Yi Yang, Doug Downey, and **Jordan Boyd-Graber**. **Efficient Methods for Incorporating Knowledge into Topic Models**. *Empirical Methods in Natural Language Processing*, 2015, 9 pages.


```
    @inproceedings{Yang:Downey:Boyd-Graber-2015,
Author = {Yi Yang and Doug Downey and Jordan Boyd-Graber},
Url = {docs/2015_emnlp_fast_priors.pdf},
Booktitle = {Empirical Methods in Natural Language Processing},
Location = {Lisbon, Portugal},
Year = {2015},
Title = {Efficient Methods for Incorporating Knowledge into Topic Models},
}
```


Links:

- Code [http://github.com/yya518/sparse-constrained-lda]

Efficient Methods for Incorporating Knowledge into Topic Models

Yi Yang, Doug Downey
Electrical Engineering and Computer Science
Northwestern University
Evanston, IL
yiyang@u.northwestern.edu
ddowney@eecs.northwetsern.edu

Jordan Boyd-Graber
Computer Science
University of Colorado
Boulder, CO
Jordan.Boyd.Graber
@colorado.edu

Abstract

Latent Dirichlet allocation (LDA) is a popular topic modeling technique for exploring hidden topics in text corpora. Increasingly, topic modeling needs to scale to larger topic spaces and use richer forms of prior knowledge, such as word correlations or document labels. However, inference is cumbersome for LDA models with prior knowledge. As a result, LDA models that use prior knowledge only work in small-scale scenarios. In this work, we propose a factor graph framework, Sparse Constrained LDA (SC-LDA), for efficiently incorporating prior knowledge into LDA. We evaluate SC-LDA's ability to incorporate word correlation knowledge and document label knowledge on three benchmark datasets. Compared to several baseline methods, SC-LDA achieves comparable performance but is significantly faster.

1 Challenge: Leveraging Prior Knowledge in Large-scale Topic Models

Topic models, such as Latent Dirichlet Allocation (Blei et al., 2003, LDA), have been successfully used for discovering hidden topics in text collections. LDA is an unsupervised model—it requires no annotation—and discovers, without any supervision, the thematic trends in a text collection. However, LDA's lack of supervision can lead to disappointing results. Often, the hidden topics learned by LDA fail to make sense to end users. Part of the problem is that the objective function of topic models does not always correlate with human judgments of topic quality (Chang et al., 2009). Therefore, it's often necessary to incorporate prior knowledge into topic models to improve the model's performance. Recent work has also shown that by interactive human feedback can improve the quality and stability of topics (Hu and Boyd-Graber, 2012; Yang et al., 2015). Information about documents (Ramage et al., 2009) or words (Boyd-Graber et al., 2007) can improve LDA's topics.

In addition to its occasional inscrutability, scalability can also hamper LDA's adoption. Conventional Gibbs sampling—the most widely used inference for LDA—scales linearly with the number of topics. Moreover, accurate training usually takes many sampling passes over the dataset. Therefore, for large datasets with millions or even billions of tokens, conventional Gibbs sampling takes too long to finish. For standard LDA, recently introduced fast sampling methods (Yao et al., 2009; Li et al., 2014; Yuan et al., 2015) enable industrial applications of topic modeling to search engines and online advertising, where capturing the "long tail" of infrequently used topics requires large topic spaces. For example, while typical LDA models in academic papers have up to 10^3 topics, industrial applications with 10^5–10^6 topics are common (Wang et al., 2014). Moreover, scaling topic models to many topics can also reveal the hierarchical structure of topics (Downey et al., 2015).

Thus, there is a need for topic models that can both benefit from rich prior information and that can scale to large datasets. However, existing methods for improving scalability focus on topic models without prior information. To rectify this, we propose a factor graph model that encodes a potential function over the hidden topic variables, encouraging topics consistent with prior knowledge. The factor model representation admits an efficient sampling algorithm that takes advantage of the model's sparsity. We show that our method achieves comparable performance but runs significantly faster than baseline methods, enabling mod-

els to discover models with many topics enriched by prior knowledge.

2 Efficient Algorithm for Incorporating Knowledge into LDA

In this section, we introduce the factor model for incorporating prior knowledge and show how to efficiently use Gibbs sampling for inference.

2.1 Background: LDA and SparseLDA

A statistical topic model represents words in documents in a collection D as mixtures of T topics, which are multinomials over a vocabulary of size V. In LDA, each document d is associated with a multinomial distribution over topics, θ_d. The probability of a word type w given topic z is $\phi_{w|z}$. The multinomial distributions θ_d and ϕ_z are drawn from Dirichlet distributions: α and β are the hyperparameters for θ and ϕ. We represent the document collection D as a sequence of words $\mathbf{w}$, and topic assignments as $\mathbf{z}$. We use symmetric priors α and β in the model and experiment, but asymmetric priors are easily encoded in the models (Wallach et al., 2009).

Discovering the latent topic assignments $\mathbf{z}$ from observed words $\mathbf{w}$ requires inferring the the posterior distribution $P(\mathbf{z}|\mathbf{w})$. Griffiths and Steyvers (2004) propose using collapsed Gibbs sampling. The probability of a topic assignment $z = t$ in document d given an observed word type w and the other topic assignments z_- is

$$P(z = t|\mathbf{z}_-, w) \propto (n_{d,t} + \alpha)\frac{n_{w,t} + \beta}{n_t + V\beta} \quad (1)$$

where $\mathbf{z}_-$ are the topic assignments of all other tokens. This conditional probability is based on cumulative counts of topic assignments: $n_{d,t}$ is the number of times topic t is used in document d, $n_{w,t}$ is the number of times word type w is used in topic t, and n_t is the marginal count of the number of tokens assigned to topic t.

Unfortunately, explicitly computing the conditional probability is quite for models with many topics. The time complexity of drawing a sample by Equation 1 is linear to the number of topics. Yao et al. (2009) propose a clever factorization of Equation 1 so that the complexity is typically sublinear by breaking the conditional probability into three "buckets":

$$\sum_t P(z = t|\mathbf{z}_-, w) = \underbrace{\sum_t \frac{\alpha\beta}{n_t + V\beta}}_{\text{s}} + \underbrace{\sum_{t,n_{d,t}>0} \frac{n_{d,t}\beta}{n_t + V\beta}}_{\text{r}} + \underbrace{\sum_{t,n_{w,t}>0} \frac{(n_{d,t} + \alpha)n_{w,t}}{n_t + V\beta}}_{\text{q}}. \quad (2)$$

The first term s is the "smoothing only" bucket—constant for all documents. The second term r is the "document only" bucket that is shared by a document's tokens. Both s and r have simple constant time updates. The last term q has to be computed specifically for each token, only for the few types with non-zero counts in a topic, due to the sparsity of word-topic count. Since q often has the largest mass and few non-zero terms, we start the sampling from bucket q.

2.2 A Factor Model for Incorporating Prior Knowledge

With SparseLDA, inferring LDA models over large topic spaces becomes tractable. However, existing methods for incorporating prior knowledge use conventional Gibbs sampling, which hinders inference. We address this limitation in this section by adding a factor graph to encode prior knowledge.

LDA assumes that the hidden topic assignment of a word is independent from other hidden topics, given the document's topic distribution θ. While this assumption facilitates computational efficiency, it loses the rich correlation between words. In many scenarios, users have external knowledge regarding word correlation, document labels, or document relations, which can reshape topic models and improve coherence.

Prior knowledge can constrain what models discover. A correlation between two words v and w indicates that they have a similar topic distribution, i.e., $p(z|v) \approx p(z|w)$.[1] Therefore, the posterior topic assignments v and w will be correlated. In contrast, if v and w are uncorrelated, nothing—other than the Dirichlet's rich get richer effect—prevents the topics from diverging. Similarly, if two documents share a label, then it is reasonable

[1] In (Andrzejewski et al., 2009) two correlated words are taken to indicate that $p(v|z) \approx p(w|z)$. However, for word types that have very different frequencies, these two quantities would never be close, and thus $p(z|v) \approx p(z|w)$ is a more intuitive constraint.

to assume that they are more likely than two random documents to share topics.

We denote the set of prior knowledge as M. Each prior knowledge $m \in M$ defines a potential function $f_m(z, w, d)$ of the hidden topic z of word type w in document d with which m is associated. Therefore, the complete prior knowledge M defines a score on the current topic assignments $\mathbf{z}$:

$$\psi(\mathbf{z}, M) = \prod_{z \in \mathbf{z}} \exp f_m(z, w, d) \tag{3}$$

If m is knowledge about word type w, then $f_m(z, w, d)$ applies to all hidden topics of word w. If m is knowledge about document d, then $f_m(z, w, d)$ applies to all topics that are in document d. The potential function assigns large values to the topics that accord with prior knowledge but penalizes the topic assignments that disagree with the prior knowledge. In an extreme case, if a prior knowledge m says word type w in document d is Topic 3, then the potential function $f_m(z, w, d)$ is zero for all topics but Topic 3.

Since the potential function ψ is a function of $\mathbf{z}$, and it is only a real-value score of current topic assignments, the potential can be factored out of the marginalized joint:

$$\begin{aligned} &P(\mathbf{w}, \mathbf{z}|\alpha, \beta, M) = P(\mathbf{w}|\mathbf{z}, \beta) P(\mathbf{z}|\alpha) \psi(\mathbf{z}, M) \qquad (4) \\ &= \int_\theta \int_\phi p(\mathbf{w}|\mathbf{z}, \phi) p(\phi|\beta) p(\mathbf{z}|\theta) p(\theta|\alpha) \psi(\mathbf{z}, M) d\theta d\phi \\ &= \psi(\mathbf{z}, M) \int_\theta \int_\phi p(\mathbf{w}|\mathbf{z}, \phi) p(\phi|\beta) p(\mathbf{z}|\theta) p(\theta|\alpha) d\theta d\phi. \end{aligned}$$

Given the joint likelihood and observed data, the goal is evaluate the posterior $P(\mathbf{z}|\mathbf{w})$. Computing $P(\mathbf{z}|\mathbf{w})$ involves evaluating a probability distribution on a large discrete state space: $P(\mathbf{z}|\mathbf{w}) = P(\mathbf{z}, \mathbf{w}) / \sum_{\mathbf{z}} P(\mathbf{z}, \mathbf{w})$. Griffiths and Steyvers (2004)—mirroring the original inspirations for Gibbs sampling (Geman and Geman, 1990)—draw an analogy to statistical physics, viewing standard LDA as a system that favors configurations $\mathbf{z}$ that compromise between having few topics per document and having few words per topic, with the terms of this compromise being set by the hyperparameters α and β. Our factor model representation of prior knowledge adds a further constraint that asks the model to also consider ensembles of topic assignments $\mathbf{z}$ that are compatible with a standard LDA model *and* the given prior knowledge.

The collapsed Gibbs Sampling for inferring topic assignment z of word w in document d is:

$$\begin{aligned} &P(z = t|w, \mathbf{z}_-, M) \qquad (5) \\ &= \frac{P(\mathbf{w}, \mathbf{z}_-, z = t|\alpha, \beta, M)}{P(\mathbf{w}, \mathbf{z}_-|\alpha, \beta, M)} \\ &= \frac{P(\mathbf{w}, \mathbf{z}_-, z = t)}{P(\mathbf{w}, \mathbf{z}_-)} \frac{\psi(\mathbf{z}_-, z = t, M)}{\psi(\mathbf{z}_-, M)} \\ &\propto \left\{ (n_{d,t} + \alpha) \frac{n_{w,t} + \beta}{n_t + W\beta} \right\} \frac{\psi(\mathbf{z}_-, z = t, M)}{\psi(\mathbf{z}_-, M)} \\ &\propto \left\{ (n_{d,t} + \alpha) \frac{n_{w,t} + \beta}{n_t + W\beta} \right\} \exp f_m(z = t, w, d). \end{aligned}$$

The first term is identical to standard LDA, and admits efficient computation using SparseLDA. However, if the second term, $\exp f_m(z, w, d)$, is dense, we still need to compute it explicitly T times (once for each topic) because we need the summation of $P(z = t)$ for sampling. Therefore, the critical part of speeding up the sampler is finding a sparse representation of the second term. In the following sections, we show that natural, sparse prior knowledge representations are possible. We first present an efficient sparse representation of word correlation prior knowledge and then one for document-label knowledge.

2.3 Word Correlation Prior Knowledge

We now illustrate how we can encode word correlation knowledge as a set of sparse constraints $f_m(z, w, d)$ in our model. In previous work (Andrzejewski et al., 2009; Hu et al., 2011; Xie et al., 2015), word correlation prior knowledge is represented as word must-link constraints and cannot-link constraints. A must-link relation between two words indicates that the two words tend to be related to the same topics, i.e. their topic probabilities are correlated. In contrast, a cannot-link relation between two words indicates that these two words are not topically similar, and they should not both be prominent within the same topic. For example, "quarterback" and "fumble" are both related to American football, so they can share a must-link relation. But "fumble" and "bank" imply two different topics, so they share a cannot-link.

Let us say word w is associated with a set of prior knowledge correlations M_w. Each prior knowledge $m \in M_w$ is a word pair (w, w'), and it has "topic preference" of w given its correlation word w'. The must-link set of w is M_w^m, and the cannot-link set of w is M_w^c, i.e., $M_w = M_w^c \bigcup M_w^m$. In the example above, $M_{fumble}^m =$

$\{quarterback\}$, and $M^c_{fumble} = \{bank\}$, so $M_{fumble} = \{quarterback, bank\}$. The topic assignment of word "fumble" has higher conditional probability for the same topics as "quarterback" but lower probability for topics containing "bank".

The potential score of sampling topic t for word type w—if M_w is not empty—is

$$f_m(z, w, d) = \sum_{u \in M^m_w} \log \max(\lambda, n_{u,z}) + \sum_{v \in M^c_w} \log \frac{1}{\max(\lambda, n_{v,z})}. \quad (6)$$

where λ is a hyperparameter, which we call the correlation strength. The intuitive explanation of Equation 6 is that the prior knowledge about the word type w will make an impact on the conditional probability of sampling the hidden topic z. Unlike standard LDA where every word's hidden topic is independent of other words given θ, Equation 6 instead increases the probability that a word w will be drawn from the same topics as those of w's must-link word set, and decreases its probability of being drawn from the same topics as those of w's cannot-link word set.

The hyperparameter λ controls the strength of each piece of prior knowledge. The smaller λ is, the stronger this correlation is. For large λ, the constraint is inactive for topics except those with the large counts. As λ decreases, the constraint becomes active for topics with lesser counts. We can adjust the value of λ for each piece of prior knowledge based on our confidence. In our experiments, for simplicity, we use the same value λ for all knowledge and set $\lambda = 1$.

From Equation 6 and Equation 5, the conditional probability of a topic z in document d given an observed word type w is:

$$P(z = t|w, \mathbf{z}_-, M) \propto \left\{ \frac{\alpha\beta}{n_t + V\beta} + \frac{n_{d,t}\beta}{n_t + V\beta} + \frac{(n_{d,t} + \alpha)n_{w,t}}{n_t + V\beta} \right\} \left\{ \prod_{u \in M^m_w} \max(\lambda, n_{u,t}) \prod_{v \in M^c_w} \frac{1}{\max(\lambda, n_{v,t})} \right\} \quad (7)$$

As explained above, λ controls the "strength" of the prior knowledge term. If λ is large, the prior knowledge has little impact on the conditional probability of topic assignments.

Let's return to the question whether Equation 6 is sparse, allowing efficient computation of Equation 7. Fortunately, $n_{u,t}$ and $n_{v,t}$, which are the topic counts for must-link word u and cannot-link word v, are often sparse. For example, in a 100-topic model trained on the NIPS dataset, 87.2% of word types have fewer than ten topics with nonzero counts. In a 500-topic model trained on a larger dataset like the New York Times News (Sandhaus, 2008), 81.9% of word types have fewer than 50 topics with nonzero counts. Moreover, the model becomes increasingly sparse with additional Gibbs iterations. Figure 1 shows the word frequency histogram of nonzero topic counts of NYT-News dataset.

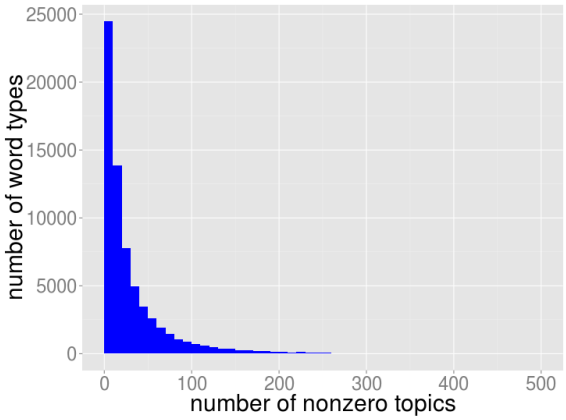

Figure 1: Histogram of nonzero topic counts for word types in NYT-News dataset after inference. 81.9% word types have fewer than 50 topics with nonzero counts. This sparsity allows our sparse constraints to speed inference.

Therefore, the computational cost of Equation 7 can be reduced. SparseLDA efficiently computes the s, r, q bins as in Equation 3. Then for words that are associated with prior knowledge, we update s, r, q with an additional potential term. We only need to compute the potential term for the topics whose counts are greater than λ. The collapsed Gibbs sampling procedure is summarized in Algorithm 1.

Algorithm 1 Gibbs Sampling for word type w in document d, given w's correlation set M_w

1: compute s_t, r_t, q_t with SparseLDA, (see Eq. 3)
2: **for** $t \leftarrow 0$ **to** T **do**
3: update s_t, r_t, q_t. $\forall u \in M_w$ if $n_{u,t} > \lambda$
4: **end for**
5: $p(t) = s_t + r_t + q_t$
6: sample new topic assignment for w from $p(t)$

2.4 Other Types of Prior Knowledge

The factor model framework can also handle other types of prior knowledge, such as document labels, sentence labels, and document link relations. We briefly describe document labels here.

Ramage et al. (2009) propose Labeled-LDA, which improves LDA with document labels. It assumes that there is a one-to-one mapping between topics and labels, and it restricts each document's topics to be sampled only from those allowed by the documents label set. Therefore, Labeled-LDA can be expressed in our model. We define

$$f_m(z, w, d) = \begin{cases} 1, & \text{if } z \in m_d \\ -\infty, & \text{else} \end{cases} \quad (8)$$

where m_d specifies document d's label set converted to corresponding topic labels. Since $f_m(z, w, d)$ is sparse, we can speed up the training as well. Sentence-level prior knowledge (e.g., for sentiment or aspect models (Paul and Girju, 2010)) can be defined in a similar way.

Documents can be associated with other useful metadata. For example, a scientific paper and the prior work it cites might have similar topics (Dietz et al., 2007) or friends in a social network might talk about the same topics (Chang and Blei, 2009). To model link relations, we can use Equation 6 and replace the word-topic counts $n_{v,z}$ with document-topic counts $n_{d,z}$. By doing so, we encourage related documents to have similar topic structures. Moreover, the document-topic count is also sparse, which fits into the efficient learning framework.

Therefore, for different types of prior knowledge, as long as we can define $\psi(\mathbf{z}, M)$ appropriately so that $f(z, w, d)$ are sparse, we are able to speed up learning.

3 Experiments

In this section, we demonstrate the effectiveness of our SC-LDA by comparing it with several baseline methods on three benchmark datasets. We first evaluate the convergence rate of each method and then evaluate the learned model parameter ϕ—the topic-word distribution—in terms of topic coherence. We show that SC-LDA can achieve results comparable to the baseline models but is significantly faster. We set up all experiments on a 8-Core 2.8GHz CPU, 16GB RAM machine.[2]

[2]Our implementation of SC-LDA is available at `https://github.com/yya518/sparse-constrained-lda`.

DATASET	DOCS	TYPE	TOKEN(APPROX)
NIPS	1,500	12,419	1,900,000
NYT-NEWS	3,000,000	102,660	100,000,000
20NG	18,828	21,514	1,946,000

Table 1: Characteristics of benchmark datasets. We use NIPS and NYT for word correlation experiments and 20NG for document label experiments.

3.1 Dataset

We use the NIPS and NYT-News datasets from the UCI bag of words data collections.[3] These two datasets have no document labels, and we use them for word correlation experiments. We also use the 20Newsgroup (20NG) dataset,[4] which has document labels, for document label experiments. Table 1 shows the characteristics of each dataset. Since NIPS and NYT-News have already been preprocessed, to ensure repeatability, we use the data "as they are" from the sources. For 20NG, we perform tokenization and stopword removal using Mallet (McCallum, 2002) and remove words that appear fewer than 10 times.

3.2 Prior Knowledge Generation

Word Correlation Prior Knowledge Previous work proposes two methods to automatically generate prior word correlation knowledge from external sources. Hu and Boyd-Graber (2012) use WordNet 3.0 to obtain synsets for word types, and then if a synset is also in the vocabulary, they add a must-link correlation between the word type and the synset. Xie et al. (2015) use a different method that takes advantage of an existing pretrained word embedding. Each word embedding is a real-valued vector capturing the word's semantic meaning based on distributional similarity. If the similarity between the embeddings of two word types in the vocabulary exceeds a threshold, they generate a must-link between the two words.

In our experiments, we adopt a hybrid method that combines the above two methods. For a noun word type, we first obtain its synsets from WordNet 3.0. We also obtain the embeddings of each word from word2vec (Mikolov et al., 2013). If the synset is also in the vocabulary, and the similarity between the synset and the word is higher than a threshold, which in our experiment is 0.2, we generate a must-link between thee words. Empir-

[3]https://archive.ics.uci.edu/ml/datasets/Bag+of+Words

[4]http://qwone.com/ jason/20Newsgroups/

ically, this hybrid method is able to obtain high quality correlated words. For example, for the NIPS dataset, the must-links we obtain for *randomness* are {noise, entropy, stochasticity}.

Document Label Prior Knowledge Since documents in the 20NG dataset are associated with labels, we use the labels directly as prior knowledge.

3.3 Baselines

The baseline methods for incorporating word correlation prior knowledge in our experiments are as follows:

DF-LDA: incorporates word must-links and cannot-links using a Dirichlet Forest prior in LDA (Andrzejewski et al., 2009). Here we use Hu and Boyd-Graber (2012)'s efficient implementation *FAST-RB-SDW* for DF-LDA.

Logic-LDA: encodes general domain knowledge as first-order logic and incorporates it in LDA (Andrzejewski et al., 2011). Logic-LDA has been used for word correlations and document label knowledge.

MRF-LDA: encodes word correlations in LDA as a Markov random field (Xie et al., 2015).

We also use Mallet's SparseLDA implementation for vanilla LDA in the topic coherence experiment. We use a symmetric Dirichlet prior for all models. We set $\alpha = 1.0$, $\beta = 0.01$. For DF-LDA, $\eta = 100$. For Logic-LDA, we use the default parameter setting in the package: a sample rate of 1.0 and step rate of 10.0. For MRF-LDA, we use the default setting with $\gamma = 1.0$. (Parameter semantics can be found in the original papers.)

3.4 Convergence

The main advantage of our method over other existing methods is efficiency. In this experiment, we show the change of our model's log likelihood over time. In topic models, the log likelihood change is a good indicator of whether a model has converged or not. Figure 2 shows the log likelihood change over time for SC-LDA and three baseline methods on NIPS and NYT-News dataset. SC-LDA converges faster than all the other methods.

We also conduct experiments on SC-LDA with varying numbers of word correlations. Table 2 shows the Gibbs sampling iteration time on the 1st, 50th, 100th and the 200th iteration. We also incorporate different numbers of word correlations in SC-LDA. SC-LDA runs faster as sampling proceeds as the sparsity increases, but additional correlations slow the model.

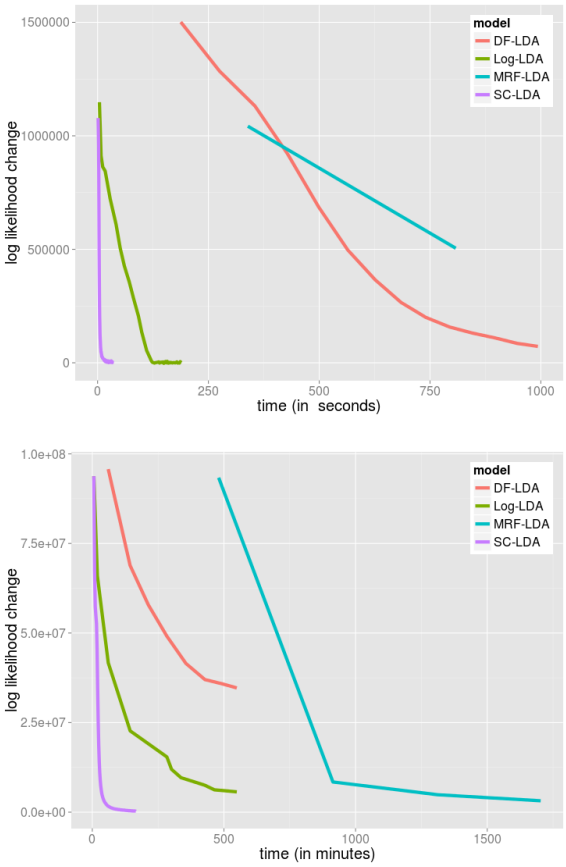

Figure 2: Models' log likelihood convergence on NIPS dataset (above) and NYT-News dataset (below). For NIPS, a 100-topic model with 100 must-links is trained. For NYT-News, a 500-topic model with 100 must-links is trained. SC-LDA reaches likelihood convergence much more rapidly than the other methods.

	# Word Correlations			
round	C0	C100	C500	C1000
1st iteration	2.02	2.14	2.30	2.50
50th iteration	0.53	0.56	0.58	0.62
100th iteration	0.48	0.50	0.53	0.56
200th iteration	0.48	0.49	0.52	0.56

Table 2: SC-LDA runtime (in seconds) in the 1st, 50th, 100th, and 200th iteration with different numbers of correlations.

3.5 Topic Coherence

Topic models are often evaluated using perplexity on held-out test data, but this evaluation is of-

ten at odds with human evaluations (Chang et al., 2009). Following Mimno et al. (2011), we employ Topic Coherence—a metric that is consistent with human judgment—to measure a topic model's quality. Topic t's coherence is defined as $C(t : V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{F(v_m^{(t)}, v_l^{(t)}) + \epsilon}{F(v_l^{(t)})}$, where $F(v)$ is the document frequency of word type v, $F(v, v')$ is the co-document frequency of word type v and v', and $V^{(t)} = (v_1^{(t)}, ..., v_M^{(t)})$ is a list of the M most probable words in topic t. In our experiments, we choose the ten words with highest probability in the topic to compute topic coherence, i.e., $M = 10$. Mimno et al. (2011) use $\epsilon = 1$, but Röder et al. (2015) show smaller ϵ (such as 10^{-12}) improves coherence stability, so we set $\epsilon = 10^{-12}$. Larger topic coherence scores imply more coherent topics.

We train a 500-topic model on the NIPS dataset with different methods and compare the average topic coherence score and the average of the top twenty topic coherence scores. Since the topics learned by topic model often contain "bad" topics (Mimno et al., 2011) which do not make sense to end users, evaluating the top twenty topics reflects the model's performance. We let each model train for one hour. Figure 3 shows the topic coherence of each method. SC-LDA has about the same average topic coherence with LDA but has higher coherence score (-36.6) for the top 20 topics than LDA (-39.1). This is because incorporating word correlation knowledge encourages correlated words to have high probability under the same topic, thus improving the coherence score. For the other methods, however, because they cannot converge within an hour, their topic coherence scores are much worse than SC-LDA and LDA. This again demonstrates the efficiency of SC-LDA over other baselines.

3.6 Document Label Prior Knowledge

SC-LDA can also handle other types of prior knowledge. We compare it with Labeled-LDA (Ramage et al., 2009). Labeled-LDA also uses Gibbs sampling for inference, allowing direct computation time comparisons.

Table 3 shows the average running time per iteration for Labeled-LDA and SC-LDA. Because document labels apply sparsity to the document-topic counts, the average running time per iteration decreases as the number of labeled document increases. SC-LDA exhibits greater speedup with more topics; when $T = 500$,[5] SC-LDA runs more than ten times faster than Labeled-LDA.

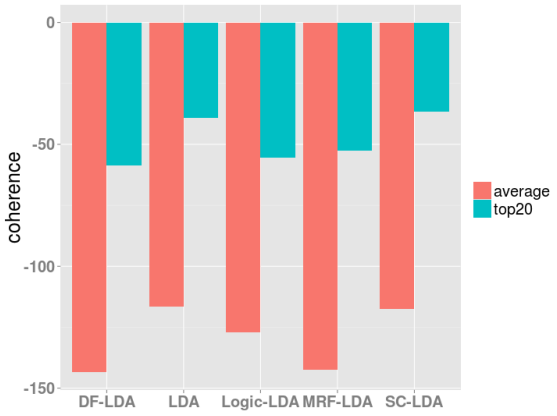

Figure 3: Average topic coherence and average top 20 topic coherence. The models are trained on NIPS dataset with 500-topic and 100 word correlations. SC-LDA achieves higher topic coherence than other methods.

# Topics				
	T50	T100	T200	T500
Labeled-LDA	0.93	1.89	3.60	8.05
SC-LDA	0.38	0.45	0.51	0.72
# Labeled Documents				
	C500	C1000	C2000	C5000
Labeled-LDA	1.95	1.88	1.75	1.48
SC-LDA	0.51	0.45	0.41	0.31

Table 3: The average running time per iteration over 100 iterations, averaged over 5 seeds, on 20NG dataset. Experiments begin with 100 topics, 1000 labeled documents, and then vary one dimension: number of topics (top), or number of labeled documents (bottom).

4 Related Work

This works brings together two lines of research: incorporating rich knowledge into probabilistic models and efficient inference of probabilistic models on large datasets. Both are common areas of interest across many machine learning formalisms: probabilistic logic (Bach et al., 2015), graph algorithms (Low et al., 2012), and probabilistic grammars (Cohen et al., 2008). However, our focus in this paper is the intersection of these lines of research with topic models.

[5]For 20NG dataset, it may overfit the data with 500 topics, but here we use it to demonstrate the scalability.

Adding knowledge and metadata to topic models makes the models richer, more understandable, and more domain-specific. A common distinction is upstream (conditioning on metadata) vs. downstream models (conditioning on variables already present in a topic model to predict metadata) (Mimno et al., 2008). Downstream models are typically better at prediction tasks such as predicting sentiment (Blei and McAuliffe, 2007), ideology (Nguyen et al., 2014a), or links in a social network (Chang and Blei, 2009). In contrast, our approach—an upstream model—is often easier to implement and leads to more interpretable topics. Upstream models at the document level have been used to understand the labels in large document collections (Ramage et al., 2009; Nguyen et al., 2014b) and capture relationships in document networks using Markov random fields (Daumé III, 2009). At the word level, Xie et al. (2015) incorporate word correlation to LDA by building a Markov Random Field regularization, similar to Newman et al. (2011), who use regularization to improve topic coherence. However, despite these exciting applications, the experiments in the above work are typically on small datasets.

In contrast, there is a huge interest in improving the scalability of topic models to large numbers of documents, numbers of topics, and vocabularies. Attempts to scale inference for topic models have started from both variational inference and Gibbs sampling—two popular learning inference techniques for topic modeling. Gibbs sampling is a popular technique because of its simplicitly and low latency. However, for large numbers of topics, Gibbs sampling can become unwieldy. Porteous et al. (2008) address this issue by creating an upper bound approximation that produces accurate results, while SparseLDA (Yao et al., 2009) present an effective factorization that speeds inference without sacrificing accuracy. Just as our model builds on SparseLDA's insights, SparseLDA has been incorporated into commercial deployments (Wang et al., 2014) and improved using alias tables (Li et al., 2014). Yuan et al. (2015) also presents an efficient constant time sampling algorithm for building big topic models. Variational inference can easily be parallelized (Nallapati et al., 2007; Zhai et al., 2012), but has high latency, which has been addressed by performing online updates (Hoffman et al., 2010) and taking stochastic gradients estimated by MCMC inference (Mimno et al., 2012). In this paper, we only focus on single-processor learning, but existing parallelization techniques (Newman et al., 2009) are applicable to our model.

At the intersection lies models that improve the scalability of upstream topic model inference. In addition to our SC-LDA, Hu and Boyd-Graber (2012) speed Gibbs sampling in tree-based topic models using SparseLDA's factorization strategy, and Hu et al. (2014) extend this approach by parallelizing global parameter updates using variational inference. Our work is more general (also encompassing document-based constraints) and is faster. In contrast to these upstream models, Zhu et al. (2013) and Nguyen et al. (2015) improve inference of downstream models.

5 Conclusion

We present a factor graph framework for incorporating prior knowledge into topic models. By expressing the prior knowledge as sparse constraints on the hidden topic variables, we are able to take advantage of the sparsity to speed up training. We demonstrate in experiments that our model runs significantly faster than the other alternative models and achieves comparable performance in terms of topic coherence. Efficient algorithms for incorporating prior knowledge with large topic models will benefit several downstream applications. For example, interactive topic modeling becomes feasible because fast model updates reduce the user's waiting time and thus improve the user experience. Personalized topic modeling is also an interesting future direction in which the model will generate a personalized topic structure based on the user's preferences or interests. For all these applications, an efficient learning algorithm is a crucial prerequisite.

Acknowledgments

We thank the anonymous reviews for their helpful comments. This research was supported in part by NSF grant IIS-1351029 and DARPA contract D11AP00268. Boyd-Graber is supported by NSF Grants CCF-1409287, IIS-1320538, and NCSE-1422492. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the International Conference of Machine Learning*.

David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. 2011. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *International Joint Conference on Artificial Intelligence*.

Stephen H. Bach, Bert Huang, Jordan Boyd-Graber, and Lise Getoor. 2015. Paired-dual learning for fast training of latent variable hinge-loss mrfs. In *Proceedings of the International Conference of Machine Learning*.

David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3.

Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jonathan Chang and David M. Blei. 2009. Relational topic models for document networks. In *Proceedings of Artificial Intelligence and Statistics*.

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

Shay B. Cohen, Kevin Gimpel, and Noah A. Smith. 2008. Logistic normal priors for unsupervised probabilistic grammar induction. In *Proceedings of Advances in Neural Information Processing Systems*.

Hal Daumé III. 2009. Markov random topic fields. In *Proceedings of Artificial Intelligence and Statistics*.

Laura Dietz, Steffen Bickel, and Tobias Scheffer. 2007. Unsupervised prediction of citation influences. In *Proceedings of the International Conference of Machine Learning*.

Doug Downey, Chandra Bhagavatula, and Yi Yang. 2015. Efficient methods for inferring large sparse topic hierarchies. In *Proceedings of the Association for Computational Linguistics*.

S. Geman and D. Geman, 1990. *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, pages 452–472. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228–5235.

Matthew Hoffman, David M. Blei, and Francis Bach. 2010. Online learning for latent Dirichlet allocation. In *Proceedings of Advances in Neural Information Processing Systems*.

Yuening Hu and Jordan Boyd-Graber. 2012. Efficient tree-based topic modeling. In *Proceedings of the Association for Computational Linguistics*.

Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. 2011. Interactive topic modeling. In *Proceedings of the Association for Computational Linguistics*.

Yuening Hu, Ke Zhai, Vlad Eidelman, and Jordan Boyd-Graber. 2014. Polylingual tree-based topic models for translation domain adaptation. In *Association for Computational Linguistics*.

Aaron Q. Li, Amr Ahmed, Sujith Ravi, and Alexander J. Smola. 2014. Reducing the sampling complexity of topic models. In *Knowledge Discovery and Data Mining*.

Yucheng Low, Danny Bickson, Joseph Gonzalez, Carlos Guestrin, Aapo Kyrola, and Joseph M. Hellerstein. 2012. Distributed graphlab: A framework for machine learning and data mining in the cloud. *Proceedings of the VLDB Endowment*, 5(8):716–727, April.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems*.

David Mimno, Hanna Wallach, and Andrew McCallum. 2008. Gibbs sampling for logistic normal topic models with graph-based priors. In *NIPS 2008 Workshop on Analyzing Graphs: Theory and Applications*.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of Empirical Methods in Natural Language Processing*.

David Mimno, Matthew Hoffman, and David Blei. 2012. Sparse stochastic inference for latent Dirichlet allocation. In *Proceedings of the International Conference of Machine Learning*.

Ramesh Nallapati, William Cohen, and John Lafferty. 2007. Parallelized variational EM for latent Dirichlet allocation: An experimental evaluation of speed and scalability. In *International Conference on Data Mining Workshops*.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. Distributed Algorithms for Topic Models. *Journal of Machine Learning Research*, pages 1801–1828.

David Newman, Edwin Bonilla, and Wray Buntine. 2011. Improving topic coherence with regularized topic models. In *Proceedings of Advances in Neural Information Processing Systems*, December.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah Cai, Jennifer Midberry, and Yuanxin Wang. 2014a. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 95:381–421.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Jonathan Chang. 2014b. Learning a concept hierarchy from multi-labeled documents. In *Neural Information Processing Systems*.

Thang Nguyen, Jordan Boyd-Graber, Jeff Lund, Kevin Seppi, and Eric Ringger. 2015. Is your anchor going up or down? Fast and accurate supervised topic models. In *North American Association for Computational Linguistics*.

Michael Paul and Roxana Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Association for the Advancement of Artificial Intelligence*.

Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Knowledge Discovery and Data Mining*.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of Empirical Methods in Natural Language Processing*.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of ACM International Conference on Web Search and Data Mining*.

Evan Sandhaus. 2008. The New York Times annotated corpus. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19.

Hanna Wallach, David Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Proceedings of Advances in Neural Information Processing Systems*.

Yi Wang, Xuemin Zhao, Zhenlong Sun, Hao Yan, Lifeng Wang, Zhihui Jin, Liubin Wang, Yang Gao, Jia Zeng, Qiang Yang, and Ching Law. 2014. Towards topic modeling for big data. *CoRR*, abs/1405.4402.

Pengtao Xie, Diyi Yang, and Eric P Xing. 2015. Incorporating word correlation knowledge into topic modeling. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Yi Yang, Shimei Pan, Yangqiu Song, Jie Lu, and Mercan Topkara. 2015. User-directed non-disruptive topic model update for effective exploration of dynamic content. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Knowledge Discovery and Data Mining*.

Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. 2015. Lightlda: Big topic models on modest computer clusters. In *Proceedings of the World Wide Web Conference*.

Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad Alkhouja. 2012. Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of the World Wide Web Conference*.

Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. 2013. Gibbs max-margin topic models with fast sampling algorithms. In *Proceedings of the International Conference of Machine Learning*.