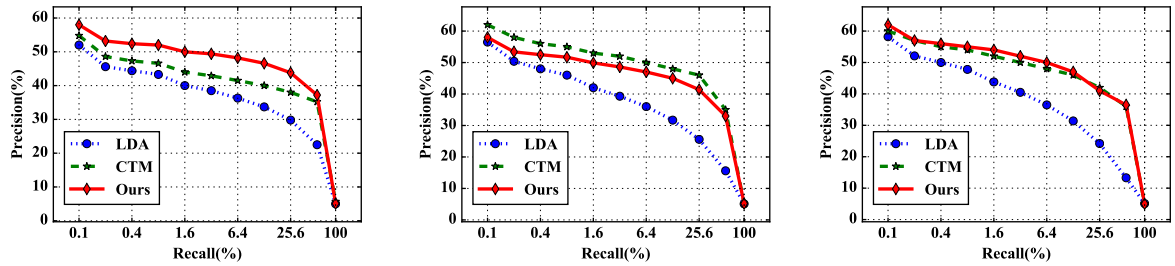


Figure 4–8: Precision-Recall curves on 20Newsgroups. Left: #topic $K = 20$. Middle: $K = 60$. Right: $K = 100$.

measured as the inner product of topic distributions, and for our model we use the inner product of document embedding vectors.

Figure 4–8 shows the retrieval results with varying number of topics, where we use the test set as query documents to retrieve similar documents from the training set, and the results are averaged over all possible queries. We observe similar patterns as in the document classification task. Our model obtains competitive performance with CTM, both of which capture topic correlations and greatly improve over LDA. This again validates our goal that the proposed method has lower modeling complexity while at the same time is as accurate and powerful as previous complicated correlation models. In addition to efficient model inference and learning, our approach based on compact document embedding vectors also enables faster document retrieval compared to conventional topic models which are based on topic distribution vectors (i.e., $M \ll K$).

4.2.4 Scalability

We now investigate the efficiency and scalability of the proposed model. Compared to topic extraction quality in which our model achieves similar or better level of performance as the conventional complicated correlated topic model, here we want our approach to tackle large problem sizes which are impossible for existing correlation methods, and to scale as efficiently as the lightweight LDA, for practical deployment.

Table 4–2 compares the total running time of model training with different sized datasets

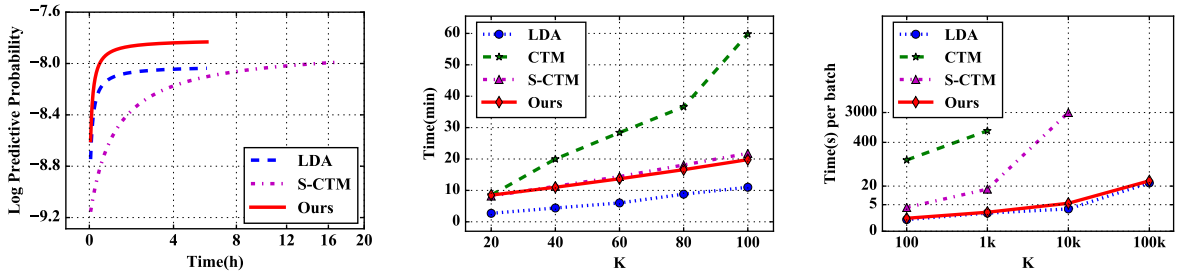


Figure 4–9: Left: Convergence on NYTimes with 1K topics. Middle: Total training time on 20Newsgroups. Right: Runtime of one inference iteration on a minibatch of 500 NYTimes articles, where the result points of CTM and S-CTM on large K are omitted as they fail to finish one iteration within 2 hours.

Table 4–2: Total training time on various datasets with different number of topics K. Entries marked with "–" indicates model training is too slow to be finished in 2 days.

Dataset	K	Running Time			
		LDA	CTM	S-CTM	Ours
20Newsgroups	100	11 min	60 min	22 min	20 min
NYTimes	100	2.5 hr	–	6.4 hr	3.5 hr
	1K	5.6 hr	–	–	5.7 hr
	10K	8.4 hr	–	–	9.2 hr
PubMed	100K	16.7 hr	–	–	19.9 hr

and models. As a common practice [61], we determine convergence of training when the difference between the test set per-word log-likelihoods of two consecutive iterations is smaller than some threshold. On small dataset like 20Newsgroups (thousands of documents) and small model (hundreds of topics), all approaches finish training in a reasonable time. However, with increasing number of documents and latent topics, we see that the vanilla CTM model (with $\mathcal{O}(K^3)$ inference complexity) and its scalable version S-CTM (with $\mathcal{O}(K^2)$ inference complexity) quickly becomes impractical, limiting their deployment in real-world scale tasks. Our proposed topic embedding method, by contrast, scales linearly with the topic size, and is capable of handling 100K topics on over 8M documents (PubMed)—a problem size several orders of magnitude larger than previously reported largest results [6] (1K topics on millions of documents). Notably, even with added model power and increased extraction performance com-

pared to LDA (as has been shown in sections 4.2.2-4.2.3), our model only imposes negligible additional training time, showing strong potential of our method for practical deployment of real-world large-scale applications as LDA does.

Figure 4–9, left panel, shows the convergence curves on NYTimes as training goes. Using similar time, our model converges to a better point (higher test likelihood) than LDA does, while S-CTM is much slower, failing to arrive convergence within the time frame.

Figure 4–9, middle panel, measures the total training time with varying number of topics. We use the small 20Newsgroups dataset since on larger data (e.g., NYTimes and PubMed) the CTM and S-CTM models are usually too slow to converge in a reasonable time. We see that the training time of CTM increases quickly as more topics are used. S-CTM works well in this small data and model scale, but, as have been shown above, it is incapable of tackling larger problems. In contrast, our approach scales as efficiently as the simpler LDA model. Figure 4–9, right panel, evaluates the runtime of one inference iteration on a minibatch of 500 documents. when the topic size grows to a large number, CTM and S-CTM fail to finish one iteration in 2 hours. Our model, by contrast, keeps as scalable as LDA and considerably speeds up over CTM and S-CTM.

4.2.5 Topic Correlation Visualization and Analysis

We qualitatively evaluate our approach by visualizing and exploring the extracted latent topics and correlation patterns.

Figure 4–10 visualizes the topic correlation graph inferred from the 20Newsgroups dataset. We can see many topics are strongly correlated to each other and exhibit clear correlation structure. For instance, the set of topics in the right upper region are mainly about astronomy and are interrelated closely, while their connections to information security topics shown in the lower part are weak. Figure 4–11 shows 100K topic embeddings and their correlations on the PubMed dataset. Related topics are close to each other in the embedding space, revealing diverse substructures of themes in the collection. Our model discovers very meaningful structures, providing insights into the semantics underlying the large text corpora and facilitating understanding of the large collection of topics.

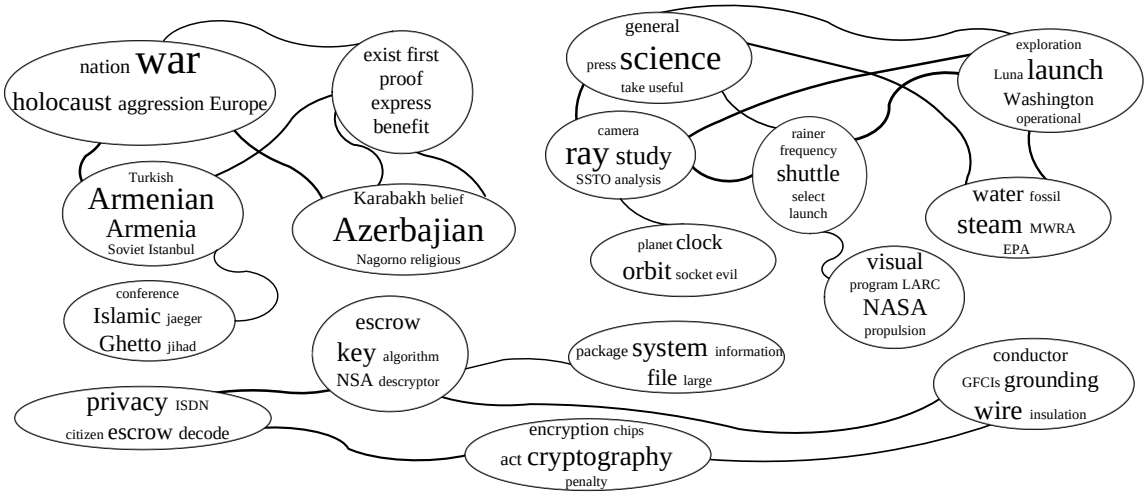


Figure 4–10: A portion of topic correlation graph learned from 20Newsgroups. Each node denotes a latent topic whose semantic meaning is characterized by the top words according to the topic's word distribution. The font size of each word is proportional to the word weight. Topics with correlation strength over some threshold are connected with edges. The thickness of the edges is proportional to the correlation strengths.

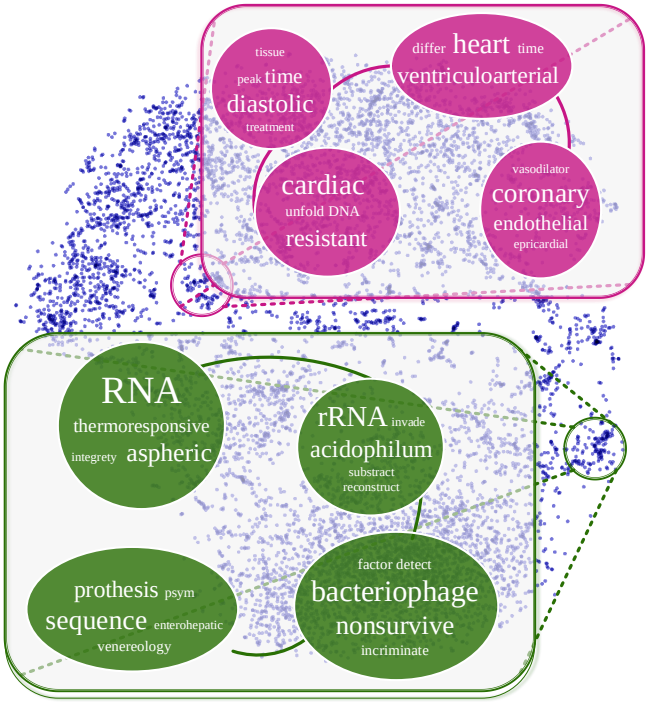


Figure 4–11: Visualization of 100K correlated topics on PubMed. See the captions of Figure 1–2 for more depictions.

Summary

In this thesis, we focus on two problems of existing topic analysis approaches for big scholarly data: (1) Previous topic models lack comprehensive insights for scholarly data on document level; (2) Existing correlated topic models cannot scale up to accommodate industry needs. We propose two different models to address them respectively.

First, we introduce a new type of topic DocTopic, which is distribution over documents, to compensate for the inadequate expressiveness of classical WordTopic. Then we present MHT, short for *Model for Heterogeneous Topic Web*, a unified generative model involving two types of topics. The relationships between the two types of topics, Word-Word relation, Doc-Doc relation and Word-Doc relation, are quantified, based on which we construct the heterogeneous web of topics. In the experiment, we build the heterogeneous topic web of AAN and CiteseerX collection and develop a corresponding prototype demo system, called *TopicAtlas* to exhibit the heterogeneous topic web and assist users' exploration. Qualitative analyses are presented to demonstrate the efficacy of TopicAtlas. Besides, MHT shows good performance as a topic model with respect to topic interpretability and held-out log likelihood.

Second, we have developed a new correlated topic model which induces distributed vector representations of latent topics, and characterizes correlations with the closeness of topic vectors in the embedding space. Such modeling scheme, along with the sparsity-aware sampling in inference, enables highly efficient model training with linear time complexity in terms of the model size. Our approach scales to unprecedentedly large data and models, while achieving strong performance in document classification and retrieval. The proposed correlation method is generally applicable to other context, such as modeling word dependencies for improved topical coherence. We are also interested in further speedup of the model inference, e.g., by incorporating variational neural Bayes techniques [54] for shared global variational updates across data examples.

Bibliography

[1] Gary Marchionini. "Exploratory Search: From Finding to Understanding". In: *Commun. ACM* 49.4 (Apr. 2006), pp. 41–46. ISSN: 0001-0782. DOI: 10.1145/1121949.1121979. URL: http://doi.acm.org/10.1145/1121949.1121979.

[2] Lauren F Klein, Jacob Eisenstein, and Iris Sun. "Exploratory Thematic Analysis for Digitized Archival Collections". In: *Digital Scholarship in the Humanities* (2015), fqv052.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent Dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[4] Rajesh Ranganath and David M Blei. "Correlated random measures". In: *Journal of the American Statistical Association* (2016).

[5] David M Blei and John D Lafferty. "A correlated topic model of science". In: *The Annals of Applied Statistics* (2007), pp. 17–35.

[6] Jianfei Chen et al. "Scalable inference for logistic-normal topic models". In: *Advances in Neural Information Processing Systems*. 2013, pp. 2445–2453.

[7] John Paisley, Chong Wang, David M Blei, et al. "The discrete infinite logistic normal distribution". In: *Bayesian Analysis* 7.4 (2012), pp. 997–1034.

[8] Duangmanee Pew Putthividhya, Hagai T Attias, and Srikantan Nagarajan. "Independent factor topic models". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM. 2009, pp. 833–840.

[9] Amr Ahmed and Eric Xing. "On tight approximate inference of the logistic-normal topic admixture model". In: *Proceedings of the 11th Tenth International Workshop on Artificial Intelligence and Statistics*. 2007.

[10] Jianfei Chen et al. "WarpLDA: a Simple and Efficient O(1) Algorithm for Latent Dirichlet Allocation". In: *International Conference on Very Large Data Bases (VLDB)*. 2016.

[11] Jinhui Yuan et al. "Lightlda: Big topic models on modest computer clusters". In: *Proceedings of the 24th International Conference on World Wide Web*. ACM. 2015, pp. 1351–1361.

[12] Yi Wang et al. "Peacock: Learning long-tail topic features for industrial applications". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 6.4 (2015), p. 47.

[13] Elena Erosheva, Stephen Fienberg, and John Lafferty. "Mixed-membership models of scientific publications". In: *Proceedings of the National Academy of Sciences* 101.suppl 1 (2004), pp. 5220–5227.

[14] Xiaolong Wang, Chengxiang Zhai, and Dan Roth. "Understanding evolution of research themes: a probabilistic generative model for citations". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2013, pp. 1115–1123.

[15] Ramesh M Nallapati et al. "Joint latent topic models for text and citations". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2008, pp. 542–550.

[16] Jonathan Chang and David M Blei. "Relational topic models for document networks". In: *International conference on artificial intelligence and statistics*. 2009, pp. 81–88.

[17] Qi He et al. "Detecting topic evolution in scientific literature: how can citations help?" In: *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM. 2009, pp. 957–966.

[18] Ramesh Nallapati, Daniel A Mcfarland, and Christopher D Manning. "TopicFlow Model: Unsupervised Learning of Topic-specific Influences of Hyperlinked Documents." In: *AISTATS*. 2011, pp. 543–551.

[19] Lilian Weng and Thomas M Lento. "Topic-Based Clusters in Egocentric Networks on Facebook." In: *ICWSM*. 2014.

[20] Chi Wang et al. “Constructing topical hierarchies in heterogeneous information networks”. In: *Knowledge and Information Systems* 44.3 (2015), pp. 529–558.

[21] Shaohua Li et al. “Generative topic embedding: a continuous representation of documents”. In: *Proceedings of The 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2016.

[22] Kayhan Batmanghelich et al. “Nonparametric Spherical Topic Modeling with Word Embeddings”. In: *ACL*. 2016.

[23] Rajarshi Das, Manzil Zaheer, and Chris Dyer. “Gaussian lda for topic models with word embeddings”. In: *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics*. 2015.

[24] Di Jiang et al. “Latent Topic Embedding”. In: *COLING*. 2016.

[25] Allison June-Barlow Chaney and David M Blei. “Visualizing Topic Models.” In: *ICWSM*. 2012.

[26] Brynjar Gretarsson et al. “Topicnets: Visual analysis of large text corpora with topic modeling”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.2 (2012), p. 23.

[27] Arun S Maiya and Robert M Rolfe. “Topic similarity networks: visual analytics for large document sets”. In: *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE. 2014, pp. 364–372.

[28] P Jahnichen et al. “Exploratory Search Through Visual Analysis of Topic Models”. In: *Digital Humanities Quarterly (special issue)* (2015).

[29] Laura Dietz, Steffen Bickel, and Tobias Scheffer. “Unsupervised prediction of citation influences”. In: *Proceedings of the 24th international conference on Machine learning*. ACM. 2007, pp. 233–240.

[30] Qiaozhu Mei et al. “Topic modeling with network regularization”. In: *Proceedings of the 17th international conference on World Wide Web*. ACM. 2008, pp. 101–110.

[31] Ramesh Nallapati and William W Cohen. "Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs." In: *ICWSM*. 2008.

[32] Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. "Topic-link LDA: joint models of topic and author community". In: *proceedings of the 26th annual international conference on machine learning*. ACM. 2009, pp. 665–672.

[33] Tuan Le and Hady W Lauw. "Probabilistic latent document network embedding". In: *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE. 2014, pp. 270–279.

[34] David Cohn and Huan Chang. "Learning to probabilistically identify authoritative documents". In: *ICML*. Citeseer. 2000, pp. 167–174.

[35] David Cohn and Thomas Hofmann. "The missing link-a probabilistic model of document content and hypertext connectivity". In: *Advances in neural information processing systems* (2001), pp. 430–436.

[36] Aaron Q Li et al. "Reducing the sampling complexity of topic models". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 891–900.

[37] Wei Li and Andrew McCallum. "Pachinko allocation: DAG-structured mixture models of topic correlations". In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 577–584.

[38] Saurabh S Kataria et al. "Entity disambiguation with hierarchical topic models". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, pp. 1037–1045.

[39] Jordan L Boyd-Graber, David M Blei, and Xiaojin Zhu. "A Topic Model for Word Sense Disambiguation." In: *EMNLP-CoNLL*. 2007, pp. 1024–1033.

[40] Kumar Dubey et al. "Dependent nonparametric trees for dynamic hierarchical clustering". In: *Advances in Neural Information Processing Systems*. 2014, pp. 1152–1160.

[41] David M Blei, Thomas L Griffiths, and Michael I Jordan. "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies". In: *Journal of the ACM (JACM)* 57.2 (2010), p. 7.

[42] Yarin Gal and Zoubin Ghahramani. "Pitfalls in the use of Parallel Inference for the Dirichlet Process." In: *ICML*. 2014, pp. 208–216.

[43] Zhiting Hu et al. "Large-scale Distributed Dependent Nonparametric Trees." In: *ICML*. 2015, pp. 1651–1659.

[44] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and Their Compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems*. 2013.

[45] Tao Lei et al. "Low-rank tensors for scoring dependency structures". In: Association for Computational Linguistics. 2014.

[46] Aditya Grover and Jure Leskovec. "node2vec: Scalable feature learning for networks". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 855–864.

[47] Jian Tang, Meng Qu, and Qiaozhu Mei. "Pte: Predictive text embedding through large-scale heterogeneous text networks". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, pp. 1165–1174.

[48] Tuan Le and Hady W Lauw. "Semantic visualization for spherical representation". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 1007–1016.

[49] Tuan Minh Van LE and Hady W Lauw. "Manifold learning for jointly modeling topic and visualization". In: (2014).

[50] Michalis K Titsias. "Variational Learning of Inducing Variables in Sparse Gaussian Processes." In: *AISTATS*. Vol. 5. 2009, pp. 567–574.

[51] Andrew G Wilson et al. "Stochastic Variational Deep Kernel Learning". In: *Advances in Neural Information Processing Systems*. 2016, pp. 2586–2594.

[52] David Mimno, Matt Hoffman, and David Blei. "Sparse stochastic inference for latent Dirichlet allocation". In: *arXiv preprint arXiv:1206.6425* (2012).

[53] Miguel Lázaro-Gredilla. "Doubly stochastic variational Bayes for non-conjugate inference". In: ICML. 2014.

[54] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[55] "Supplementary material". In: (2017). URL: www.cs.cmu.edu/~zhitingh/kddsupp.

[56] Dragomir R Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. "The ACL anthology network corpus". In: *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*. Association for Computational Linguistics. 2009, pp. 54–61.

[57] David Mimno et al. "Optimizing semantic coherence in topic models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011, pp. 262–272.

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

[59] Jonathan Chang et al. "Reading tea leaves: How humans interpret topic models". In: *Advances in neural information processing systems*. 2009, pp. 288–296.

[60] David Newman et al. "Automatic evaluation of topic coherence". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2010, pp. 100–108.

[61] Matthew D Hoffman et al. "Stochastic variational inference." In: *Journal of Machine Learning Research* 14.1 (2013), pp. 1303–1347.

[62] Geoffrey E Hinton and Ruslan R Salakhutdinov. "Replicated softmax: an undirected topic model". In: *Advances in neural information processing systems*. 2009, pp. 1607–1614.

Acknowledgement

First, I want to thank Prof. Xinbing Wang for his great help and advising in the past three years. I worked with Prof. Xinbing Wang since my sophomore year, and he provided me with excellent platform to improve myself, where I met a lot of talented and diligent friends. In the third year, I published my first paper under the instruction of Prof. Xinbing Wang. Such experience enables my strong learning ability and comprehensive understanding of probabilistic graphical model, which helps me get into the famous labs of UIUC and CMU in the next year. In addition to academics, Prof. Xinbing Wang often talks with me about his insights on study, work, and life, benefiting me quite a lot and playing a pivot role in the key moment of my life. He keeps supporting my development along the three years with his enthusiasm and advises me to complete the this work on big scholarly data as my bachelor thesis.

Then I feel very fortunate to enjoy the courses of some great professors. Prof. Guonian Shao and Prof. Zhaotai Qiu cultivate my logical thinking ability through Mathematical Analysis course. More importantly, as senior professors, their serious attitude toward mathematics and responsibility for their work, for us, impress me deeply. Besides, Prof. Xiangzhong Fang, Prof. Xinbao Gong, Prof. Min Tang and Prof. Xiaoying Gan are all great professors, they have their education ideal and are making efforts to achieve it. Their enthusiasm about work influences me and lets me know what a good teacher is like.

Next I want to express my gratitude to my excellent peer friends, including but not limited to Changfeng Liu, Hanxiao He, Jiaming Shen, Zhenyu Song, Yuning Mao, Lequn Wang, and etc. They are indispensable in my development along this way, we survive difficulties together, share pleasures together, and are stepping into success together.

Finally, the biggest thanks are going to my family and my girlfriend Luling Han. They are generous to support me as possible as they could, no matter whether I succeed or fail. My graduation, along with this thesis, would credit to them mostly.

Publication

[1] JUNXIAN HE, et al. Efficient Correlated Topic Modeling with Topic Embedding. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2017.

[2] JUNXIAN HE, et al. Text Network Exploration via Heterogeneous Web of Topics. IEEE, International Conference on Data Mining Workshops, 2017.