# A Variational Approximation for Topic Modeling of Hierarchical Corpora

**Do-kyum Kim** DOK027@CS.UCSD.EDU
**Geoffrey M. Voelker** VOELKER@CS.UCSD.EDU
**Lawrence K. Saul** SAUL@CS.UCSD.EDU
Department of Computer Science and Engineering, University of California, San Diego

## Abstract

We study the problem of topic modeling in corpora whose documents are organized in a multi-level hierarchy. We explore a parametric approach to this problem, assuming that the number of topics is known or can be estimated by cross-validation. The models we consider can be viewed as special (finite-dimensional) instances of hierarchical Dirichlet processes (HDPs). For these models we show that there exists a simple variational approximation for probabilistic inference. The approximation relies on a previously unexploited inequality that handles the conditional dependence between Dirichlet latent variables in adjacent levels of the model's hierarchy. We compare our approach to existing implementations of nonparametric HDPs. On several benchmarks we find that our approach is faster than Gibbs sampling and able to learn more predictive models than existing variational methods. Finally, we demonstrate the large-scale viability of our approach on two newly available corpora from researchers in computer security—one with 350,000 documents and over 6,000 internal subcategories, the other with a *five-level* deep hierarchy.

## 1. Introduction

In the last decade, probabilistic topic models have emerged as a leading framework for analyzing and organizing large collections of text (Blei & Lafferty, 2009). These models represent documents as "bags of words" and explain frequent co-occurrences of words as evidence of topics that run throughout the corpus. The first properly Bayesian topic model was latent Dirichlet allocation (LDA) (Blei et al., 2003). A great deal of subsequent work has investigated hierarchical extensions of LDA, much of it stemming from interest in nonparametric Bayesian methods (Teh et al., 2006). In these models, topics are shared across different but related corpora (or across different parts of a single, larger corpus). One challenge of topic models is that exact inference is intractable. Thus, it remains an active area of research to devise practical approximations for computing the statistics of their latent variables.

In this paper we are interested in the topic modeling of corpora whose documents are organized in a multi-level hierarchy. Often such structure arises from prior knowledge of a corpus's subject matter and readership. For example, news articles appear in different sections of the paper (e.g., business, politics), and these sections are sometimes further divided into subcategories (e.g., domestic, international). Our goal is to explore the idea that prior knowledge of this form, though necessarily imperfect and incomplete, should inform the discovery of topics.

We explore a parametric model of such corpora, assuming for simplicity that the number of topics is known or can be estimated by (say) cross-validation. The models that we consider assign topic proportions to each node in a corpus's hierarchy—not only the leaf nodes that represent documents, but also the ancestor nodes that reflect higher-level categories. Conditional dependence ensures that nearby nodes in the hierarchy have similar topic proportions. In particular, the topic proportions of lower-level categories are on average the same as their parent categories. However, useful variations naturally arise as one descends the hierarchy. As we discuss later, these models can also be viewed as special (finite-dimensional) instances of hierarchical Dirichlet processes (HDPs) (Teh et al., 2006).

*Proceedings of the 30$^{th}$ International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. 

Our main contributions are two. First, we devise a new variational approximation for inference in these models. Based on a previously unexploited inequality, the approximation enables us to compute a rigorous lower bound on the likelihood in Bayesian networks where Dirichlet random variables appear as the children of other Dirichlet random variables. We believe that this simple inequality will be of broad interest.

Our second contribution is to demonstrate the large-scale viability of our approach. Our interest in this subject arose from the need to analyze two sprawling, real-world corpora from the field of computer security. The first is a seven-year collection of over 350,000 job postings from *Freelancer.com*, a popular Web site for crowdsourcing. We view this corpus as a three-layer tree in which leaf nodes represent the site's job postings and interior nodes represent the active buyers (over 6,000 of them) on the site; see Fig. 1. The second corpus is derived from the BlackHatWorld Internet forum, in which users create and extend threads in a deep, content-rich hierarchy of pre-defined subcategories; see Fig. 2. Our results break new ground for hierarchical topic models in terms of both the breadth (i.e., number of interior nodes) and depth (i.e., number of levels) of the corpora that we consider. Moreover, it is our experience that sampling-based approaches for HDPs (Teh et al., 2006) do not easily scale to corpora of this size, while other variational approaches (Teh et al., 2008; Wang et al., 2011; Wang & Blei, 2012; Bryant & Sudderth, 2012) have not been demonstrated (or even fully developed) for hierarchies of this depth.

The organization of this paper is as follows. In section 2, we describe our probabilistic models for hierarchical corpora and review related work. In section 3, we develop the variational approximation for inference and parameter estimation in these models. In section 4, we evaluate our approach on several corpora and compare the results to existing implementations of HDPs. Finally, in section 5, we conclude and discuss possible extensions of interest. The supplementary material for our paper contains a full proof of the key inequality for variational inference, a brief description of our large-scale (parallelized) implementation, and additional background and results on the Freelancer and BlackHatWorld corpora.

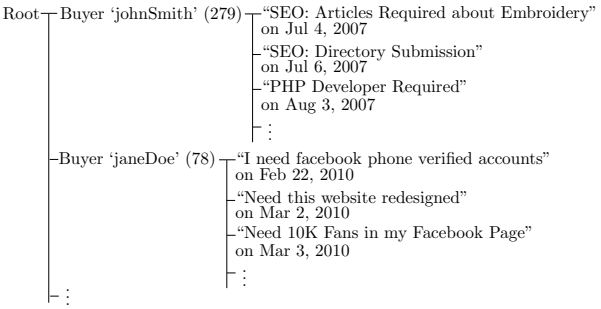


*Figure 1.* The hierarchy of buyers and job advertisements on Freelancer.com. The number of ads per buyer is indicated in parentheses. For brevity, only titles and dates of ads are shown.

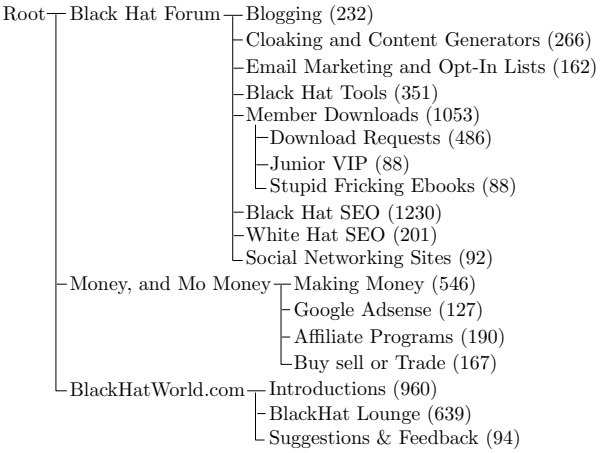


*Figure 2.* The hierarchy of subforums in the BlackHatWorld Internet forum. The number of threads in each subforum is indicated in parentheses.

## 2. Model and Related Work

Figs. 1 and 2 illustrate the types of structure we seek to model in hierarchical corpora. This structure is most easily visualized as a tree in which the root node represents the corpus as a whole, the children of the root node represent top-level categories, the interior nodes represent subcategories of their parents, and the leaf nodes represent individual documents. In this section we describe a probabilistic generative model for corpora of this form and discuss its relation to previous work in topic modeling.

### 2.1. Model

Our model is essentially an extension of LDA to account for the tree structure in Figs. 1 and 2. In LDA, each document $d$ is modeled by topic proportions $\theta_d$, which are mixture weights over a finite set of $K$ topics. In our approach, we model not only the documents in this way—the leaves of the tree—but also the categories and subcategories that appear at higher levels in the tree. Thus for each (sub)category $t$, we model its topic proportions by a latent Dirichlet random variable $\theta_t$, and we associate one of these variables to each non-leaf node in the tree. We use $\theta_0$ to denote the topic proportions of the root node in the tree (i.e. the

corpus-wide topic proportions), and we sample these from a symmetric Dirichlet prior $\gamma$.

The topic proportions of the corpus, its categories, subcategories, and documents are the latent Dirichlet variables in our model. It remains to specify how these variables are related—in particular, how topic proportions are inherited from parent to child as one traverses the trees in Figs. 1 and 2. We parameterize this conditional dependence by associating a (scalar) concentration parameter $\alpha_t$ to each category $t$. The parameter $\alpha_t$ governs how closely the topic proportions of category $t$ are inherited by its subcategories and documents; in particular, small values of $\alpha_t$ allow for more variance, and large values for less. More formally, let $\pi(t)$ denote the parent category of the category $t$. Then we stipulate:

$$\theta_t \sim \text{Dirichlet}\left(\alpha_{\pi(t)}\theta_{\pi(t)}\right). \tag{1}$$

Likewise, documents inherit their topic proportions from parent categories in the same way:

$$\theta_d \sim \text{Dirichlet}\left(\alpha_{\pi(d)}\theta_{\pi(d)}\right), \tag{2}$$

where $\pi(d)$ in the above equation denotes the parent category of document $d$.

The final assumption of our model is one of conditional independence: namely, that the topic proportions of subcategories are conditionally independent of their "ancestral" categories given the topic proportions of their parent categories. With this assumption, we obtain the simple generative model of hierarchical corpora shown in Fig. 3. To generate an individual document, we begin by recursively sampling the topic proportions of its (sub)categories conditioned on those of their parent categories. Finally, we sample the words of the document, conditioned on its topic proportions, in the same manner as LDA. In what follows we refer to this model as tree-informed LDA, or simply tiLDA.

In general, it is a bit unwieldy to depict the Bayesian network for topic models of this form. However, a special case occurs when the corpus hierarchy has uniform depth—that is, when all documents are attached to subcategories at the same level. Fig. 4 shows the graphical model when all documents in the corpus are attached (for example) to third-level nodes.

### 2.2. Related Work

Our model can be viewed as a generalization of certain previous approaches and a special instance of others. Consider, for example, the special case of tiLDA for a "flat" corpus, where all the documents are attached directly to its "root." This case of tiLDA corresponds to LDA with an asymmetric Dirichlet prior over topic proportions. Wallach et al. (2009a) showed how to perform Gibbs sampling in such models and demonstrated their advantages over LDA with a symmetric Dirichlet prior.

**Procedure** Main()
1. Draw topics $\beta_k \sim \text{Dirichlet}(\eta)$ for $k \in \{1, ..., K\}$
2. Draw topic proportions $\theta_0 \sim \text{Dirichlet}(\gamma)$
3. Call GenerateCategory(0)

**Procedure** GenerateCategory($t$)
1. For each subcategory $c$ of $t$:
   (a) Draw topic proportions $\theta_c \sim \text{Dirichlet}(\alpha_t\theta_t)$
   (b) Call GenerateCategory($c$)
2. For each document $d$ of $t$:
   (a) Draw topic proportions $\theta_d \sim \text{Dirichlet}(\alpha_t\theta_t)$
   (b) Call GenerateDocument($d$)

**Procedure** GenerateDocument($d$)
1. For each word $w_{dn} \in \{1, 2, \ldots, V\}$ of $d$,
   (a) Draw a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$
   (b) Draw a word $w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$

*Figure 3.* The generative process of our topic model for hierarchical corpora. The process begins in the Main procedure, sampling topic-word profiles and topic proportions from symmetric Dirichlet distributions. Then it recursively executes the GenerateCategory procedure for each internal node of the corpus and the GenerateDocument procedure for each leaf node.

Our approach also draws on inspiration from hierarchical Dirichlet processes (HDPs) (Teh et al., 2006). In tiLDA, as in HDPs, the sample from one Dirichlet distribution serves as the base measure for another Dirichlet distribution. HDPs are a nonparametric generalization of LDA in which the number of topics is potentially unbounded and can be learned from data. We can view the generative model of tiLDA as a special case of multi-level HDPs whose base measure is finite (thus only allowing for a finite number of topics). Though tiLDA does not possess the full richness of HDPs, our results will show that for some applications it is a compelling alternative.

Gibbs sampling is perhaps the most popular strategy for inference and learning in hierarchical topic models. The seminal work by Teh et al. (2006) developed a Gibbs sampler for HDPs of arbitrary depth and used it to learn a three-level hierarchical model of 160 papers from two distinct tracks of the NIPS conference. We note also that Du et al. (2010) developed a collapsed Gibbs sampling algorithm for a three-level hierarchical model similar in spirit to ours. The drawback to Gibbs sampling is its slowness; it is not currently a viable

approach for large corpora.

Many researchers have pursued variational inference in HDPs as a faster, cheaper alternative to Gibbs sampling. Teh et al. (2008) developed a framework for collapsed variational inference in two-level (but not arbitrarily deep) HDPs, and later Sato et al. (2012) proposed a related but simplified approach. Yet another framework for variational inference was developed by Wang et al. (2011), who achieved speedups with online updates. While the first variational methods for HDPs truncated the number of possible topics, two recent papers have investigated online approaches with dynamically varying levels of truncation (Wang & Blei, 2012; Bryant & Sudderth, 2012). There have been many successful applications of variational HDPs to large corpora; however, we are unaware of any actual applications to hierarchical corpora (i.e., involving HDPs that are three or more levels deep). It seems fair to say that variational inference in nonparametric Bayesian models involves many complexities (e.g., auxiliary variables, stick-breaking constructions, truncation schemes) beyond those in parametric models. We note that even for two-level HDPs, the variational approximations already require a good degree of cleverness (sometimes just to identify the latent variables).

The above considerations suggest regimes where an approach such as tiLDA may compete favorably with nonparametric HDPs. In this paper, we are interested in topic models of large corpora with known hierarchical structure, sometimes many levels deep. In addition, the corpora are static, not streaming; thus we are not attempting to model the introduction of new (or a potentially unbounded number of) topics over time. We seek a model richer than LDA, one that can easily corporate prior knowledge in the form of Figs. 1 and 2, but with a minimum of additional complexity. (Here it bears reminding that LDA—in its most basic form—still remains a wildly popular and successful model.) We shall see that tiLDA fits the bill perfectly in this regime.

## 3. Algorithms

In this section we develop the algorithms for inference and learning in tiLDA. Our large-scale implementation is described in the paper's supplementary material.

### 3.1. Variational Inference

The problem of inference in tiLDA is to compute the posterior distribution over the model's latent variables given the observed words in the corpus. In tiLDA, the latent variables are the topic proportions $\theta_t$ of each category (or subcategory) $t$, the topic proportions $\theta_d$ of each document $d$, the topic $z_{dn}$ associated with each word $w_{dn}$, and the multinomial parameters $\beta_k$ for each topic. Exact inference is not possible; approximations are required. Here we pursue a variational method for approximate inference (Jordan et al., 1999) that generalizes earlier approaches to LDA (Blei et al., 2003).

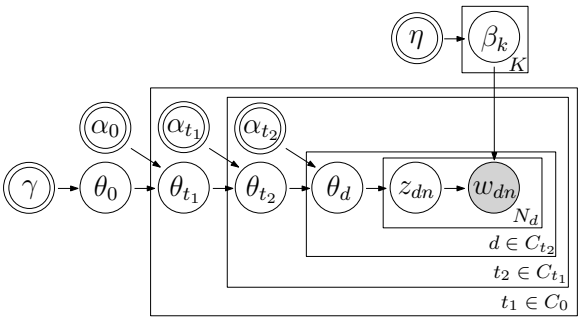


*Figure 4.* Graphical model for tiLDA in which all documents of a hierarchical corpus are attached to third-level nodes. Here $C_t$ denotes the set of indexes for the subcategories and documents of category $t$, and $N_d$ denotes the length of the document $d$.

The variational method is based on a parameterized approximation to the posterior distribution over the model's latent variables. The approximation takes the fully factorized form:

$$q(\theta, z, \beta|\nu, \rho, \lambda) = \Big[\prod_k q(\beta_k|\lambda_k)\Big]\Big[\prod_t q(\theta_t|\nu_t)\Big] \Big[\prod_d q(\theta_d|\nu_d)\prod_n q(z_{dn}|\rho_{dn})\Big], \quad (3)$$

where the parameters $\nu_t$, $\nu_d$, $\rho_{dn}$, and $\lambda_k$ are varied to make the approximation as accurate as possible. The component distributions in this variational approximation are the exponential family distributions:

$$\theta_t \sim \text{Dirichlet}(\nu_t), \qquad \theta_d \sim \text{Dirichlet}(\nu_d),$$
$$z_{dn} \sim \text{Multinomial}(\rho_{dn}), \qquad \beta_k \sim \text{Dirichlet}(\lambda_k).$$

Figs. 4 and 5 contrast the graphical models for the true posterior and its variational approximation.

The variational parameters $\nu_t$, $\nu_d$, $\rho_{dn}$, and $\lambda_k$ are found by attempting to minimize the Kullback-Leibler divergence between the approximation in eq. (3) and the true posterior distribution of the model. It can be shown that this is equivalent to maximizing a lower bound $\mathscr{L} \leq \log p(w|\gamma, \alpha, \eta)$ on the marginal log-likelihood of the corpus. This lower bound is given by:

$$\mathscr{L} = \mathrm{E}_q\left[\log p(\theta, z, w, \beta|\gamma, \alpha, \eta)\right] + H(q), \quad (4)$$

where $\mathrm{E}_q$ denotes the expectation with respect to the variational distribution and $H(q)$ denotes its entropy.

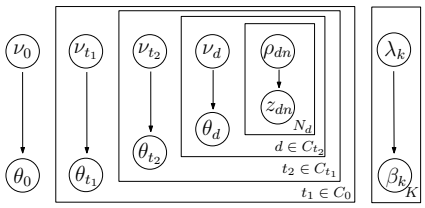


*Figure 5.* Variational approximation to the posterior distribution for the graphical model in Fig. 4.

So far we have developed the variational approximation for our model by following exactly the same approach used in LDA. The lower bound in eq. (4), however, cannot be computed analytically, even for the simple factorized distribution in eq. (3). In particular, new terms arise from the expectation $\mathrm{E}_q\left[\log p(\theta, z, w, \beta|\gamma, \alpha, \eta)\right]$ that are not present in the variational approximation for LDA.

Let us see where these terms arise. Consider the model's *prior* distribution over latent topic proportions for each subcategory $t$ and document $d$ in the corpus:

$$p(\theta|\alpha,\gamma) \;\propto\; \prod_{t>0} p(\theta_t|\alpha_{\pi(t)}\theta_{\pi(t)}) \prod_d p(\theta_d|\alpha_{\pi(d)}\theta_{\pi(d)}). \tag{5}$$

In eq. (5), we have again used $\pi(t)$ and $\pi(d)$ to denote the parent categories of $t$ and $d$, respectively, in the tree. Note that both terms in this prior distribution express conditional dependencies between Dirichlet variables at adjacent levels in the tree. In eq. (4), they give rise to averages such as $\mathrm{E}_q[\log p(\theta_t|\alpha_{\pi(t)}\theta_{\pi(t)})]$ that cannot be analytically computed.

In this paper, we do not have space for a complete derivation of the log-likelihood bound in our model. However, the extra steps beyond LDA are essentially applications of the following theorem.

**Theorem 3.1.** *Let $\theta \sim \mathrm{Dirichlet}(\nu)$, and let $\alpha > 0$. As shorthand, let $\nu_0 = \sum_i \nu_i$. Then:*

$$\begin{aligned}\mathrm{E}\left[\log\Gamma(\alpha\theta_i)\right] \;\leq\; &\log\Gamma(\alpha\mathrm{E}[\theta_i]) \;+\alpha(1-\mathrm{E}[\theta_i])/\nu_0 \\ &+\; (1-\alpha\mathrm{E}[\theta_i])\left[\log\mathrm{E}[\theta_i]+\Psi(\nu_0)-\Psi(\nu_i)\right],\end{aligned}$$

*where $\mathrm{E}[\theta_i] = \nu_i/\nu_0$ and $\Gamma(\cdot)$ and $\Psi(\cdot)$ are respectively the gamma and digamma functions.*

A proof of this theorem is given in the paper's supplement. Note especially the direction of the bound. The function $\log\Gamma(\cdot)$ is convex, and hence a naive application of Jensen's inequality to the left hand side of the equation yields the *lower* bound $\mathrm{E}[\log\Gamma(\alpha\theta_i)] \geq \log\Gamma(\alpha\mathrm{E}[\theta_i])$. It is the *additional* terms on the right hand side of the equation that establish the theorem's *upper* bound. The direction of inequality is crucial in the context of variational inference, where the upper bound is needed to maintain an overall lower bound on the log-likelihood. Thus it can be used to compute a looser (but still rigorous) lower bound $\mathscr{L}' \leq \mathscr{L}$ on the log-likelihood in terms of the model's variational parameters. We shall see that this surrogate bound remains highly effective for inference and learning.

We obtain the best approximation in the form of eq. (3) by maximizing $\mathscr{L}'$ with respect to the variational parameters $\nu$, $\rho$ and $\lambda$. In practice, we perform the optimization by coordinate ascent in repeated bottom-up sweeps through the corpus hierarchy. Each sweep begins by updating the parameters $\nu_d$ and $\rho_d$ attached to individual documents; these updates take essentially the same form as in LDA. Then, once these parameters have converged, we turn to updating the variational parameters $\nu_t$ attached to different-level categories; these maximizations are performed using variants of Newton's method. The bottom-up sweep continues through the different levels of the corpus until we reach the root of the corpus. Finally, the whole procedure repeats until $\mathscr{L}'$ converges.

### 3.2. Variational Learning

We can either fix the model parameters $\gamma$, $\alpha$ and $\eta$ or learn them from data. For the latter, we use the lower bound from section 3.1 as a surrogate for maximum likelihood estimation. The variational EM algorithm alternates between computing the best factorized approximation in eq. (3) and updating the model parameters to maximize the lower bound $\mathscr{L}'$. The first of these steps is the variational E-step; the second is the variational M-step. In the M-step we update the model parameters by block coordinate ascent. In particular, we use Newton's method to update the concentration parameter $\alpha_t$ associated to each category $t$ as well as $\gamma$ and $\eta$ at the root of the corpus.

It is useful to view the variational EM algorithm as a double-optimization over both the variational parameters (E-step) and the model parameters (M-step). This view naturally suggests an interweaving of the two steps, and, in fact, this is how we implement the algorithm in practice; see Algorithm 1.

## 4. Experiments

In this section we evaluate tiLDA on several corpora and compare its results where possible to existing implementations of HDPs. We followed more or less standard procedures in training. The variational EM algorithms for tiLDA and HDPs were iterated until conver-

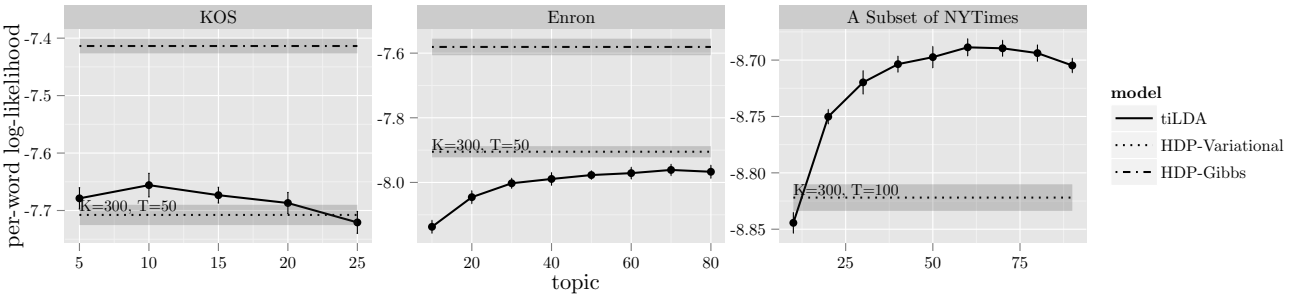


*Figure 6.* Predictive log-likelihood from two-level models of tiLDA and HDPs. See text for details.

**Algorithm 1** The variational EM algorithm for tiLDA. The algorithm begins in main, then invokes **OPT_SUBTREE** recursively for each category. At the deepest level of recursion, **OPT_DOCUMENT** infers the hidden variables of documents given their words and prior on topic proportions (just as in LDA).

**main** ()
  initialize $\gamma$, $\eta$ and $\lambda$
  **OPT_SUBTREE**(0)

**function OPT_SUBTREE**($t$)
  initialize $\alpha_t$ and $\nu_t$
  **while** $\mathscr{L}'$ increases **do**
    **for all** subcategory $c$ of $t$ **do**
      **OPT_SUBTREE**($c$)
    **for all** document $d$ of $t$ **do**
      **OPT_DOCUMENT**($d$)
    Update $\nu_t$ and $\alpha_t$
    **if** t = 0 **then**
      Update $\lambda$, $\eta$ and $\gamma$

gence in their log-likelihood bounds. HDPs estimated by Gibbs sampling were trained for 5,000 iterations.

Since the log-likelihood of held-out data cannot be computed exactly in topic models, we use a method known as *document completion* (Wallach et al., 2009b) to evaluate each model's predictive power. First, for each trained model, we estimate a set of topics $\beta$ and category topic proportions $\theta_t$ on this set of topics. Then we split each document in the held-out set into two parts; on the first part, we estimate document topic proportions $\theta_d$, and on the second part, we use these proportions to compute a per-word likelihood. This approach permits a fair comparison of different (or differently trained) models.

The topic proportions of held-out documents were computed as follows. For variational inference, we simply estimated $\theta_d$ by $\mathrm{E}_q[\theta_d]$. In the HDPs trained by Gibbs sampling, we sampled topic assignments $z_{dn}$ for each word in the first part of the document and computed $\theta^s_{dk} = \frac{\alpha_{\pi(d)}\theta_{\pi(d)k} + N^s_{dk}}{\alpha_{\pi(d)} + N_d}$, where $N^s_{dk}$ is the number of tokens assigned to $k$th topic in $s$th sample. Finally, we averaged $\theta^s_{dk}$ over 2,000 samples after 500 iterations of burn-in.

### 4.1. Comparison to HDPs

HDPs have been evaluated on several "flat" corpora, which in the manner of Figs. 1–2 we can visualize as *two-level* trees in which all documents are directly attached to a single root node. In this section we compare the results from tiLDA and HDPs on three such corpora from the UCI Machine Learning Repository (Frank & Asuncion, 2010). These corpora are: (1) KOS—a collection of 3,430 blog articles with 467,714 tokens and a 6,906-term vocabulary; (2) Enron—a collection of 39,861 email messages with roughly 6 million tokens and a 28,102-term vocabulary; (3) NYTimes—a collection of 300K news articles with a 102,660-term vocabulary. The full NYTimes corpus was too large for our experiments on (batch) HDPs so we extracted a subset of 80K articles with 26 million tokens.

On the KOS, Enron, and NYTimes corpora we compared tiLDA to two publicly available batch implementations[1,2] of HDPs, one based on Gibbs sampling (Teh et al., 2006), the other based on variational methods (Wang et al., 2011). We denote the former by HDP-Gibbs and the latter by HDP-Variational. For all algorithms we used the same hyperparameters ($\gamma = \alpha_0 = 1$) and the same symmetric Dirichlet prior on topics. We initialized HDP-Gibbs with 100 topics, and we experimented with three settings of the truncation parameters $(K, T)$ in HDP-Variational, where $K$ is the number of topics per corpus and $T$ is the number of topics per document. These settings were $(K = 150, T = 15)$ as reported in previous

[1] http://www.stats.ox.ac.uk/~teh/software.html
[2] http://www.cs.cmu.edu/~chongw/resource.html

work (Wang et al., 2011) as well as ($K=300, T=50$) and ($K=300, T=100$). For each corpus we only report the results from HDP-Variational for the *best* of these settings. In our experience, however, HDP-Variational was sensitive to these settings, exhibiting the same or more variance than tiLDA over widely different choices for its fixed number of topics.

Figure 6 summarizes our experimental results. The error bars for tiLDA show the standard deviation in per-word log-likelihood over five different folds of each corpus. (In each experiment, one fold was held out for testing while the other four were used for training.) Also shown are the range of results on these folds for HDP-Gibbs and HDP-Variational. On the smaller KOS and Enron corpora, we obtain our best results[3] with HDP-Gibbs; however, we emphasize that HDP-Gibbs was too slow to train even on our subset of the NYTimes corpus. Comparing tiLDA and HDP-Variational, we find that the former does significantly better on the KOS and NYTimes corpora. On Enron, the corpus which appears to contain the most topics, the order is reversed (but only provided that one explores the space of truncation parameters for HDP-Variational). Though one cannot conclude too much from three corpora, these results certainly establish the viability and scalability of tiLDA. We now turn to the sorts of applications for which tiLDA was explicitly conceived.

### 4.2. Hierarchical Corpora

In this section we demonstrate the benefits of tiLDA when it can exploit known hierarchical structure in corpora. We experimented on three corpora with such structure. These corpora are: (1) NIPS—a collection[4] of 1567 NIPS papers from 9 subject categories, with over 2 million tokens and a 13,649-term vocabulary; (2) Freelancer—a collection of 355,386 job postings by 6,920 advertisers, with over 16M tokens and a 27,600-term vocabulary, scraped from a large crowdsourcing site; (3) BlackHatWorld—a collection of 7,143 threads from a previously underground Internet forum, with roughly 1.4M tokens and a 7,056-term vocabulary. More details and results on the Freelancer and BlackHatWorld corpora can be found in the supplementary material. We previously analyzed the Freelancer corpus using LDA (Kim et al., 2011), but this earlier work did not attempt to model the authorship of job postings as we do here. The BlackHatWorld corpora was collected as part of a larger effort (Motoyama et al., 2011) to examine the social networks among distrustful parties in underground forums.

We evaluated three-level tiLDA models on the NIPS and Freelancer corpora (Fig. 1) and five-level tiLDA models on the BlackHatWorld corpus (Fig. 2). For comparison we also evaluated two-level models of tiLDA that ignored the internal structure of these corpora. We adopted the same settings as in the previous section except that we also learned the models' concentration parameters $\alpha$. Note that we do not have comparative results for multi-level HDPs on these corpora. We know of no Gibbs samplers for HDPs that would scale to corpora of this size and depth. Likewise we know of no variational HDPs that have been implemented for general (multi-level) hierarchies.

Figure 7 shows the results of these evaluations. The plot for the Freelancer corpus (*middle*) shows the average and standard deviation of the per-word log-likelihood over five folds. The plots for the NIPS and BlackHatWorld corpora (*left* and *right*) show the average and standard deviation over five runs, where each run averaged the test results over folds. (We did this because the NIPS and BlackHatWorld corpora are much smaller, and the folds themselves exhibit large variance regardless of the settings.)

The results in Figure 7 paint a consistent picture over a wide range of choices for the number of topics, $K$. In every set of experiments, the deep tiLDA models of hierarchical corpora outperform the flat ones. Overall the results support the notion that deep tiLDA generalizes better for two reasons: first, because it can model different categories with different topic proportions, and second, because it shares information across different categories. These abilities guard, respectively, against the challenges of underfitting and overfitting the data.

We also examined the topics learned by the deep tiLDA models with the highest held-out likelihoods. On the Freelancer corpus, which consists of job postings, these topics can be interpreted as different job types (Kim et al., 2011). Table 1 shows four of the more pernicious job types on Freelancer.com identified by discovered topics. The "OSN (Online Social Network) Linking" topic describes jobs to generate friends and fans on sites such as Facebook and Twitter. The "Ad Posting" topic describes jobs to post classified ads on sites such as Craigslist. Many other jobs are related to search engine optimization (SEO). The "SEO Content Generation" topic describes jobs to generate keyword-rich articles that drive traffic from search engines. Likewise, the "SEO Link Building" topic de-

[3] It has been suggested that the careful selection of hyperparameters may reduce the gap between Gibbs sampling and variational methods in topic models (Asuncion et al., 2009); we did not explore that here.

[4] http://www.stats.ox.ac.uk/~teh/data.html

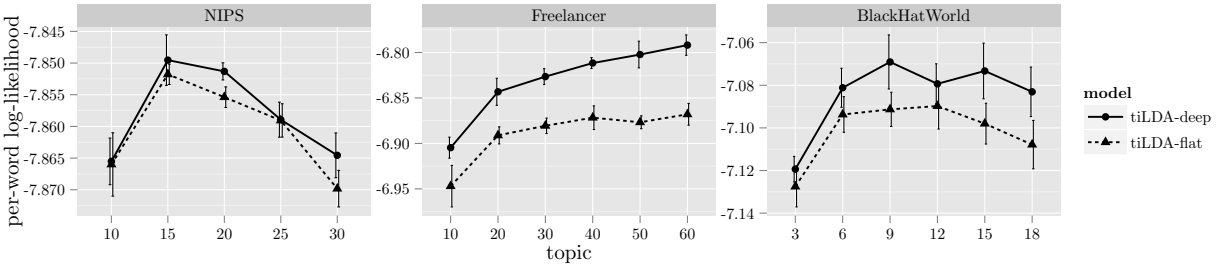


*Figure 7.* Predictive likelihood on the NIPS, Freelancer, and BlackHatWorld corpora from deep (multi-level) and flat (two-level) models of tiLDA, with varying numbers of topics.

*Table 1.* Four examples of the $K = 60$ topics discovered by tiLDA on the Freelancer corpus; training time was 60 hours. Shown are the six most probable words for each topic. Capitalized terms indicate project keywords.

| **"OSN Linking"** | **"Ad Posting"** | **"SEO Content Generation"** | **"SEO Link Building"** |
|---|---|---|---|
| facebook | ad | articl | post |
| fan | post | keyword | blog |
| friend | craigslist | word | forum |
| page | day | topic | comment |
| twitter | poster | write | link |
| Facebook | section | written | site |

*Table 2.* Four examples of the $K = 9$ topics discovered by tiLDA on the BlackHatWorld corpus; training time was 30 minutes. Shown are the six most probable words for each topic. We replaced dollar amounts by the token DOLLAR.

| **"Email" Marketing"** | **"Google Adsense"** | **"Affiliate Program"** | **"Blogging"** |
|---|---|---|---|
| email | site | DOLLAR | blog |
| list | traffic | make | forum |
| proxy | googl | money | learn |
| ip | adsens | affili | post |
| send | domain | market | black |
| server | ad | product | hat |

scribes jobs to increase a Web site's PageRank (Brin & Page, 1998) by adding links from blogs and forums.

On the BlackHatWorld corpus, the topics discovered by tiLDA relate to different types of Internet marketing. Table 2 shows four particularly interpretable topics. The "Email Marketing" topic describes strategies for bulk emailing (spam). The "Google Adsense" topic describes ways for online publishers (e.g., bloggers) to earn money by displaying ads suggested by Google on their Web sites. The "Affiliate Program" topic describes ways to earn commissions by marketing on behalf of other merchants. Finally, the "Blogging" topic describes the use of blogs for Internet marketing.

## 5. Conclusion

In this paper we have explored a generalization of LDA for hierarchical corpora. The parametric model that we introduce can also be viewed as a finite-dimensional HDP. Our main technical contribution is theorem 3.1, which has many potential uses in graphical models with latent Dirichlet variables. Our main empirical contribution is a parallel implementation that scales to very large corpora and deep hierarchies. Our results on the Freelancer and BlackHatWorld corpora illustrate two real-world applications of our approach.

Unlike tiLDA, nonparametric topic models can infer the number of topics from data and grow this number as more data becomes available. But this advantage of HDPs does not come without various complexities. Variational inference in tiLDA does not require stick-breaking constructions or truncation schemes, and it generalizes easily to hierarchies of arbitrary depth. For many applications, we believe that tiLDA provides a compelling alternative to the full generality of HDPs. The approximations we have developed for tiLDA may also be useful for truncated versions of nonparametric models (Kurihara et al., 2007).

We note one potential direction for future work. In this paper, we have studied a batch framework for variational inference. Online approaches, like those recently explored for LDA (Hoffman et al., 2010) and HDPs (Wang et al., 2011; Wang & Blei, 2012; Bryant & Sudderth, 2012), also seem worth exploring for tiLDA. Such approaches may facilitate even larger and more diverse applications.

## Acknowledgments

We thank the reviewers for helpful comments. This work was supported in part by ONR MURI grant N000140911081 and NSF grant NSF-1237264.

## References


Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. On smoothing and inference for topic models. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

Blei, D. M. and Lafferty, J. Topic models. In *Text Mining: Theory and Applications*. Taylor and Francis, London, UK, 2009.

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

Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on the World Wide Web*, pp. 107–117, 1998.

Bryant, M. and Sudderth, E. Truly nonparametric online variational inference for hierarchical Dirichlet processes. In Bartlett, P., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 2708–2716. 2012.

Du, L., Buntine, W., and Jin, H. A segmented topic model based on the two-parameter Poisson-Dirichlet process. *Machine Learning*, 81:5–19, 2010.

Frank, A. and Asuncion, A. UCI machine learning repository, 2010.

Hoffman, M., Blei, D., and Bach, F. Online learning for latent Dirichlet allocation. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 856–864. 2010.

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

Kim, D.-k., Motoyama, M., Voelker, G. M., and Saul, L. K. Topic modeling of freelance job postings to monitor Web service abuse. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence (AISec-11)*, pp. 11–20, 2011.

Kurihara, K., Welling, M., and Teh, Y. W. Collapsed variational Dirichlet process mixture models. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, 2007.

Motoyama, M., McCoy, D., Levchenko, K., Savage, S., and Voelker, G. M. An analysis of underground forums. In *Proceedings of the 2011 ACM SIGCOMM Internet Measurement Conference (IMC-11)*, pp. 71–80, 2011.

Sato, I., Kurihara, K., and Nakagawa, H. Practical collapsed variational Bayes inference for hierarchical Dirichlet process. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-12)*, August 2012.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, December 2006.

Teh, Y. W., Kurihara, K., and Welling, M. Collapsed variational inference for HDP. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems 20*, pp. 1481–1488. 2008.

Wallach, H., Mimno, D., and McCallum, A. Rethinking LDA: why priors matter. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 22*, pp. 1973–1981. 2009a.

Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. Evaluation methods for topic models. In Bottou, L. and Littman, M. (eds.), *Proceedings of the 26th International Conference on Machine Learning (ICML-09)*, pp. 1105–1112, Montreal, June 2009b. Omnipress.

Wang, C., Paisley, J., and Blei, D. Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

Wang, C. and Blei, D. Truncation-free online variational inference for Bayesian nonparametric models. In Bartlett, P., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 422–430. 2012.