Dirichlet belief networks for topic structure learning

He Zhao[1], Lan Du[1]* **Wray Buntine[1], and Mingyuan Zhou[2]***
[1]Faculty of Information Technology, Monash University, Australia
[2]McCombs School of Business, The University of Texas at Austin, USA

Abstract

Recently, considerable research effort has been devoted to developing deep architectures for topic models to learn topic structures. Although several deep models have been proposed to learn better topic proportions of documents, how to leverage the benefits of deep structures for learning word distributions of topics has not yet been rigorously studied. Here we propose a new multi-layer generative process on word distributions of topics, where each layer consists of a set of topics and each topic is drawn from a mixture of the topics of the layer above. As the topics in all layers can be directly interpreted by words, the proposed model is able to discover interpretable topic hierarchies. As a self-contained module, our model can be flexibly adapted to different kinds of topic models to improve their modelling accuracy and interpretability. Extensive experiments on text corpora demonstrate the advantages of the proposed model.

1 Introduction

Understanding text has been an important task in machine learning, natural language processing, and data mining. Text is discrete, unstructured, and often highly sparse. A popular way of analysing texts is to represent them as a set of latent factors via topic modelling or matrix factorisation. With great success in modelling text, probabilistic topic models discover a set of latent topics from a collection of documents. Those topics, as latent factors, can be interpreted by distributions over words and used to derive low dimensional representations of the documents. Specifically, most existing topic models are built on top of the following generative process: Each topic is a distribution over the words (i.e., *word distribution*, WD) in the vocabulary; each document is associated with a *topic proportion* (TP) vector; and a word in a document is generated by first drawing a topic according to the document's TP, then sampling the word according to the topic's WD.

In a Bayesian setting, TPs and WDs are both imposed on prior distributions. For example, one commonly-used prior for TP and WD is a Dirichlet distribution, as in Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Recently, deep hierarchical priors, especially imposed on TPs, have been developed to generate hierarchical document representations as well as discover interpretable topic hierarchies. For example, there are hierarchical tree-structured constructions based on the Dirichlet Process (DP) or Chinese Restaurant Process (CRP), such as the nested CRP (nCRP) (Blei et al., 2010) and the nested hierarchical DP (Paisley et al., 2015); deep constructions based on restricted Boltzmann machines and neural networks such as the Replicated Softmax Model (RSM) (Hinton and Salakhutdinov, 2009), the Neural Autoregressive Density Estimator (NADE) (Larochelle and Lauly, 2012), and the Over-replicated Softmax Model (OSM) (Srivastava et al., 2013); models based on variational autoencoders (VAE) including Srivastava and Sutton (2017); Miao et al. (2017); Zhang et al. (2018). Recently, models that generalise the sigmoid belief network (Hinton et al., 2006) have been proposed, such as Deep Poisson Factor Analysis (DPFA) (Gan et al., 2015), Deep Exponential Families (DEF) (Ranganath et al., 2015), Deep Poisson Factor Modelling (DPFM) (Henao et al., 2015), and Gamma Belief Networks (GBNs) (Zhou et al., 2016).

---

*Corresponding authors

Compared with the considerable interest in deep models on TPs, to our knowledge, the counterparts on WDs have not been fully investigated. In this paper, we propose a new multi-layer generative process on WDs, as a self-contained module and an alternative to the single-layer Dirichlet prior. In the proposed model, WDs are the output units of the bottom layer in a DBN with hidden layers parameterised by Dirichlet-distributed hidden units and connected with gamma-distributed weights. Specifically, each Dirichlet unit in a hidden layer is a probability distribution over the words in the vocabulary and can be view as a "hidden" topic. In each layer, the Dirichlet prior of a topic is a mixture of the topics in the layer above. As the hidden units are drawn from Dirichlet, the proposed model is named the Dirichlet Belief Network, hereafter referred to as *DirBN*[2].

Compared with existing related deep models, DirBN has the following appealing properties: **1) Interpretability of hidden units**: Every hidden unit in every layer of DirBN is a probability distribution over the words, making them real topics that can be directly interpreted. **2) Discovering topic hierarchies**: The mixture structure of DirBN enables the model to enjoy a straightforward way of discovering semantic correlations of topics in two adjacent layers, which further form topic hierarchies with the multi-layer construction of the model. Due to the intrinsic abstraction effect of DBN, the topics in the higher layers are more abstract and can be treated as the generalisation of the ones in the lower layers. **3) Better modelling accuracy**: It is known that TPs are local variables (specific to individual document), while WDs are global variables over the target corpus. Unlike many other hierarchical parallels on TP, DirBN imposes a deep structure on WD, which "absorbs the information" from the entire corpus. It makes DirBN be able to get better modelling accuracy especially in the case of sparse texts such as tweets and news abstracts, where the context information of an individual document is not enough to learn a good model using existing approaches. **4) Adaptability**: As many sophisticated models on TPs usually use a simple Dirichlet prior on WDs, including the well-known ones such as Supervised Topic Model (Mcauliffe and Blei, 2008) and Author Topic Model (Rosen-Zvi et al., 2004), our DirBN can be easily adapted to them to further improve modelling accuracy and interpretability.

In conclusion, the contributions of this paper include: **1)** We propose DirBN, a deep structure that can be used as an advanced alternative to the Dirichlet prior on WDs with better modelling performance and interpretability. **2)** We demonstrate our model's adaptability by applying DirBN with several well-developed models, including Poisson Factor Analysis (PFA) (Zhou et al., 2012), MetaLDA (Zhao et al., 2017a), and GBN (Zhou et al., 2016). **3)** With proper data augmentation and marginalisation techniques, DirBN enjoys full local conjugacy, which facilitates the derivation of a simple and effective inference algorithm.

2 The proposed DirBN

In this section, we introduce the details of the generative and inference processes of DirBN.

2.1 Generative process

We first define the essential notation and review the basic framework of topic modelling, followed by the details of the proposed DirBN. Assume that the bag-of-words of document d in a corpus with N documents and V unique words in the vocabulary are stored in a count vector $\boldsymbol{x}_d \in \mathbb{N}_0^V$. A topic model with K topics is composed of the TP vector $\boldsymbol{\theta}_d \in \mathbb{R}_+^K$ for each document d and the WD vector $\boldsymbol{\phi}_k \in \mathbb{R}_+^V$ for each topic k ($k \in \{1, \cdots, K\}$). To generate a word in document d, one can first sample a topic according to its TP, and then sample the word type according to the topic's WD. Given this framework, many prior constructions of TPs have been proposed, such as the Dirichlet distribution in LDA, logistic normal distributions for modelling topic correlations in Correlated Topic Model (CTM) (Lafferty and Blei, 2006), nonparametric priors like the Hierarchical Dirichlet Process (Teh et al., 2012), and recently-proposed deep models like DPFA (Gan et al., 2015), DPFM (Henao et al., 2015), and GBN (Zhou et al., 2016). Unlike the extensive choices for constructing TP, the symmetric Dirichlet distribution on WDs still dominates in many advanced topic models. Here DirBN is a new hierarchical approach of constructing WDs, detailed as follows.

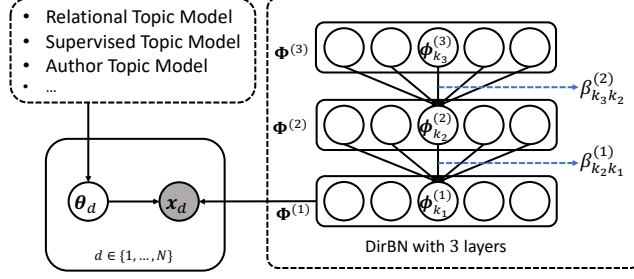[2]Code available at `https://github.com/ethanhezhao/DirBN`

Figure 1: Demonstration of the generative process of DirBN with three layers.

A DirBN with T layers leaves the TPs of the basic framework untouched and draws ϕ_k according to the following generative process:

$$\phi_{k_T}^{(T)} \sim \text{Dir}_V(\eta),$$
$$\cdots$$
$$\phi_{k_t}^{(t)} \sim \text{Dir}_V(\psi_{k_t}^{(t)}), \psi_{k_t}^{(t)} = \sum_{k_{t+1}}^{K_{t+1}} \phi_{k_{t+1}}^{(t+1)} \beta_{k_{t+1}k_t}^{(t)}, \beta_{k_{t+1}k_t}^{(t)} \sim \text{Ga}(\gamma_{k_{t+1}}^{(t)}, 1/c^{(t)}),$$
$$\cdots$$
$$\phi_{k_1}^{(1)} \sim \text{Dir}_V(\psi_{k_1}^{(1)}), \ \psi_{k_1}^{(1)} = \sum_{k_2}^{K_2} \phi_{k_2}^{(2)} \beta_{k_2 k_1}^{(1)}, \beta_{k_2 k_1}^{(1)} \sim \text{Ga}(\gamma_{k_2}^{(1)}, 1/c^{(1)}), \tag{1}$$

where 1) $\text{Ga}(-,-)$ is the gamma distribution with shape and scale parameters and $\text{Dir}_V(-)$ is the Dirichlet distribution[3]; 2) The superscript with a bracket over a variable indicates which layer it belongs to and $k_t \in \{1, \cdots, K_t\}$ is the topic index in the t-th layer; 3) The output of DirBN is $\phi_{k_1}^{(1)}$, which corresponds to ϕ_k in the basic framework and hereafter, we use $\phi_{k_1}^{(1)}$ instead; 4) We further impose gamma priors on the following variables: $\eta \sim \text{Ga}(a_0, 1/b_0)$, $\gamma_{k_{t+1}}^{(t)} \sim \text{Ga}(\gamma_0^{(t)}/K_t, 1/c_0^{(t)})$, $\gamma_0^{(t)} \sim \text{Ga}(e_0, f_0)$, $c_0^{(t)} \sim \text{Ga}(g_0, 1/h_0)$, and $c^{(t)} \sim \text{Ga}(g_0, 1/h_0)$. The generative process of a topic model equipped with DirBN is demonstrated in Figure 1.

The idea of our DirBN can be summarised as follows:

1. From a bottom-up view, DirBN is a multi-layer matrix factorisation, which factorises the matrix of the WDs in the t-th layer as: $\Phi^{(t)} \sim \text{Dir}(\Phi^{(t+1)}\mathbf{B}^{(t)})$. Here we define $\Phi^{(t)} \in \mathbb{R}_+^{V \times K_t}$ ($\phi_{k_t}^{(t)}$ is the k_t-th column) and $\mathbf{B}^{(t)} \in \mathbb{R}_+^{K_{t+1} \times K_t}$ ($\beta_{k_t}^{(t)}$ is the k_t-th column). From a top-down view, the model can be considered as a stochastic feedforward network (Tang and Salakhutdinov, 2013), where the input matrix in $\Phi^{(T)}$, the output matrix is $\Phi^{(1)}$, and the stochastic units are drawn from the Dirichlet distribution.

2. As DirBN is a Bayesian probabilistic model, consider a DirBN with only two layers as an example: each first-layer topic $\phi_{k_1}^{(1)}$ is drawn from a Dirichlet with the topic-specific asymmetric parameter $\psi_{k_1}^{(1)}$, which is a mixture of the second-layer topics. So the statistical strength is shared via the mixture, which plays an important role in handling sparse texts.

3. In DirBN, not only in the bottom layer, but also in any other layer t, each hidden unit is a distribution over the vocabulary and can be viewed as *real topic* directly interpreted by words. Although the bottom layer serves as the actual WDs for generating the words, the topics in the higher layers are involved with the belief prorogation in the network.

4. The weight $\beta_{k_{t+1}k_t}^{(t)}$ is drawn from a hierarchical gamma prior (i.e., the shape parameter $\gamma_{k_{t+1}}^{(t)}$ of the gamma prior on $\beta_{k_{t+1}k_t}^{(t)}$ is also drawn from a gamma). It allows topics in the $(t + 1)$-th layer to contribute differently to those in the t-th layer. In addition, the hierarchical structure on $\beta_{k_{t+1}k_t}^{(t)}$ is similar to the one in Zhou (2015), which provides an

---

[3] $-$ can be a vector as a set of asymmetric parameters or a scalar as a symmetric parameter of Dirichlet

intrinsic shrinkage mechanism on $\boldsymbol{\beta}_{k_t}^{(t)}$. In other words, each k_t is expected to be sparsely connected by a subset of k_{t+1}. We will demonstrate the shrinkage effect of DirBN in the experiments.

2.2 Inference process

The learning of DirBN can be done by the inference of its latent variables, i.e., $\boldsymbol{\Phi}^{(t)}$ and $\mathbf{B}^{(t)}$ for all t. With several data augmentation techniques, we are able to derive a layer-wise Gibbs sampling algorithm facilitated by local conjugacy. Given θ and ϕ (despite their constructions), a topic model usually samples the topic assignment of each word in the corpus. After that, each topic k_1 is associated with a vector of word counts, denoted as $\boldsymbol{x}_{k_1}^{(1)} = [x_{1k_1}^{(1)}, \cdots, x_{Vk_1}^{(1)}]$, which encodes the semantic information of topic k_1 and is one of the *input count vectors* of DirBN in the inference process. Given the input vectors, the inference of DirBN involves two key steps: **1)** propagating the semantic information of the input vectors up to the top layer *via* latent counts; **2)** updating $\boldsymbol{\Phi}^{(t)}$ and $\mathbf{B}^{(t)}$ down to the bottom given the latent counts. Without loss of generality, we illustrate the inference details with a two-layer DirBN as follows[4]:

Propagating the latent counts from the bottom up By integrating $\phi_{k_1}^{(1)}$ out from its multinomial likelihood, we can get the likelihood of $\boldsymbol{\psi}_{k_1}^{(1)}$ as:

$$\mathcal{L}\left(\boldsymbol{\psi}_{k_1}^{(1)}\right) \propto \frac{\Gamma(\psi_{\cdot k_1}^{(1)})}{\Gamma(\psi_{\cdot k_1}^{(1)} + x_{\cdot k_1}^{(1)})} \prod_v^V \frac{\Gamma(\psi_{vk_1}^{(1)} + x_{vk_1}^{(1)})}{\Gamma(\psi_{vk_1}^{(1)})}, \tag{2}$$

where $\Gamma(-)$ is the gamma function, $\psi_{\cdot k_1}^{(1)} = \sum_v^V \psi_{vk_1}^{(1)}$, and $x_{\cdot k_1}^{(1)} = \sum_v^V x_{vk_1}^{(1)}$. By integrating $\phi_{k_1}^{(1)}$ out and introducing two auxiliary variables $q_{k_1}^{(1)}$ and $y_{vk_1}^{(1)}$, Eq. (2) can be augmented as (Zhao et al., 2017a):

$$\mathcal{L}\left(\boldsymbol{\psi}_{k_1}^{(1)}, q_{k_1}^{(1)}, y_{vk_1}^{(1)}\right) \propto \prod_v^V \left(q_{k_1}^{(1)}\right)^{\psi_{vk_1}^{(1)}} \left(\psi_{vk_1}^{(1)}\right)^{y_{vk_1}^{(1)}}, \tag{3}$$

where $q_{k_1}^{(1)} \sim \text{Beta}(\psi_{\cdot k_1}^{(1)}, x_{\cdot k_1}^{(1)})$ and $y_{vk_1}^{(1)} \sim \text{CRT}\left(x_{vk_1}^{(1)}, \psi_{vk_1}^{(1)}\right)$. Here CRT stands for the Chinese Restaurant Table distribution (Zhou and Carin, 2015; Zhao et al., 2017b). Now we can define $\boldsymbol{y}_{k_1}^{(1)} = [y_{1k_1}^{(1)}, \cdots, y_{Vk_1}^{(1)}]$, the *latent count vector* derived from the input count vector $\boldsymbol{x}_{k_1}^{(1)}$.

With $\psi_{vk_1}^{(1)} = \sum_{k_2}^{K_2} \phi_{vk_2}^{(2)} \beta_{k_2 k_1}^{(1)}$, we can then distribute the latent count $y_{vk_1}^{(1)}$ on $\psi_{vk_1}^{(1)}$ to each second layer topic k_2 by:

$$\left(z_{v1k_1}^{(1)}, \cdots, z_{vK_2 k_1}^{(1)}\right) \sim \text{Mult}\left(y_{vk_1}^{(1)}, \frac{\phi_{v1}^{(2)} \beta_{1k_1}^{(1)}}{\psi_{vk_1}^{(1)}}, \cdots, \frac{\phi_{vK_2}^{(2)} \beta_{K_2 k_1}^{(1)}}{\psi_{vk_1}^{(1)}}\right), \tag{4}$$

where $z_{vk_2 k_1}^{(1)}$ is the latent count allocated to k_2 and $\sum_{k_2}^{K_2} z_{vk_2 k_1}^{(1)} = y_{vk_1}^{(1)}$.

We now note $\boldsymbol{x}_{k_2}^{(2)} = [x_{1k_2}^{(2)}, \cdots, x_{Vk_2}^{(2)}]$ where $x_{vk_2}^{(2)} = \sum_{k_1}^{K_1} z_{vk_2 k_1}^{(1)}$. $\boldsymbol{x}_{k_2}^{(2)}$ can be viewed as one of the *output count vectors* of the first layer and also the input count vector of the second layer topic k_2.

In conclusion, to propagate the semantic information from the first to the second layer, we fist derive $y_{vk_1}^{(1)}$ from $x_{vk_1}^{(1)}$, then distribute $y_{vk_1}^{(1)}$ to all the second layer topics (i.e., $z_{vk_2 k_1}^{(1)}$), and finally aggregate $z_{vk_2 k_1}^{(1)}$ into $x_{vk_2}^{(2)}$.

Updating the latent variables from the top down After the latent counts are propagated, we start updating the latent variables from the top layer (i.e. the second layer here). Given $\boldsymbol{x}_{k_2}^{(2)}$, $\boldsymbol{\phi}_{k_2}^{(2)}$ is easy to sample from its Dirichlet posterior. With $z_{vk_2 k_1}^{(1)}$ and $\sum_v^V \phi_{k_2,v}^{(2)} = 1$, we can sample $\beta_{k_2 k_1}^{(1)}$ from its gamma posterior given the following likelihood:

$$\mathcal{L}\left(\beta_{k_2 k_1}^{(1)}\right) \propto e^{-\beta_{k_2 k_1}^{(1)}(-\log q_{k_1}^{(1)})}(\beta_{k_2 k_1}^{(1)})^{z_{\cdot k_2 k_1}^{(1)}}, \tag{5}$$

---

[4] Omitted details of inference as well as the overall algorithm are given in the supplementary materials.

where $z^{(1)}_{\cdot k_2 k_1} = \sum_v^V z^{(1)}_{v k_2 k_1}$. Given the newly sampled $\phi^{(2)}_{k_2}$ and $\beta^{(1)}_{k_2 k_1}$, we can recompute $\psi^{(1)}_{k_1}$ and sample $\phi^{(1)}_{k_1}$ from its Dirichlet posterior. Now the inference of a two-layer DirBN is done.

3 Using DirBN in topic modelling

DirBN is a self-contained module on ϕ, leaving θ untouched. Therefore, it can be used as an alternative to the simple Dirichlet prior on ϕ in many existing models. The adaptability of DirBN enables us to easily apply it to advanced models so that those models can benefit from the advantages of DirBN. To demonstrate this, we adapt the proposed DirBN structure to the following models:

PFA+DirBN Poisson Factor Analysis (PFA) is a popular framework for topic analysis (DPFA (Gan et al., 2015), DPFM (Henao et al., 2015), GBN (Zhou et al., 2016) can be viewed as a deep extension to PFA). Specifically, we use the Bayesian nonparametric version of PFA named BGGPFA (Zhou et al., 2012), where θ_d is constructed from a negative binomial process and ϕ_k is drawn from a Dirichlet distribution. Note that there are close relationships between PFA and LDA, and between BGGPFA and HDP (Teh et al., 2012), analysed in Zhou (2018). Here we replace the Dirichlet construction on ϕ with DirBN, yielding a model named PFA+DirBN.

MetaLDA+DirBN MetaLDA (Zhao et al., 2017a, 2018a) is a supervised topic model that is able to incorporate document labels to inform the learning of θ_d. Keeping the structure on θ untouched, we replace the MetaLDA's structure on ϕ with our DirBN to get a combined model that discovers the topic hierarchies informed by the document labels. The proposed model is able to discover the correlations between labels and topic hierarchies.

GBN+DirBN Recall that GBN (Zhou et al., 2015, 2016) imposes a hierarchical structure on θ, which is able to learn multi-layer document representations and topic hierarchies. Here we combine DirBN and GBN together to yield a "dual" deep model, where the GBN part is on θ and the DirBN part is on ϕ. Both parts discover topic hierarchies and the bottom-layer topics are shared by the two parts/hierarchies. It would be interesting to see how the two deep structures interact with each other.

4 Related work

As the proposed model introduces a hierarchical architecture on WDs (i.e., ϕ) in topic models, we first review various priors on ϕ, starting with the ones on sampling/optimising the Dirichlet parameters in topic models. The Dirichlet parameters in topic models were studied comprehensively in Wallach et al. (2009), which showed that Dirichlet with a symmetric parameter sampled from an uninformative gamma is the best choice. Actually, our DirBN can be reduced to this choice if $T = 1$ (i.e., DirBN-1, with one layer only). However, unlike the sampling/optimising approaches used in Wallach et al. (2009), DirBN-1 uses a negative binomial augmentation shown in Eq. (3), which leads to a simpler inference scheme. Recently, models like Zhao et al. (2017a,c, 2018b) construct informative and asymmetric Dirichlet priors by taking into account some external knowledge like word embeddings. Whereas DirBN learns the asymmetric priors purely based on the context of the target corpus.

Instead of Dirichlet, the Pitman-Yor process (PYP) has been used on WDs to model the power-law distribution of words, as in Sato and Nakagawa (2010); Buntine and Mishra (2014). Chen et al. (2015) used a transformed PYP prior on ϕ to model multiple document collections. Lindsey et al. (2012) imposed a hierarchical PYP prior on ϕ to discover word phrases. Besides PYP, the Indian Buffet Process (IBP) has been used as a prior on ϕ to introduce word focusing on topics, as in Archambeau et al. (2015). In general, existing models use different priors on ϕ for modelling various linguistic phenomena, which have different purposes to DirBN. The deep structures induced by DirBN on WDs have not yet been rigorously studied.

To our knowledge, most existing models explore the structure of topics by imposing a deep/hierarchical prior on θ. For example, hierarchical PYPs were used for domain adaptation in language models (Wood and Teh, 2009) and topic models (Du et al., 2012). nCRP (Blei et al., 2010) models topic hierarchies by introducing a tree-structured prior. Paisley et al. (2015); Kim et al. (2012); Ahmed et al. (2013) extended nCRP by either softening its constraints or applying it to different problems. Li and McCallum (2006) proposed the Pachinko Allocation model (PAM), which
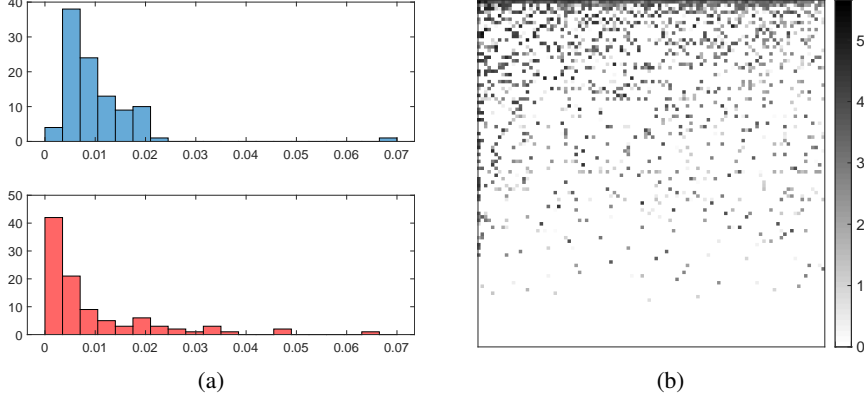
Figure 2: (a): Histograms of the normalised (latent) words counts. (b): $\mathbf{B}^{(1)}$.

captures the topic correlations with a directed acyclic graph. Recently, several deep extensions of PFA on θ have been proposed, including DPFA (Gan et al., 2015), DPFM (Henao et al., 2015), and GBN (Zhou et al., 2016). DPFM and GBN are the most related models to ours, which are also able to discover topic hierarchies. In DPFM and GBN, the higher-layer topics are not distributions over words but distributions over the topics in the layer below (they are called "meta-topics" in DPFM). To interpret those meta-topics, one needs to project them all the way down to the bottom-layer topics with matrix multiplication. Whereas in our model, the topics on all the layers are directly interpretable.

5 Experiments

The experiments were conducted on three real-world datasets, detailed as follows: **1)** Web Snippets (WS), containing 12,237 web search snippets labelled with 8 categories. The vocabulary contains 10,052 word types. **2)** Tag My News (TMN), consisting of 32,597 RSS news labelled with 7 categories. Each document contains a title and a description. There are 13,370 word types in the vocabulary. **3)** Twitter, extracted in 2011 and 2012 microblog tracks at Text REtrieval Conference (TREC)[5]. It has 11,109 tweets in total. The vocabulary size is 6,344.

With the framework of PFA, we compared three options of constructing ϕ: (1) The default setting of PFA, where ϕ is drawn from a symmetric Dirichlet distribution with parameter 0.05, i.e., $\phi_k \sim \text{Dir}_V(0.05)$; (2) PFA+Mallet, where $\phi_k \sim \text{Dir}_V(\alpha_0)$ and α_0 is sampled by Mallet [6]; (3) PFA+DirBN, the proposed model, where ϕ_k is drawn from an asymmetric Dirichlet distribution specific to k, the parameter of which is constructed with the higher-layer topics. Note that Wallach et al. (2009) tested the option using specific asymmetric Dirichlet parameter, i.e., $\phi_k \sim \text{Dir}_V([\alpha_1, \cdots, \alpha_V])$, but the performance is not as good as the symmetric parameter (the second one above). In addition, following a similar routine, we compared MetaLDA (Zhao et al., 2017a), and GBN (Zhou et al., 2016) with/without DirBN. Note that PFA is a widely used Bayesian topic model, MetaLDA is the state-of-the-art topic model capable of handling sparse texts, and GBN is reported Cong et al. (2017) to outperform many other deep models including DPFA (Gan et al., 2015), DPFM (Henao et al., 2015), nHDP (Paisley et al., 2015), and RSM (Hinton and Salakhutdinov, 2009).

For all the models, we ran 3,000 MCMC iterations with 1,500 burnin. For DirBN, we set $a_0 = b_0 = g_0 = h_0 = 1.0$ and $e_0 = f_0 = 0.01$. For PFA, MetaLDA, and GBN, we used their original implementations and settings, except that ϕ is drawn from DirBN in the combined models. For all the models, the number of topics in each layer of DirBN was set to 100, i.e., $K_T = \cdots = K_1 = 100$. For GBN and GBN+DirBN, we set the number of topics in each layer of GBN to 100 as well. Due to the shrinkage mechanisms of PFA, GBN, and DirBN, the number of active topics will be adjusted according to the data. In all the experiments, we varied the number of layers of DirBN T from 1 to 3. For GBN+DirBN, the dual deep model, we fixed the number of layers of GBN as 3.

---

[5]http://trec.nist.gov/data/microblog.html
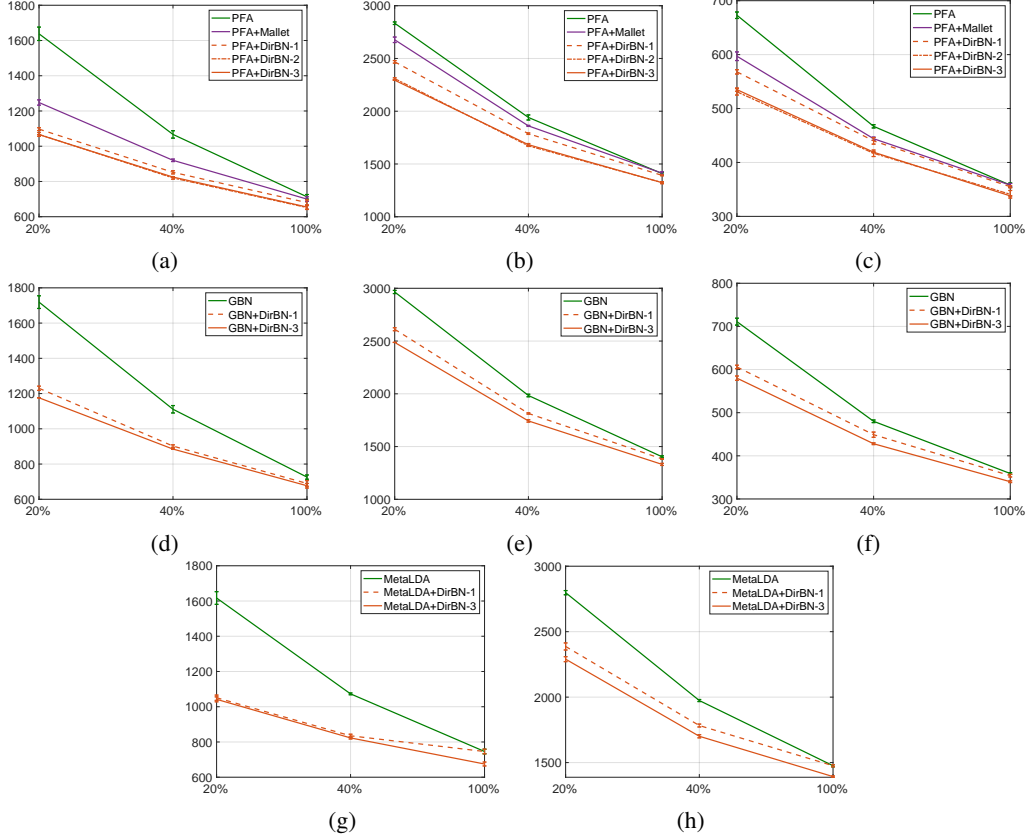[6]http://mallet.cs.umass.edu

Figure 3: Perplexity (the vertical axis) with varied proportion (the horizontal axis) of the words for training in the training documents. (a-c): Results of the models based on PFA on WS, TMN, Twitter. (d-f): Results of the models based on GBN on WS, TMN, Twitter. (g,h): Results of the models based on MetaLDA on WS and TMN. The errorbars indicate the standard deviations of five runs. The number of a model indicates the number of layers used in DirBN. The results of MetaLDA and document classification on Twitter are not reported due to the unavailability of labels.

Demonstration of DirBN's shrinkage effect As previously discussed, DirBN has an intrinsic shrinkage mechanism that is able to automatically learn the number of active topics in each layer (i.e., the network width). We empirically demonstrate the shrinkage effect in Figure 2, with the results of PFA+DirBN-3 on the TMN dataset. Figure 2a plots the histograms of the normalised (latent) words counts $\sum_v x_{vk_t}^{(t)} / \sum_{vk_t} x_{vk_t}^{(t)}$ for all k_t where $x_{vk_t}^{(t)}$ is the word count for topic k_t. The blue and red bars are for the first- ($t = 1$) and the second-layer ($t = 2$) topics, respectively. The histogram indicates the number of topics (the vertical axis) that are with a specific word count (the horizontal axis). A topic with a larger word count is more important. The shrinkage effect is that large proportion of the topics are with very small word counts, indicating that the number of effective topics is less than the truncation (i.e., $K_t = 100$). This is more obvious, in the second layer. Moreover, we display $\log \mathbf{B}^{(1)}$ as an image in Figure 2b. The vertical and horizontal axes are for the second- and first-layer topics, respectively. We ranked the first- and second-layer topics by their word counts. The sparsity of $\mathbf{B}^{(1)}$ indicates that the first- and second-layer topics are sparsely connected. This also demonstrates the shrinkage effect of the model.

Quantitative results We report the per-heldout-word perplexity and topic coherence results. To compute perplexity, we randomly selected 80% of the documents in each dataset to train the models and 20% for testing. For each testing document, we randomly used one half of its words to infer its TP, and the other half to calculate perplexity. Topic coherence measures the semantic coherence in the most significant words (top words) of a topic. Here we used the Normalized Pointwise Mutual Information (NPMI) (Aletras and Stevenson, 2013; Lau et al., 2014) to calculate topic coherence

Table 1: Topic coherence with varied proportion of the words for training in the training documents. $\pm$ indicates the standard deviation of five runs. The best result in each column is in boldface.

Training words	WS			TMN			Twitter		
	20%	40%	100%	20%	40%	100%	20%	40%	100%
PFA	-0.070±0.010	0.008±0.002	0.062±0.011	-0.059±0.008	0.064±0.009	0.103±0.006	-0.003±0.003	0.031±0.003	0.046±0.002
PFA+Mallet	0.008±0.004	0.049±0.005	0.063±0.003	0.035±0.006	0.083±0.005	0.108±0.005	0.022±0.003	0.037±0.002	0.045±0.003
PFA+DirBN-1	0.013±0.003	0.052±0.004	0.060±0.006	0.031±0.003	0.080±0.001	0.108±0.008	0.019±0.004	0.037±0.004	0.049±0.007
PFA+DirBN-3	**0.021**±0.005	0.059±0.002	0.068±0.004	0.046±0.003	0.090±0.003	0.111±0.004	0.024±0.001	0.038±0.002	0.049±0.002
GBN	-0.072±0.013	0.007±0.005	0.069±0.009	-0.065±0.008	0.063±0.006	0.106±0.004	-0.005±0.003	0.032±0.002	0.047±0.00
GBN+DirBN-1	0.015±0.005	0.057±0.002	0.069±0.005	0.032±0.002	0.086±0.002	0.112±0.007	0.021±0.004	0.040±0.005	0.050±0.005
GBN+DirBN-3	0.018±0.006	**0.061**±0.004	**0.075**±0.002	0.048±0.003	**0.094**±0.004	**0.113**±0.004	**0.025**±0.003	**0.040**±0.002	**0.051**±0.003

Table 2: Topic hierarchy comparison in GBN+DirBN. Each row in boldface is the top 10 words in a first-layer topic. Each of these topics is associated with three most correlated topics in the second layer of DirBN (left) and GBN (right), respectively. The number associated with a second-layer topic is its (normalised) link weight to the first-layer topic.

police arrested man charged woman authorities death year found accused			
0.13	case charges accused trial courtattorney investigation judge allegations criminal	0.38	police arrested man charged year accused found charges woman death
0.13	police official killing attack deaddeath army security man family	0.19	police prison man china years arrested charges charged year chinese
0.11	woman men drug suicide girl sexual death found human york	0.15	china police chinese bomb fire people blast city artist officials
heat miami james lebron game nba finals celtics bulls wade			
0.43	season team game play run night star series fans career	0.97	heat miami james game nba finals lebron bulls mavericks dallas
0.15	nba playoffs court brink seeds defeated berth seed opponent semifinals	0.00	trial rajaratnam insider trading fund hedge raj anthony galleon case
0.10	win victory beat lead winning top fourth loss straight beating	0.00	music album lady gaga justin star pop band rock tour
facebook google internet social twitter online web media site search			
0.18	phone plan video technology mobile devices computer tech ceo content	0.22	facebook social internet google online twitter chief executive media web
0.14	company million buy billion corp industry sales companies consumers products	0.19	court lawsuit case facebook judge social federal internet google online
0.12	government report country nation pressure official state move released public	0.18	facebook social internet google online twitter world web media site
study cancer drug risk heart patients women researchers disease people			
0.12	rising percent high higher economic increase low growth strong recovery	0.91	study cancer drug risk researchers heart people patients health women
0.12	reactions periods technique method declared important realized treatment peril scores	0.04	world war years family oil dies year energy women american
0.10	study experience finding recent security kids challenges millions report special	0.18	facebook social internet google online twitter world web media site
nuclear japan plant power radiation crisis japanese fukushima crippled tokyo			
0.17	government united states officials state report country group official agency	0.53	nuclear japan plant power radiation crisis japanese fukushima earthquake tokyo
0.14	safety water nearby land found caused sea believed center parts	0.44	nuclear japan plant power radiation crisis japanese fukushima water tokyo
0.13	work plans part future system rules program bring offers decision	0.01	theater review broadway play york musical stage life time love

score from the top 10 words of each topic and reported scores averaged over top 50 topics with highest NPMI, where "rubbish" topics are eliminated, following Yang et al. (2015)[7]. In the training documents, we further varied the proportion of the words used in training to mimic the case of sparse texts. All the models ran five times with different random seeds and we reported the averaged value with standard deviations.

The results of perplexity and topic coherence are shown in Figure 3 and Table 1, respectively. We have the following remarks on the results: **(1)** In general, for the models with DirBN, the performance is significantly improved compared with the counterparts without DirBN, especially in terms of perplexity and topic coherence and with low proportion of the training words. **(2)** In terms of all the measures, DirBN-2/3 always has better results than DirBN-1. Whereas if we compare GBN with PFA, its perplexity is worse than PFA's, especially for sparse texts. This demonstrates that hierarchical structures on θ (i.e., GBN) may not perform as well as hierarchical structures on ϕ (i.e., DirBN) on sparse texts. **(3)** Although PFA+DirBN-1 and PFA+Mallet both impose a symmetric Dirichlet on ϕ, the former usually has better perplexity. **(4)** The dual deep model (GBN+DirBN-3) usually performs the best on topic coherence, which demonstrates the benefits of the deep structures.

Qualitative analysis on topic hierarchies [8] GBN+DirBN is a dual deep model that discovers two sets of hierarchies, one induced by GBN on θ and the other induced by DirBN on ϕ. The topics in the

---

[7]We used the Palmetto package (`http://palmetto.aksw.org`) with a large Wikipedia dump.

[8]More visualisations of topic hierarchies are shown in the supplementary material.
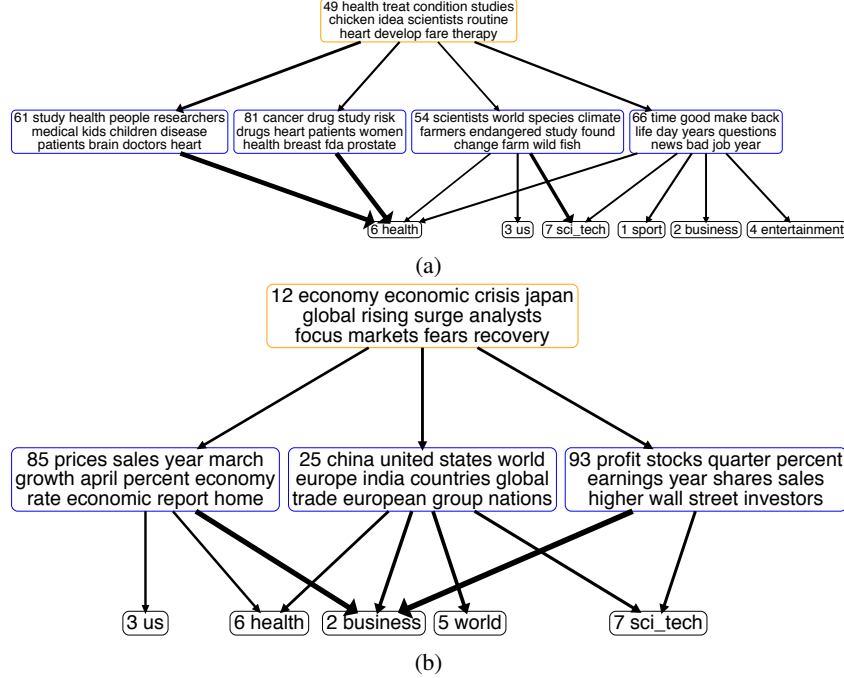
Figure 4: Topic hierarchies discovered by MetaLDA+DirBN. The topics in the yellow and blue rectangles are the second and first layer topics in DirBN and the correlated labels to the first-layer topics are shown at the bottom of each figure. Thicker arrows indicate stronger correlations.

first layer of DirBN connect the two sets of hierarchies. In Table 2, we show the first-layer topics and the correlated second-layer topics in the two hierarchies. It is interesting to see that the second-layer topics of DirBN are more abstract. For example, the second topic is about teams and player in NBA, while its correlated second-layer topics are more general words for sports. Moreover, DirBN is able to discover layer-wise semantically meaningful topic correlations with fewer overlapping top words. This is because GBN combines the words in the first-layer topics to form the second-layer topics, whereas DirBN decomposes the first-layer topics into the second-layer ones.

In MetaLDA+DirBN, the MetaLDA part is able to use document labels to construct TPs (Zhao et al., 2017a), by learning a correlation matrix between the labels and topics, while the DirBN part learns the topic hierarchy. The first-layer topics of DirBN link the correlation matrix and the topic hierarchy together. Figure 4 shows the sample linkages between topic hierarchies and labels on TMN, where the documents are labelled with 7 categories: 1 sport, 2 business, 3 us, 4 entertainment, 5 world, 6 health, 7 sci-tech. One can observe that there is a well correspondence between the topic hierarchies and the labels.

6 Conclusions

We have presented DirBN, a multi-layer process generating word distributions of topics. With real topics in each layer, DirBN is able to discover interpretable topic hierarchies. As a flexible module, DirBN can be adapted to other advanced topic models and improve the performance and interpretability, especially on sparse texts. We have demonstrated DirBN's advantages by equipping PFA, MetaLDA, and GBN, with DirBN. With the help of data augmentation, the inference of DirBN can be done by a layer-wise Gibbs sampling, as a full conjugate model.

Future directions include deriving alternative inference algorithms, such as variational inference (Hoffman et al., 2013), conditional density filtering (Guhaniyogi et al., 2018), and stochastic gradient-based approaches (Chen et al., 2014; Ding et al., 2014; Welling and Teh, 2011).

9

References

D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *Journal of the ACM (JACM)*, vol. 57, no. 2, p. 7, 2010.

J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan, "Nested hierarchical Dirichlet processes," *TPAMI*, vol. 37, no. 2, pp. 256–270, 2015.

G. E. Hinton and R. R. Salakhutdinov, "Replicated softmax: An undirected topic model," in *NIPS*, 2009, pp. 1607–1614.

H. Larochelle and S. Lauly, "A neural autoregressive topic model," in *NIPS*, 2012, pp. 2708–2716.

N. Srivastava, R. Salakhutdinov, and G. Hinton, "Modeling documents with a deep Boltzmann machine," in *UAI*, 2013, pp. 616–624.

A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," 2017.

Y. Miao, E. Grefenstette, and P. Blunsom, "Discovering discrete latent topics with neural variational inference," in *ICML*, 2017, pp. 2410–2419.

H. Zhang, B. Chen, D. Guo, and M. Zhou, "WHAI: Weibull hybrid autoencoding inference for deep topic modeling," 2018.

G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

Z. Gan, C. Chen, R. Henao, D. Carlson, and L. Carin, "Scalable deep Poisson factor analysis for topic modeling," in *ICML*, 2015, pp. 1823–1832.

R. Ranganath, L. Tang, L. Charlin, and D. Blei, "Deep exponential families," in *AISTATS*, 2015, pp. 762–771.

R. Henao, Z. Gan, J. Lu, and L. Carin, "Deep Poisson factor modeling," in *NIPS*, 2015, pp. 2800–2808.

M. Zhou, Y. Cong, and B. Chen, "Augmentable gamma belief networks," *JMLR*, vol. 17, no. 163, pp. 1–44, 2016.

J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *NIPS*, 2008, pp. 121–128.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *UAI*, 2004, pp. 487–494.

M. Zhou, L. Hannah, D. B. Dunson, and L. Carin, "Beta-negative binomial process and Poisson factor analysis," in *AISTATS*, 2012, pp. 1462–1471.

H. Zhao, L. Du, W. Buntine, and G. Liu, "Metalda: A topic model that efficiently incorporates meta information," in *ICDM*, 2017, pp. 635–644.

J. D. Lafferty and D. M. Blei, "Correlated topic models," in *NIPS*, 2006, pp. 147–154.

Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2012.

Y. Tang and R. R. Salakhutdinov, "Learning stochastic feedforward neural networks," in *NIPS*, 2013, pp. 530–538.

M. Zhou, "Infinite edge partition models for overlapping community detection and link prediction," in *AISTATS*, 2015, pp. 1135—-1143.

M. Zhou and L. Carin, "Negative binomial process count and mixture modeling," *TPAMI*, vol. 37, no. 2, pp. 307–320, 2015.

H. Zhao, L. Du, and W. Buntine, "Leveraging node attributes for incomplete relational data," in *ICML*, 2017, pp. 4072–4081.

M. Zhou, "Nonparametric Bayesian negative binomial factor analysis," *Bayesian Analysis*, 2018.

H. Zhao, L. Du, W. Buntine, and G. Liu, "Leveraging external information in topic modelling," *KAIS*, pp. 1–33, 2018.

M. Zhou, Y. Cong, and B. Chen, "The Poisson gamma belief network," in *NIPS*, 2015, pp. 3043–3051.

H. M. Wallach, D. M. Mimno, and A. McCallum, "Rethinking LDA: Why priors matter," in *NIPS*, 2009, pp. 1973–1981.

H. Zhao, L. Du, and W. Buntine, "A word embeddings informed focused topic model," in *ACML*, 2017, pp. 423–438.

H. Zhao, L. Du, W. Buntine, and M. Zhou, "Inter and intra topic structure learning with word embeddings," in *ICML*, 2018, pp. 5887–5896.

I. Sato and H. Nakagawa, "Topic models with power-law using Pitman-Yor process," in *SIGKDD*, 2010, pp. 673–682.

W. L. Buntine and S. Mishra, "Experiments with non-parametric topic models," in *SIGKDD*, 2014, pp. 881–890.

C. Chen, W. Buntine, N. Ding, L. Xie, and L. Du, "Differential topic models," *TPAMI*, vol. 37, no. 2, pp. 230–242, 2015.

R. V. Lindsey, W. P. Headden III, and M. J. Stipicevic, "A phrase-discovering topic model using hierarchical Pitman-Yor processes," in *EMNLP*, 2012, pp. 214–222.

C. Archambeau, B. Lakshminarayanan, and G. Bouchard, "Latent IBP compound Dirichlet allocation," *TPAMI*, vol. 37, no. 2, pp. 321–333, 2015.

F. Wood and Y. W. Teh, "A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation," in *AISTATS*, 2009, pp. 607–614.

L. Du, W. Buntine, and H. Jin, "Modelling sequential text with an adaptive topic model," in *EMNLP*, 2012, pp. 535–545.

J. H. Kim, D. Kim, S. Kim, and A. Oh, "Modeling topic hierarchies with the recursive chinese restaurant process," in *CIKM*, 2012, pp. 783–792.

A. Ahmed, L. Hong, and A. Smola, "Nested Chinese restaurant franchise process: Applications to user tracking and document modeling," in *ICML*, 2013, pp. 1426–1434.

W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *ICML*, 2006, pp. 577–584.

Y. Cong, B. Chen, H. Liu, and M. Zhou, "Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC," in *ICML*, 2017, pp. 864–873.

N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proc. of the 10th Intnl. Conf. on Computational Semantics*, 2013, pp. 13–22.

J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *EACL*, 2014, pp. 530–539.

Y. Yang, D. Downey, and J. Boyd-Graber, "Efficient methods for incorporating knowledge into topic models," in *EMNLP*, 2015, pp. 308–317.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *JMLR*, vol. 14, no. 1, pp. 1303–1347, 2013.

R. Guhaniyogi, S. Qamar, and D. B. Dunson, "Bayesian conditional density filtering," *Journal of Computational and Graphical Statistics*, no. just-accepted, 2018.

T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient Hamiltonian Monte Carlo," in *ICML*, 2014, pp. 1683–1691.

N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, and H. Neven, "Bayesian sampling using stochastic gradient thermostats," in *NIPS*, 2014, pp. 3203–3211.

M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *ICML*, 2011, pp. 681–688.

Supplementary materials for "Dirichlet belief networks for topic structure learning"

He Zhao[1], Lan Du[1], Wray Buntine[1], and Mingyuan Zhou[2]
[1]Faculty of Information Technology, Monash University, Australia
[2]McCombs School of Business, The University of Texas at Austin, USA

1 Details of the inference

Given the latent counts $\boldsymbol{x}_{k_t}^{(t)}$, the details of inference of the the t-th ($t < T$) layer of DirBN are as follows:

$$\boldsymbol{\phi}_{k_t}^{(t)} \sim \text{Dir}\left(\boldsymbol{\psi}_{k_t}^{(t)} + \boldsymbol{x}_{k_t}^{(t)}\right), \tag{1}$$

$$q_{k_1}^{(1)} \sim \text{Beta}\left(\psi_{\cdot k_1}^{(1)}, x_{\cdot k_1}^{(1)}\right), \tag{2}$$

$$y_{vk_t}^{(t)} \sim \text{CRT}\left(x_{vk_t}^{(t)}, \psi_{vk_t}^{(t)}\right), \tag{3}$$

$$\left(z_{v1k_t}^{(t)}, \cdots, z_{vK_{t+1}k_t}^{(t)}\right) \sim \text{Mult}\left(y_{vk_t}^{(t)}, \frac{\phi_{v1}^{(t+1)}\beta_{1k_t}^{(t)}}{\psi_{vk_t}^{(t)}}, \cdots, \frac{\phi_{vK_{t+1}}^{(t+1)}\beta_{K_{t+1}k_t}^{(t)}}{\psi_{vk_t}^{(t)}}\right), \tag{4}$$

$$\beta_{k_{t+1}k_t}^{(t)} \sim \text{Ga}\left(\gamma_{k_{t+1}}^{(t)} + z_{\cdot k_{t+1}k_t}^{(t)}, 1.0\right) / \left(c^{(t)} - \log q_{k_t}^{(t)}\right), \tag{5}$$

$$m_{k_{t+1}k_t}^{(t)} \sim \text{CRT}\left(z_{\cdot k_{t+1}k_t}^{(t)}, \gamma_{k_{t+1}}^{(t)}\right), \tag{6}$$

$$\gamma_{k_{t+1}}^{(t)} \sim \text{Ga}\left(\gamma_0^{(t)}/K_{t+1} + \sum_{k_t}^{K_t} m_{k_{t+1}k_t}^{(t)}, 1.0\right) / \left(c_0^{(t)} + n^{(t)}\right), \tag{7}$$

$$c^{(t)} \sim \text{Ga}\left(g_0 + K_t \sum_{k_{t+1}}^{K_{t+1}} \gamma_{k_{t+1}}^{(t)}, 1.0\right) / \left(h_0 + \sum_{k_{t+1},k_t}^{K_{t+1},K_t} \beta_{k_{t+1}k_t}^{(t)}\right), \tag{8}$$

$$p_{k_{t+1}}^{(t)} \sim \text{CRT}\left(\sum_{k_t}^{K_t} m_{k_{t+1}k_t}^{(t)}, \gamma_0^{(t)}/K_{t+1}\right), \tag{9}$$

$$\gamma_0^{(t)} \sim \text{Ga}\left(e_0 + \sum_{k_{t+1}}^{K_{t+1}} p_{k_{t+1}}^{(t)}, 1.0\right) / \left(f_0 + \log \frac{n^{(t)} + c_0^{(t)}}{c_0^{(t)}}\right), \tag{10}$$

$$c_0^{(t)} \sim \text{Ga}\left(g_0 + \gamma_0^{(t)}, 1.0\right) / \left(h_0 + \sum_{k_{t+1}} \gamma_{k_{t+1}}^{(t)}\right), \tag{11}$$

where $n^{(t)} = \sum_{k_t}^{K_t} \log \frac{c^{(t)} - \log q_{k_t}^{(t)}}{c^{(t)}}$.

In the top layer $t = T$, we have:

$$\phi_{k_T}^{(T)} \sim \text{Dir}\left(\eta + \boldsymbol{x}_{k_T}^{(T)}\right), \tag{12}$$

$$s_{vk_T} \sim \text{CRT}\left(x_{vk_T}^{(T)}, \eta\right), \tag{13}$$

$$\eta \sim \text{Ga}\left(a_0 + \sum_{v,k_T}^{V,K_T} s_{vK_T}, 1.0\right) / \left(b_0 - \sum_{k_T}^{K_T} \log q_{k_T}^{(T)}\right). \tag{14}$$

The inference process of DirBN is in Algorithm 1. Note that in different models, after the topic assignments of words are obtained, the inference of DirBN is the same.

2 Details of the combined models

PFA+DirBN The generative process of PFA+DirBN is shown as follows:

$$p_k \sim \text{Beta}\left(c\epsilon, c(1-\epsilon)\right), r_k \sim \text{Ga}(c_0 r_0, 1/c_0), \theta_{kd} \sim \text{Ga}\left(r_k, \frac{p_k}{1-p_k}\right),$$

$$\phi_k \sim \text{DirBN}(T), x_{vd} = \sum_k^K x_{vdk}, x_{vdk} \sim \text{Pois}(\phi_{vk}\theta_{kd}), \tag{15}$$

where $\text{DirBN}(T)$ stands for the generative process of DirBN with T layers.

MetaLDA+DirBN The generative process of MetaLDA+DirBN is as follows:

$$\lambda_{lk} \sim \text{Ga}(a_0, 1/b_0), \alpha_{kd} = \prod_l^L (\lambda_{lk})^{f_{ld}}, \boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha}_d), \phi_k \sim \text{DirBN}(T),$$

$$z_{id} \sim \text{Categorical}(\boldsymbol{\theta}_d), w_{id} \sim \text{Categorical}(\phi_{z_{id}}), \tag{16}$$

where L is the number of unique document labels, $l \in \{1, \cdots, L\}$, $f_{ld} \in \{0, 1\}$ indicates whether document d has label l, $w_{id} = v$ is the i-th word in document d, and $z_{id} = k$ is the topic assignment of w_{id}.

GBN+DirBN The generative process of GBN+DirBN is as follows:

$$\boldsymbol{\theta}_d^{(S)} \sim \text{Ga}\left(\boldsymbol{r}, 1/c_j^{(T+1)}\right), \cdots, \boldsymbol{\theta}_d^{(s)} \sim \text{Ga}\left(\widetilde{\boldsymbol{\Phi}}^{(s+1)}\boldsymbol{\theta}_d^{(s+1)}, 1/c_d^{(s+1)}\right),$$

$$\cdots$$

$$\boldsymbol{\theta}_d^{(1)} \sim \text{Ga}\left(\widetilde{\boldsymbol{\Phi}}^{(2)}\boldsymbol{\theta}_d^{(2)}, p_d^{(2)}/(1-p_d^{(2)})\right), \phi_k \sim \text{DirBN}(T),$$

$$x_{vd} = \sum_k^K x_{vdk}, x_{vdk} \sim \text{Pois}(\phi_{vk}\theta_{kd}^{(1)}), \tag{17}$$

where $s \in \{1, \cdots, S\}$ is the index of the s layer in GBN.

3 More results

For document classification, the TPs were used as input features for a L_2 regularized logistic regression using the LIBLINEAR package to predict the document labels. We used the same train/test splits as in perplexity evaluation, except that all the words in a test document were used to infer its TP. The results on WS and TMN are shown in Table 1.

4 Visualisation of topic hierarchies

Shown in Figure 2 to 5.

Table 1: Document classification

Training words	WS			TMN		
	20%	40%	100%	20%	40%	100%
PFA	67.58±5.73	81.08±0.83	82.29±0.73	73.02±1.43	78.68±0.28	80.00±0.51
PFA+mallet	73.97±1.12	79.64±0.89	82.75±0.89	72.84±0.40	78.02±0.89	79.92±0.66
PFA+DirBN-1	77.11±0.55	81.69±0.53	82.26±0.48	73.08±0.33	78.40±0.31	79.77±0.56
PFA+DirBN-3	76.74±0.57	82.04±0.28	83.68±1.04	74.41±0.60	78.99±0.46	79.91±0.56
MetaLDA	67.94±3.00	**83.26**±1.21	84.18±1.10	74.02±0.62	78.88±0.27	80.04±0.49
MetaLDA + DirBN-1	76.67±0.88	81.38±1.02	83.07±0.70	74.10±0.22	79.67±0.67	80.63±0.10
MetaLDA + DirBN-3	77.84±1.06	82.53±0.46	83.97±1.09	75.03±0.26	79.37±0.63	80.99±0.22
GBN	68.87±4.67	82.97±0.49	84.35±0.91	72.88±1.08	79.28±0.41	**81.44**±0.21
GBN+DirBN-1	76.73±0.70	82.54±0.81	83.18±0.40	74.42±0.32	79.59±0.30	80.87±0.68
GBN+DirBN-3	**78.17**±1.88	82.82±1.08	**84.28**±1.12	**75.36**±0.60	**79.79**±0.48	81.10±0.34

Require: $\boldsymbol{x}_{k_1}^{(1)}$ for all $k_1, T(T > 1), a_0, b_0, e_0, f_0, g_0, h_0$ $MaxIteration$
Ensure: $\boldsymbol{\beta}_{k_t}^{(t)}, \boldsymbol{\phi}_{k_t}^{(t)}$ for all k_t

1: Randomly initialise all the latent variables according to the generative process
2: **for** $iter \leftarrow 1$ **to** $MaxIteration$ **do**
3:    $/ *$ Propagating the latent counts from the bottom up $* /$
4:    **for** $t \leftarrow 1$ **to** T **do**
5:      **for all** k_t and v **do**
6:        Sample $y_{vk_t}^{(t)}$ by Eq. (3)
7:        **for all** k_{t+1} **do**
8:          Sample $z_{vk_{t+1}k_t}^{(t)}$ by Eq. (4)
9:        **end for**
10:     **end for**
11:   **end for**
12:   $/ *$ Updating the latent variables from the top down $* /$
13:   **for** $t \leftarrow T$ **to** 1 **do**
14:     **if** $t = T$ **then**
15:       **for all** k_T and v **do**
16:         Sample s_{vk_T} by Eq. (13)
17:       **end for**
18:       Sample η by Eq. (14)
19:       **for all** k_T **do**
20:         Sample $\boldsymbol{\phi}_{k_T}^{(T)}$ by Eq. (12)
21:       **end for**
22:     **else**
23:       **for all** k_t **do**
24:         Compute $\boldsymbol{\psi}_{k_t}^{(t)}$ by $\boldsymbol{\psi}_{k_t}^{(t)} = \sum_{k_{t+1}}^{K_{t+1}} \phi_{k_{t+1}}^{(t+1)} \beta_{k_{t+1}k_t}^{(t)}$
25:       **end for**
26:       **for all** k_t **do**
27:         Sample $q_{k_t}^{(t)}$ by Eq. (2)
28:       **end for**
29:       **for all** k_t and k_{t+1} **do**
30:         Sample $m_{k_{t+1}k_t}^{(t)}$ by Eq. (6)
31:       **end for**
32:       **for all** k_{t+1} **do**
33:         Sample $\gamma_{k_{t+1}}^{(t)}, p_{k_{t+1}}^{(t)}$ by Eq. (7,9)
34:       **end for**
35:       Sample $c^{(t)}, \gamma_0^{(t)}, c_0^{(t)}$ by Eq. (8,10,11)
36:       **for all** k_t and k_{t+1} **do**
37:         Sample $\beta_{k_{t+1}k_t}^{(t)}$ by Eq. (5)
38:       **end for**
39:       **for all** k_t **do**
40:         Sample $\boldsymbol{\phi}_{k_t}^{(t)}$ by Eq. (1)
41:       **end for**
42:     **end if**
43:   **end for**
44: **end for**

Figure 1: Infernece algorithm for DirBN

38 party election presidential poll leader voted vote term democratic elected president ensure

14 election party vote elections president government opposition political leader state ruling presidential

17 obama president barack election peru presidential poll haiti humala wing run wednesday

39 state governor law republican abortion bill senate jersey campaign house medical john

5 world   6 health   3 us

94 billion financial banks funds finance debt bank cash money capital assets private

52 bank financial banks mortgage report goldman billion street wall regulators investment securities

69 billion deal buy group stake sell chrysler million company shares firm private

25 china united states world europe india countries global trade european group nations

2 business   7 sci_tech   5 world   6 health

49 file household management large granted animals vehicle license shook recalls owners caused

42 investigator animals sheriff poison killer searched spray serial fleeing national bars prosecution

60 crash drivers vehicle motor crashes car greatly fuel loaded travel driver truck

74 anthony trial year casey daughter mother jackson day doctor michael thai killed

59 gay clinton sex marriage dog state secretary pet dogs chavez venezuela couples

41 man police prison death charged arrested guilty sentenced years woman charges year

49 mexico border drug mexican police found bodies mass authorities missing killed people

78 bus crash york police city fire people killed driver injured safety car

98 air plane flight jet airport france crash airlines traffic crashed boeing pilot

42 nascar race busch sprint cup win kyle wins pole earnhardt racing edwards

62 cars fifa ford general car auto toyota president motors motor fuel plant

6 health   4 entertainment   5 world   3 us   1 sport   2 business

79 teams field complex heading toronto nba west room injury keeping opener bash

44 team player won field sports winner professional event draw drew top briefing

70 baseball mets yankees phillies giants manager athletics list diamondbacks lineup rodriguez starter

75 game season coach team play games players nba scored conference forward loss

66 time good make back life day years questions news bad job year

71 world olympic champion cup title london hockey championship race marathon games olympics

36 sports day replay photo yankees baseball jeter mets april york team derek

65 mets injury list season baseball left disabled miss league team weeks surgery

67 yankees sox win mets run red hit inning game beat home phillies

84 rangers playoff capitals playoffs win scored victory game goal season team washington

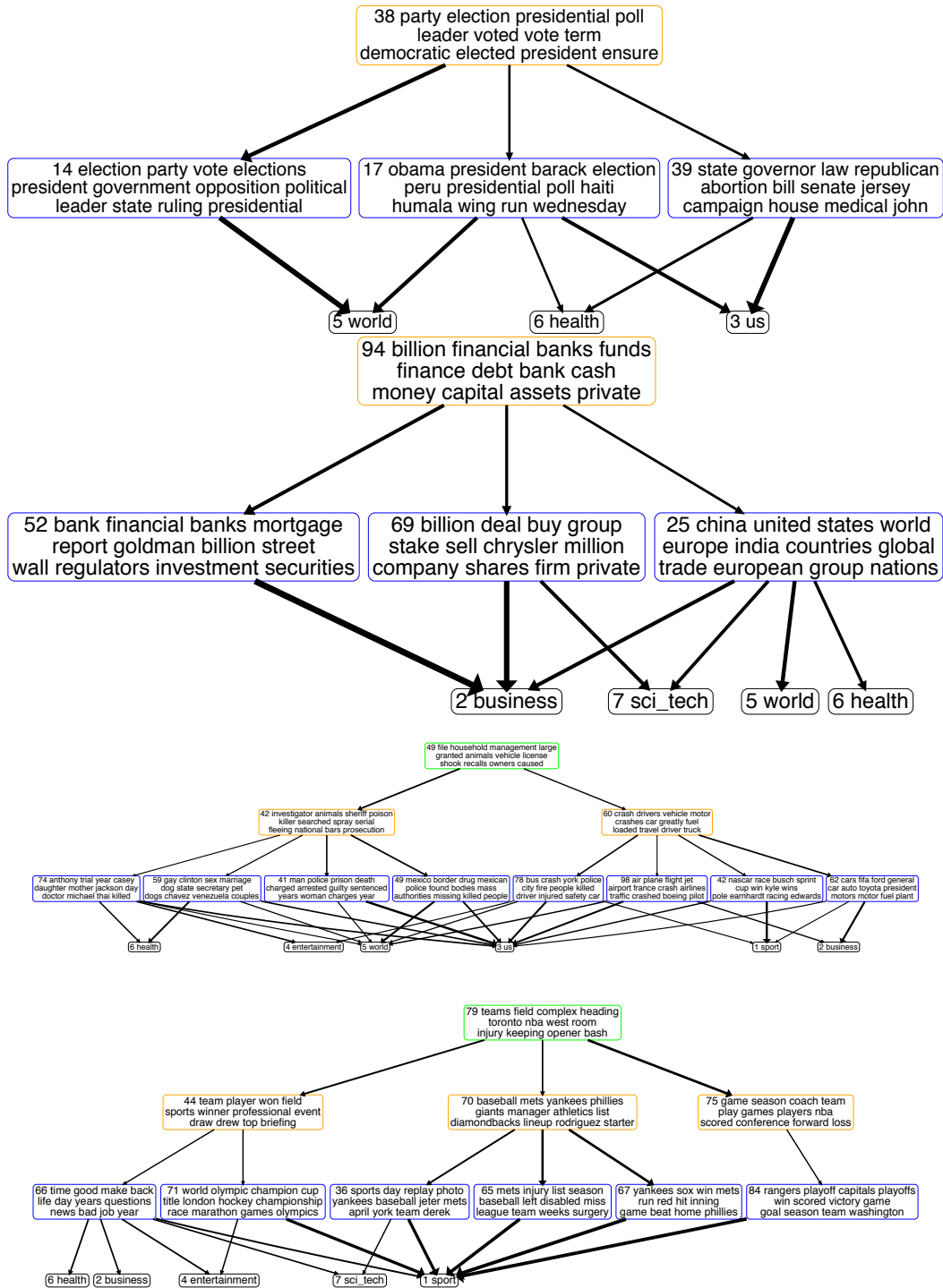6 health   2 business   4 entertainment   7 sci_tech   1 sport

Figure 2: Topic hierarchies discovered by MetaLDA+DirBN on TMN. The topics in the green, yellow, and blue rectangles are the third, second, and first layer topics in DirBN and the correlated document labels are shown on the bottom of each figure. Thicker arrows indicate stronger correlations.
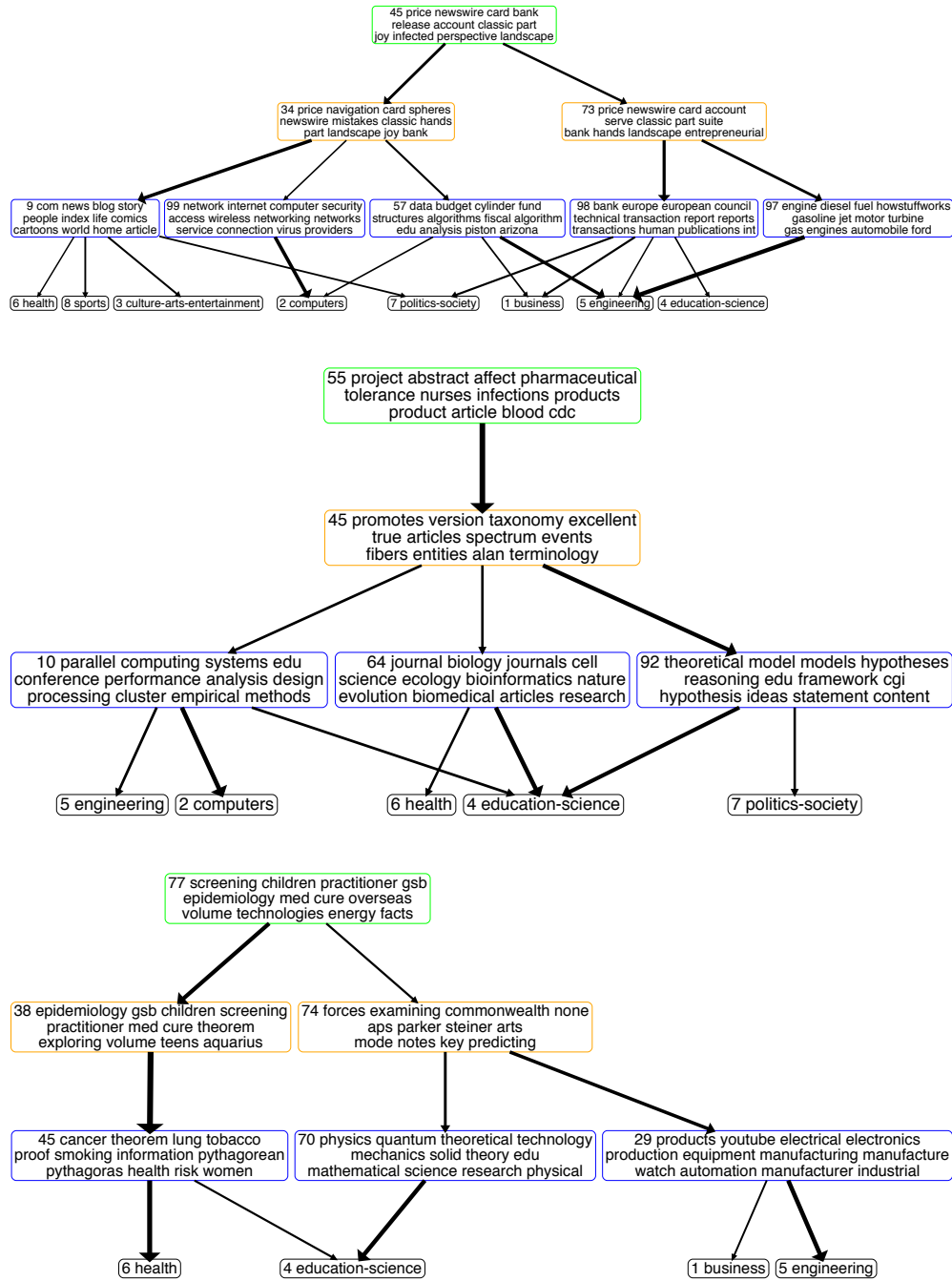
Figure 3: Topic hierarchies discovered by MetaLDA+DirBN on WS. The topics in the green, yellow, and blue rectangles are the third, second, and first layer topics in DirBN and the correlated document labels are shown on the bottom of each figure. Thicker arrows indicate stronger correlations.
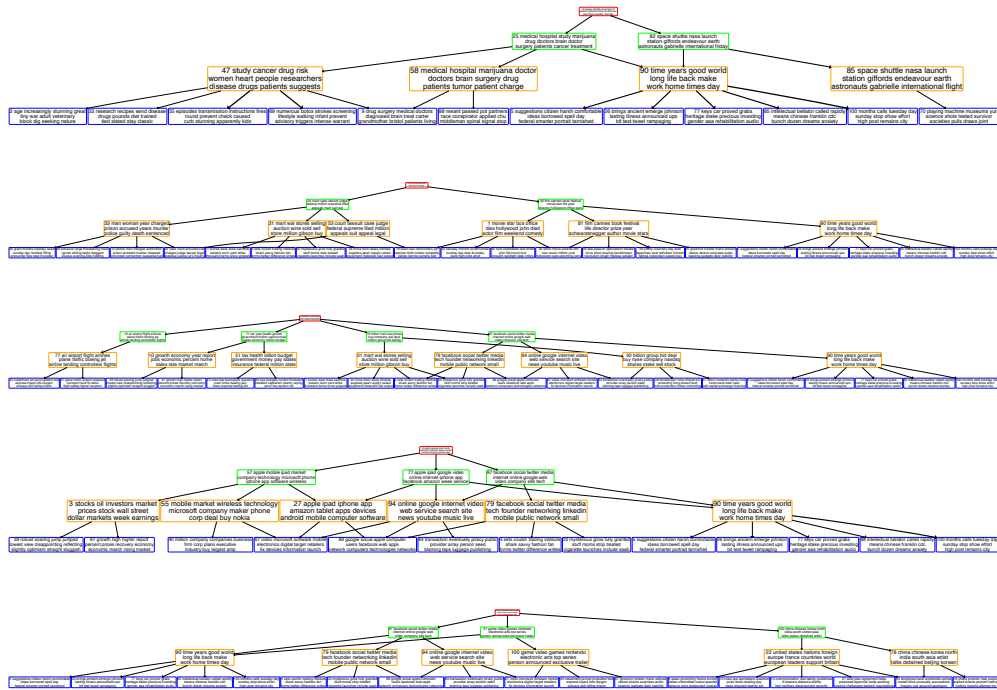
6

Figure 4: Topic hierarchies discovered by GBN+DirBN on TMN. The topics in the red and green rectangles are the third and second-layer topics discovered by GBN on TPs. The topics in the blue rectangles are the second-layer topics discovered by DirBN on WDs. The topics in the yellow rectangles are the first-layer topics connecting the higher-layer topics of GBN and DirBN.
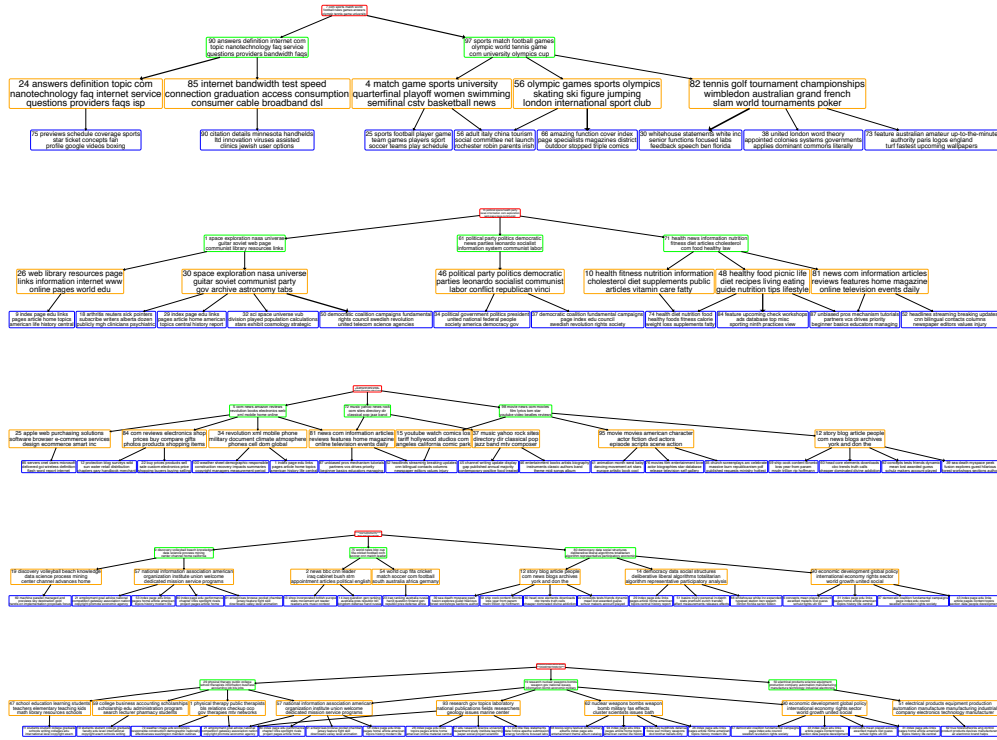
Figure 5: Topic hierarchies discovered by GBN+DirBN on WS. The topics in the red and green rectangles are the third and second-layer topics discovered by GBN on TPs. The topics in the blue rectangles are the second-layer topics discovered by DirBN on WDs. The topics in the yellow rectangles are the first-layer topics connecting the higher-layer topics of GBN and DirBN.